CHAPTER

1

Elementary Materials Science Concepts[1]

Understanding the basic building blocks of matter has been one of the most intriguing endeavors of humankind. Our understanding of interatomic interactions has now reached a point where we can quite comfortably explain the macroscopic properties of matter, based on quantum mechanics and electrostatic interactions between electrons and ionic nuclei in the material. There are many properties of materials that can be explained by a classical treatment of the subject. In this chapter, as well as Chapter 2, we treat the interactions in a material from a classical perspective and introduce a number of elementary concepts. These concepts do not invoke any quantum mechanics, which is a subject of modern physics and is introduced in Chapter 3. Although many useful engineering properties of materials can be treated with hardly any quantum mechanics, it is impossible to develop the science of electronic materials and devices without modern physics.

1.1 ATOMIC STRUCTURE AND ATOMIC NUMBER

The model of the atom that we must use to understand the atom's general behavior involves quantum mechanics, a topic we will study in detail in Chapter 3. For the present, we will simply accept the following facts about a simplified, but intuitively satisfactory, atomic model called the **shell model**, based on the **Bohr model** (1913).

The mass of the atom is concentrated at the nucleus, which contains protons and neutrons. Protons are positively charged particles, whereas neutrons are neutral particles, and both have about the same mass. Although there is a Coulombic repulsion between the protons, all the protons and neutrons are held together in the nucleus by the

[1] This chapter may be skipped by readers who have already been exposed to an elementary course in materials science.

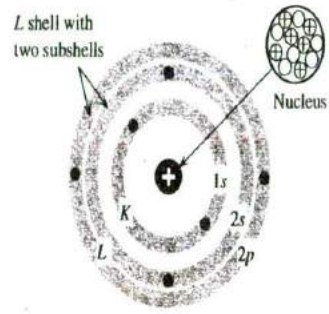


Figure 1.1 The shell model of the carbon atom, in which the electrons are confined to certain shells and subshells within shells.

strong force, which is a powerful, fundamental, natural force between particles. This force has a very short range of influence, typically less than 10^{-15} m. When the protons and neutrons are brought together very closely, the strong force overcomes the electrostatic repulsion between the protons and keeps the nucleus intact. The number of protons in the nucleus is the **atomic number** Z of the element.

The electrons are assumed to be orbiting the nucleus at very large distances compared to the size of the nucleus. There are as many orbiting electrons as there are protons in the nucleus. An important assumption in the Bohr model is that only certain orbits with fixed radii are stable around the nucleus. For example, the closest orbit of the electron in the hydrogen atom can only have a radius of 0.053 nm. Since the electron is constantly moving around an orbit with a given radius, over a long time period (perhaps $\sim 10^{-12}$ seconds on the atomic time scale), the electron would appear as a spherical negative-charge cloud around the nucleus and not as a single dot representing a finite particle. We can therefore view the electron as a charge contained within a spherical **shell** of a given radius.

Due to the requirement of stable orbits, the electrons therefore do not randomly occupy the whole region around the nucleus. Instead, they occupy various well-defined spherical regions. They are distributed in various shells and **subshells** within the shells, obeying certain occupation (or seating) rules.[2] The example for the carbon atom is shown in Figure 1.1.

The shells and subshells that define the whereabouts of the electrons are labeled using two sets of integers, n and ℓ. These integers are called the **principal** and **orbital angular momentum quantum numbers,** respectively. (The meanings of these names are not critical at this point.) The integers n and ℓ have the values $n = 1, 2, 3, \ldots$, and $\ell = 0, 1, 2, \ldots, n-1$, and $\ell < n$. For each choice of n, there are n values of ℓ, so higher-order shells contain more subshells. The shells corresponding to $n = 1, 2, 3, 4, \ldots$

[2] In Chapter 3, in which we discuss the quantum mechanical model of the atom, we will see that these shells and subshells are spatial regions around the nucleus where the electrons are most likely to be found.

Table 1.1 Maximum possible number of electrons in the shells and subshells of an atom

		Subshell			
		$\ell = 0$	1	2	3
n	Shell	s	p	d	f
1	K	2			
2	L	2	6		
3	M	2	6	10	
4	N	2	6	10	14

are labeled by the capital letters $K, L, M, N, \ldots$, and the subshells denoted by $\ell = 0, 1, 2, 3, \ldots$ are labeled $s, p, d, f, \ldots$. The subshell with $\ell = 1$ in the $n = 2$ shell is thus labeled the $2p$ subshell, based on the standard notation $n\ell$.

There is a definite rule to filling up the subshells with electrons; we cannot simply put all the electrons in one subshell. The number of electrons a given subshell can take is fixed by nature to be[3] $2(2\ell + 1)$. For the s subshell ($\ell = 0$), there are two electrons, whereas for the p subshell, there are six electrons, and so on. Table 1.1 summarizes the most number of electrons that can be put into various subshells and shells of an atom. Obviously, the larger the shell, the more electrons it can take, simply because it contains more subshells. The shells and subshells are filled starting with those closest to the nucleus as explained next.

The number of electrons in a subshell is indicated by a superscript on the subshell symbol, so the electronic structure, or configuration, of the carbon atom (atomic number 6) shown in Figure 1.1 becomes $1s^2 2s^2 2p^2$. The K shell has only one subshell, which is full with two electrons. This is the structure of the inert element He. We can therefore write the electronic configuration more simply as $[\text{He}]2s^2 2p^2$. The general rule is put the nearest previous inert element, in this case He, in square brackets and write the subshells thereafter.

The electrons occupying the outer subshells are the farthest away from the nucleus and have the most important role in atomic interactions, as in chemical reactions, because these electrons are the first to interact with outer electrons on neighboring atoms. The outermost electrons are called **valence electrons** and they determine the **valency** of the atom. Figure 1.1 shows that carbon has four valence electrons in the L shell.

When a subshell is full of electrons, it cannot accept any more electrons and it is said to have acquired a stable configuration. This is the case with the inert elements at the right-hand side of the Periodic Table, all of which have completely filled subshells and are rarely involved in chemical reactions. The majority of such elements are gases inasmuch as the atoms do not bond together easily to form a

[3] We will actually show this in Chapter 3 using quantum mechanics.

liquid or solid. They are sometimes used to provide an inert atmosphere instead of air for certain reactive materials.

In an atom such as the Li atom, there are two electrons in the 1*s* subshell and one electron in the 2*s* subshell. The atomic structure of Li is $1s^2 2s^1$. The third electron is in the 2*s* subshell, rather than any other subshell, because this is the arrangement of the electrons that results in the lowest overall energy for the whole atom. It requires energy (work) to take the third electron from the 2*s* to the 2*p* or higher subshells as will be shown in Chapter 3. Normally the zero energy reference corresponds to the electron being at infinity, that is, isolated from the atom. When the electron is inside the atom, its energy is negative, which is due to the attraction of the positive nucleus. An electron that is closer to the nucleus has a lower energy. The electrons nearer the nucleus are more closely bound and have higher binding energies. The $1s^2 2s^1$ configuration of electrons corresponds to the lowest energy structure for Li and, at the same time, obeys the occupation rules for the subshells. If the 2*s* electron is somehow excited to another outer subshell, the energy of the atom increases, and the atom is said to be **excited.**

The smallest energy required to remove a single electron from a neutral atom and thereby create a positive ion (*cation*) and an isolated electron is defined as the **ionization energy** of the atom. The Na atom has only a single valence electron in its outer shell, which is the easiest to remove. The energy required to remove this electron is 5.1 eV, which is the Na atom's ionization energy. The **electron affinity** represents the energy that is needed, or released, when we add an electron to a neutral atom to create a negative ion (*anion*). Notice that the ionization term implies the generation of a positive ion, whereas the electron affinity implies that we have created a negative ion. Certain atoms, notably the halogens (such as F, Cl, Br, I), can actually attract an electron to form a negative ion. Their electron affinities are negative. When we place an electron into a Cl atom, we find that an energy of 3.6 eV is *released*. The Cl^- ion has a lower energy than the Cl atom, which means that it is energetically favorable to form a Cl^- ion by introducing an electron into the Cl atom.

There is a very useful theorem in physics, called the **Virial theorem,** that allows us to relate the average kinetic energy $\overline{KE}$, average potential energy $\overline{PE}$, and average total or overall energy $\overline{E}$ of an electron in an atom, or electrons and nuclei in a molecule, through remarkably simple relationships,[4]

Virial theorem

$$\overline{E} = \overline{KE} + \overline{PE} \qquad \text{and} \qquad \overline{KE} = -\frac{1}{2}\overline{PE} \tag{1.1}$$

For example, if we define zero energy for the H atom as the H^+ ion and the electron infinitely separated, then the energy of the electron in the H atom is −13.6 electron volts (eV). It takes 13.6 eV to ionize the H atom. The average $\overline{PE}$ of the electron, due to its Coulombic interaction with the positive nucleus, is −27.4 eV. Its average $\overline{KE}$ turns out to be 13.6 eV. Example 1.1 uses the Virial theorem to calculate the radius of the hydrogen atom, the velocity of the electron, and its frequency of rotation.

[4] While the final result stated in Equation 1.1 is elegantly simple, the actual proof is quite involved and certainly not trivial. As stated here, the Virial theorem applies to a system of charges that interact through electrostatic forces only.

EXAMPLE 1.1

VIRIAL THEOREM AND THE BOHR ATOM Consider the hydrogen atom in Figure 1.2 in which the electron is in the stable 1*s* orbit with a radius r_o. The ionization energy of the hydrogen atom is 13.6 eV.

a. It takes 13.6 eV to ionize the hydrogen atom, i.e., to remove the electron to infinity. If the condition when the electron is far removed from the hydrogen nucleus defines the zero reference of energy, then the total energy of the electron within the H atom is −13.6 eV. Calculate the average *PE* and average *KE* of the electron.

b. Assume that the electron is in a stable orbit of radius r_o around the positive nucleus. What is the Coulombic *PE* of the electron? Hence, what is the radius r_o of the electron orbit?

c. What is the velocity of the electron?

d. What is the frequency of rotation (oscillation) of the electron around the nucleus?

SOLUTION

a. From Equation 1.1 we obtain

$$\overline{E} = \overline{PE} + \overline{KE} = \tfrac{1}{2}\overline{PE}$$

or

$$\overline{PE} = 2\overline{E} = 2 \times (-13.6\ \text{eV}) = -27.2\ \text{eV}$$

The average kinetic energy is

$$\overline{KE} = -\tfrac{1}{2}\overline{PE} = 13.6\ \text{eV}$$

b. The Coulombic *PE* of interaction between two charges Q_1 and Q_2 separated by a distance r_o, from elementary electrostatics, is given by

$$PE = \frac{Q_1 Q_2}{4\pi\varepsilon_o r_o} = \frac{(-e)(+e)}{4\pi\varepsilon_o r_o} = -\frac{e^2}{4\pi\varepsilon_o r_o}$$

where we substituted $Q_1 = -e$ (electron's charge), and $Q_2 = +e$ (charge of nucleus). Thus the radius r_o is

$$r_o = -\frac{(1.6 \times 10^{-19}\ \text{C})^2}{4\pi(8.85 \times 10^{-12}\ \text{F m}^{-1})(-27.2\ \text{eV} \times 1.6 \times 10^{-19}\ \text{J/eV})}$$

$$= 5.29 \times 10^{-11}\ \text{m} \quad \text{or} \quad 0.0529\ \text{nm}$$

which is called the **Bohr radius** (also denoted a_o).

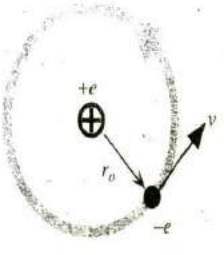


Figure 1.2 The planetary model of the hydrogen atom in which the negatively charged electron orbits the positively charged nucleus.

c. Since $KE = \frac{1}{2}m_e v^2$, the average velocity is

$$v = \sqrt{\frac{\overline{KE}}{\frac{1}{2}m_e}} = \sqrt{\frac{13.6\text{ eV} \times 1.6 \times 10^{-19}\text{ J/eV}}{\frac{1}{2}(9.1 \times 10^{-31}\text{ kg})}} = 2.19 \times 10^6\text{ m s}^{-1}$$

d. The period of orbital rotation T is

$$T = \frac{2\pi r_o}{v} = \frac{2\pi(0.0529 \times 10^{-9}\text{ m})}{2.19 \times 10^6\text{ m s}^{-1}} = 1.52 \times 10^{-16}\text{ seconds}$$

The orbital frequency $\nu = 1/T = 6.59 \times 10^{15}\text{ s}^{-1}$ (Hz).

1.2 ATOMIC MASS AND MOLE

We had defined the atomic number Z as the number of protons in the nucleus of an atom. The **atomic mass number** A is simply the total number of protons and neutrons in the nucleus. It may be thought that we can use the atomic mass number A of an atom to gauge its atomic mass, but this is done slightly differently to account for the existence of different isotopes of an element; isotopes are atoms of a given element that have the same number of protons but a different number of neutrons in the nucleus. The **atomic mass unit** (amu) u is a convenient atomic mass unit that is equal to $\frac{1}{12}$ of the mass of a neutral carbon atom which has a mass number $A = 12$ (6 protons and 6 neutrons). It has been found that $u = 1.66054 \times 10^{-27}$ kg.

The **atomic mass** or **relative atomic mass** or simply **atomic weight** M_{at} of an element is the average atomic mass, in atomic mass units, of all the naturally occurring isotopes of the element. Atomic masses are listed in the Periodic Table. **Avogadro's number** N_A is the number of atoms in exactly 12 grams of carbon-12, which is 6.022×10^{23} to three decimal places. Since the atomic mass M_{at} is defined as $\frac{1}{12}$ of the mass of the carbon-12 atom, it is straightforward to show that N_A number of atoms of any substance has a mass equal to the atomic mass M_{at} in grams.

A **mole** of a substance is that amount of the substance which contains exactly Avogadro's number N_A of atoms or molecules that make up the substance. One mole of a substance has a mass as much as its atomic (molecular) mass in grams. For example, 1 mole of copper contains 6.022×10^{23} number of copper atoms and has a mass of 63.55 grams. Thus, an amount of an element which has 6.022×10^{23} atoms has a mass in grams equal to the atomic mass. This means we can express the atomic mass as grams per unit mole (g mol^{-1}). The atomic mass of Au is 196.97 amu or g mol^{-1}. Thus, a 10 gram bar of gold has (10 g) / (196.97 g mol^{-1}) or 0.0507 moles.

Frequently we have to convert the composition of a substance from atomic percentage to weight percentage, and vice versa. Compositions in materials engineering generally use weight percentages, whereas chemical formulas are given in terms of atomic composition. Suppose that a substance (an alloy or a compound) is composed of two elements, A and B. Let the *weight fractions* of A and B be w_A and w_B, respectively. Let n_A and n_B be the *atomic* or *molar fractions* of A and B; that is, n_A represents the fraction of type A atoms, n_B represents the fraction of type B atoms in the whole

substance, and $n_A + n_B = 1$. Suppose that the atomic masses of A and B are M_A and M_B. Then n_A and n_B are given by

$$n_A = \frac{w_A/M_A}{w_A/M_A + w_B/M_B} \quad \text{and} \quad n_B = 1 - n_A \qquad [1.2]$$

Weight to atomic percentage

where $w_A + w_B = 1$. Equation 1.2 can be readily rearranged to obtain w_A and w_B in terms of n_A and n_B.

COMPOSITIONS IN ATOMIC AND WEIGHT PERCENTAGES Consider a Pb–Sn solder that is 38.1 wt.% Pb and 61.9 wt.% Sn (this is the eutectic composition with the lowest melting point). What are the atomic fractions of Pb and Sn in this solder? **EXAMPLE 1.2**

SOLUTION

For Pb, the weight fraction and atomic mass are respectively $w_A = 0.381$ and $M_A = 207.2$ g mol^{-1} and for Sn, $w_B = 0.619$ and $M_B = 118.71$ g mol^{-1}. Thus, Equation 1.2 gives

$$n_A = \frac{w_A/M_A}{w_A/M_A + w_B/M_B} = \frac{(0.381)/(207.2)}{0.381/207.2 + 0.619/118.71}$$

$$= 0.261 \quad \text{or} \quad 26.1 \text{ at.}\%$$

and

$$n_B = \frac{w_B/M_B}{w_A/M_A + w_B/M_B} = \frac{(0.619)/(118.71)}{0.381/207.2 + 0.619/118.71}$$

$$= 0.739 \quad \text{or} \quad 73.9 \text{ at.}\%$$

Thus the alloy is 26.1 at.% Pb and 73.9 at.% Sn which can be written as $Pb_{0.261}Sn_{0.739}$.

1.3 BONDING AND TYPES OF SOLIDS

1.3.1 MOLECULES AND GENERAL BONDING PRINCIPLES

When two atoms are brought together, the valence electrons interact with each other and with the neighbor's positively charged nucleus. The result of this interaction is often the formation of a bond between the two atoms, producing a molecule. The formation of a bond means that the energy of the system of two atoms together must be less than that of the two atoms separated, so that the molecule formation is energetically favorable, that is, more stable. The general principle of molecule formation is illustrated in Figure 1.3a, showing two atoms brought together from infinity. As the two atoms approach each other, the atoms exert attractive and repulsive forces on each other as a result of mutual electrostatic interactions. Initially, the attractive force F_A dominates over the repulsive force F_R. The net force F_N is the sum of the two,

$$F_N = F_A + F_R$$

Net force

and this is initially attractive, as indicated in Figure 1.3a.

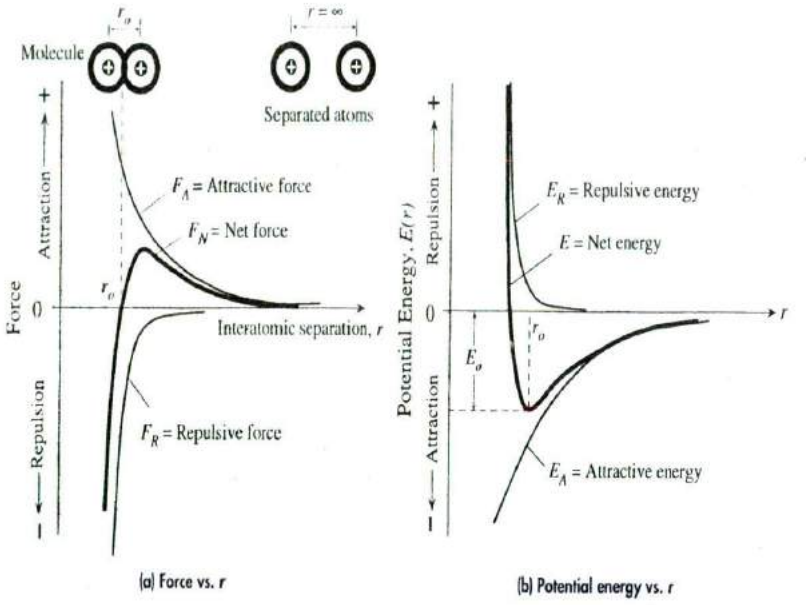


Figure 1.3 (a) Force versus interatomic separation and (b) potential energy versus interatomic separation.

Net force and potential energy

The potential energy $E(r)$ of the two atoms can be found from[5]

$$F_N = \frac{dE}{dr}$$

by integrating the net force F_N. Figure 1.3a and b shows the variation of the net force $F_N(r)$ and the overall potential energy $E(r)$ with the interatomic separation r as the two atoms are brought together from infinity. The lowering of energy corresponds to an attractive interaction between the two atoms.

The variations of F_A and F_R with distance are different. Force F_A varies slowly, whereas F_R varies strongly with separation and is strongest when the two atoms are very close. When the atoms are so close that the individual electron shells overlap, there is a very strong electron-to-electron shell repulsion and F_R dominates. An equilibrium will be reached when the attractive force just balances the repulsive force and the net force is zero, or

Net force in bonding between atoms

$$F_N = F_A + F_R = 0 \qquad [1.3]$$

In this state of equilibrium, the atoms are separated by a certain distance r_o, as shown in Figure 1.3. This distance is called the equilibrium separation and is effectively the **bond length.** On the energy diagram, $F_N = 0$ means $dE/dr = 0$, which means that the equilibrium of two atoms corresponds to the potential energy of the

[5] Remember that the change dE in the PE is the work done against the force, $dE = F_N\,dr$.

system acquiring its minimum value. Consequently, the molecule will only be formed if the energy of the two atoms as they approach each other can attain a minimum. This minimum energy also defines the bond energy of the molecule, as depicted in Figure 1.3b. An energy of E_0 is required to separate the two atoms, and this represents the **bond energy.**

Although we considered only two atoms, similar arguments also apply to bonding between many atoms, or between millions of atoms as in a typical solid. Although the actual details of F_A and F_R will change from material to material, the general principle that there is a bonding energy E_0 per atom and an equilibrium interatomic separation r_0 will still be valid. Even in a solid in the presence of many interacting atoms, we can still identify a general potential energy curve $E(r)$ per atom similar to the type shown in Figure 1.3b. We can also use the curve to understand the properties of the solid, such as the thermal expansion coefficient and elastic and bulk moduli.

1.3.2 COVALENTLY BONDED SOLIDS: DIAMOND

Two atoms can form a bond with each other by sharing some or all of their valence electrons and thereby reducing the overall potential energy of the combination. The covalent bond results from the sharing of valence electrons to complete the subshells of each atom. Figure 1.4 shows the formation of a covalent bond between two hydrogen atoms as they come together to form the H_2 molecule. When the 1*s* subshells overlap, the electrons are shared by both atoms and each atom now has a complete subshell. As illustrated in Figure 1.4, electrons 1 and 2 must now orbit both atoms; they therefore cross the overlap region more frequently, indeed twice as often. Thus, electron sharing,

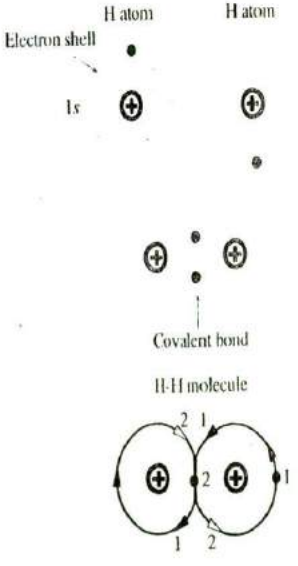


Figure 1.4 Formation of a covalent bond between two H atoms, leading to the H_2 molecule. Electrons spend a majority of their time between the two nuclei, which results in a net attraction between the electrons and the two nuclei, which is the origin of the covalent bond.

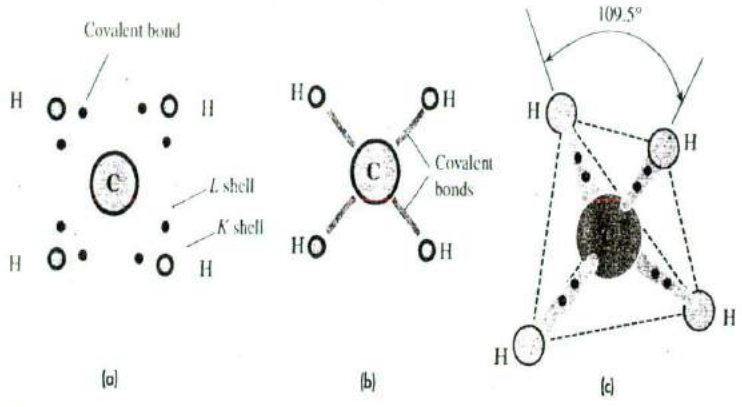


Figure 1.5

(a) Covalent bonding in methane, CH_4, which involves four hydrogen atoms sharing electrons with one carbon atom.

Each covalent bond has two shared electrons. The four bonds are identical and repel each other.

(b) Schematic sketch of CH_4 on paper.

(c) In three dimensions, due to symmetry, the bonds are directed toward the corners of a tetrahedron.

on average, results in a greater concentration of negative charge in the region between the two nuclei, which keeps the two nuclei bonded to each other. Furthermore, by synchronizing their motions, electrons 1 and 2 can avoid crossing the overlap region at the same time. For example, when electron 1 is at the far right (or left), electron 2 is in the overlap region; later, the situation is reversed.

The electronic structure of the carbon atom is $[He]2s^2 2p^2$ with four empty seats in the $2p$ subshell. The $2s$ and $2p$ subshells, however, are quite close. When other atoms are in the vicinity, as a result of interatomic interactions, the two subshells become indistinguishable and we can consider only the shell itself, which is the L shell with a capacity of eight electrons. It is clear that the C atom with four vacancies in the L shell can readily share electrons with four H atoms, as depicted in Figure 1.5, whereby the C atom and each of the H atoms attain complete shells. This is the CH_4 molecule, which is the gas methane. The repulsion between the electrons in one bond and the electrons in a neighboring bond causes the bonds to spread as far out from each other as possible, so that in three dimensions, the H atoms occupy the corners of an imaginary tetrahedron and the CH bonds are at an angle of 109.5° to each other, as sketched in Figure 1.5.

The C atom can also share electrons with other C atoms, as shown in Figure 1.6. Each neighboring C atom can share electrons with other C atoms, leading to a three-dimensional network of a covalently bonded structure. This is the structure of the precious diamond crystal, in which all the carbon atoms are covalently bonded to each other, as depicted in the figure. The **coordination number (CN)** is the number of nearest neighbors for a given atom in the solid. As is apparent in Figure 1.6, the coordination number for a carbon atom in the diamond crystal structure is 4.

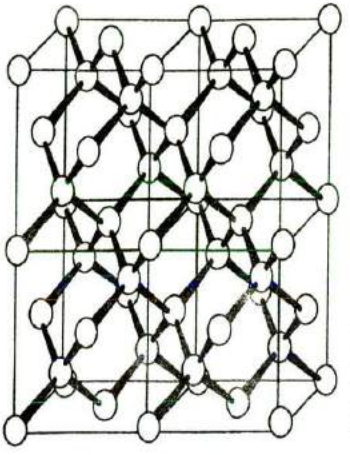


Figure 1.6 The diamond crystal is a covalently bonded network of carbon atoms. Each carbon atom is bonded covalently to four neighbors, forming a regular three-dimensional pattern of atoms that constitutes the diamond crystal.

Due to the strong Coulombic attraction between the shared electrons and the positive nuclei, the covalent bond energy is usually the highest for all bond types, leading to very high melting temperatures and very hard solids: diamond is one of the hardest known materials.

Covalently bonded solids are also insoluble in nearly all solvents. The directional nature and strength of the covalent bond also make these materials nonductile (or nonmalleable). Under a strong force, they exhibit brittle fracture. Further, since all the valence electrons are locked in the bonds between the atoms, these electrons are not free to drift in the crystal when an electric field is applied. Consequently, the electrical conductivity of such materials is very poor.

1.3.3 METALLIC BONDING: COPPER

Metal atoms have only a few valence electrons, which are not very difficult to remove. When many metal atoms are brought together to form a solid, these valence electrons are lost from individual atoms and become collectively shared by all the ions. The valence electrons therefore become **delocalized** and form an **electron gas** or **electron cloud,** permeating the space between the ions, as depicted in Figure 1.7. The attraction between the negative charge of this electron gas and the metal ions more than compensates for the energy initially required to remove the valence electrons from the individual atoms. Thus, the bonding in a metal is essentially due to the attraction between the stationary metal ions and the freely wandering electrons between the ions.

The bond is a **collective sharing** of electrons and is therefore nondirectional. Consequently, the metal ions try to get as close as possible, which leads to **close-packed crystal** structures with high coordination numbers, compared to covalently bonded solids. In the particular example shown in Figure 1.7, Cu^+ ions are packed as closely as possible by the gluing effect of the electrons between the ions, forming a crystal

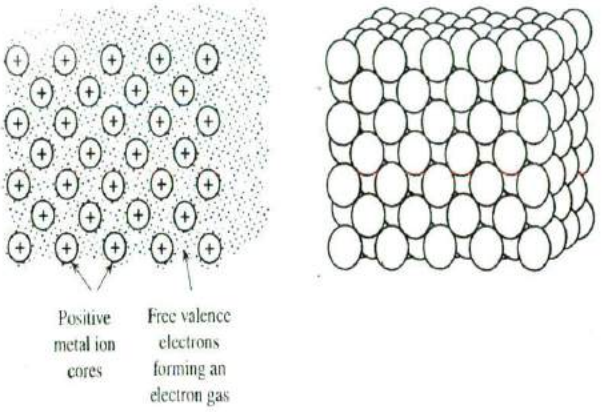


Figure 1.7 In metallic bonding, the valence electrons from the metal atoms form a "cloud of electrons," which fills the space between the metal ions and "glues" the ions together through Coulombic attraction between the electron gas and the positive metal ions.

structure called the **face-centered cubic (FCC)**. The FCC crystal structure, as explained later in Section 1.8, has Cu^+ ions at the corners of a cube and a Cu^+ at the center of each cube-face. (See Figure 1.31.)

The results of this type of bonding are dramatic. First, the nondirectional nature of the bond means that under an applied force, metal ions are able to move with respect to each other, especially in the presence of certain crystal defects (such as dislocations). Thus, metals tend to be ductile. Most importantly, however, the "free" valence electrons in the electron gas can respond readily to an applied electric field and drift along the force of the field, which is the reason for the high electrical conductivity of metals. Furthermore, if there is a temperature gradient along a metal bar, the free electrons can also contribute to the energy transfer from the hot to the cold regions, since they frequently collide with the metal ions and thereby transfer energy. Metals therefore, typically, also have good thermal conductivities; that is, they easily conduct heat. This is why when you touch your finger to a metal it feels cold because it conducts heat "away" from the finger to the ambient (making the fingertip "feel" cold).

1.3.4 IONICALLY BONDED SOLIDS: SALT

Common table salt, NaCl, is a classic example of a solid in which the atoms are held together by ionic bonding. Ionic bonding is frequently found in materials that normally have a metal and a nonmetal as the constituent elements. Sodium (Na) is an alkaline metal with only one valence electron that can easily be removed to form an Na^+ ion with complete subshells. The ion Na^+ looks like the inert element Ne, but with a positive charge. Chlorine has five electrons in its $3p$ subshell and can readily accept

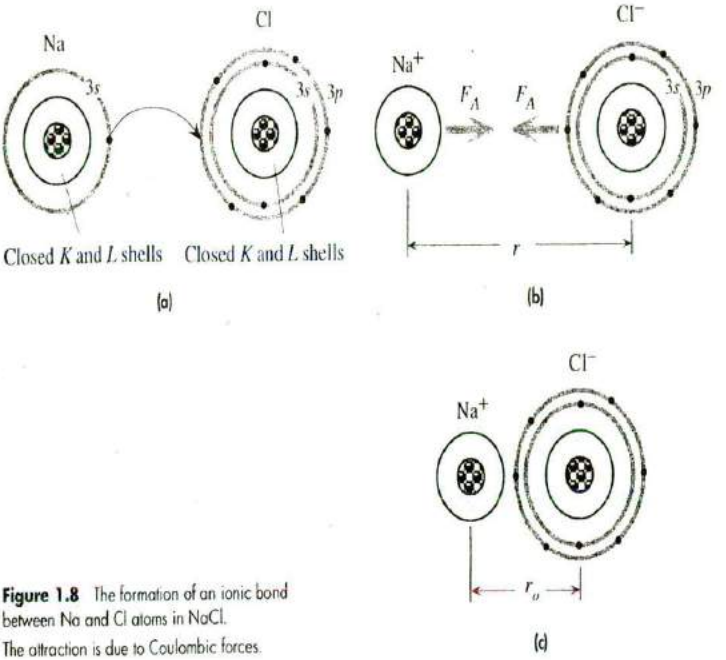


Figure 1.8 The formation of an ionic bond between Na and Cl atoms in NaCl. The attraction is due to Coulombic forces.

one more electron to close this subshell. By taking the electron given up by the Na atom, the Cl atom becomes negatively charged and looks like the inert element Ar with a net negative charge. Transferring the valence electron of Na to Cl thus results in two oppositely charged ions, Na^+ and Cl^-, which are called the **cation** and **anion,** respectively, as shown in Figure 1.8. As a result of the Coulombic force, the two ions pull each other until the attractive force is just balanced by the repulsive force between the closed electron shells. Initially, energy is needed to remove the electron from the Na atom; this is the **energy of ionization.** However, this is more than compensated for by the energy of Coulombic attraction between the two resulting oppositely charged ions, and the net effect is a lowering of the potential energy of the Na^+ and Cl^- ion pair.

When many Na and Cl atoms are ionized and brought together, the resulting collection of ions is held together by the Coulombic attraction between the Na^+ and Cl^- ions. The solid thus consists of Na^+ cations and Cl^- anions holding each other through the Coulombic force, as depicted in Figure 1.9. The Coulombic force around a charge is nondirectional; also, it can be attractive or repulsive, depending on the polarity of the interacting ions. There are also repulsive Coulombic forces between the Na^+ ions themselves and between the Cl^- ions themselves. For the solid to be stable, each Na^+ ion must therefore have Cl^- ions as nearest neighbors and vice versa so that like-ions are not close to each other.

The ions are in equilibrium and the solid is stable when the net potential energy is minimum, or $dE/dr = 0$. Figure 1.10 illustrates the variation of the net potential

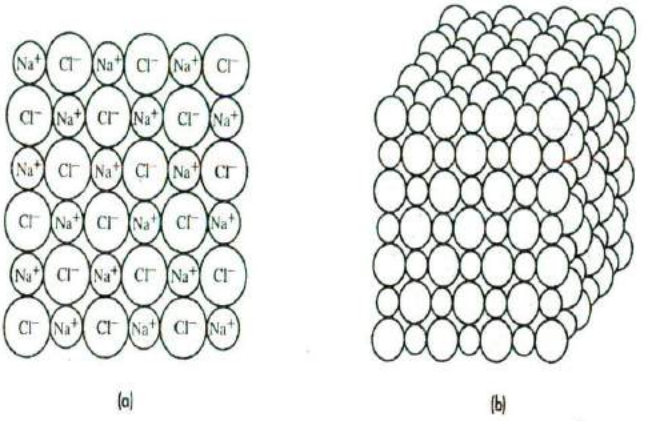


Figure 1.9

(a) A schematic illustration of a cross section from solid NaCl. Solid NaCl is made of Cl^- and Na^+ ions arranged alternatingly, so the oppositely charged ions are closest to each other and attract each other. There are also repulsive forces between the like-ions. In equilibrium, the net force acting on any ion is zero.

(b) Solid NaCl.

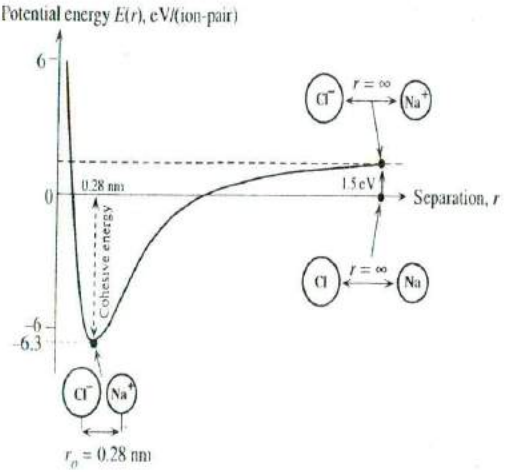


Figure 1.10 Sketch of the potential energy per ion pair in solid NaCl. Zero energy corresponds to neutral Na and Cl atoms infinitely separated.

energy for a pair of ions as the interatomic distance r is reduced from infinity to less than the equilibrium separation, that is, as the ions are brought together from infinity. Zero energy corresponds to separated Na and Cl atoms. Initially, about 1.5 eV is required to transfer the electron from the Na to Cl atom and thereby form Na^+ and Cl^- ions. Then, as the ions come together, the energy is lowered, until it reaches a

minimum at about 6.3 eV below the energy of the separated Na and Cl atoms. When $r = 0.28$ nm, the energy is minimum and the ions are in equilibrium. The bonding energy per ion in solid NaCl is thus 6.3/2 or 3.15 eV, as is apparent in Figure 1.10. The energy required to take solid NaCl apart into individual Na and Cl atoms is the **atomic cohesive energy** of the solid, which is 3.15 eV per atom.

In solid NaCl, the Na^+ and Cl^- ions are thus arranged with each one having oppositely charged ions as its neighbors, to attain a minimum of potential energy. Since there is a size difference between the ions and since we must avoid like-ions getting close to each other, if we want to achieve a stable structure, each ion can have only six oppositely charged ions as nearest neighbors. Figure 1.9b shows the packing of Na^+ and Cl^- ions in the solid. The number of nearest neighbors, that is, the **coordination number,** for both cations and anions in the NaCl crystal is 6.

A number of solids consisting of metal–nonmetal elements follow the NaCl example and have ionic bonding. They are called **ionic crystals** and, by virtue of their ionic bonding characteristics, share many physical properties. For example, LiF, MgO (magnesia), CsCl, and ZnS are all ionic crystals. They are strong, brittle materials with high melting temperatures compared to metals. Most become soluble in polar liquids such as water. Since all the electrons are within the rigidly positioned ions, there are no free or loose electrons to wander around in the crystal as in metals. Therefore, ionic solids are typically electrical insulators. Compared to metals and covalently bonded solids, ionically bonded solids have lower thermal conductivity since ions cannot readily pass vibrational kinetic energy to their neighbors.

EXAMPLE 1.3

IONIC BONDING AND LATTICE ENERGY The potential energy E per Na^+–Cl^- pair within the NaCl crystal depends on the interionic separation r as

Energy per ion pair in an ionic crystal

$$E(r) = -\frac{e^2M}{4\pi\varepsilon_0 r} + \frac{B}{r^m} \qquad [1.4]$$

where the first term is the *attractive* and the second term is the *repulsive* potential energy, and M, B, and m are constants explained in the following. If we were to consider the potential energy PE of one ion pair in isolation from all others, the first term would be a simple Coulombic interaction energy for the Na^+–Cl^- pair, and M would be 1. Within the NaCl crystal, however, a given ion, such as Na^+, interacts not only with its nearest six Cl^- neighbors (Figure 1.9b), but also with its twelve second neighbors (Na^+), eight third neighbors (Cl^-), and so on, so the total or effective PE has a factor M, called the *Madelung constant*, that takes into account all these different Coulombic interactions. M depends only on the geometrical arrangement of ions in the crystal, and hence on the particular crystal structure; for the FCC crystal structure, $M = 1.748$. The Na^+–Cl^- ion pair also have a repulsive PE that is due to the repulsion between the electrons in filled electronic subshells of the ions. If the ions are pushed toward each other, the filled subshells begin to overlap, which results in a strong repulsion. The repulsive PE decays rapidly with distance and can be modeled by a short-range PE of the form B/r^m as in the second term in Equation 1.4 where for Na^+–Cl^-, $m = 8$ and $B = 6.972 \times 10^{-96}$ J m^8. Find the equilibrium separation (r_o) of the ions in the crystal and the ionic bonding energy, defined as $-E(r_o)$. Given the *ionization energy* of Na (the energy to remove an electron) is 5.14 eV and the *electron affinity* of Cl (energy released when an electron is added) is 3.61 eV, calculate the *atomic cohesive energy* of the NaCl crystal as joules per mole.

SOLUTION

Bonding occurs when the potential energy $E(r)$ is a minimum at $r = r_o$, corresponding to the equilibrium separation between the Na^+ and Cl^- ions. We differentiate $E(r)$ and set it to zero at $r = r_o$,

$$\frac{dE(r)}{dr} = \frac{e^2M}{4\pi\varepsilon_o r^2} - \frac{mB}{r^{m+1}} = 0 \qquad \text{at } r = r_o$$

Solving for r_o,

Equilibrium ionic separation

$$r_o = \left[\frac{4\pi\varepsilon_o Bm}{e^2M}\right]^{1/(m-1)} \qquad [1.5]$$

Thus,

$$r_o = \left[\frac{4\pi(8.85 \times 10^{-12}\ \text{F m}^{-1})(6.972 \times 10^{-96}\ \text{J m}^8)(8)}{(1.6 \times 10^{-19}\text{C})^2(1.748)}\right]^{1/(8-1)}$$

$$= 0.281 \times 10^{-9}\text{m} \qquad \text{or} \qquad 0.28\ \text{nm}$$

The minimum energy E_{min} per ion pair is $E(r_o)$ and can be simplified further by substituting for B in terms of r_o:

Minimum PE at bonding

$$E_{min} = -\frac{e^2M}{4\pi\varepsilon_o r_o} + \frac{B}{r_o^m} = -\frac{e^2M}{4\pi\varepsilon_o r_o}\left(1 - \frac{1}{m}\right) \qquad [1.6]$$

Thus,

$$E_{min} = -\frac{(1.6 \times 10^{-19}\text{C})^2(1.748)}{4\pi(8.85 \times 10^{-12}\ \text{F m}^{-1})(2.81 \times 10^{-10}\ \text{m})}\left(1 - \frac{1}{8}\right)$$

$$= -1.256 \times 10^{-18}\ \text{J} \qquad \text{or} \qquad -7.84\ \text{eV}$$

This is the energy with respect to two isolated Na^+ and Cl^- ions. We need 7.84 eV to break up a Na^+–Cl^- pair into isolated Na^+ and Cl^- ions, which represents the *ionic cohesive energy*. Some authors call this ionic cohesive energy simply the **lattice energy**. To take the crystal apart into its neutral atoms, we have to transfer the electron from the Cl^- ion to the Na^+ ion to obtain neutral Na and Cl atoms. It takes 3.61 eV to remove the electron from the Cl^- ion, but 5.14 eV is released when it is put into the Na^+ ion. Thus, we need 7.84 eV + 3.61 eV but get back 5.14 eV.

$$\text{Bond energy per Na–Cl pair} = 7.84\ \text{eV} + 3.61\ \text{eV} - 5.14\ \text{eV} = 6.31\ \text{eV}$$

The *atomic cohesive energy* in terms of joules per mole is

$$E_{\text{cohesive}} = (6.31\ \text{eV})(1.6022 \times 10^{-19}\ \text{J/eV})(6.022 \times 10^{23}\text{mol}^{-1}) = 608\ \text{kJ mol}^{-1}$$

1.3.5 SECONDARY BONDING

Covalent, ionic, and metallic bonds between atoms are known as **primary bonds**. It may be thought that there should be no such bonding between the atoms of the inert elements as they have full shells and therefore cannot accept or lose any electrons, nor share any electrons. However, the fact that a solid phase of argon exists at low temperatures, below −189 °C, means that there must be some bonding mechanism between the Ar atoms. The magnitude of this bond cannot be strong because above −189 °C solid argon melts. Although each water molecule H_2O is neutral overall, these molecules nonetheless attract each other to form the liquid state below 100 °C and the solid state below 0 °C. Between all atoms and molecules, there exists a weak type of attraction, the

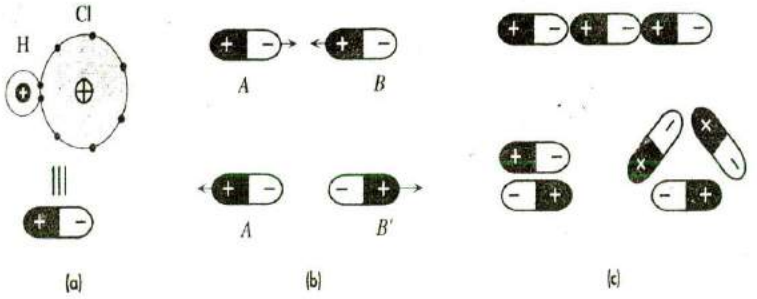


Figure 1.11

(a) A permanently polarized molecule is called an electric dipole moment.

(b) Dipoles can attract or repel each other depending on their relative orientations.

(c) Suitably oriented dipoles attract each other to form van der Waals bonds.

so-called van der Waals–London force, which is due to a net electrostatic attraction between the electron distribution of one atom and the positive nucleus of the other.

In many molecules the concentrations of negative and positive charges do not coincide. As apparent in the HCl molecule in Figure 1.11a, the electrons spend most of their time around the Cl nucleus, so the positive nucleus of the H atom is exposed (H has effectively donated its electron to the Cl atom) and the Cl-region acquires more negative charge than the H-region. An **electric dipole moment** occurs whenever a negative and a positive charge of equal magnitude are separated by a distance as in the H^+–Cl^- molecule in Figure 1.11a. Such molecules are **polar,** and depending on their relative orientations, they can attract or repel each other as depicted in Figure 1.11b. Two dipoles arranged head to tail attract each other because the closest separation between charges on A and B is between the negative charge on A and the positive charge on B, and the *net* result is an electrostatic attraction. The magnitude of the *net force* between two dipoles A and B, however, does not depend on their separation r as $1/r^2$ because there are both attractions and repulsions between the charges on A and charges on B and the net force is only *weakly* attractive. (In fact, the net force depends on $1/r^4$.) If the dipoles are arranged head to head or tail to tail, then, by similar arguments, the two dipoles repel each other. Suitably arranged dipoles can attract each other and form van **der Waals bonds** as illustrated in Figure 1.11c. The energies of such dipole arrangements as in Figure 1.11c are less than that of totally isolated dipoles and therefore encourage "bonding." Such bonds are weaker than primary bonds and are called **secondary bonds.**

The water molecule H_2O is also polar and has a net dipole moment as shown in Figure 1.12a. The attractions between the positive charges on one molecule and the negative charges on a neighboring molecule lead to van der Waals bonding between the H_2O molecules in water as illustrated in Figure 1.12b. When the positive charge of a dipole as in H_2O arises from an exposed H nucleus, van der Waals bonding is referred to as **hydrogen bonding.** In ice, the H_2O molecules, again attracted by van der Waals forces, bond to form a regular pattern and hence a crystal structure.

Van der Waals attraction also occurs between neutral atoms and nonpolar molecules. Consider the bonding between Ne atoms at low temperatures. Each has closed (or full) electron shells. The center of mass of the electrons in the closed shells, when

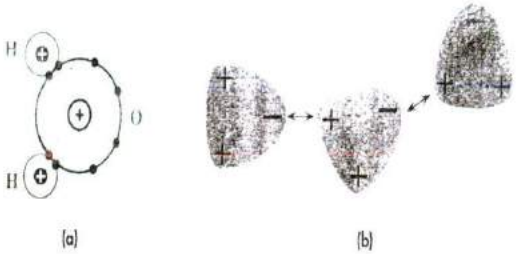


Figure 1.12 The origin of van der Waals bonding between water molecules. (a) The H_2O molecule is polar and has a net permanent dipole moment. (b) Attractions between the various dipole moments in water give rise to van der Waals bonding.

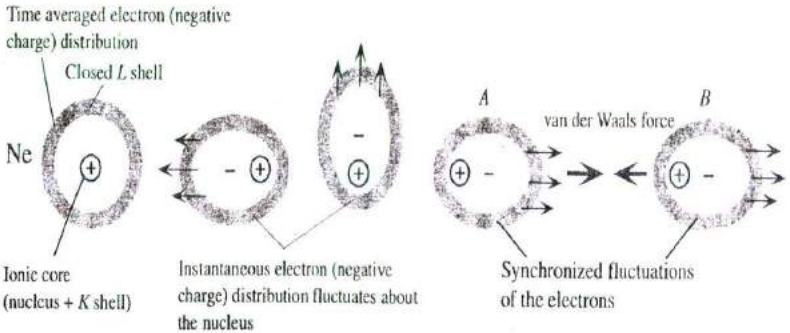


Figure 1.13 Induced-dipole–induced-dipole interaction and the resulting van der Waals force.

averaged over time, coincides with the location of the positive nucleus. At any one instant, however, the center of mass is displaced from the nucleus due to various motions of the individual electrons around the nucleus as depicted in Figure 1.13. In fact, the center of mass of all the electrons fluctuates with time about the nucleus. Consequently, the electron charge distribution is not static around the nucleus but fluctuates asymmetrically, giving rise to an instantaneous dipole moment.

When two Ne atoms, *A* and *B*, approach each other, the rapidly fluctuating negative charge distribution on one affects the motion of the negative charge distribution on the other. A lower energy configuration (*i.e.*, attraction) is produced when the fluctuations are synchronized so that the negative charge distribution on *A* gets closer to the nucleus of the other, *B*, while the negative distribution on *B* at that instant stays away from that on *A* as shown in Figure 1.13. The strongest electrostatic interaction arises from the closest charges which are the displaced electrons in *A* and the nucleus in *B*. This means that there will be a *net* attraction between the two atoms and hence a lowering of the net energy which in turn leads to bonding.

This type of attraction between two atoms is due to induced synchronization of the electronic motions around the nuclei, and we refer to this as *induced-dipole–induced-*

Table 1.2 Comparison of bond types and typical properties (general trends)

Bond Type	Typical Solids	Bond Energy (eV/atom)	Melt. Temp. (°C)	Elastic Modulus (GPa)	Density (g cm^{-3})	Typical Properties
Ionic	NaCl (rock salt)	3.2	801	40	2.17	Generally electrical insulators. May become conductive at high temperatures.
	MgO (magnesia)	10	2852	250	3.58	High elastic modulus. Hard and brittle but cleavable. Thermal conductivity less than metals.
Metallic	Cu	3.1	1083	120	8.96	Electrical conductor.
	Mg	1.1	650	44	1.74	Good thermal conduction. High elastic modulus. Generally ductile. Can be shaped.
Covalent	Si	4	1410	190	2.33	Large elastic modulus. Hard and brittle.
	C (diamond)	7.4	3550	827	3.52	Diamond is the hardest material. Good electrical insulator. Moderate thermal conduction, though diamond has exceptionally high thermal conductivity.
van der Waals: hydrogen bonding	PVC (polymer)		212	4	1.3	Low elastic modulus. Some ductility.
	H_2O (ice)	0.52	0	9.1	0.917	Electrical insulator. Poor thermal conductivity. Large thermal expansion coefficient.
van der Waals: induced dipole	Crystalline argon	0.09	−189	8	1.8	Low elastic modulus. Electrical insulator. Poor thermal conductivity. Large thermal expansion coefficient.

dipole. It is weaker than permanent dipole interactions and at least an order of magnitude less than primary bonding. This is the reason why the inert elements Ne and Ar solidify at temperatures below 25 K (−248 °C) and 84 K (−189 °C). Induced dipole–induced dipole interactions also occur between nonpolar molecules such as H_2, I_2, CH_4, etc. Methane gas (CH_4) can be solidified at very low temperatures. Solids in which constituent molecules (or atoms) have been bonded by van der Waals forces are known as **molecular solids;** ice, solidified CO_2 (dry ice), O_2, H_2, CH_4, and solid inert gases, are typical examples.

Van der Waals bonding is responsible for holding the carbon chains together in polymers. Although the C-to-C bond in a C-chain is due to covalent bonding, the interaction between the C-chains arises from van der Waals forces and the interchain bonding is therefore of secondary nature. These bonds are weak and can be easily stretched or broken. Polymers therefore have substantially lower elastic moduli and melting temperatures than metals and ceramics.

Table 1.2 compares the energies involved in the five types of bonding found in materials. It also lists some important properties of these materials to show the correlation

with the bond type and its energy. The greater is the bond energy, for example, the higher is the melting temperature. Similarly, strong bond energies lead to greater elastic moduli and smaller thermal expansion coefficients. Metals generally have the greatest electrical conductivity since only this type of bonding allows a very large number of free charges (conduction electrons) to wander in the solid and thereby contribute to electrical conduction. Electrical conduction in other types of solid may involve the motion of ions or charged defects from one fixed location to another.

1.3.6 MIXED BONDING

In many solids, the bonding between atoms is generally not just of one type; rather, it is a mixture of bond types. We know that bonding in the silicon crystal is totally covalent, because the shared electrons in the bonds are equally attracted by the neighboring positive ion cores and are therefore equally shared. When there is a covalent-type bond between two different atoms, the electrons become unequally shared, because the two neighboring ion cores are different and hence have different electron-attracting abilities. The bond is no longer purely covalent; it has some ionic character, because the shared electrons spend more time close to one of the ion cores. Covalent bonds that have an ionic character, due to an unequal sharing of electrons, are generally called **polar bonds.** Many technologically important semiconductor materials, such as III–V compounds (*e.g.*, GaAs) have polar covalent bonds. In GaAs, for example, the electrons in a covalent bond spend slightly more time around the As^{+5} ion core than the Ga^{+3} ion core.

Electronegativity is a relative measure of the ability of an atom to attract the electrons in a bond it forms with another atom. The *Pauling scale of electronegativity* assigns an electronegativity value X, a pure number, to various elements, the highest being 4 for F, and the lowest values being for the alkali metal atoms, for which X are less than 1. In this scheme, the difference $X_A - X_B$ in the electronegativities of two atoms A and B is a measure of the polar or ionic character of the bond A–B between A and B. There is obviously no electronegativity difference for a covalent bond. While it is possible to calculate the fractional ionicity of a single bond between two different atoms using $X_A - X_B$, inside the crystal the overall ionic character can be substantially higher because ions can interact with distant ions further away than just the nearest neighbors, as we have found out in NaCl. Many technologically important semiconductor materials, such as III-V compounds (*e.g.*, GaAs) have polar covalent bonds. In GaAs, for example, the bond in the crystal is about 30 percent ionic in character ($X_{As} - X_{Ga} = 2.18 - 1.81 = 0.37$). In the ZnSe crystal, an important II-VI semiconductor, the bond is 63 percent ionic ($X_{Se} - X_{Zn} = 2.55 - 1.65 = 0.85$).[6]

Ceramic materials are compounds that generally contain metallic and nonmetallic elements. They are well known for their brittle mechanical properties, hardness, high

[6] Chemists use "Ionicity $= 1 - \exp[0.24(X_A - X_B)]^2$" to calculate the *ionicity* of the bond between A and B. While this is undoubtedly useful in identifying the trend, it substantially underestimates the actual ionicity of bonding within the crystal itself. (It is left as an exercise to show this fact from the above X_A and X_B values.) The quoted ionicity percentages are from J. C. Phillips' book *Bonds and Bands in Semiconductors*, New York: Academic Press, 1973. By the way, the units of X are sometimes quoted as Pauling units, after its originator Linus Pauling.

melting temperatures, and electrical insulating properties. The type of bonding in a ceramic material may be covalent, ionic, or a mixture of the two, in which the bond between the atoms involves some electron sharing and, to some extent, the partial formation of cations and anions; the shared electrons spend more time with one type of atom, which then becomes a partial anion while the other becomes a partial cation. Silicon nitride (Si_3N_4), magnesia (MgO), and alumina (Al_2O_3) are all ceramics, but they have different types of bonding: Si_3N_4 has covalent, MgO has ionic, and Al_2O_3 has a mixture of ionic and covalent bonding. All three are brittle, have high melting temperatures, and are electrical insulators.

EXAMPLE 1.4

ENERGY OF SECONDARY BONDING Consider the van der Waals bonding in solid argon. The potential energy as a function of interatomic separation can generally be modeled by the **Lennard–Jones 6–12 potential energy curve,** that is,

$$E(r) = -Ar^{-6} + Br^{-12}$$

where A and B are constants. Given that $A = 8.0 \times 10^{-77}$ J m^6 and $B = 1.12 \times 10^{-133}$ J m^{12}, calculate the bond length and bond energy (in eV) for solid argon.

SOLUTION

Bonding occurs when the potential energy is at a minimum. We therefore differentiate the Lennard–Jones potential $E(r)$ and set it to zero at $r = r_o$, the interatomic equilibrium separation or

$$\frac{dE}{dr} = 6Ar^{-7} - 12Br^{-13} = 0 \qquad \text{at } r = r_o$$

that is,

$$r_o^6 = \frac{2B}{A}$$

or

$$r_o = \left[\frac{2B}{A}\right]^{1/6}$$

Substituting $A = 8.0 \times 10^{-77}$ and $B = 1.12 \times 10^{-133}$ and solving for r_o, we find

$$r_o = 3.75 \times 10^{-10} \text{ m} \qquad \text{or} \qquad 0.375 \text{ nm}$$

When $r = r_o = 3.75 \times 10^{-10}$ m, the potential energy is at a minimum and corresponds to $-E_{bond}$, so

$$E_{bond} = \left|-Ar_o^{-6} + Br_o^{-12}\right| = \left|-\frac{8.0 \times 10^{-77}}{(3.75 \times 10^{-10})^6} + \frac{1.12 \times 10^{-133}}{(3.75 \times 10^{-10})^{12}}\right|$$

that is,

$$E_{bond} = 1.43 \times 10^{-20} \text{ J} \qquad \text{or} \qquad 0.089 \text{ eV}$$

Notice how small this energy is compared to primary bonding.

EXAMPLE 1.5

ELASTIC MODULUS The elastic modulus, or Young's modulus Y, of a solid indicates its ability to deform elastically. The greater is the elastic modulus, the more effort is required for the same amount of elastic deformation given a constant sample geometry. When a solid is subjected to tensile forces F acting on two opposite faces, as in Figure 1.14a, it experiences a **stress** σ defined as the *force per unit area* F/A, where A is the area on which F acts. If the original length of the specimen is L_o, then the applied stress σ stretches the solid by an amount δL. The **strain** ε is the fractional increase in the length of the solid $\delta L/L_o$. As long as the applied force displaces the atoms in the solid by a small amount from their equilibrium positions, the deformation is elastic and recoverable when the forces are removed. The applied stress σ and the resulting elastic strain ε are related by the **elastic modulus** Y by

Definition of elastic modulus

$$\sigma = Y\varepsilon \qquad [1.7]$$

The applied stress causes two neighboring atoms along the direction of force to be further separated. Their displacement $\delta r (= r - r_o)$ results in a net attractive force δF_N between two neighboring atoms as indicated in Figure 1.14b (which is the same as Figure 1.3a) where F_N is the net interatomic force. δF_N attempts to restore the separation to equilibrium. This force δF_N, however, is balanced by a portion of the applied force acting on these atoms as in Figure 1.14a. If we were to proportion the area A in Figure 1.14a among all the atoms on this area, each atom would have an area roughly r_o^2. (If there are N atoms on A, $Nr_o^2 = A$.) The force δF_N is therefore σr_o^2. The strain ε is $\delta r / r_o$. Thus, Equation 1.7 gives

$$\frac{\delta F_N}{r_o^2} = \sigma = Y\frac{\delta r}{r_o}$$

Clearly, Y depends on the gradient of the F_N versus r curve at r_o, or the curvature of the minimum of E versus r at r_o.

Elastic modulus and bonding

$$Y = \frac{1}{r_o}\left[\frac{dF_N}{dr}\right]_{r=r_o} = \frac{1}{r_o}\left[\frac{d^2E}{dr^2}\right]_{r=r_o} \qquad [1.8]$$

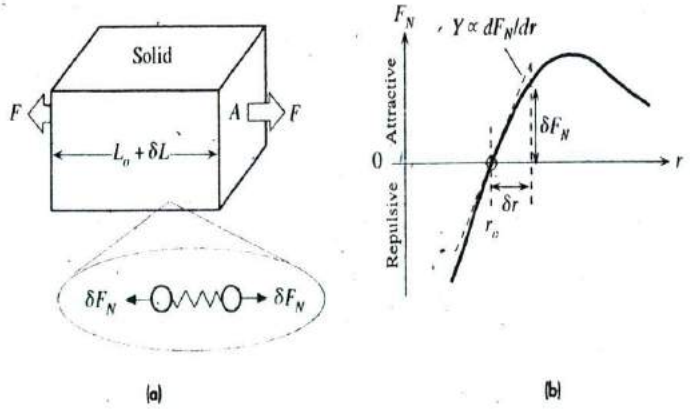


Figure 1.14

(a) Applied forces F stretch the solid elastically from L_o to $L_o + \delta L$. The force is divided among chains of atoms that make the solid. Each chain carries a force δF_N.

(b) In equilibrium, the applied force is balanced by the net force δF_N between the atoms as a result of their increased separation.

The bonding energy E_{bond} is the minimum of E versus r at r_o (Figure 1.3b) and can be related to the curvature of E versus r which leads to[7]

Elastic modulus and bond energy

$$Y \approx f\frac{E_{bond}}{r_o^3} \qquad [1.9]$$

where f is a numerical factor (constant) that depends on the crystal structure and the type of bond (of the order of unity). The well-known Hooke's law for a spring expresses the magnitude of the net force δF_N in terms of the displacement δr by $\delta F_N = \beta|\delta r|$ where β is the spring constant. Thus $Y = \beta/r_o$.

Solids with higher bond energies therefore tend to have higher elastic moduli as apparent in Table 1.2. Secondary bonding has both a smaller E_{bond} and a larger r_o than primary bonding and Y is much smaller. For NaCl, from Figure 1.10, $E_{bond} = 6.3$ eV, $r_o = 0.28$ nm, and Y is of the order of ~50 GPa using Equation 1.9 and $f \approx 1$; and not far out from the value in Table 1.2.

1.4 KINETIC MOLECULAR THEORY

1.4.1 MEAN KINETIC ENERGY AND TEMPERATURE

The kinetic molecular theory of matter is a classical theory that can explain such seemingly diverse topics as the pressure of a gas, the heat capacity of metals, the average speed of electrons in a semiconductor, and electrical noise in resistors, among many interesting phenomena. We start with the kinetic molecular theory of gases, which considers a collection of gas molecules in a container and applies the classical equations of motion from elementary mechanics to these molecules. We assume that the collisions between the gas molecules and the walls of the container result in the gas pressure P. Newton's second law, dp/dt = force, where $p = mv$ is the momentum, is used to relate the pressure P (force per unit area) to the mean square velocity $\overline{v^2}$, and the number of molecules per unit volume N/V. The result can be stated simply as

Kinetic molecular theory for gases

$$PV = \tfrac{1}{3}Nm\overline{v^2} \qquad [1.10]$$

where m is the mass of the gas molecule. Comparing this theoretical derivation with the experimental observation that

$$PV = \left(\frac{N}{N_A}\right)RT$$

where N_A is **Avogadro's number** and R is the gas constant, we can relate the mean kinetic energy of the molecules to the temperature. Our objective is to derive Equation 1.10; to do so, we make the following assumptions:

1. The molecules are in constant random motion. Since we are considering a large number of molecules, perhaps 10^{20} m^{-3}, there are as many molecules traveling in one direction as in any other direction, so the center of mass of the gas is at rest.

[7] The mathematics and a more rigorous description may be found in the textbook's CD.

2. The range of intermolecular forces is short compared to the average separation of the gas molecules. Consequently,
   a. Intermolecular forces are negligible, except during a collision.
   b. The volume of the gas molecules (all together) is negligible compared to the volume occupied by the gas (that is, the container).
3. The duration of a collision is negligible compared to the time spent in free motion between collisions.
4. Each molecule moves with uniform velocity between collisions, and the acceleration due to the gravitational force or other external forces is neglected.
5. On average, the collisions of the molecules with one another and with the walls of the container are perfectly elastic. Collisions between molecules result in exchanges of kinetic energy.
6. Newtonian mechanics can be applied to describe the motion of the molecules.

We consider a collection of N gas molecules within a cubic container of side a. We focus our attention on one of the molecules moving toward one of the walls. The velocity can be decomposed into two components, one directly toward the wall v_x, and the other parallel to the wall v_y, as shown in Figure 1.15. Clearly, the collision of the molecule, which is perfectly elastic, does not change the component v_y along the wall, but reverses the perpendicular component v_x. The change in the momentum of the molecule following its collision with the wall is

$$\Delta p = 2mv_x$$

where m is the mass of the molecule. Following its collision, the molecule travels back across the box, collides with the opposite face B, and returns to hit face A again. The time interval Δt is the time to traverse twice the length of the box, or $\Delta t = 2a/v_x$. Thus, every Δt seconds, the molecule collides with face A and changes its momentum by $2mv_x$. To find the force F exerted by this molecule on face A, we need the rate of

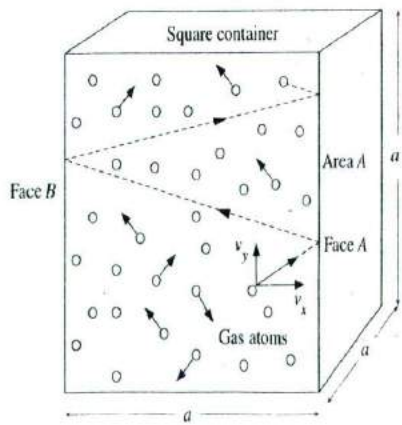


Figure 1.15 The gas molecules in the container are in random motion.

change of momentum, or

$$F = \frac{\Delta p}{\Delta t} = \frac{2mv_x}{(2a/v_x)} = \frac{mv_x^2}{a}$$

The total pressure P exerted by N molecules on face A, of area a^2, is due to the sum of all individual forces F, or

$$P = \frac{\text{Total force}}{a^2} = \frac{mv_{x1}^2 + mv_{x2}^2 + \cdots + mv_{xN}^2}{a^3}$$

$$= \frac{m}{a^3}\left(v_{x1}^2 + v_{x2}^2 + \cdots + v_{xN}^2\right)$$

that is,

$$P = \frac{mN\overline{v_x^2}}{V}$$

where $\overline{v_x^2}$ is the average of v_x^2 for all the molecules and is called the *mean square velocity*, and V is the volume a^3.

Since the molecules are in random motion and collide randomly with each other, thereby exchanging kinetic energy, the mean square velocity in the x direction is the same as those in the y and z directions, or

$$\overline{v_x^2} = \overline{v_y^2} = \overline{v_z^2}$$

For any molecule, the velocity v is given by

$$\overline{v^2} = \overline{v_x^2} + \overline{v_y^2} + \overline{v_z^2} = 3\overline{v_x^2}$$

The relationship between the pressure P and the mean square velocity of the molecules is therefore

Gas pressure in the kinetic theory

$$P = \frac{Nm\overline{v^2}}{3V} = \frac{1}{3}\rho\overline{v^2} \qquad [1.11]$$

where ρ is the density of the gas, or Nm/V. By using elementary mechanical concepts, we have now related the pressure exerted by the gas to the number of molecules per unit volume and to the mean square of the molecular velocity.

Equation 1.11 can be written explicitly to show the dependence of PV on the mean kinetic energy of the molecules. Rearranging Equation 1.11, we obtain

$$PV = \frac{2}{3}N\left(\frac{1}{2}m\overline{v^2}\right)$$

where $\frac{1}{2}m\overline{v^2}$ is the average kinetic energy $\overline{KE}$ per molecule. If we consider one mole of gas, then N is simply N_A, Avogadro's number.

Experiments on gases lead to the empirical gas equation

$$PV = \left(\frac{N}{N_A}\right)RT$$

where R is the universal gas constant. Comparing this equation with the kinetic theory equation shows that the average kinetic energy per molecule must be proportional to

the temperature.

Mean kinetic energy per atom

$$\overline{KE} = \frac{1}{2}m\overline{v^2} = \frac{3}{2}kT \qquad [1.12]$$

where $k = R/N_A$ is called the **Boltzmann constant.** Thus, the mean square velocity is proportional to the absolute temperature. This is a major conclusion from the kinetic theory, and we will use it frequently.

When heat is added to a gas, its internal energy and, by virtue of Equation 1.12, its temperature both increase. The rise in the internal energy per unit temperature is called the **heat capacity.** If we consider 1 mole of gas, then the heat capacity is called the **molar heat capacity** C_m. The total internal energy U of 1 mole of monatomic gas (i.e., a gas with only one atom in each molecule) is

$$U = N_A\left(\frac{1}{2}m\overline{v^2}\right) = \frac{3}{2}N_AkT$$

so, from the definition of C_m, at constant volume, we have

Molar heat capacity at constant volume

$$C_m = \frac{dU}{dT} = \frac{3}{2}N_Ak = \frac{3}{2}R \qquad [1.13]$$

Thus, the heat capacity per mole of a monatomic gas at constant volume is simply $\frac{3}{2}R$. By comparison, we will see later that the heat capacity of metals is twice this amount. The reason for considering constant volume is that the heat added to the system then increases the internal energy without doing mechanical work by expanding the volume.

There is a useful theorem called **Maxwell's principle of equipartition of energy,** which assigns an average of $\frac{1}{2}kT$ to each independent energy term in the expression for the total energy of a system. A monatomic molecule can only have translational kinetic energy, which is the sum of kinetic energies in the x, y, and z directions. The total energy is therefore

$$E = \frac{1}{2}mv_x^2 + \frac{1}{2}mv_y^2 + \frac{1}{2}mv_z^2$$

Each of these terms represents an independent way in which the molecule can be made to absorb energy. Each method by which a system can absorb energy is called a **degree of freedom.** A monatomic molecule has only three degrees of freedom. According to Maxwell's principle, for a collection of molecules in thermal equilibrium, each degree of freedom has an average energy of $\frac{1}{2}kT$, so the average kinetic energy of the monatomic molecule is $3(\frac{1}{2}kT)$.

A rigid diatomic molecule (such as an O_2 molecule) can acquire energy as translational motion and rotational motion, as depicted in Figure 1.16. Assuming the moment of inertia I_x about the molecular axis (along x) is negligible, the energy of the molecule is

$$E = \frac{1}{2}mv_x^2 + \frac{1}{2}mv_y^2 + \frac{1}{2}mv_z^2 + \frac{1}{2}I_y\omega_y^2 + \frac{1}{2}I_z\omega_z^2$$

where I_y and I_z are moments of inertia about the y and z axes and ω_y and ω_z are angular velocities about the y and z axes (Figure 1.16).

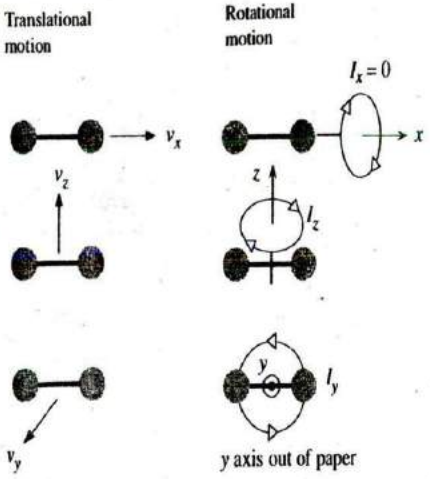


Figure 1.16 Possible translational and rotational motions of a diatomic molecule. Vibrational motions are neglected.

This molecule has five degrees of freedom and hence an average energy of $5(\frac{1}{2}kT)$. Its molar heat capacity is therefore $\frac{5}{2}R$.

The atoms in the molecule will also vibrate by stretching or bending the bond, which behaves like a "spring." At room temperature, the addition of heat only results in the translational and rotational motions becoming more energetic (excited), whereas the molecular vibrations remain the same and therefore do not absorb energy. This occurs because the vibrational energy of the molecule can only change in finite steps; in other words, the vibrational energy is quantized. For many molecules, the energy required to excite a more energetic vibration is much more than the energy possessed by the majority of molecules. Therefore, energy exchanges via molecular collisions cannot readily excite more energetic vibrations; consequently, the contribution of molecular vibrations to the heat capacity is negligible.

In a solid, the atoms are bonded to each other and can only move by vibrating about their equilibrium positions. In the simplest view, a typical atom in a solid is joined to its neighbors by "springs" that represent the bonds, as depicted in Figure 1.17. If we consider a given atom, its potential energy as a function of displacement from the equilibrium position is such that if it is displaced slightly in any direction, it will experience a restoring force proportional to the displacement. Thus, this atom can acquire energy by vibrations in three directions. The energy associated with the x direction, for example, is the kinetic energy of vibration plus the potential energy of the "spring," or $\frac{1}{2}mv_x^2 + \frac{1}{2}K_x x^2$, where v_x is the velocity, x is the extension of the spring, and K_x is the spring constant, all along the x direction. Clearly, there are similar energy terms in the y and z directions, so there are six energy terms in the total energy equation:

$$E = \frac{1}{2}mv_x^2 + \frac{1}{2}mv_y^2 + \frac{1}{2}mv_z^2 + \frac{1}{2}K_x x^2 + \frac{1}{2}K_y y^2 + \frac{1}{2}K_z z^2$$

We know that for simple harmonic motion, the average KE is equal to the average PE. Since, by virtue of the equipartition of energy principle, each average KE term has

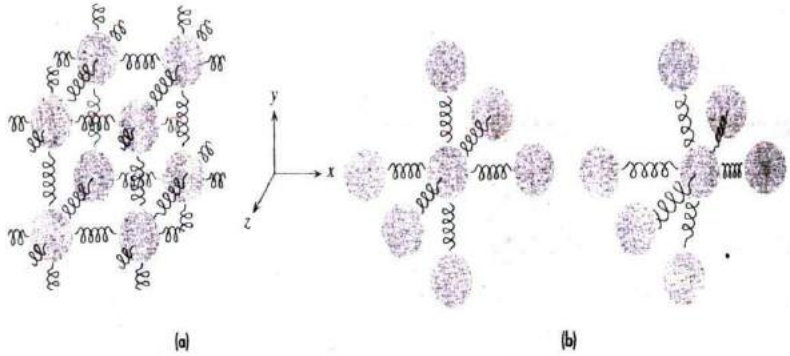


Figure 1.17

(a) The ball-and-spring model of solids, in which the springs represent the interatomic bonds. Each ball (atom) is linked to its neighbors by springs. Atomic vibrations in a solid involve three dimensions.

(b) An atom vibrating about its equilibrium position. The atom stretches and compresses its springs to its neighbors and has both kinetic and potential energy.

an energy of $\frac{1}{2}kT$, the average total energy per atom is $6(\frac{1}{2}kT)$. The internal energy U per mole is

$$U = N_A 6\left(\frac{1}{2}kT\right) = 3RT$$

The molar heat capacity then becomes

Dulong–Petit rule

$$C_m = \frac{dU}{dT} = 3R = 25\ \text{J K}^{-1}\ \text{mol}^{-1}$$

This is the **Dulong–Petit rule.**

The kinetic molecular theory of matter is one of the successes of classical physics, with a beautiful simplicity in its equations and predictions. Its failures, however, are numerous. For example, the theory fails to predict that, at low temperatures, the heat capacity increases as T^3 and that the resistivity of a metal increases linearly with the absolute temperature. We will explain the origins of these phenomena in Chapter 4.

EXAMPLE 1.6 **SPEED OF SOUND IN AIR** Calculate the root mean square (rms) velocity of nitrogen molecules in atmospheric air at 27 °C. Also calculate the root mean square velocity in one direction ($v_{\text{rms},x}$). Compare the speed of propagation of sound waves in air, 350 m s^{-1}, with $v_{\text{rms},x}$ and explain the difference.

SOLUTION

From the kinetic theory

$$\frac{1}{2}m v_{\text{rms}}^2 = \frac{3}{2}kT$$

so that

$$v_{rms} = \sqrt{\frac{3kT}{m}}$$

where m is the mass of the nitrogen molecule N_2. The atomic mass of nitrogen is M_{at} = 14 g mol^{-1}, so that in kilograms

$$m = \frac{2M_{at}(10^{-3})}{N_A}$$

Thus

$$v_{rms} = \left[\frac{3kN_AT}{2M_{at}(10^{-3})}\right]^{1/2} = \left[\frac{3RT}{2M_{at}(10^{-3})}\right]^{1/2}$$

$$= \left[\frac{3(8.314 \text{ J mol}^{-1} \text{ K}^{-1})(300 \text{ K})}{2(14 \times 10^{-3} \text{ kg mol}^{-1})}\right]^{1/2} = 517 \text{ m s}^{-1}$$

Consider an rms velocity in one direction. Then

$$v_{rms,x} = \sqrt{\overline{v_x^2}} = \sqrt{\frac{1}{3}\overline{v^2}} = \frac{1}{\sqrt{3}}v_{rms} = 298 \text{ m s}^{-1}$$

which is slightly less than the velocity of sound in air (350 m s^{-1}). The difference is due to the fact that the propagation of a sound wave involves rapid compressions and rarefactions of air, and the result is that the propagation is not isothermal. Note that accounting for oxygen in air lowers $v_{rms,x}$. (Why?)

SPECIFIC HEAT CAPACITY Estimate the heat capacity of copper per unit gram, given that its atomic mass is 63.6. **EXAMPLE 1.7**

SOLUTION

From the Dulong–Petit rule, $C_m = 3R$ for N_A atoms. But N_A atoms have a mass of M_{at} grams, so the heat capacity per gram, the specific heat capacity c_s, is

$$c_s = \frac{3R}{M_{at}} = \frac{25 \text{ J mol}^{-1} \text{ K}^{-1}}{63.6 \text{ g mol}^{-1}}$$

$$\approx 0.39 \text{ J g}^{-1} \text{ K}^{-1} \qquad \text{(The experimental value is 0.38 J g}^{-1} \text{ K}^{-1}\text{.)}$$

1.4.2 THERMAL EXPANSION

Nearly all materials expand as the temperature increases. This phenomenon is due to the asymmetric nature of the interatomic forces and the increase in the amplitude of atomic vibrations with temperature as expected from the kinetic molecular theory.

The potential energy curve $U(r)$ for two atoms separated by a distance r is shown in Figure 1.18. In equilibrium the PE is a minimum at $U_{min} = -U_0$ and the bonding energy is simply U_0. The atoms are separated by the equilibrium separation r_0. However,

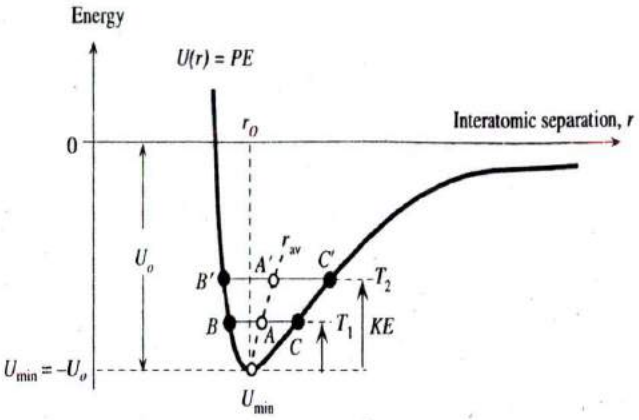


Figure 1.18 The potential energy *PE* curve has a minimum when the atoms in the solid attain the interatomic separation at $r = r_o$.

Because of thermal energy, the atoms will be vibrating and will have vibrational kinetic energy. At $T = T_1$, the atoms will be vibrating in such a way that the bond will be stretched and compressed by an amount corresponding to the *KE* of the atoms. A pair of atoms will be vibrating between *B* and *C*. Their average separation will be at *A* and greater than r_o.

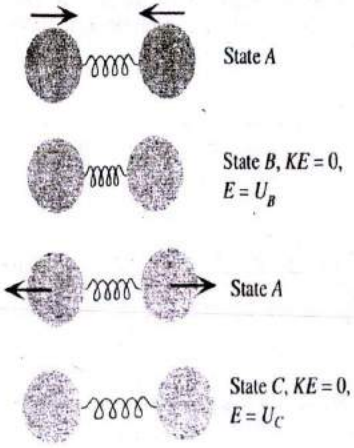


Figure 1.19 Vibrations of atoms in the solid. We consider for simplicity a pair of atoms. Total energy is $E = PE + KE$, and this is constant for a pair of vibrating atoms executing simple harmonic motion. At *B* and *C*, *KE* is zero (atoms are stationary and about to reverse direction of oscillation) and *PE* is maximum.

according to the kinetic molecular theory, atoms are vibrating about their equilibrium positions with a mean vibrational kinetic energy that increases with the temperature as $\frac{3}{2}kT$. At any instant the total energy E of the pair of atoms is $U + KE$, and this is constant inasmuch as no external forces are being applied. The atoms will be vibrating about their equilibrium positions, stretching and compressing the bond, as depicted in Figure 1.19. At positions *B* and *C*, U is maximum and the *KE* is zero; the atoms are stationary and about to reverse their direction of oscillation. Thus at *B* and *C* the total

energy $E = U_B = U_C$ and the *PE* has increased from its minimum value U_{min} by an amount equal to *KE*. The line *BC* corresponds to the total energy *E*. The atoms are confined to vibrate between *B* and *C*, executing simple harmonic motion and hence maintaining $E = U + KE = \text{constant}$.

But the *PE* curve $U(r)$ is *asymmetric*. $U(r)$ is broader in the $r > r_o$ region. Thus, the atoms spend more time in the $r > r_o$ region, that is, more time stretching the bond than compressing the bond (with respect to the equilibrium length r_o). The average separation corresponds to point *A*,

$$r_{av} = \tfrac{1}{2}(r_B + r_C)$$

which is clearly greater than r_o. As the temperature increases, *KE* increases, the total energy *E* increases, and the atoms vibrate between wider extremes of the $U(r)$ curve, between B' and C'. The new average separation at A' is now greater than that at A: $r_{A'} > r_A$. Thus as the temperature increases, the average separation between the atoms also increases, which leads to the phenomenon of **thermal expansion.** If the *PE* curve were symmetric, then there would be no thermal expansion as the atoms would spend equal times in the $r < r_o$ and $r > r_o$ regions.

When the temperature increases by a small amount δT, the energy per atom increases by $C_{atom}\,\delta T$ where C_{atom} is the heat capacity per atom (molar heat capacity divided by N_A). If $C_{atom}\,\delta T$ is large, then the line $B'C'$ in Figure 1.18 will be higher up on the energy curve and the average separation A' will therefore be larger. Thus, the increase δr_{av} in the average separation is proportional to δT. If the total length L_o is made up of *N* atoms, $L_o = N r_{av}$, then the change δL in L_o is proportional to $N\,\delta T$ or $L_o\,\delta T$. The proportionality constant is the **thermal coefficient of linear expansion,** or simply, **thermal expansion coefficient** λ, which is defined as the fractional change in length per unit temperature,

Definition of thermal expansion coefficient

$$\lambda = \frac{1}{L_o} \cdot \frac{\delta L}{\delta T} \qquad [1.14]$$

If L_o is the original length at temperature T_o, then the length *L* at temperature *T*, from Equation 1.14, is

Thermal expansion

$$L = L_o[1 + \lambda(T - T_o)] \qquad [1.15]$$

We note that λ is a material property that depends on the nature of the bond. The variation of r_{av} with *T* in Figure 1.18 depends on the shape of the *PE* curve $U(r)$. Typically, λ is larger for metallic bonding than for covalent bonding.

We can use a mathematical procedure (known as a Taylor expansion) to describe the $U(r)$ versus *r* curve in terms of its minimum value U_{min}, plus correction terms that depend on the powers of the *displacement* $(r - r_o)$ from r_o,

Potential energy of an atom

$$U(r) = U_{min} + a_2(r - r_o)^2 + a_3(r - r_o)^3 + \cdots \qquad [1.16]$$

where a_2 and a_3 are coefficients that are related to the second and third derivatives of *U* at r_o. The term $a_1(r - r_o)$ is missing because $dU/dr = 0$ at $r = r_o$ where $U = U_{min}$. The U_{min} and $a_2(r - r_o)^2$ terms in Equation 1.16 give a parabola about U_{min} which is a symmetric curve around r_o and therefore does not lead to thermal expansion. The average

location at any energy on a symmetric curve at r_o is always at r_o. It is the a_3 term that gives the expansion because it leads to asymmetry. Thus, λ depends on the amount of asymmetry, that is, a_3/a_2. The asymmetric *PE* curve in Figure 1.18 which has a finite cubic a_3 term as in Equation 1.16 does not lead to a perfect simple harmonic (sinusoidal) vibration about r_o because the restoring force is not proportional to the displacement alone. Such oscillations are **unharmonic,** and the *PE* curve is said to possess an **unharmonicity** (terms such as a_3). Thermal expansion is an **unharmonic effect.**

The thermal expansion coefficient normally depends on the temperature, $\lambda = \lambda(T)$, and typically increases with increasing temperature, except at the lowest temperatures. We can always expand $\lambda(T)$ about some useful temperature such as T_o to obtain a polynomial series in temperature terms up to the most significant term, usually the T^2 containing term. Thus, Equation 1.14 becomes

Thermal expansion coefficient and temperature

$$\frac{dL}{L_o\,dT} = \lambda(T) = A + B(T - T_o) + C(T - T_o)^2 + \cdots \qquad [1.17]$$

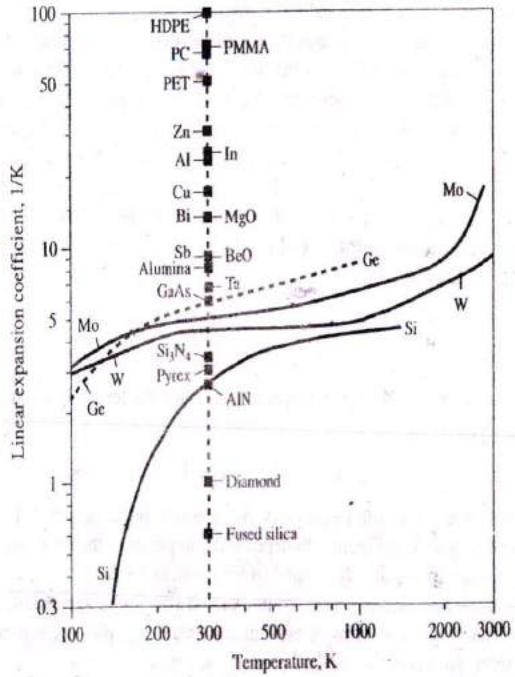


Figure 1.20 Dependence of the linear thermal expansion coefficient λ [K^{-1}] on temperature T [K] on a log-log plot.

HDPE, high-density polyethylene; PMMA, polymethylmethacrylate (acrylic); PC, polycarbonate; PET, polyethylene terephthalate (polyester); fused silica, SiO_2; alumina, Al_2O_3.

SOURCE: Data extracted from various sources including G. A. Slack and S. F. Bartram, *J. Appl. Phys.*, **46**, 89, 1975.

where A, B, and C are temperature-independent constants, and the expansion is about T_o. To find the total fractional change in the length $\Delta L/L_o$ from T_o to T, we have to integrate $\lambda(T)$ with respect to temperature from T_o to T. We can still employ Equation 1.15 provided that we use a properly defined mean value for the expansion coefficient from T_o to T,

$$L = L_o[1 + \bar{\lambda}(T - T_o)] \qquad [1.18]$$

Thermal expansion

where

$$\bar{\lambda} = \frac{1}{(T - T_o)} \int_{T_o}^{T} \lambda(T)\, dT \qquad [1.19]$$

Mean thermal expansion coefficient

Figure 1.20 shows the temperature dependence of λ for various materials. In very general terms, except at very low (typically below 100 K) and very high temperatures (near the melting temperature), for most metals λ does not depend strongly on the temperature; many engineers take λ for a metal to be approximately temperature independent. There is a simple relationship between the linear expansion coefficient and the heat capacity of a material, which is discussed in Chapter 4.

EXAMPLE 1.8

VOLUME EXPANSION COEFFICIENT Suppose that the volume of a solid body at temperature T_o is V_o. The volume expansion coefficient α_V of a solid body characterizes the change in its volume from V_o to V due to a temperature change from T_o to T by

$$V = V_o[1 + \alpha_V(T - T_o)] \qquad [1.20]$$

Volume expansion

Show that α_V is given by

$$\alpha_V = 3\lambda \qquad [1.21]$$

Volume expansion coefficient

Aluminum has a density of 2.70 g cm^{-3} at 25 °C. Its thermal expansion coefficient is 24×10^{-6} °C^{-1}. Calculate the density of Al at 350 °C.

SOLUTION

Consider the solid body in the form of a rectangular parallelepiped with sides x_o, y_o, and z_o. Then at T_o,

$$V_o = x_o y_o z_o$$

and at T,

$$V = [x_o(1 + \lambda\,\Delta T)][y_o(1 + \lambda\,\Delta T)][z_o(1 + \lambda\,\Delta T)]$$
$$= x_o y_o z_o(1 + \lambda\,\Delta T)^3$$

that is

$$V = x_o y_o z_o[1 + 3\lambda\,\Delta T + 3\lambda^2(\Delta T)^2 + \lambda^3(\Delta T)^3]$$

We can now substitute for V from Equation 1.20, use $V_o = x_o y_o z_o$, and neglect the $\lambda^2(\Delta T)^2$ and $\lambda^3(\Delta T)^3$ terms compared with the $\lambda\,\Delta T$ term ($\lambda \ll 1$) to obtain,

$$V = V_o[1 + 3\lambda(T - T_o)] = V_o[1 + \alpha_V(T - T_o)]$$

Since density ρ is mass/volume, volume expansion leads to a density reduction. Thus,

$$\rho = \frac{\rho_o}{1 + \alpha_V(T - T_o)} \approx \rho_o[1 - \alpha_V(T - T_o)]$$

For Al, the density at 350 °C is

$$\rho = 2.70[1 - 3(24 \times 10^{-6})(350 - 25)] = 2.637 \text{ g cm}^{-3}$$

EXAMPLE 1.9

Thermal expansion coefficient of Si

EXPANSION OF Si The expansion coefficient of silicon over the temperature range 120–1500 K is given by Okada and Tokumaru (1984) as

$$\lambda = 3.725 \times 10^{-6}[1 - e^{-5.88 \times 10^{-3}(T - 124)}] + 5.548 \times 10^{-10}T \qquad [1.22]$$

where λ is in K^{-1} (or $°C^{-1}$) and T is in kelvins. At a room temperature of 20 °C, the above gives $\lambda = 2.51 \times 10^{-6}\ K^{-1}$. Calculate the fractional change $\Delta L/L_o$ in the length L_o of the Si crystal from 20 to 320 °C, by (*a*) assuming a constant λ equal to the room temperature value and (*b*) assuming the above temperature dependence. Calculate the mean $\bar{\lambda}$ for this temperature range.

SOLUTION

Assuming a constant we have

$$\frac{\Delta L}{L_o} = \lambda(T - T_0) = (2.51 \times 10^{-6}\ °C^{-1})(320 - 20) = 0.753 \times 10^{-3} \quad \text{or} \quad 0.075\%$$

With a temperature-dependent $\lambda(T)$,

$$\frac{\Delta L}{L_o} = \int_{T_o}^{T} \lambda(T)\, dT$$

$$= \int_{20+273}^{320+273} \{3.725 \times 10^{-6}[1 - e^{-5.88 \times 10^{-3}(T - 124)}] + 5.548 \times 10^{-10}T\}\, dT$$

The integration can either be done numerically or analytically (both left as an exercise) with the result that

$$\frac{\Delta L}{L_o} = 1.00 \times 10^{-3} \quad \text{or} \quad 0.1\%$$

which is substantially more than when using a constant λ. The mean $\bar{\lambda}$ over this temperature range can be found from

$$\frac{\Delta L}{L_o} = \bar{\lambda}(T - T_o) \quad \text{or} \quad 1.00 \times 10^{-3} = \bar{\lambda}(320 - 20)$$

which gives $\bar{\lambda} = 3.33 \times 10^{-6}\ °C^{-1}$. A 0.1 percent change in length means that a 1 mm chip would expand by 1 micron.

1.5 MOLECULAR VELOCITY AND ENERGY DISTRIBUTION

Although the kinetic theory allows us to determine the root mean square velocity of the gas molecules, it says nothing about the distribution of velocities. Due to random collisions between the molecules and the walls of the container and between the molecules themselves, the molecules do not all have the same velocity. The velocity distribution of molecules can be determined experimentally by the simple scheme illustrated in Figure 1.21. Gas molecules are allowed to escape from a small aperture of a hot oven in which the substance is vaporized. Two blocking slits allow only those molecules that are moving along the line through the two slits to pass through, which results in a **collimated beam.** This beam is directed toward two rotating disks, which have slightly displaced

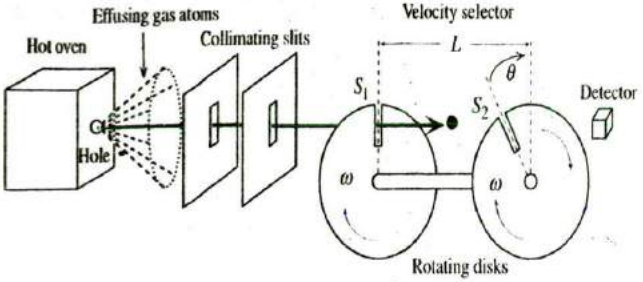


Figure 1.21 Schematic diagram of a Stern-type experiment for determining the distribution of molecular speeds.

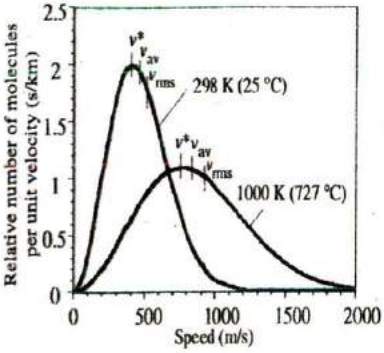


Figure 1.22 Maxwell–Boltzmann distribution of molecular speeds in nitrogen gas at two temperatures. The ordinate is $dN/(N\,dv)$, the fractional number of molecules per unit speed interval in $(km/s)^{-1}$.

slits. The molecules that pass through the first slit can only pass through the second if they have a certain speed; that is, the exact speed at which the second slit lines up with the first slit. Thus, the two disks act as a speed selector. The speed of rotation of the disks determines which molecular speeds are allowed to go through. The experiment therefore measures the number of molecules ΔN with speeds in the range v to $(v + \Delta v)$.

It is generally convenient to describe the number of molecules dN with speeds in a certain range v to $(v + dv)$ by defining a **velocity density function** n_v as follows:

$$dN = n_v\, dv$$

where n_v is the number of molecules per unit velocity that have velocities in the range v to $(v + dv)$. This number represents the velocity distribution among the molecules and is a function of the molecular velocity $n_v = n_v(v)$. From the experiment, we can easily obtain n_v by $n_v = \Delta N/\Delta v$ at various velocities. Figure 1.22 shows the velocity density function n_v of nitrogen gas at two temperatures. The average (v_{av}), most probable (v^*), and rms (v_{rms}) speeds are marked to show their relative positions. As expected, these speeds all increase with increasing temperature. From various experiments of the type shown in Figure 1.21, the velocity distribution function n_v has been widely studied and found to obey the following equation:

Maxwell–Boltzmann distribution for molecular speeds

$$n_v = 4\pi N \left(\frac{m}{2\pi kT}\right)^{3/2} v^2 \exp\left(-\frac{mv^2}{2kT}\right) \qquad [1.23]$$

where N is the total number of molecules and m is the molecular mass. This is the **Maxwell–Boltzmann distribution function,** which describes the statistics of particle velocities in thermal equilibrium. The function assumes that the particles do not interact with each other while in motion and that all the collisions are elastic in the sense that they involve an exchange of kinetic energy. Figure 1.22 clearly shows that molecules move around randomly, with a variety of velocities ranging from nearly zero to almost infinity. The kinetic theory speaks of their rms value only.

What is the energy distribution of molecules in a gas? In the case of a monatomic gas, the total energy E is purely translational kinetic energy, so we can use $E = \frac{1}{2}mv^2$. To relate an energy range dE to a velocity range dv, we have $dE = mv\,dv$. Suppose that n_E is the number of atoms per unit volume per unit energy at an energy E. Then $n_E\,dE$ is the number of atoms with energies in the range E to $(E + dE)$. These are also the atoms with velocities in the range v to $(v + dv)$, because an atom with a velocity v has an energy E. Thus,

$$n_E dE = n_v\,dv$$

i.e.,

$$n_E = n_v\left(\frac{dv}{dE}\right)$$

Maxwell–Boltzmann distribution for translational kinetic energies

If we substitute for n_v and (dv/dE), we obtain the expression for n_E as a function of E:

$$n_E = \frac{2}{\pi^{1/2}}N\left(\frac{1}{kT}\right)^{3/2}E^{1/2}\exp\left(-\frac{E}{kT}\right) \qquad [1.24]$$

Thus, the total internal energy is distributed among the atoms according to the Maxwell–Boltzmann distribution in Equation 1.24. The exponential factor $\exp(-E/kT)$ is called the **Boltzmann factor.** Atoms have widely differing kinetic energies, but a mean energy of $\frac{3}{2}kT$. Figure 1.23 shows the Maxwell–Boltzmann energy distribution among the gas atoms in a tank at two temperatures. As the temperature increases, the distribution extends to higher energies. The area under the curve is the total number of molecules, which remains the same for a closed container.

Equation 1.24 represents the energy distribution among the N gas atoms at any time. Since the atoms are continually colliding and exchanging energies, the energy of one

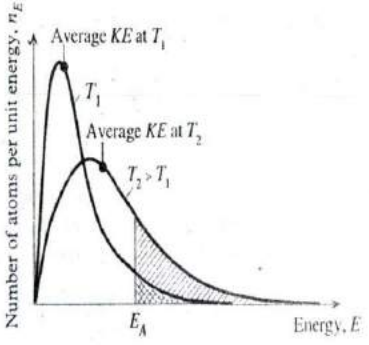


Figure 1.23 Energy distribution of gas molecules at two different temperatures. The shaded area shows the number of molecules that have energies greater than E_A. This area depends strongly on the temperature as $\exp(-E_A/kT)$.

atom will sometimes be small and sometimes be large, but averaged over a long time, this energy will be $\frac{3}{2}kT$ as long as all the gas atoms are in thermal equilibrium (*i.e.*, the temperature is the same everywhere in the gas). Thus, we can also use Equation 1.24 to represent all possible energies an atom can acquire over a long period. There are a total of N atoms, and $n_E\,dE$ of them have energies in the range E to $(E + dE)$. Thus,

$$\text{Probability of energy being in } E \text{ to } (E + dE) = \frac{n_E\,dE}{N} \qquad [1.25]$$

When the probability in Equation 1.25 is integrated (*i.e.*, summed) for all energies ($E = 0$ to ∞), the result is unity, because the atom must have an energy somewhere in the range of zero to infinity.

What happens to the Maxwell–Boltzmann energy distribution law in Equation 1.24 when the total energy is not simply translational kinetic energy? What happens when we do not have a monatomic gas? Suppose that the total energy of a molecule (which may simply be an atom) in a system of N molecules has vibrational and rotational kinetic energy contributions, as well as potential energy due to intermolecular interactions. In all cases, the number of molecules per unit energy n_E turns out to contain the Boltzmann factor, and the energy distribution obeys what is called the **Boltzmann energy distribution:**

Boltzmann energy distribution

$$\frac{n_E}{N} = C\exp\left(-\frac{E}{kT}\right) \qquad [1.26]$$

where E is the total energy ($KE + PE$), N is the total number of molecules in the system, and C is a constant that relates to the specific system (*e.g.*, a monatomic gas or a liquid). The constant C may depend on the energy E, as in Equation 1.24, but not as strongly as the exponential term. Equation 1.26 is the **probability per unit energy** that a molecule in a given system has an energy E. Put differently, $(n_E\,dE)/N$ is the fraction of molecules in a small energy range E to $E + dE$.

EXAMPLE 1.10

MEAN AND RMS SPEEDS OF MOLECULES Given the Maxwell–Boltzmann distribution law for the velocities of molecules in a gas, derive expressions for the mean speed (v_{av}), most probable speed (v^*), and rms velocity (v_{rms}) of the molecules and calculate the corresponding values for a gas of noninteracting electrons.

SOLUTION

The number of molecules with speeds in the range v to $(v + dv)$ is

$$dN = n_v\,dv = 4\pi N\left(\frac{m}{2\pi kT}\right)^{3/2} v^2\exp\left(-\frac{mv^2}{2kT}\right)dv$$

By definition, then, the mean speed is given by

$$v_{av} = \frac{\int v\,dN}{\int dN} = \frac{\int v n_v\,dv}{\int n_v\,dv} = \sqrt{\frac{8kT}{\pi m}}$$

Mean speed

where the integration is over all speeds ($v = 0$ to ∞). The mean square velocity is given by

$$\overline{v^2} = \frac{\int v^2\,dN}{\int dN} = \frac{\int v^2 n_v\,dv}{\int n_v\,dv} = \frac{3kT}{m}$$

Root mean square velocity

so the rms velocity is

$$v_{rms} = \sqrt{\frac{3kT}{m}}$$

where n_v/N is the probability per unit speed that a molecule has a speed in the range v to $(v + dv)$. Differentiating n_v with respect to v and setting this to zero, $dn_v/dv = 0$, gives the position of the peak of n_v versus v, and thus the most probable speed v^*,

Most probable speed

$$v^* = \left[\frac{2kT}{m}\right]^{1/2}$$

Substituting $m = 9.1 \times 10^{-31}$ kg for electrons and using $T = 300$ K, we find $v^* = 95.3$ km s^{-1}, $v_{av} = 108$ km s^{-1}, and $v_{rms} = 117$ km s^{-1}, all of which are close in value. We often use the term **thermal velocity** to describe the mean speed of particles due to their thermal random motion. Also, the integrations shown are not trivial and they involve substitution and integration by parts.

1.6 HEAT, THERMAL FLUCTUATIONS, AND NOISE

Generally, thermal equilibrium between two objects implies that they have the same temperature, where temperature (from the kinetic theory) is a measure of the mean kinetic energy of the molecules. Consider a solid in a monatomic gas atmosphere such as He gas, as depicted in Figure 1.24. Both the gas and the solid are at the same temperature. The gas molecules move around randomly, with a mean kinetic energy given by $\frac{1}{2}m\overline{v^2} = \frac{3}{2}kT$, where m is the mass of the gas molecule. We also know that the atoms in the solid vibrate with a mean kinetic energy given by $\frac{1}{2}M\overline{V^2} = \frac{3}{2}kT$, where M is the mass of the solid atom and V is the velocity of vibration. The gas molecules will collide with the atoms on the surface of the solid and will thus exchange energy with those solid atoms. Since both are at the same temperature, the solid atoms and gas molecules have the same mean kinetic energy, which means that over a long time, there will be no net transfer of energy from one to the other. This is basically what we mean by **thermal equilibrium.**

If, on the other hand, the solid is hotter than the gas, $T_{solid} > T_{gas}$, and thus $\frac{1}{2}M\overline{V^2} > \frac{1}{2}m\overline{v^2}$, then when an average gas molecule and an average solid atom collide,

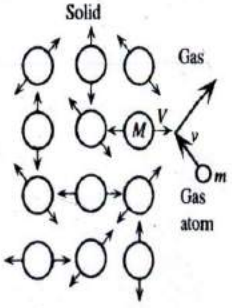


Figure 1.24 Solid in equilibrium in air. During collisions between the gas and solid atoms, kinetic energy is exchanged.

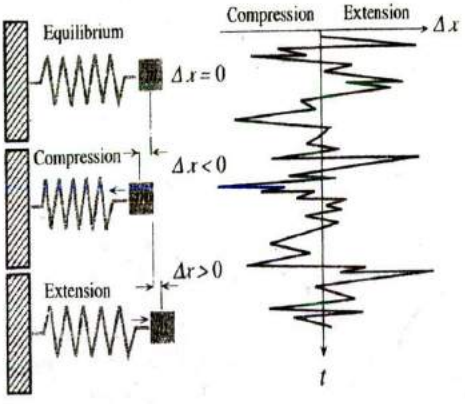


Figure 1.25 Fluctuations of a mass attached to a spring, due to random bombardment by air molecules.

energy will be transferred from the solid atom to the gas molecule. As many more gas molecules collide with solid atoms, more and more energy will be transferred, until the mean kinetic energy of atoms in each substance is the same and they reach the same temperature: the bodies have **equilibrated.** The amount of energy transferred from the kinetic energy of the atoms in the hot solid to the kinetic energy of the gas molecules is called **heat.** Heat represents the energy transfer from the hot body to the cold body by virtue of the *random* motions and collisions of the atoms and molecules.

Although, over a long time, the energy transferred between two systems in thermal equilibrium is certainly zero, this does not preclude a net energy transfer from one to the other at one instant. For example, at any one instant, an average solid atom may be hit by a fast gas molecule with a speed at the far end of the Maxwell–Boltzmann distribution. There will then be a transfer of energy from the gas molecule to the solid atom. At another instant, a slow gas molecule hits the solid, and the reverse is true. Thus, although the mean energy transferred from one atom to the other is zero, the instantaneous value of this energy is not zero and varies randomly about zero.

As an example, consider a small mass attached to a spring, as illustrated in Figure 1.25. The gas or air molecules will bombard and exchange energy with the solid atoms. Some air molecules will be fast and some will be slow, which means that there will be an instantaneous exchange of energy. Consequently, the spring will be compressed when the bombarding air molecules are fast (more energetic) and extended when they are less energetic. This leads to a mechanical fluctuation of the mass about its equilibrium position, as depicted in Figure 1.25. These fluctuations make the measurement of the exact position of the mass uncertain, and it is futile to try to measure the position more accurately than these fluctuations permit.

If the mass m compresses the spring by Δx, then at time t, the energy stored as potential energy in the spring is

$$PE(t) = \frac{1}{2}K(\Delta x)^2 \qquad [1.27]$$

where K is the spring constant. At a later instant, this energy will be returned to the gas by the spring. The spring will continue to fluctuate because of the fluctuations in the velocity of the bombarding air molecules. Over a long period, the average value of PE will be the same as KE and, by virtue of the Maxwell equipartition of energy theorem, it will be given by

$$\overline{\frac{1}{2}K(\Delta x)^2} = \frac{1}{2}kT \qquad [1.28]$$

Root mean square fluctuations of a body attached to a spring of stiffness K

Thus, the rms value of the fluctuations of the mass about its equilibrium position is

$$(\Delta x)_{\text{rms}} = \sqrt{\frac{\overline{kT}}{K}} \qquad [1.29]$$

To understand the origin of electrical noise, for example, we consider the thermal fluctuations in the instantaneous local electron concentration in a conductor, such as that shown in Figure 1.26. Because of fluctuations in the electron concentration at any one instant, end A of the conductor can become more negative with respect to end B, which will give rise to a voltage across the conductor. This fluctuation in the electron concentration is due to more electrons at that instant moving toward end A than toward B. At a later instant, the situation reverses and more electrons move toward B than toward A, resulting in end B becoming more negative and leading to a reversal of the voltage between A and B. Clearly, there will therefore be voltage fluctuations across the conductor, even though the mean voltage across it over a long period is always zero. If the conductor is connected to an amplifier, these voltage fluctuations will be amplified and recorded as noise at the output. This noise corrupts the actual signal at the amplifier input and is obviously undesirable. As engineers, we have to know how to calculate the magnitude of this noise. Although the mean voltage due to thermal fluctuations is zero, the rms value is not. The average voltage from a power outlet is zero, but the rms value is 120 V. We use the rms value to calculate the amount of average power available.

Consider a conductor of resistance R. To derive the noise voltage generated by R we place a capacitor C across this conductor, as in Figure 1.27, and we assume that both are at the same temperature; they are in thermal equilibrium. The capacitor is placed as a *convenient device* to obtain or derive the noise voltage generated by R. It should be emphasized that C itself does not contribute to the source of the fluctuations (it generates no noise) but is inserted into the circuit to impose a finite *bandwidth* over which we will calculate the noise voltage. The reason is that all practical electric circuits have some kind of bandwidth, and the noise voltage we will derive depends on this bandwidth. Even if we remove the capacitor, there will still be stray capacitances; and if we short the conductor, the shorting wires will have some inductance that will also impose a bandwidth. As we mentioned previously, thermal fluctuations in the conductor give rise to voltage fluctuations across R. There is only so much average energy available in these thermal fluctuations, and this is the energy that is used to charge and discharge the external capacitor C. The voltage across the capacitor depends on how much energy that can be stored on it, which in turn depends on the thermal fluctuations in the conductor. Charging a capacitor to a voltage v implies that an energy $E = \frac{1}{2}Cv^2$ is stored on the capacitor. The mean stored energy $\overline{E}$ in a thermal equilibrium system can only be

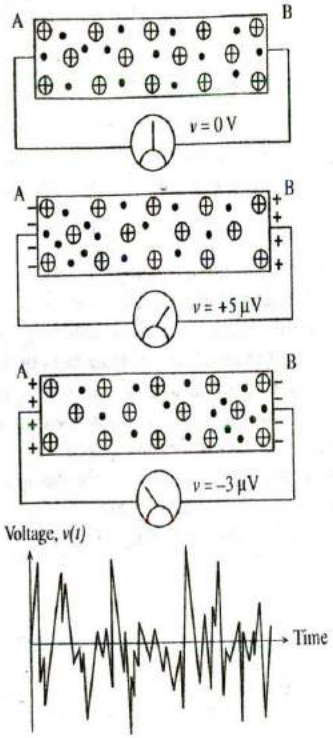


Figure 1.26 Random motion of conduction electrons in a conductor, resulting in electrical noise.

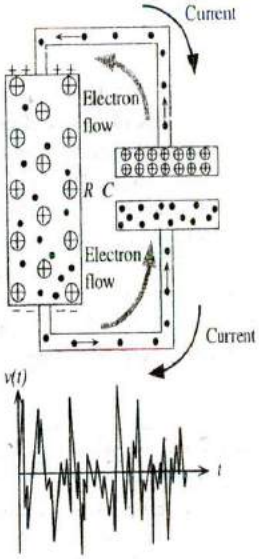


Figure 1.27 Charging and discharging of a capacitor by a conductor, due to the random thermal motions of the conduction electrons.

$\frac{1}{2}kT$, according to the Maxwell energy equipartition theorem. Thus $\overline{E(t)}$, the mean energy stored on C due to thermal fluctuations, is given by

$$\overline{E(t)} = \tfrac{1}{2}C\overline{v(t)^2} = \tfrac{1}{2}kT$$

We see that the mean square voltage across the capacitor is given by

$$\overline{v(t)^2} = \frac{kT}{C} \qquad [1.30]$$

Interestingly, the rms noise voltage across an RC network seems to be independent of the resistance. However, the origin of the noise voltage arises from the electron fluctuations in the conductor and we must somehow reexpress Equation 1.30 to reflect this fact; that is, we must relate the electrical fluctuations to R.

The voltage fluctuations across the network will have many sinusoidal components, but only those below the cutoff frequency of the RC network will contribute to the mean

square voltage (that is, we effectively have a low-pass filter). If B is the bandwidth of the RC network,[8] then $B = 1/(2\pi RC)$ and we can eliminate C in Equation 1.30 to obtain

$$\overline{v(t)^2} = 2\pi kTRB$$

Root mean square noise voltage across a resistance

This is the key equation for calculating the mean square noise voltage from a resistor over a bandwidth B. A more rigorous derivation makes the numerical factor 4 rather than 2π. For a network with a bandwidth B, the **rms noise voltage** is therefore

$$v_{rms} = \sqrt{4kTRB} \qquad [1.31]$$

Equation 1.31 is known as the **Johnson resistor noise equation,** and it sets the lower limit of the magnitude of small signals that can be amplified. Note that Equation 1.31 basically tells us the rms value of the voltage fluctuations within a given bandwidth (B) and not the origin and spectrum (noise voltage vs. frequency) of the noise. The origin of noise is attributed to the random motions of electrons in the conductor (resistor), and Equation 1.31 is the fundamental description of electrical fluctuations; that is, the fluctuations in the conductor's instantaneous local electron concentration that charges and discharges the capacitor. To determine the rms noise voltage across a network with an impedance $Z(j\omega)$, all we have to do is find the real part of Z, which represents the resistive part, and use this for R in Equation 1.31.

EXAMPLE 1.11 **NOISE IN AN *RLC* CIRCUIT** Most radio receivers have a tuned parallel-resonant circuit, which consists of an inductor L, capacitor C, and resistance R in parallel. Suppose L is 100 μH; C is 100 pF; and R, the equivalent resistance due to the input resistance of the amplifier and to the loss in the coil (coil resistance plus ferrite losses), is about 200 kΩ. What is the minimum rms radio signal that can be detected?

SOLUTION

Consider the bandwidth of this tuned RLC circuit, which can be found in any electrical engineering textbook:

$$B = \frac{f_o}{Q}$$

where $f_o = 1/[2\pi\sqrt{LC}]$ is the resonant frequency and $Q = 2\pi f_o CR$ is the quality factor. Substituting for L, C, and R, we get, $f_o = 10^7/2\pi = 1.6 \times 10^6$ Hz and $Q = 200$, which gives $B = 10^7/[2\pi(200)]$ Hz, or 8 kHz. The rms noise voltage is

$$v_{rms} = [4kTRB]^{1/2} = [4(1.38 \times 10^{-23}\ \text{J K}^{-1})(300\ \text{K})(200 \times 10^3\ \Omega)(8 \times 10^3\ \text{Hz})]^{1/2}$$
$$= 5.1 \times 10^{-6}\ \text{V} \quad \text{or} \quad 5.1\ \mu\text{V}$$

This rms voltage is within a bandwidth of 8 kHz centered at 1.6 MHz. This last information is totally absent in Equation 1.31. If we attempt to use

$$v_{rms} = \left[\frac{kT}{C}\right]^{1/2}$$

[8] A low-pass filter allows all signal frequencies up to the cutoff frequency B to pass. B is $1/[2\pi RC]$.

we get

$$v_{rms} = \left[\frac{(1.38 \times 10^{-23}\ \text{J K}^{-1})(300\ \text{K})}{100 \times 10^{-12}\ \text{F}}\right]^{1/2} = 6.4\ \mu\text{V}$$

However, Equation 1.30 was derived using the *RC* circuit in Figure 1.27, whereas we now have an *LCR* circuit. The correct approach uses Equation 1.31, which is generally valid, and the appropriate bandwidth B.

1.7 THERMALLY ACTIVATED PROCESSES

1.7.1 ARRHENIUS RATE EQUATION

Many physical and chemical processes strongly depend on temperature and exhibit what is called an **Arrhenius type behavior,** in which the rate of change is proportional to $\exp(-E_A/kT)$, where E_A is a characteristic energy parameter applicable to the particular process. For example, when we store food in the refrigerator, we are effectively using the Arrhenius rate equation: cooling the food diminishes the rate of decay. Processes that exhibit an Arrhenius type temperature dependence are referred to as **thermally activated.**

For an intuitive understanding of a thermally activated process, consider a vertical filing cabinet that stands in equilibrium, with its center of mass at A, as sketched in Figure 1.28. Tilting the cabinet left or right increases the potential energy PE and requires external work. If we could supply this energy, we could move the cabinet over its edge and lay it flat, where its PE would be lower than at A. Clearly, since the PE at B is lower, this is a more stable position than A. Further, in going from A to B, we had to overcome a **potential energy barrier** of amount E_A, which corresponds to the cabinet standing on its edge with the center of mass at the highest point at A*. To topple the cabinet, we must first provide energy[9] equal to E_A to take the center of mass to A*, from which point the cabinet, with the slightest encouragement, will fall spontaneously to B to attain the lowest PE. At the end of the whole tilting process, the internal energy change for the cabinet, ΔU, is due to the change in the PE ($=mgh$) from A to B, which is negative; B has lower PE than A.

Suppose, for example, a person with an average energy less than E_A tries to topple the cabinet. Like everyone else, that person experiences energy fluctuations as a result of interactions with the environment (e.g., what type of day the person had). During one of those high-energy periods, he can topple the cabinet, even though most of the time he cannot do so because his average energy is less then E_A. The rate at which the cabinet is toppled depends on the number of times (frequency) the person tries and the probability that he possesses energy greater than E_A.

As an example of a thermally activated process, consider the diffusion of impurity atoms in a solid, one of which is depicted in Figure 1.29. In this example, the impurity atom is at an interatomic void A in the crystal, called an **interstitial site.** For the impurity atom to move from A to a neighboring void B, the atom must push the host neighbors apart as it moves across. This requires energy in much the same way

[9] According to the conservation of energy principle, the increase in the PE from A to A* must come from the external work.

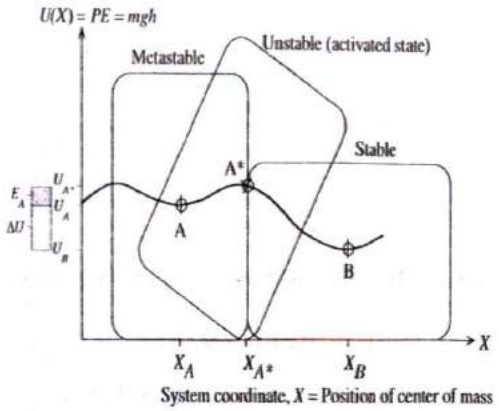


Figure 1.28 Tilting a filing cabinet from state A to its edge in state A* requires an energy E_A.

After reaching A*, the cabinet spontaneously drops to the stable position B. The *PE* of state B is lower than A, and therefore state B is more stable than A.

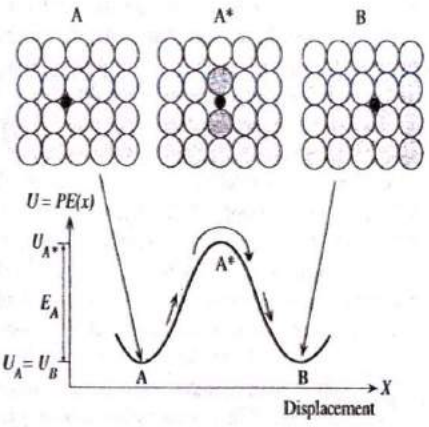


Figure 1.29 Diffusion of an interstitial impurity atom in a crystal from one void to a neighboring void.

The impurity atom at position A must possess an energy E_A to push the host atoms away and move into the neighboring void at B.

as does toppling the filing cabinet. There is a potential energy barrier E_A to the motion of this atom from A to B.

Both the host and the impurity atoms in the solid vibrate about their equilibrium positions, with a distribution of energies, and they also continually exchange energies, which leads to energy fluctuations. In thermal equilibrium, at any instant, we can expect the energy distribution of the atoms to obey the Boltzmann distribution law (see Equation 1.26). The average kinetic energy per atom is vibrational and is $\frac{3}{2}kT$, which will not allow the impurity simply to overcome the *PE* barrier E_A, because typically $E_A \gg \frac{3}{2}kT$.

The rate of jump, called the **diffusion,** of the impurity from A to B depends on two factors. The first is the number of times the atom tries to go over the potential barrier, which is the vibrational frequency v_o, in the AB direction. The second factor is the probability that the atom has sufficient energy to overcome the *PE* barrier. Only during those times when the atom has an energy greater than the potential energy barrier $E_A = U_{A^*} - U_A$ will it jump across from A to B. During this diffusion process, the atom attains an **activated state,** labeled A* in Figure 1.29, with an energy E_A above U_A, so the crystal internal energy is higher than U_A. E_A is called the **activation energy.**

Suppose there are N impurity atoms. At any instant, according to the Boltzmann distribution, $n_E\,dE$ of these will have kinetic energies in the range E to $(E + dE)$, so the probability that an impurity atom has an energy E greater than E_A is

$$\text{Probability } (E > E_A) = \frac{\text{Number of impurities with } E > E_A}{\text{Total number of impurities}}$$

$$= \frac{\int_{E_A}^{\infty} n_E\,dE}{N} = A\exp\left(-\frac{E_A}{kT}\right)$$

where A is a dimensionless constant that has only a weak temperature dependence. The rate of jumps, jumps per seconds, or simply the **frequency of jumps** ϑ from void to void is

Rate for a thermally activated process

$$\vartheta = (\text{Frequency of attempt along AB})(\text{Probability of } E > E_A)$$

$$= A\nu_a\exp\left(-\frac{E_A}{kT}\right) \qquad E_A = U_{A*} - U_A \tag{1.32}$$

Equation 1.32 describes the rate of a thermally activated process, for which increasing the temperature causes more atoms to be energetic and hence results in more jumps over the potential barrier. Equation 1.32 is the well-known **Arrhenius rate equation** and is generally valid for a vast number of transformations, both chemical and physical.

1.7.2 ATOMIC DIFFUSION AND THE DIFFUSION COEFFICIENT

Consider the motion of the impurity atom in Figure 1.29. For simplicity, assume a two-dimensional crystal in the plane of the paper, as in Figure 1.30. The impurity atom has four neighboring voids into which it can jump. If θ is the angle with respect to the x axis, then these voids are at directions $\theta = 0°$, 90°, 180°, and 270°, as depicted in Figure 1.30. Each jump is in a random direction along one of these four angles. As the impurity atom jumps from void to void, it leaves its original location at O, and after N jumps, after time t, it has been displaced from O to O′.

Let a be the closest void-to-void separation. Each jump results in a displacement along x which is equal to $a\cos\theta$, with $\theta = 0°$, 90°, 180°, or 270°. Thus, each jump

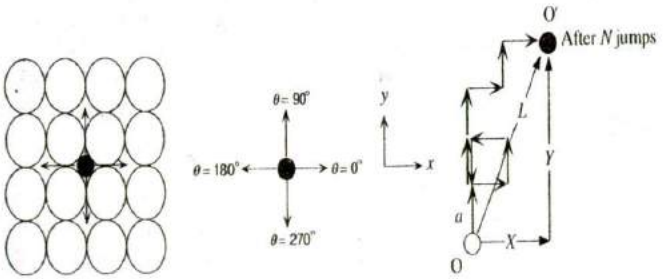


Figure 1.30 An impurity atom has four site choices for diffusion to a neighboring interstitial vacancy. After N jumps, the impurity atom would have been displaced from the original position at O.

results in a displacement along x which can be a, 0, $-a$, or 0, corresponding to the four possibilities. After N jumps, the mean displacement along x will be close to zero, just as the mean voltage of the ac voltage from a power outlet is zero, even though it has an rms value of 120 V. We therefore consider the square of the displacements. The total square displacement, denoted X^2, is

$$X^2 = a^2\cos^2\theta_1 + a^2\cos^2\theta_2 + \cdots + a^2\cos^2\theta_N$$

Clearly, $\theta = 90°$ and $270°$ give $\cos^2\theta = 0$. Of all N jumps, $\frac{1}{2}N$ are $\theta = 0$ and $180°$, each of which gives $\cos^2\theta = 1$. Thus,

$$X^2 = \frac{1}{2}a^2N$$

There will be a similar expression for Y^2, which means that after N jumps, the total square distance L^2 from O to O′ in Figure 1.30 is

$$L^2 = X^2 + Y^2 = a^2N$$

The rate of jumping (frequency of jumps) is given by Equation 1.32

$$\vartheta = v_o A \exp\left(-\frac{E_A}{kT}\right)$$

so the time per jump is $1/\vartheta$. Time t for N jumps is N/ϑ. Thus, $N = \vartheta t$ and

Mean square displacement

$$L^2 = a^2\vartheta t = 2Dt \qquad [1.33]$$

where, by definition, $D = \frac{1}{2}a^2\vartheta$, which is a constant that depends on the diffusion process, as well as the temperature, by virtue of ϑ. This constant is generally called the **diffusion coefficient.** Substituting for ϑ, we find

$$D = \frac{1}{2}a^2 v_o A \exp\left(-\frac{E_A}{kT}\right)$$

or

Diffusion coefficient

$$D = D_o \exp\left(-\frac{E_A}{kT}\right) \qquad [1.34]$$

where D_o is a constant. The root square displacement L in time t, from Equation 1.33, is given by $L = [2Dt]^{1/2}$. Since L^2 is evaluated from X^2 and Y^2, L is known as the **root mean square (rms) displacement.**

The preceding specific example considered the diffusion of an impurity in a void between atoms in a crystal; this is a simple way to visualize the diffusion process. An impurity, indeed any atom, at a regular atomic site in the crystal can also diffuse around by various other mechanisms. For example, such an impurity can simultaneously exchange places with a neighbor. But, more significantly, if a neighboring atomic site has a *vacancy* that has been left by a missing host atom, then the impurity can simply jump into this vacancy. (Vacancies in crystals are explained in detail in Section 1.9.1; for the present, they simply correspond to missing atoms in the crystal.) The activation energy E_A in Equation 1.34 is a measure of the difficulty of the diffusion process. It may be as simple as the energy (or work) required for an impurity atom to deform (or strain) the crystal around it as it jumps from one interstitial site to a neighboring interstitial site, as in Figure 1.29; or it may be more complicated, for example, involving vacancy creation.

Various Si semiconductor devices are fabricated by doping a single Si crystal with impurities (dopants) at high temperatures. For example, doping the Si crystal with phosphorus (P) gives the crystal a higher electrical conductivity. The P atoms substitute directly for Si atoms in the crystal. These dopants migrate from high to low dopant concentration regions in the crystal by diffusion, which occurs efficiently only at sufficiently high temperatures.

EXAMPLE 1.12

DIFFUSION OF DOPANTS IN SILICON The diffusion coefficient of P atoms in the Si crystal follows Equation 1.34 with $D_o = 10.5\ \text{cm}^2\,\text{s}^{-1}$ and $E_A = 3.69$ eV. What is the diffusion coefficient at a temperature of 1100 °C at which dopants such as P are diffused into Si to fabricate various devices? What is the rms distance diffused by P atoms in 5 minutes? Estimate, as an order of magnitude, how many jumps the P atom makes in 1 second if you take the jump distance to be roughly the mean interatomic separation, ~0.27 nm.

SOLUTION

From Equation 1.34,

$$D = D_o \exp\left(-\frac{E_A}{kT}\right) = (10.5\ \text{cm}^2\,\text{s}^{-1}) \exp\left[-\frac{(3.69\ \text{eV})(1.602 \times 10^{-19}\ \text{J eV}^{-1})}{(1.381 \times 10^{-23}\ \text{J K}^{-1})(1100 + 273\ \text{K})}\right]$$

$$= 3.0 \times 10^{-13}\ \text{cm}^2\,\text{s}^{-1}$$

The rms distance L diffused in a time $t = 5\ \text{min} = 5 \times 60$ seconds is

$$L = \sqrt{2Dt} = [2(3.0 \times 10^{-13}\ \text{cm}^2\,\text{s}^{-1})(5 \times 60\ \text{s})]^{1/2} = 1.3 \times 10^{-5}\ \text{cm} \quad \text{or} \quad 13\ \mu\text{m}$$

Equation 1.33 was derived for a two-dimensional crystal as in Figure 1.30, and for an impurity diffusion. Nonetheless, we can still use it to estimate how many jumps a P atom makes in 1 second. From Equation 1.33, $\vartheta \approx 2D/a^2 \approx 2(3.0 \times 10^{-17}\ \text{m}^2\,\text{s}^{-1})/(0.27 \times 10^{-9}\ \text{m})^2 = 823$ jumps per second. It takes roughly 1 ms to make one jump. It is left as an exercise to show that at room temperature it will take a P atom 10^{46} years to make a jump! (Scientists and engineers know how to use thermally activated processes.)

1.8 THE CRYSTALLINE STATE

1.8.1 TYPES OF CRYSTALS

A **crystalline solid** is a solid in which the atoms bond with each other in a regular pattern to form a periodic collection (or array) of atoms, as shown for the copper crystal in Figure 1.31. The most important property of a crystal is **periodicity,** which leads to what is termed **long-range order.** In a crystal, the local bonding geometry is repeated many times at regular intervals, to produce a periodic array of atoms that constitutes the crystal structure. The location of each atom is well known by virtue of periodicity. There is therefore a long-range order, since we can always predict the atomic arrangement anywhere in the crystal. Nearly all metals, many ceramics and semiconductors, and various polymers are crystalline solids in the sense that the atoms or molecules are positioned on a **periodic array of points in space.**

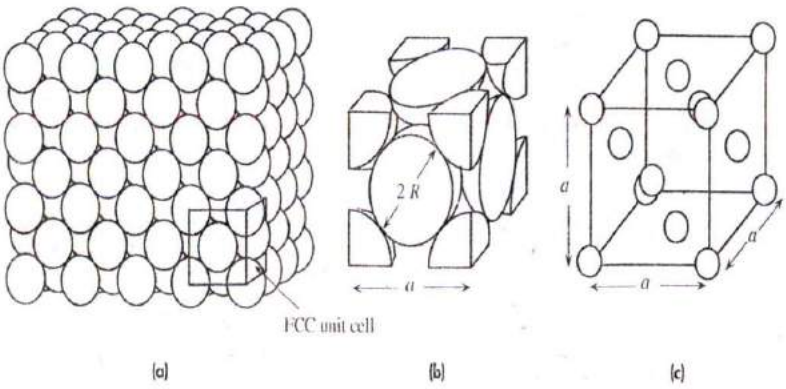


Figure 1.31

(a) The crystal structure of copper which is face-centered cubic (FCC). The atoms are positioned at well-defined sites arranged periodically, and there is a long-range order in the crystal.

(b) An FCC unit cell with close-packed spheres.

(c) Reduced-sphere representation of the FCC unit cell.

Examples: Ag, Al, Au, Ca, Cu, γ-Fe (>912 °C), Ni, Pd, Pt, Rh.

All crystals can be described in terms of a lattice and a basis.[10] A **lattice** is an infinite *periodic* array of geometric points in space, without any atoms. When we place an identical group of atoms (or molecules), called a **basis,** at each lattice point, we obtain the actual **crystal structure.** The crystal is thus a lattice plus a basis at each lattice point. In the copper crystal in Figure 1.31a, each lattice point has one Cu atom and the basis is a single Cu atom. As apparent from Figure 1.31a, the lattice of the copper crystal has **cubic symmetry** and is one of many possible lattices.

Since the crystal is essentially a periodic repetition of a small volume (or cell) of atoms in three dimensions, it is useful to identify the repeating unit so that the crystal properties can be described through this unit. The **unit cell** is the most convenient small cell in the crystal structure that carries the properties of the crystal. The repetition of the unit cell in three dimensions generates the whole crystal structure, as is apparent in Figure 1.31a for the copper crystal.

The unit cell of the copper crystal is cubic with Cu atoms at its corners and one Cu atom at the center of each face, as indicated in Figure 1.31b. The unit cell of Cu is thus said to have a **face-centered cubic (FCC)** structure. The Cu atoms are shared with neighboring unit cells. Effectively, then, only one-eighth of a corner atom is in the unit cell and one-half of the face-centered atom belongs to the unit cell, as shown in Figure 1.31b. This means there are effectively four atoms in the unit cell. The length of the cubic unit cell is termed the **lattice parameter** a of the crystal structure. For Cu, for example, a is 0.362 nm, whereas the radius R of the Cu atom in the crystal is 0.128 nm.

[10] Lattice is a purely imaginary geometric concept whose only requirement is that the infinite array of points has periodicity. In many informal discussions, the term lattice or crystal lattice is used to mean the crystal structure itself. These concepts are further developed in Section 1.13 under Additional Topics.

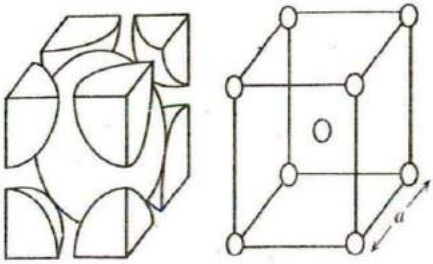


Figure 1.32 Body-centered cubic (BCC) crystal structure.
(a) A BCC unit cell with close-packed hard spheres representing the Fe atoms.
(b) A reduced-sphere unit cell.

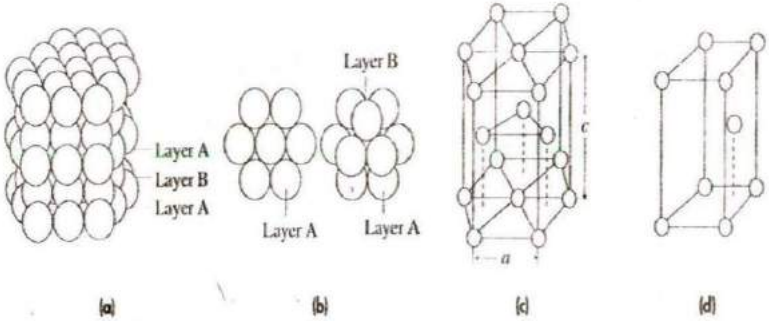


Figure 1.33 The hexagonal close-packed (HCP) crystal structure.
(a) The hexagonal close-packed (HCP) structure. A collection of many Zn atoms. Color difference distinguishes layers (stacks).
(b) The stacking sequence of closely packed layers is ABAB.
(c) A unit cell with reduced spheres.
(d) The smallest unit cell with reduced spheres.

Assuming the Cu atoms are spheres that touch each other, we can geometrically relate a and R. For clarity, it is often more convenient to draw the unit cell with the spheres reduced, as in Figure 1.31c.

The FCC crystal structure of Cu is known as a **close-packed crystal structure** because the Cu atoms are packed as closely as possible, as is apparent in Figure 1.31a and b. The volume of the FCC unit cell is 74 percent full of atoms, which is the maximum packing possible with identical spheres. By comparison, iron has a **body-centered cubic (BCC)** crystal structure and its unit cell is shown in Figure 1.32. The BCC unit cell has Fe atoms at its corners and one Fe atom at the center of the cell. The volume of the BCC unit cell is 68 percent full of atoms, which is lower than the maximum possible packing.

The FCC crystal structure is only one way to pack the atoms as closely as possible. For example, in zinc, the atoms are arranged as closely as possible in a hexagonal symmetry, to form the **hexagonal close-packed (HCP) structure** shown in Figure 1.33a. This structure corresponds to packing spheres as closely as possible first as one layer A, as shown in Figure 1.33b. You can visualize this by arranging six pennies as closely as possible on a table top. On top of layer A we can place an identical layer B, with the

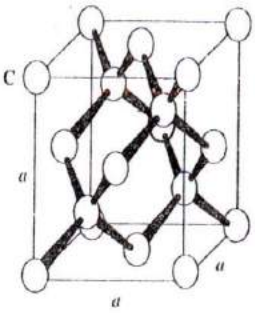


Figure 1.34 The diamond unit cell which is cubic. The cell has eight atoms.

Gray Sn (α-Sn) and the elemental semiconductors Ge and Si have this crystal structure.

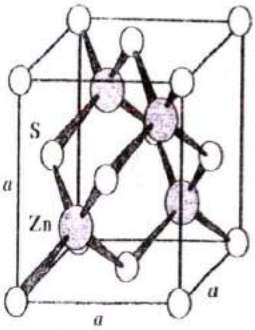


Figure 1.35 The zinc blende (ZnS) cubic crystal structure.

Many important compound crystals have the zinc blende structure. Examples: AlAs, GaAs, GaP, GaSb, InAs, InP, InSb, ZnS, ZnTe.

spheres taking up the voids on layer A, as depicted in Figure 1.33b. The third layer can be placed on top of B and lined up with layer A. The stacking sequence is therefore ABAB. . . . A unit cell for the HCP structure is shown in Figure 1.33c, which shows that this is not a cubic structure. The unit cell shown, although convenient, is not the smallest unit cell. The smallest unit cell for the HCP structure is shown in Figure 1.33d and is called the **hexagonal unit cell.** The repetition of this unit cell will generate the whole HCP structure. The atomic packing density in the HCP crystal structure is 74 percent, which is the same as that in the FCC structure.

Covalently bonded solids, such as silicon and germanium, have a diamond crystal structure brought about by the directional nature of the covalent bond, as shown in Figure 1.34 (see also Figure 1.6). The rigid local bonding geometry of four Si–Si bonds in the tetrahedral configuration forces the atoms to form what is called the **diamond cubic crystal structure.** The unit cell in this case can be identified with the cubic structure. Although there are atoms at each corner and at the center of each face, indicating an FCC-like structure, there are four atoms within the cell as well. Thus, there are eight atoms in the unit cell. The diamond unit cell can actually be described in terms of an FCC lattice (a geometric arrangement of points) with each lattice point having a basis of two Si atoms. If we place the two Si atoms at each site appropriately, for example, one right at the lattice point, and the other displaced from it by a quarter lattice distance $a/4$ along the cube edges, we can easily generate the diamond unit cell. In the copper crystal, each FCC lattice point has one Cu atom, whereas in the Si crystal each lattice point has two Si atoms; thus there are $4 \times 2 = 8$ atoms in the diamond unit cell.

In the GaAs crystal, as in the silicon crystal, each atom forms four directional bonds with its neighbors. The unit cell looks like a diamond cubic, as indicated in Figure 1.35 but with the Ga and As atoms alternating positions. This unit cell is termed

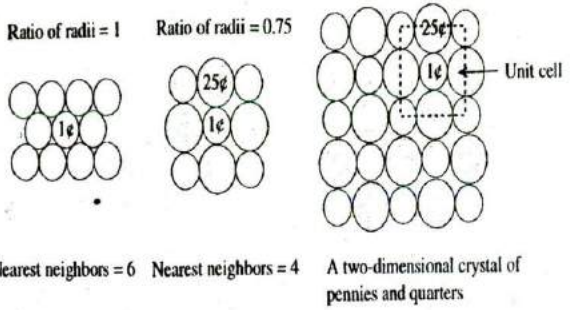


Figure 1.36 Packing of coins on a table top to build a two-dimensional crystal.

the **zinc blende** structure after ZnS, which has this type of unit cell. Many important compound semiconductors have this crystal structure, GaAs being the most commonly known. The zinc blende unit cell can also be described in terms of a fundamental FCC lattice and a basis that has two atoms, Zn and S (or Ga and As). For example, we can place one Zn at each lattice point and one S atom displaced from the Zn by $a/4$ along the cube edges.

In ionic solids, the cations (*e.g.*, Na^+) and the anions (Cl^-) attract each other nondirectionally. The crystal structure depends on how closely the opposite ions can be brought together and how the same ions can best avoid each other while maintaining long-range order, or maintaining symmetry. These depend on the relative charge and relative size per ion.

To demonstrate the importance of the size effect in two dimensions, consider identical coins, say pennies (1-cent coins). At most, we can make six pennies touch one penny, as shown in Figure 1.36. On the other hand, if we use quarters[11] (25-cent coins) to touch one penny, at most only five quarters can do so. However, this arrangement cannot be extended to the construction of a two-dimensional crystal with periodicity. To fulfill the long-range symmetry requirement for crystals, we can only use four quarters to touch the penny and thereby build a two-dimensional "penny–quarter" crystal, which is shown in the figure. In the two-dimensional crystal, a penny has four quarters as nearest neighbors; similarly, a quarter has four pennies as nearest neighbors. A convenient unit cell is a square cell with one-quarter of a penny at each corner and a full penny at the center (as shown in the figure).

The three-dimensional equivalent of the unit cell of the penny–quarter crystal is the **NaCl unit cell** shown in Figure 1.37. The Na^+ ion is about half the size of the Cl^- ion, which permits six nearest neighbors while maintaining long-range order. The repetition of this unit cell in three dimensions generates the whole NaCl crystal, which was depicted in Figure 1.9b.

A similar unit cell with Na^+ and Cl^- interchanged is also possible and equally convenient. We can therefore describe the whole crystal with two interpenetrating FCC

[11] Although many are familiar with the United States coinage, any two coins with a size ratio of about 0.75 would work out the same.

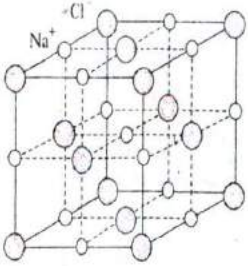


Figure 1.37 A possible reduced-sphere unit cell for the NaCl (rock salt) crystal.

An alternative unit cell may have Na^+ and Cl^- interchanged. Examples: AgCl, CaO, CsF, LiF, LiCl, NaF, NaCl, KF, KCl, MgO.

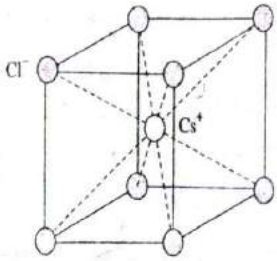


Figure 1.38 A possible reduced-sphere unit cell for the CsCl crystal.

An alternative unit cell may have Cs^+ and Cl^- interchanged. Examples: CsCl, CsBr, CsI, TlCl, TlBr, TlI.

unit cells, each having oppositely charged ions at the corners and face centers. Many ionic solids have the rock salt (NaCl) crystal structure.

When the cation and anions have equal charges and are about the same size, as in the CsCl crystal, the unit cell is called the **CsCl structure,** which is shown in Figure 1.38. Each cation is surrounded by eight anions (and vice versa), which are at

Table 1.3 Properties of some important crystal structures

Crystal Structure	a and R (R is the Radius of the Atom)	Coordination Number (CN)	Number of Atoms per Unit Cell	Atomic Packing Factor	Examples
Simple cubic	$a = 2R$	6	1	0.52	No metals (Except Po)
BCC	$a = \frac{4R}{\sqrt{3}}$	8	2	0.68	Many metals: α-Fe, Cr, Mo, W
FCC	$a = \frac{4R}{\sqrt{2}}$	12	4	0.74	Many metals: Ag, Au, Cu, Pt
HCP	$a = 2R$ $c = 1.633a$	12	2	0.74	Many metals: Co, Mg, Ti, Zn
Diamond	$a = \frac{8R}{\sqrt{3}}$	4	8	0.34	Covalent solids: Diamond, Ge, Si, α-Sn
Zinc blende		4	8	0.34	Many covalent and ionic solids. Many compound semiconductors. ZnS, GaAs, GaSb, InAs, InSb
NaCl		6	4 cations	0.67	Ionic solids such as NaCl, AgCl, LiF, MgO, CaO
			4 anions	(NaCl)	Ionic packing factor depends on relative sizes of ions.
CsCl		8	1 cation 1 anion		Ionic solids such as CsCl, CsBr, CsI

the corners of a cube. This is not a true BCC unit cell because the atoms at various BCC lattice points are different. (As discussed in Section 1.13, CsCl has a simple cubic lattice with a basis that has one Cl^- ion and one Na^+ ion.)

Table 1.3 summarizes some of the important properties of the main crystal structures considered in this section.

EXAMPLE 1.13

THE COPPER (FCC) CRYSTAL Consider the FCC unit cell of the copper crystal shown in Figure 1.39.

a. How many atoms are there per unit cell?

b. If R is the radius of the Cu atom, show that the lattice parameter a is given by $a = R2\sqrt{2}$.

c. Calculate the **atomic packing factor** (APF) defined by

$$\text{APF} = \frac{\text{Volume of atoms in unit cell}}{\text{Volume of unit cell}}$$

d. Calculate the **atomic concentration** (number of atoms per unit volume) in Cu and the density of the crystal given that the atomic mass of Cu is 63.55 g mol^{-1} and the radius of the Cu atom is 0.128 nm.

SOLUTION

a. There are four atoms per unit cell. The Cu atom at each corner is shared with eight other adjoining unit cells. Each Cu atom at the face center is shared with the neighboring unit cell. Thus, the number of atoms in the unit cell = 8 corners ($\frac{1}{8}$ atom) + 6 faces ($\frac{1}{2}$ atoms) = 4 atoms.

b. Consider the unit cell shown in Figure 1.39 and one of the cubic faces. The face is a square of side a and the diagonal is $\sqrt{a^2 + a^2}$ or $a\sqrt{2}$. The diagonal has one atom at the center of diameter $2R$, which touches two atoms centered at the corners. The diagonal, going from corner to corner, is therefore $R + 2R + R$. Thus, $4R = a\sqrt{2}$ and $a = 4R/\sqrt{2} = R2\sqrt{2}$. Therefore, $a = 0.3620$ nm.

c. $$\text{APF} = \frac{(\text{Number of atoms in unit cell}) \times (\text{Volume of atom})}{\text{Volume of unit cell}}$$

$$= \frac{4 \times \frac{4}{3}\pi R^3}{a^3} = \frac{\frac{4^2}{3}\pi R^3}{(R2\sqrt{2})^3} = \frac{4^2\pi}{3(2\sqrt{2})^3} = 0.74$$

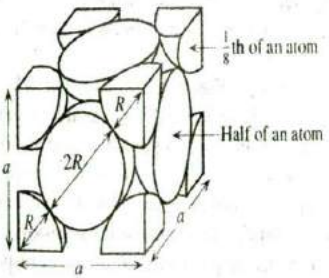


Figure 1.39 The FCC unit cell. The atomic radius is R and the lattice parameter is a.

d. In general, if there are x atoms in the unit cell, the atomic concentration is

$$n_{at} = \frac{\text{Number of atoms in unit cell}}{\text{Volume of unit cell}} = \frac{x}{a^3}$$

Thus, for Cu

$$n_{at} = \frac{4}{(0.3620 \times 10^{-7}\ \text{cm})^3} = 8.43 \times 10^{22}\ \text{cm}^{-3}$$

There are x atoms in the unit cell, and each atom has a mass of M_{at}/N_A grams. The density ρ is

$$\rho = \frac{\text{Mass of all atoms in unit cell}}{\text{Volume of unit cell}} = \frac{x\left(\frac{M_{at}}{N_A}\right)}{a^3}$$

that is, $$\rho = \frac{n_{at}M_{at}}{N_A} = \frac{(8.43 \times 10^{22}\ \text{cm}^{-3})(63.55\ \text{g mol}^{-1})}{6.022 \times 10^{23}\ \text{mol}^{-1}} = 8.9\ \text{g cm}^{-3}$$

The expression $\rho = (n_{at}M_{at})/N_A$ is independent of the crystal structure.

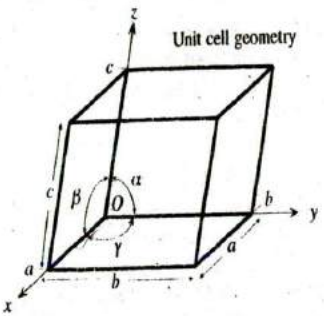
1.8.2 CRYSTAL DIRECTIONS AND PLANES

There can be a number of possibilities for choosing a unit cell for a given crystal structure, as is apparent in Figure 1.33c and d for the HCP crystal. As a convention, we generally represent the **geometry of the unit cell** as a parallelepiped with sides a, b, and c and angles α, β, and γ, as depicted in Figure 1.40a. The sides a, b, and c and angles α, β, and γ are referred to as the **lattice parameters.** To establish a reference frame and to apply three-dimensional geometry, we insert an xyz coordinate system. The x, y, and z axes follow the edges of the parallelepiped and the origin is at the lower-left rear corner of the cell. The unit cell extends along the x axis from 0 to a, along y from 0 to b, and along z from 0 to c.

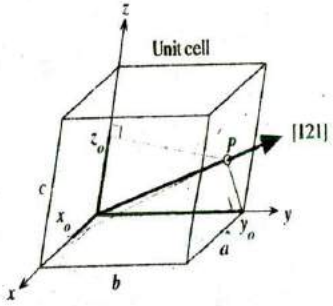
For Cu and Fe, the unit-cell geometry has $a = b = c$, $\alpha = \beta = \gamma = 90°$, and cubic symmetry. For Zn, the unit cell has hexagonal geometry, with $a = b \neq c$, $\alpha = \beta = 90°$, and $\gamma = 120°$, as shown in Figure 1.33d.

In explaining crystal properties, we must frequently specify a direction in a crystal, or a particular plane of atoms. Many properties, for example, the elastic modulus, electrical resistivity, magnetic susceptibility, etc., are directional within the crystal. We use the convention described here for labeling crystal directions based on three-dimensional geometry.

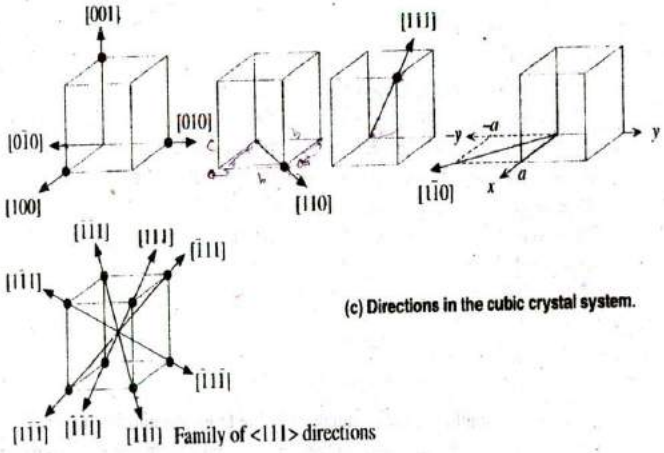
All parallel vectors have the same indices. Therefore, the direction to be labeled can be moved to pass through the origin of the unit cell. As an example, Figure 1.40b shows a direction whose indices are to be determined. A point P on the vector can be expressed by the coordinates x_o, y_o, z_o, where x_o, y_o, and z_o are projections from point P onto the x, y, and z axes, respectively, as shown in Figure 1.40b. It is generally convenient to place P where the line cuts a surface (though this is not necessary). We can express these coordinates in terms of the lattice parameters a, b, and c, respectively. We then have three coordinates, say x_1, y_1, and z_1, for point P in terms of a, b, and c.



(a) A parallelepiped is chosen to describe the *geometry* of a unit cell. We line the *x*, *y*, and *z* axes with the edges of the parallelepiped taking the lower-left rear corner as the origin.



(b) Identification of a direction in a crystal.



(c) Directions in the cubic crystal system.

Figure 1.40

For example, if

$$x_o, y_o, z_o \quad \text{are} \quad \frac{1}{2}a, b, \frac{1}{2}c$$

then P is at

$$x_1, y_1, z_1 \quad i.e., \quad \frac{1}{2}, 1, \frac{1}{2}$$

We then multiply or divide these numbers until we have the smallest integers (which may include 0). If we call these integers u, v, and w, then the direction is written in square brackets without commas as $[uvw]$. If any integer is a negative number, we use a bar on top of that integer. For the particular direction in Figure 1.40b, we therefore have [121].

Some of the important directions in a cubic lattice are shown in Figure 1. 40c. For example, the x, y, and z directions in the cube are [100], [010], and [001], as shown. Reversing a direction simply changes the sign of each index. The negative x, y, and z directions are $[\bar{1}00]$, $[0\bar{1}0]$, and $[00\bar{1}]$, respectively.

Certain directions in the crystal are equivalent because the differences between them are based only on our arbitrary decision for labeling x, y, and z directions. For example, [100] and [010] are different simply because of the way in which we labeled the x and y axes. Indeed, directional properties of a material (*e.g.*, elastic modulus, and dielectric susceptibility) along the edge of the cube [100] are invariably the same as along the other edges, for example, [010] and [001]. All of these directions along the edges of the cube constitute a **family of directions,** which is any set of directions considered to be equivalent. We label a family of directions, for example, [100], [010], [001], ..., by using a common notation, triangular brackets. Thus, ⟨100⟩ represents the family of six directions, [100], [010], [001], $[\bar{1}00]$, $[0\bar{1}0]$, and $[00\bar{1}]$ in a cubic crystal. Similarly, the family of diagonal directions in the cube, shown in Figure 1.40c, is denoted ⟨111⟩.

We also frequently need to describe a particular plane in a crystal. Figure 1.41 shows a general unit cell with a plane to be labeled. We use the following convention, called the **Miller indices of a plane,** for this purpose.

We take the intercepts x_o, y_o, and z_o of the plane on the x, y, and z axes, respectively. If the plane passes through the origin, we can use another convenient parallel plane, or simply shift the origin to another point. All planes that have been shifted by a lattice parameter have identical Miller indices.

We express the intercepts x_o, y_o, and z_o in terms of the lattice parameters a, b, and c, respectively, to obtain x_1, y_1, and z_1. We then invert these numbers. Taking the reciprocals, we obtain

$$\frac{1}{x_1}, \frac{1}{y_1}, \frac{1}{z_1}$$

We then clear all fractions, without reducing to lowest integers, to obtain a set of integers, say h, k, and ℓ. We then put these integers into parentheses, without commas, that is, $(hk\ell)$. For the plane in Figure 1.41a, we have

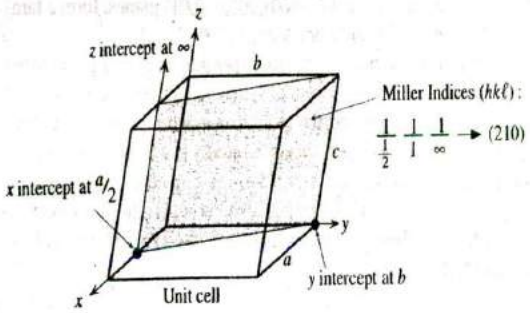
Intercepts x_o, y_o, and z_o are $\frac{1}{2}a$, $1b$, and ∞c.

Intercepts x_1, y_1, and z_1, in terms of a, b, and c, are $\frac{1}{2}$, 1, and ∞.

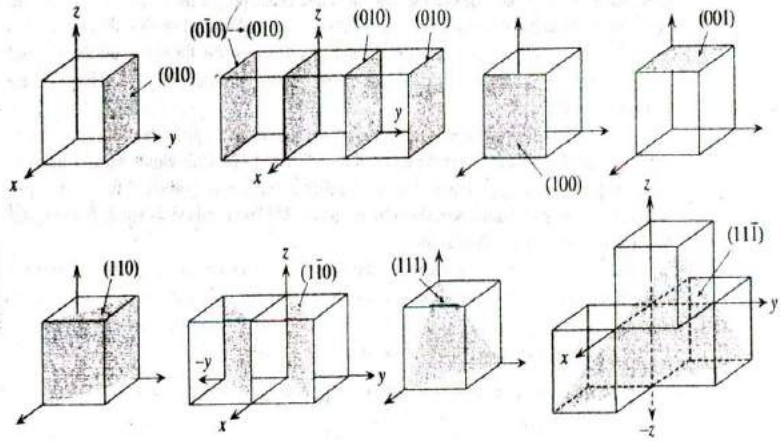
Reciprocals $1/x_1$, $1/y_1$, and $1/z_1$ are $1/\frac{1}{2}$, $1/1$, $1/\infty = 2, 1, 0$.

This set of numbers does not have fractions, so it is not necessary to clear fractions. Hence, the Miller indices $(hk\ell)$ are (210).

If there is a negative integer due to a negative intercept, a bar is placed across the top of the integer. Also, if parallel planes differ only by a shift that involves a multiple number of lattice parameters, then these planes may be assigned the same Miller indices. For example, the plane $(0\bar{1}0)$ is the xz plane that cuts the y axis at $-b$. If we shift the plane along y by two lattice parameters ($2b$), it will cut the y axis at b and the Miller indices will become (010). In terms of the unit cell, the $(0\bar{1}0)$ plane is the same as the (010) plane, as shown in Figure 1.41b. Note that not all parallel planes are



(a) Identification of a plane in a crystal



(b) Various planes in the cubic lattice.

Figure 1.41 Labeling of crystal planes and typical examples in the cubic lattice.

identical. Planes can have the same Miller indices *only* if they are separated by a multiple of the lattice parameter. For example, the (010) plane is not identical to the (020) plane, even though they are geometrically parallel. In terms of the unit cell, plane (010) is a face of the unit cell cutting the y axis at b, whereas (020) is a plane that is halfway inside the unit cell, cutting the y axis at $\frac{1}{2}b$. The planes contain different numbers of atoms. The (020) plane cannot be shifted by the lattice parameter b to coincide with plane (010).

It is apparent from Figure 1.41b that in the case of the cubic crystal, the $[hk\ell]$ direction is always perpendicular to the $(hk\ell)$ plane.

Certain planes in the crystal belong to a **family of planes** because their indices differ only as a consequence of the arbitrary choice of axis labels. For example, the indices of the (100) plane become (010) if we switch the x and y axes. All the (100),

(010), and (001) planes, and hence the parallel $(\bar{1}00)$, $(0\bar{1}0)$, $(00\bar{1})$ planes, form a family of planes, conveniently denoted by curly brackets as {100}.

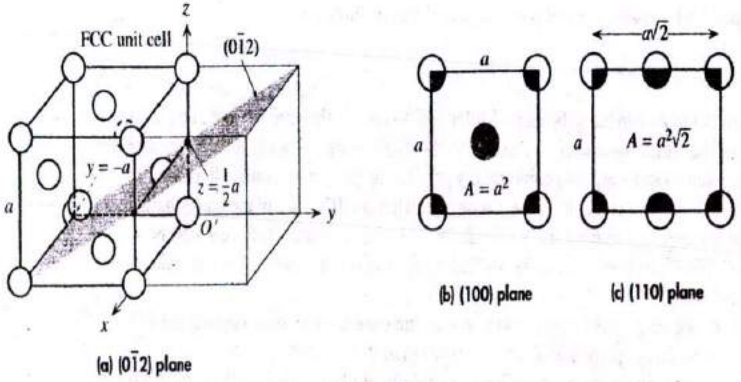
Frequently we need to know the number of atoms per unit area on a given plane $(hk\ell)$. For example, if the surface concentration of atoms is high on one plane, then that plane may encourage oxide growth more rapidly than another plane where there are less atoms per unit area. **Planar concentration of atoms** is the number of atoms per unit area, that is, the surface concentration of atoms, on a given plane in the crystal. Among the {100}, {110}, and {111}, planes in FCC crystals, the most densely packed planes, those with the highest planar concentration, are {111} planes and the least densely packed are {110}.

EXAMPLE 1.14 **MILLER INDICES AND PLANAR CONCENTRATION** Consider the plane shown in Figure 1.42a, which passes through one side of a face and the center of an opposite face in the FCC lattice. The plane passes through the origin at the lower-left rear corner. We therefore shift the origin to say point O' at the lower-right rear corner of the unit cell. In terms of a, the plane cuts the x, y, and z axes at $\infty, -1, \frac{1}{2}$, respectively. We take the reciprocals to obtain, $0, -1, 2$. Therefore, the Miller indices are $(0\bar{1}2)$.

To calculate the planar concentration $n_{(hk\ell)}$ on a given $(hk\ell)$ plane, we consider a bound area A of the $(hk\ell)$ plane within the unit cell as in Figure 1.42b. Only atoms whose centers lie on A are involved in $n_{(hk\ell)}$. For each atom, we then evaluate what portion of the atomic cross section (a circle in two dimensions) cut by the plane $(hk\ell)$ is contained within A. Consider the Cu FCC crystal with $a = 0.3620$ nm.

The (100) plane corresponds to a cube face and has an area $A = a^2$. There is one full atom at the center; that is, the (100) plane cuts through one full atom, one full circle in two dimensions, at the face center as in Figure 1.42b. However, not all corner atoms are within A. Only a quarter of a circle is within the bound area A in Figure 1.42b.

$$\text{Number of atoms in } A = (4 \text{ corners}) \times \left(\tfrac{1}{4} \text{ atom}\right) + 1 \text{ atom at face center} = 2$$



(a) (0$\bar{1}$2) plane

(b) (100) plane

(c) (110) plane

Figure 1.42 The (0$\bar{1}$2) plane and planar concentrations in an FCC crystal.

Planar concentration $n_{(100)}$ of (100) is

$$n_{(100)} = \frac{4\left(\frac{1}{4}\right)+1}{a^2} = \frac{2}{a^2} = \frac{2}{(0.3620 \times 10^{-9}\ \text{m})^2} = 15.3 \text{ atoms nm}^{-2}$$

Consider the (110) plane as in Figure 1.42c. The number of atoms in the area $A = (a)(a\sqrt{2})$ defined by two face diagonals and two cube sides is

$$(4 \text{ corners}) \times \left(\tfrac{1}{4} \text{ atom}\right) + (2 \text{ face diagonals}) \times \left(\tfrac{1}{2} \text{ atom at diagonal center}\right) = 2$$

Planar concentration on (110) is

$$n_{(110)} = \frac{4\left(\frac{1}{4}\right)+2\left(\frac{1}{2}\right)}{(a)(a\sqrt{2})} = \frac{2}{a^2\sqrt{2}} = 10.8 \text{ atoms nm}^{-2}$$

Similar for the (111) plane, $n_{(111)}$ is 17.0 atoms nm^{-2}. Clearly the (111) planes are the most and (110) planes are the least densely packed among the (100), (110), and (111) planes.

1.8.3 ALLOTROPY AND CARBON

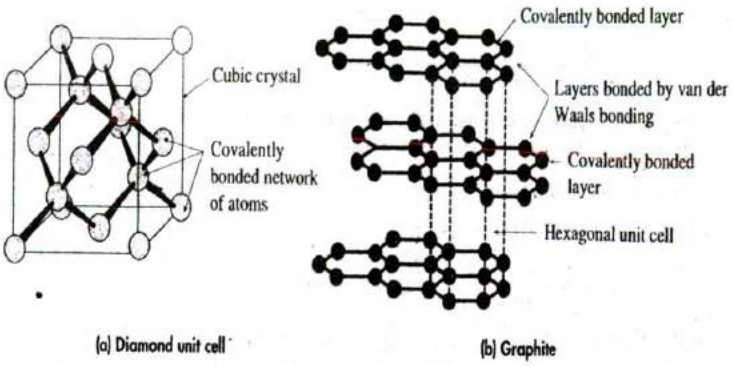
Certain substances can have more than one crystal structure, iron being one of the best-known examples. This characteristic is termed **polymorphism** or **allotropy.** Below 912 °C, iron has the BCC structure and is called α-Fe. Between 912 °C and 1400 °C, iron has the FCC structure and is called γ-Fe. Above 1400 °C, iron again has the BCC structure and is called δ-Fe. Since iron has more than one crystal structure, it is called **polymorphic.** Each iron crystal structure is an allotrope or a polymorph.

The allotropes of iron are all metals. Furthermore, one allotrope changes to another at a well-defined temperature called a **transition temperature,** which in this case is 912 °C.

Many substances have allotropes that exhibit widely different properties. Moreover, for some polymorphic substances, the transformation from one allotrope to another cannot be achieved by a change of temperature, but requires the application of pressure, as in the transformation of graphite to diamond.

Carbon has three important crystalline allotropes: diamond, graphite, and the newly discovered **buckminsterfullerene.** These crystal structures are shown in Figure 1.43a, b and c, respectively, and their properties are summarized in Table 1.4. Graphite is the carbon form that is stable at room temperature. Diamond is the stable form at very high pressures. Once formed, diamond continues to exist at atmospheric pressures and below about 900 °C, because the transformation rate of diamond to graphite is virtually zero under these conditions. Graphite and diamond have widely differing properties, which lead to diverse applications. For example, graphite is an electrical conductor, whereas diamond is an insulator. Diamond is the hardest substance known. On the other hand, the carbon layers in graphite can readily slide over each other under shear stresses, because the layers are only held together by weak secondary bonds (van der Waals bonds). This is the reason for graphite's lubricating properties.

Buckminsterfullerene is another polymorph of carbon. In the buckminsterfullerene molecule (called the "buckyball"), 60 carbon atoms bond with each other to form a



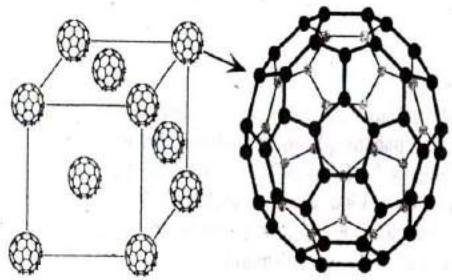


Figure 1.43 The three allotropes of carbon.

perfect soccer ball–type molecule. The C_{60} molecule has 12 pentagons and 20 hexagons joined together to form a spherical molecule, with each C atom at a corner, as depicted in Figure 1.43c. The molecules are produced in the laboratory by a carbon arc in a partial atmosphere of an inert gas (He); they are also found in the soot of partial combustion. The crystal form of buckminsterfullerene has the FCC structure, with each C_{60} molecule occupying a lattice point and being held together by van der Waals forces, as shown in Figure 1.43c. The Buckminsterfullerene crystal is a semiconductor, and its compounds with alkali metals, such as K_3C_{60}, exhibit superconductivity at low temperatures (below 18 K). Mechanically, it is a soft material.

Diamond, graphite, and the fullerene crystals are not the only crystalline allotropes of carbon, and neither are they the only structural forms of carbon. For example, **lonsdaleite,** which is another crystalline allotrope, is *hexagonal diamond*

Table 1.4 Crystalline allotropes of carbon (ρ is the density and Y is the elastic modulus or Young's modulus)

	Graphite	Diamond	Buckminsterfullerene Crystal
Structure	Covalent bonding within layers. Van der Waals bonding between layers. Hexagonal unit cell.	Covalently bonded network. Diamond crystal structure.	Covalently bonded C_{60} spheroidal molecules held in an FCC crystal structure by van der Waals bonding.
Electrical and thermal properties	Good electrical conductor. Thermal conductivity comparable to metals.	Very good electrical insulator. Excellent thermal conductor, about five times more than silver or copper.	Semiconductor. Compounds with alkali metals (e.g., K_3C_{60}) exhibit superconductivity.
Mechanical properties	Lubricating agent. Machinable. Bulk graphite: $Y \approx 27$ GPa $\rho = 2.25$ g cm^{-3}	The hardest material. $Y = 827$ GPa $\rho = 3.25$ g cm^{-3}	Mechanically soft. $Y \approx 18$ GPa $\rho = 1.65$ g cm^{-3}
Comment	Stable allotrope at atmospheric pressure	High-pressure allotrope.	Laboratory synthesized. Occurs in the soot of partial combustion.
Uses, potential uses	Metallurgical crucibles, welding electrodes, heating elements, electrical contacts, refractory applications.	Cutting tool applications. Diamond anvils. Diamond film coated drills, blades, bearings, etc. Jewelry. Heat conductor for ICs. Possible thin-film semiconductor devices, as the charge carrier mobilities are large.	Possible future semiconductor or superconductivity applications.

in which each C atom covalently bonds to four neighbors, as in diamond, but the crystal structure has hexagonal symmetry. (It forms from graphite on meteors when the meteors impact the Earth; currently it is only found in Arizona.) **Amorphous carbon** has no crystal structure (no long-range order), so it is not a crystalline allotrope, but many scientists define it as a form or phase of carbon, or as a structural "allotrope." The recently discovered **carbon nanotubes** are thin and long carbon tubes, perhaps 10 to 100 microns long but only several nanometers in diameter, hence the name *nanotube*. They are tubes made from rolling a graphite sheet into a tube and then capping the ends with hemispherical buckyballs. The carbon tube is really a single macromolecule rather than a crystal in its traditional sense[12]; it is a structural form of carbon. Carbon nanotubes have many interesting and remarkable properties and offer much potential for various applications in electronics; the most topical currently being carbon nanotube field emission devices. (Chapter 4 has an example.)

[12] It is possible to define a unit cell on the surface of a carbon nanotube and apply various crystalline concepts, as some scientists have done. To date, however, there seems to be no single crystal of carbon nanotubes in the same way that there is a fullerene crystal in which the C_{60} molecules are bonded to form an FCC structure.

1.9 CRYSTALLINE DEFECTS AND THEIR SIGNIFICANCE

By bringing all the atoms together to try to form a perfect crystal, we lower the total potential energy of the atoms as much as possible for that particular structure. What happens when the crystal is grown from a liquid or vapor; do you always get a perfect crystal? What happens when the temperature is raised? What happens when impurities are added to the solid?

There is no such thing as a perfect crystal. We must therefore understand the types of defects that can exist in a given crystal structure. Quite often, key mechanical and electrical properties are controlled by these defects.

1.9.1 POINT DEFECTS: VACANCIES AND IMPURITIES

Above the absolute zero temperature, all crystals have atomic vacancies or atoms missing from lattice sites in the crystal structure. The vacancies exist as a requirement of thermal equilibrium and are called **thermodynamic defects.** Vacancies introduce disorder into the crystal by upsetting the perfect periodicity of atomic arrangements.

We know from the kinetic molecular theory that all the atoms in a crystal vibrate about their equilibrium positions with a distribution of energies, a distribution that closely resembles the Boltzmann distribution. At some instant, there may be one atom with sufficient energy to break its bonds and jump to an adjoining site on the surface, as depicted in Figure 1.44. This leaves a vacancy behind, just below the surface. This vacancy can then diffuse into the bulk of the crystal, because a neighboring atom can diffuse into it.

This latter process of vacancy creation has been shown to be a sequence of events in Figure 1.44. Suppose that E_v is the average energy required to create such a vacancy. Then only a fraction, $\exp(-E_v/kT)$, of all the atoms in the crystal can have

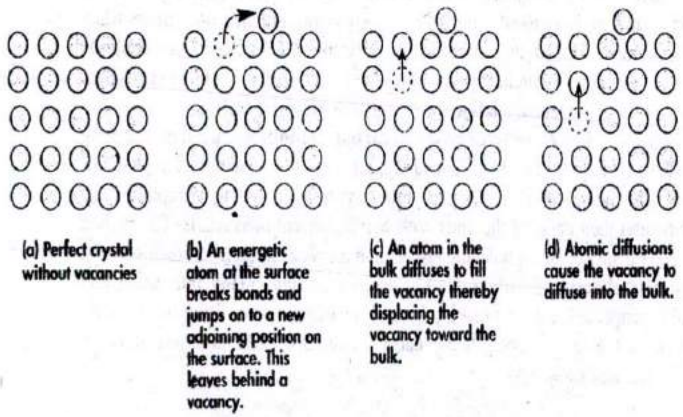
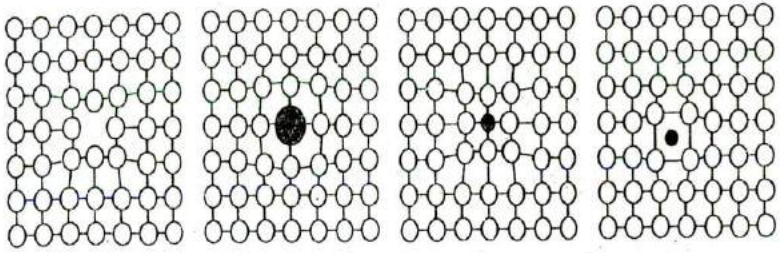


Figure 1.44 Generation of a vacancy by the diffusion of an atom to the surface and the subsequent diffusion of the vacancy into the bulk.



(a) A vacancy in the crystal.

(b) A substitutional impurity in the crystal. The impurity atom is larger than the host atom.

(c) A substitutional impurity in the crystal. The impurity atom is smaller than the host atom.

(d) An interstitial impurity in the crystal. The impurity occupies an empty space between host atoms.

Figure 1.45 Point defects in the crystal structure.
The regions around the point defect become distorted; the lattice becomes strained.

sufficient energy to create vacancies. If the number of atoms per unit volume in the crystal is N, then the vacancy concentration n_v is given by[13]

Equilibrium concentration of vacancies

$$n_v = N \exp\left(-\frac{E_v}{kT}\right) \qquad [1.35]$$

At all temperatures above absolute zero, there will always be an equilibrium concentration of vacancies, as dictated by Equation 1.35. Although we considered only one possible vacancy creation process in Figure 1.44 there are other processes that also create vacancies. Furthermore, we have shown the vacancy to be the same size in the lattice as the missing atom, which is not entirely true. The neighboring atoms around a vacancy close in to take up some of the slack, as shown in Figure 1.45a. This means that the crystal lattice around the vacancy is distorted from the perfect arrangement over a few atomic dimensions. The vacancy volume is therefore smaller than the volume of the missing atom.

Vacancies are only one type of **point defect** in a crystal structure. Point defects generally involve lattice changes or distortions of a few atomic distances, as depicted in Figure 1.45. The crystal structure may contain impurities, either naturally or as a consequence of intentional addition, as in the case of silicon crystals grown for microelectronics. If the impurity atom substitutes directly for the host atom, the result is called a **substitutional impurity** and the resulting crystal structure is that of a **substitutional solid solution,** as shown in Figure 1.45b and c. When a Si crystal is "doped" with small amounts of arsenic (As) atoms, the As atoms substitute directly for the Si atoms in the Si crystal; that is, the arsenic atoms are substitutional impurities. The impurity atom can also place itself in an interstitial site, that is, in a void between the host atoms, as

[13] The proper derivation of the vacancy concentration involves considering thermodynamics and equilibrium concepts. In the actual thermodynamic expression, the pre-exponential term in Equation 1.35 is not unity but a factor that depends on the change in the entropy of the crystal upon vacancy creation. For nearly all practical purposes Equation 1.35 is sufficient.

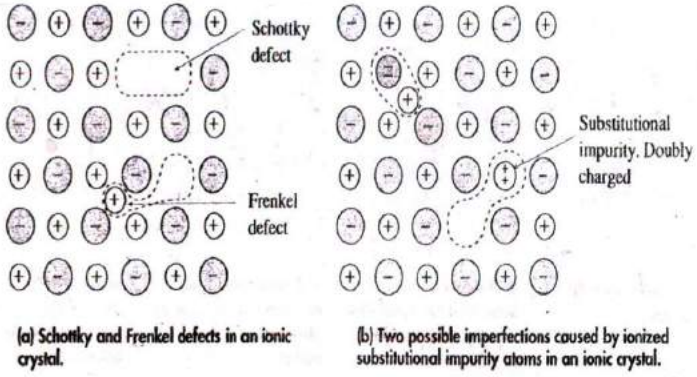


Figure 1.46 Point defects in ionic crystals.

carbon does in the BCC iron crystal. In that case, the impurity is called an **interstitial impurity,** as shown in Figure 1.45d.

In general, the impurity atom will have both a different valency and a different size. It will therefore distort the lattice around it. For example, if a substitutional impurity atom is larger than the host atom, the neighboring host atoms will be pushed away, as in Figure 1.45b. The crystal region around an impurity is therefore distorted from the perfect periodicity and the lattice is said to be **strained around a point defect.** A smaller substitutional impurity atom will pull in the neighboring atoms, as in Figure 1.45c. Typically, interstitial impurities tend to be small atoms compared to the host atoms, a typical example being the small carbon atom in the BCC iron crystal.

In an ionic crystal, such as NaCl, which consists of anions (Cl^-) and cations (Na^+), one common type of defect is called a **Schottky defect.** This involves a missing cation–anion pair (which may have migrated to the surface), so the neutrality is maintained, as indicated in Figure 1.46a. These Schottky defects are responsible for the major optical and electrical properties of alkali halide crystals. Another type of defect in the ionic crystal is the **Frenkel defect,** which occurs when a host ion is displaced into an interstitial position, leaving a vacancy at its original site. The interstitial ion and the vacancy pair constitute the Frenkel defect, as identified in Figure 1.46a. For the AgCl crystal, which has predominantly Frenkel defects, an Ag^+ is in an interstitial position. The concentration of such Frenkel defects is given by Equation 1.35, with an appropriate defect creation energy E_{defect} instead of E_v.

Ionic crystals can also have substitutional and interstitial impurities that become ionized in the lattice. Overall, the ionic crystal must be neutral. Suppose that an Mg^{2+} ion substitutes for an Na^+ ion in the NaCl crystal, as depicted in Figure 1.46b. Since the overall crystal must be neutral, either one Na^+ ion is missing somewhere in the crystal, or an additional Cl^- ion exists in the crystal. Similarly, when a doubly charged negative ion, such as O^{2-}, substitutes for Cl^-, there must either be an additional cation (usually in an interstitial site) or a missing Cl^- somewhere in order to maintain charge

neutrality in the crystal. The most likely type of defect depends on the composition of the ionic solid and the relative sizes and charges of the ions.

EXAMPLE 1.15

VACANCY CONCENTRATION IN A METAL The energy of formation of a vacancy in the aluminum crystal is about 0.70 eV. Calculate the fractional concentration of vacancies in Al at room temperature, 300 K, and very close to its melting temperature 660 °C. What is the vacancy concentration at 660 °C given that the atomic concentration in Al is about 6.0×10^{22} cm^{-3}?

SOLUTION

Using Equation 1.35, the fractional concentration of vacancies are as follows:
At 300 °C,

$$\frac{n_v}{N} = \exp\left(-\frac{E_v}{kT}\right) = \exp\left[-\frac{(0.70\text{ eV})(1.6 \times 10^{-19}\text{ J eV}^{-1})}{(1.38 \times 10^{-23}\text{ J K}^{-1})(300\text{ K})}\right]$$

$$= 1.7 \times 10^{-12}$$

At 660 °C or 933 K,

$$\frac{n_v}{N} = \exp\left(-\frac{E_v}{kT}\right) = \exp\left[-\frac{(0.70\text{ eV})(1.6 \times 10^{-19}\text{ J eV}^{-1})}{(1.38 \times 10^{-23}\text{ J K}^{-1})(933\text{ K})}\right]$$

$$= 1.7 \times 10^{-4}$$

That is, almost 1 in 6000 atomic sites is a vacancy. The atomic concentration N in Al is about 6.0×10^{22} cm^{-3}, which means that the vacancy concentration n_v at 660 °C is

$$n_v = (6.0 \times 10^{22}\text{ cm}^{-3})(1.7 \times 10^{-4}) = 1.0 \times 10^{19}\text{ cm}^{-3}$$

The mean vacancy separation (on the order of $n_v^{-1/3}$) at 660 °C is therefore roughly 5 nm. The mean atomic separation in Al is ~ 0.3 nm ($\sim N^{-1/3}$), so the mean separation between vacancies is only about 20 atomic separations! (A more accurate version of Equation 1.35, with an entropy term, shows that the vacancy concentration is even higher than the estimate in this example.) The increase in the linear thermal expansion coefficient of a metal with temperature near its melting temperature, as shown for Mo in Figure 1.20, has been attributed to the generation of vacancies in the crystal.

EXAMPLE 1.16

VACANCY CONCENTRATION IN A SEMICONDUCTOR The energy of vacancy formation in the Ge crystal is about 2.2 eV. Calculate the fractional concentration of vacancies in Ge at 938 °C, just below its melting temperature. What is the vacancy concentration given that the atomic mass M_{at} and density ρ of Ge are 72.64 g mol^{-1} and 5.32 g cm^{-3}, respectively? Neglect the change in the density with temperature which is small compared with other approximations in Equation 1.35.

SOLUTION

Using Equation 1.34, the fractional concentration of vacancies at 938 °C or 1211 K is

$$\frac{n_v}{N} = \exp\left(-\frac{E_v}{kT}\right) = \exp\left[-\frac{(2.2\text{ eV})(1.6 \times 10^{-19}\text{ J eV}^{-1})}{(1.38 \times 10^{-23}\text{ J K}^{-1})(1211\text{ K})}\right] = 7.0 \times 10^{-10}$$

which is orders of magnitude less than that for Al at its melting temperature in Example 1.15; vacancies in covalent crystals cost much more energy than those in metals.

The number of Ge atoms per unit volume is

$$N = \frac{\rho N_A}{M_{at}} = \frac{(5.32 \text{ g cm}^{-3})(6.022 \times 10^{23} \text{ g mol}^{-1})}{72.64 \text{ g mol}^{-1}} = 4.41 \times 10^{22} \text{ cm}^{-3}$$

so that at 938 °C,

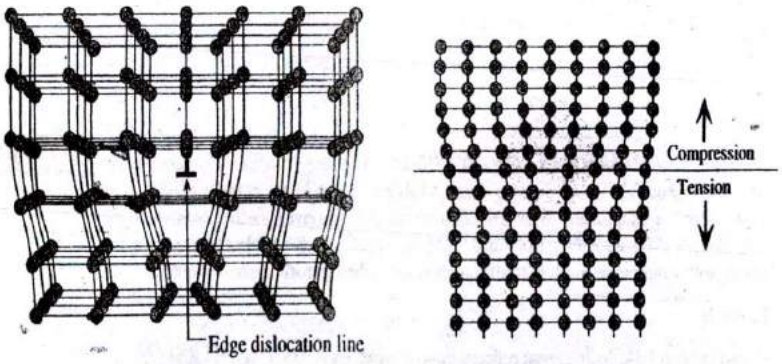
$$n_v = (4.4 \times 10^{22} \text{ cm}^{-3})(7.0 \times 10^{-10}) = 3.1 \times 10^{13} \text{ cm}^{-3}$$

Only 1 in 10^9 atoms is a vacancy.

1.9.2 LINE DEFECTS: EDGE AND SCREW DISLOCATIONS

A line defect is formed in a crystal when an atomic plane terminates within the crystal instead of passing all the way to the end of the crystal, as depicted in Figure 1.47a. The edge of this short plane of atoms is therefore like a line running inside the crystal. The planes neighboring (*i.e.*, above) this short plane are dislocated (displaced) with respect to those below the line. We therefore call this type of defect an **edge dislocation** and use an inverted T symbol. The vertical line corresponds to the half-plane of atoms in the crystal, as illustrated in Figure 1.47a. It is clear that the atoms around the dislocation line have been effectively displaced from their perfect-crystal equilibrium positions, which results in atoms being out of registry above and below the dislocation. The atoms above the dislocation line are pushed together, whereas those below it are pulled apart, so there are regions of compression and tension above and below the dislocation line, respectively, as depicted by the shaded region around the dislocation line in Figure 1.47b. Therefore, around a dislocation line, we have a **strain field** due to the stretching or compressing of bonds.

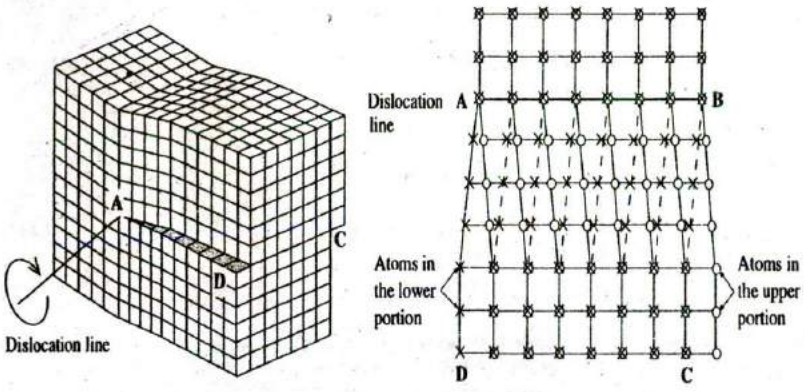
The energy required to create a dislocation is typically in the order of 100 eV per nm of dislocation line. On the other hand, it takes only a few eV to form a point defect,



(a) Dislocation is a line defect. The dislocation shown runs into the paper.

(b) Around the dislocation there is a strain field as the atomic bonds have been compressed above and stretched below the dislocation line.

Figure 1.47 Dislocation in a crystal. This is a line defect, which is accompanied by lattice distortion and hence a lattice strain around it.



(a) A screw dislocation in a crystal

(b) The screw dislocation in (a) as viewed from above

Figure 1.48 A screw dislocation, which involves shearing one portion of a perfect crystal with respect to another, on one side of a line (AB).

which is a few nanometers in dimension. In other words, forming a number of point defects is energetically more favorable than forming a dislocation. Dislocations are not equilibrium defects. They normally arise when the crystal is deformed by stress, or when the crystal is actually being grown.

Another type of dislocation is the **screw dislocation,** which is essentially a shearing of one portion of the crystal with respect to another, by one atomic distance, as illustrated in Figure 1.48a. The displacement occurs on either side of the **screw dislocation line.** The circular arrow around the line symbolizes the screw dislocation. As we move away from the dislocation line, the atoms in the upper portion become more out of registry with those below; at the edge of the crystal, this displacement is one atomic distance, as illustrated in Figure 1.48b.

Both edge and screw dislocations are generally created by stresses resulting from thermal and mechanical processing. A line defect is not necessarily either a pure edge or a pure screw dislocation; it can be a mixture, as depicted in Figure 1.49. Screw dislocations frequently occur during crystal growth, which involves atomic stacking on the surface of a crystal. Such dislocations aid crystallization by providing an additional "edge" to which the incoming atoms can attach, as illustrated in Figure 1.50. To explain, if an atom arrives at the surface of a perfect crystal, it can only attach to one atom in the plane below. However, if there is a screw dislocation, the incoming atom can attach to an edge and thereby form more bonds; hence, it can lower its potential energy more than anywhere else on the surface. With incoming atoms attaching to the edges, the growth occurs spirally around the screw dislocation, and the final crystal surface reflects this spiral growth geometry.

The phenomenon of **plastic** or **permanent deformation** of a metal depends totally on the presence and motions of dislocations, as discussed in elementary books on the mechanical properties of materials. In the case of electrical properties of metals,

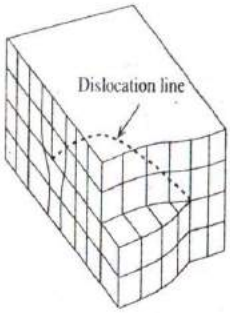


Figure 1.49 A mixed dislocation.

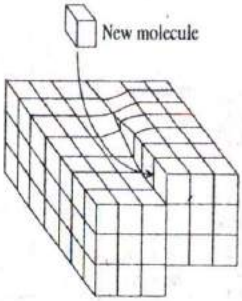
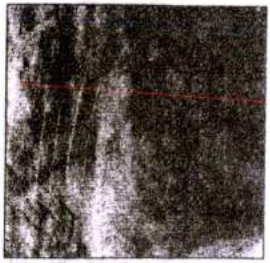


Figure 1.50 Screw dislocation aids crystal growth because the newly arriving atom can attach to two or three atoms instead of one atom and thereby form more bonds.



Growth spiral on the surface of a polypropylene crystal due to screw dislocation aided crystal growth.

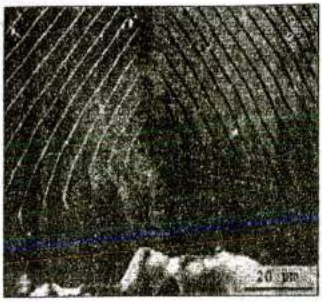
SOURCE: Photo by Phillip Geil, Courtesy of Case Western Reserve University.

we will see in Chapter 2 that dislocations increase the resistivity of materials, cause significant leakage current in a *pn* junction, and give rise to unwanted noise in various semiconductor devices. Fortunately, the occurrence of dislocations in semiconductor crystals can be controlled and nearly eliminated. In a metal interconnection line on a chip, there may be an average of 10^4–10^5 dislocation lines per mm^2 of crystal, whereas a silicon crystal wafer that is carefully grown may typically have only 1 dislocation line per mm^2 of crystal.

1.9.3 PLANAR DEFECTS: GRAIN BOUNDARIES

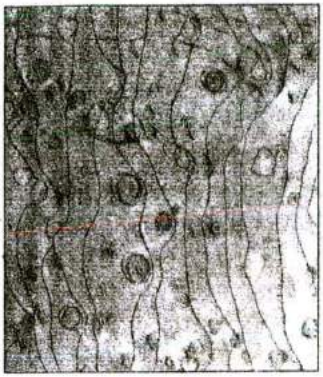
Many materials are polycrystalline; that is, they are composed of many small crystals oriented in different directions. In fact, the growth of a flawless single crystal from what is called the **melt** (liquid) requires special skills, in addition to scientific knowledge. When a liquid is cooled to below its freezing temperature, solidification does not occur at every point; rather, it occurs at certain sites called **nuclei**, which are small crystal-like structures containing perhaps 50 to 100 atoms. Figure 1.51a to c depicts a typical solidification process from the melt. The liquid atoms adjacent to a nucleus diffuse into the nucleus, thereby causing it to grow in size to become a small crystal, or a crystallite, called a **grain**. Since the nuclei are randomly oriented when they are formed, the grains have random crystallographic orientations during crystallite growth. As the liquid between the grains is consumed, some grains meet and obstruct each other. At the end of solidification, therefore, the whole structure has grains with irregular shapes and orientations, as shown in Figure 1.51c.

It is apparent from Figure 1.51c that in contrast to a single crystal, a polycrystalline material has grain boundaries where differently oriented crystals meet. As indicated in Figure 1.52, the atoms at the grain boundaries obviously cannot follow their



The photograph of the surface of a synthetic diamond grown on the {111} surface of natural diamond from sodium carbonate solvent at 5.5 GPa and 1600 °C.

SOURCE: Courtesy of Dr. Hisao Kanda, National Institute for Materials Science, Ibaraki, Japan.



Dislocations can be seen by examining a thin slice of the sample under a transmission electron microscope (TEM). They appear as dark lines and loops as shown here in a Ni–Si alloy single crystal. The loop dislocations are around Ni_3Si particles inside the crystal. The sample had been mechanically deformed, which generates dislocations.

SOURCE: Courtesy of Professor John Humphreys, UMIST, England. (J. Humphreys and V. Ramaswamy in *High Voltage Electron Microscopy*, ed. P. R. Swann, C. J. Humphreys and M. J. Goringe, New York: Academic Press, 1974, p. 26.)





Left: A polycrystalline diamond film on the {100} surface of a single crystal silicon wafer. The film thickness is 6 microns and the SEM magnification is 6000.

Right: A 6-micron-thick CVD diamond film grown on a single crystal silicon wafer. SEM magnification is 8000.

SOURCE: Courtesy of Dr. Paul May, The School of Chemistry, University of Bristol, England.

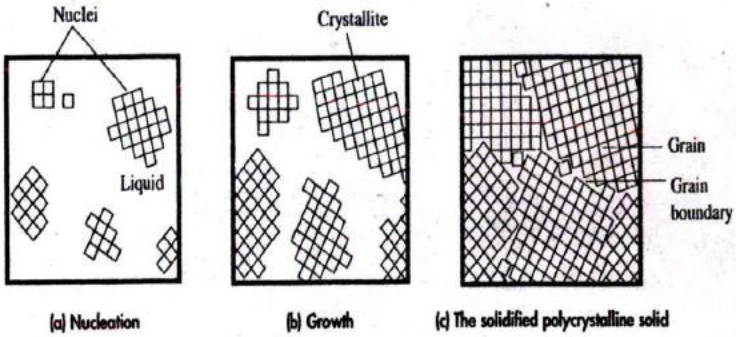


Figure 1.51 Solidification of a polycrystalline solid from the melt. For simplicity, cubes represent atoms.

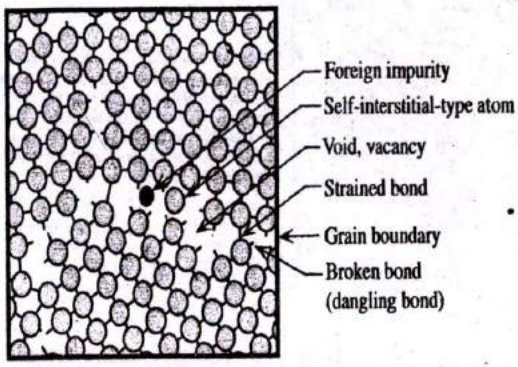


Figure 1.52 The grain boundaries have broken bonds, voids, vacancies, strained bonds, and interstitial-type atoms.
The structure of the grain boundary is disordered, and the atoms in the grain boundaries have higher energies than those within the grains.

natural bonding habits, because the crystal orientation suddenly changes across the boundary. Therefore, there are both voids at the grain boundary and stretched and broken bonds. In addition, in this region, there are misplaced atoms that do not follow the crystalline pattern on either side of the boundary. Consequently, the grain boundary represents a high-energy region per atom with respect to the energy per atom within the bulk of the grains themselves. The atoms can diffuse more easily along a grain boundary because (a) less bonds need to be broken due to the presence of voids and (b) the bonds are strained and easily broken anyway. In many polycrystalline materials, impurities therefore tend to congregate in the grain boundary region. We generally refer to the atomic arrangement in the grain boundary region as being **disordered** due to the presence of the voids and misplaced atoms.

Since the energy of an atom at the grain boundary is greater than that of an atom within the grain, these grain boundaries are nonequilibrium defects; consequently, they try to reduce in size to give the whole structure a lower potential energy. At or around room temperature, the atomic diffusion process is slow; thus, the reduction in the grain boundary is insignificant. At elevated temperatures, however, atomic diffusion allows big grains to grow, at the expense of small grains, which leads to **grain coarsening (grain growth)** and hence to a reduction in the grain boundary area.

Mechanical engineers have learned to control the grain size, and hence the mechanical properties of metals to suit their needs, through various thermal treatment cycles. For electrical engineers, the grain boundaries become important when designing electronic devices based on polysilicon or any polycrystalline semiconductor. For example, in highly polycrystalline materials, particularly thin-film semiconductors (*e.g.*, polysilicon), the resistivity is invariably determined by polycrystallinity, or grain size, of the material, as discussed in Chapter 2.

1.9.4 CRYSTAL SURFACES AND SURFACE PROPERTIES

In describing crystal structures, we assume that the periodicity extends to infinity which means that the regular array of atoms is not interrupted anywhere by the presence of real surfaces of the material. In practice, we know that all substances have real surfaces. When the crystal lattice is abruptly terminated by a surface, the atoms at the surface cannot fulfill their bonding requirements as illustrated in Figure 1.53. For simplicity, the figure shows a Si crystal schematically sketched in two dimensions where each atom in the bulk of the crystal has four covalent bonds, each covalent bond

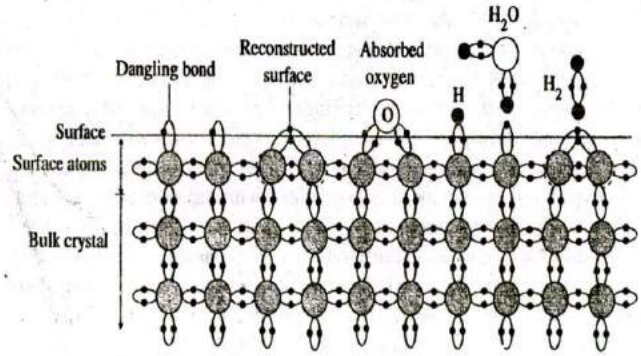


Figure 1.53 At the surface of a hypothetical two-dimensional crystal, the atoms cannot fulfill their bonding requirements and therefore have broken, or dangling, bonds.

Some of the surface atoms bond with each other; the surface becomes reconstructed. The surface can have physisorbed and chemisorbed atoms.

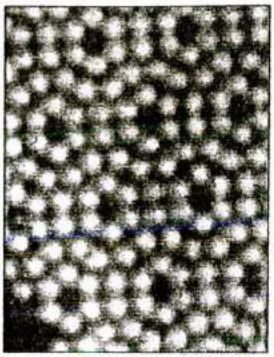
having two electrons.[14] The atoms at the surface are left with **dangling bonds,** bonds that are half full, only having one electron. These dangling bonds are looking for atoms to which they can bond. Two neighboring surface atoms can share each other's dangling bond electrons, that is, form a surface bond with each other. This bonding between surface atoms causes a slight displacement of the surface atoms and leads to a surface that has been **reconstructed.**

Atoms from the environment can also bond with the atoms on the crystal surface. For example, a hydrogen atom can be captured by a dangling bond at the surface to form a chemical bond as a result of which hydrogen becomes **absorbed.** Primary bonding of foreign atoms to a crystal surface is called **chemisorption.** The H atom in Figure 1.53 forms a covalent bond with a Si atom and hence becomes **chemisorbed.** However, the H_2O molecule cannot form a covalent bond, but, because of hydrogen bonding, it can form a secondary bond with a surface Si atom and become **adsorbed.** Secondary bonding of foreign atoms or molecules to a crystal surface is called **physisorption** (*physical adsorption*). Water molecules in the air can readily become adsorbed at the surface of a crystal. Although the figure also shows a physisorbed H_2 molecule as an example, this normally occurs at very low temperatures where crystal vibrations are too weak to quickly dislodge the H_2 molecule. It should be remarked that in many cases, atoms or molecules from the environment become adsorbed at the surface for only a certain period of time; they have a certain sticking or dwell time. For example, at room temperature, inert gases stick to a metal surface only for a duration of the order of microseconds, which is extremely long compared with the vibrational period of the crystal atoms ($\sim 10^{-12}$ seconds). A dangling bond can capture a free electron from the environment if one is available in its vicinity. The same idea applies to a dangling bond at a grain boundary as in Figure 1.52.

At sufficiently high temperatures, some of the absorbed foreign surface atoms can diffuse into the crystal volume to become bulk impurities. Many substances have a natural oxide layer on the surface that starts with the chemical bonding of oxygen atoms to the surface atoms and the subsequent growth of the oxide layer. For example, aluminum surfaces always have a thin aluminum oxide layer. In addition, the surface of the oxide often has adsorbed organic species of atoms usually from machining and handling. The surface condition of a Si crystal wafer in microelectronics is normally controlled by first etching the surface and then oxidizing it at a high temperature to form a SiO_2 **passivating layer** on the crystal surface. This oxide layer is an excellent barrier against the diffusion of impurity atoms into the crystal. (It is also an excellent electrical insulator.)

Figure 1.53 shows only some of the possibilities at the surface of a crystal. Generally the surface structure depends greatly on the mode of surface formation, which invariably involves thermal and mechanical processing, and previous environmental history. One visualization of a crystal surface is based on the **terrace-ledge-kink model,** the so-called **Kossel model,** as illustrated in Figure 1.54. The surface has ledges, kinks, and various imperfections such as holes and dislocations, as well as impurities which can diffuse to and from the surface. The dimensions of the various imperfections (*e.g.*, the step size) depend on the process that generated the surface.

[14] Not all possibilities shown in Figure 1.53 occur in practice, their occurrences depend on the preparation method of the crystal.



Atomic arrangements on a reconstructed (111) surface of a Si crystal as seen by a surface tunneling microscope.

SOURCE: Courtesy of Burleigh Instruments, Inc.

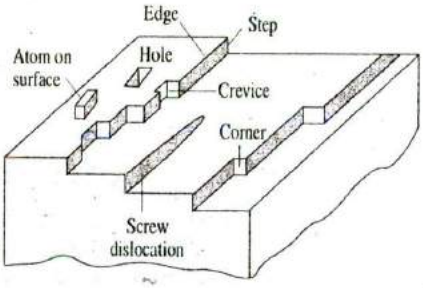
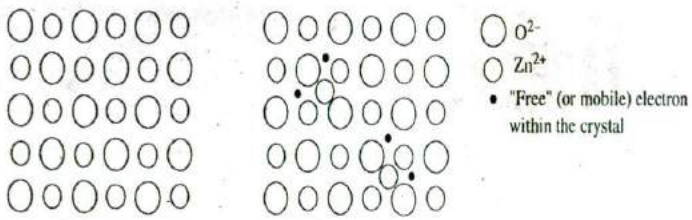


Figure 1.54 Typically, a crystal surface has many types of imperfections, such as steps, ledges, kinks, crevices, holes, and dislocations.

1.9.5 STOICHIOMETRY, NONSTOICHIOMETRY, AND DEFECT STRUCTURES

Stoichiometric compounds are those that have an integer ratio of atoms, for example, as in CaF_2 where two F atoms bond with one Ca atom. Similarly, in the compound ZnO, if there is one O atom for every Zn atom, the compound is stoichiometric, as schematically illustrated in Figure 1.55a. Since there are equal numbers of O^{2-} anions and Zn^{2+} cations, the crystal overall is neutral. It is also possible to have a nonstoichiometric ZnO in which there is excess zinc. This may result if, for example, there is insufficient oxygen during the preparation of the compound. The Zn^{2+} ion has a radius of 0.074 nm, which is about 1.9 times smaller than the O^{2-} anion (radius of 0.14 nm), so it is much easier for a Zn^{2+} ion to enter an interstitial site than the O^{2-} ion or the Zn atom itself, which has a radius of 0.133 nm. Excess Zn atoms therefore occupy interstitial sites as Zn^{2+} cations. Even though the excess zinc atoms are still ionized within the crystal, their lost electrons cannot be taken by oxygen atoms, which are all



(a) Stoichiometric ZnO crystal with equal number of anions and cations and no free electrons

(b) Nonstoichiometric ZnO crystal with excess Zn in interstitial sites as Zn^{2+} cations

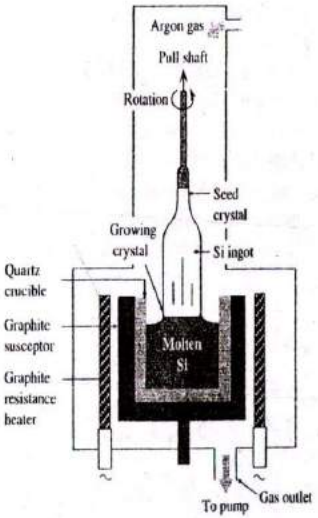
Figure 1.55 Stoichiometry and nonstoichiometry and the resulting defect structure.

O^{2-} anions, as indicated in Figure 1.55b. Thus, the nonstoichiometric ZnO with excess Zn has Zn^{2+} cations in interstitial sites and mobile electrons within the crystal, which can contribute to the conduction of electricity. Overall, the crystal is neutral, as the number of Zn^{2+} ions is equal to the number of O^{2-} ions plus two electrons from each excess Zn. The structure shown in Figure 1.55b is a defect structure, since it deviates from the stoichiometry.

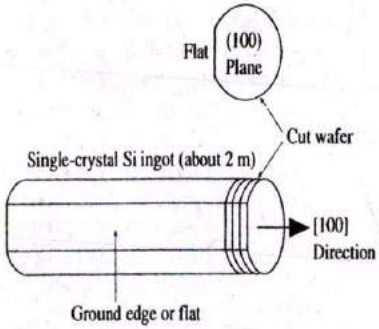
1.10 SINGLE-CRYSTAL CZOCHRALSKI GROWTH

The fabrication of discrete and integrated circuit (IC) solid-state devices requires semiconductor crystals with impurity concentrations as low as possible and crystals that contain very few imperfections. A number of laboratory techniques are available for growing high-purity semiconductor crystals. Generally, they involve either solidification from the melt or condensation of atoms from the vapor phase. The initial process in IC fabrication requires large single-crystal wafers that are typically 15 cm in diameter and 0.6 mm thick. These wafers are cut from a long, cylindrical single Si crystal (typically, 1–2 m in length).

Large, single Si crystals for IC fabrication are often grown by the **Czochralski method**, which involves growing a single-crystal ingot from the melt, using solidification on a seed crystal, as schematically illustrated in Figure 1.56a. Molten Si is held in a quartz (crystalline SiO_2) crucible in a graphite susceptor, which is either heated by

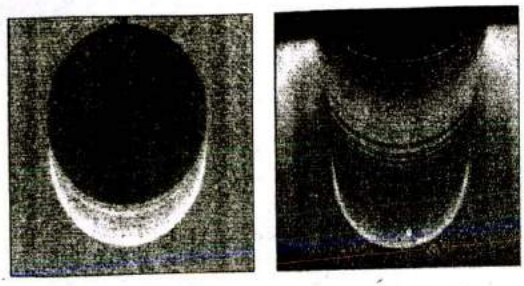


(a) Schematic illustration of the growth of a single-crystal Si ingot by the Czochralski technique.



(b) The crystallographic orientation of the silicon ingot is marked by grounding a flat. The ingot can be as long as 2 m. Wafers are cut using a rotating annulo diamond saw. Typical wafer thickness is 0.6–0.7 mm.

Figure 1.56



Silicon ingot being pulled from the melt in a Czochralski crystal drawer.
SOURCE: Courtesy of MEMC Electronic Materials, Inc.

a graphite resistance heater or by a radio frequency induction coil (a process called **RF heating**).[15] A small dislocation-free crystal, called a **seed,** is lowered to touch the melt and then slowly pulled out of the melt; a crystal grows by solidifying on the seed crystal. The seed is rotated during the pulling stage, to obtain a cylindrical ingot. To suppress evaporation from the melt and prevent oxidation, argon gas is passed through the system.

Initially, as the crystal is withdrawn, its cross-sectional area increases; it then reaches a constant value determined by the temperature gradients, heat losses, and the rate of pull. As the melt solidifies on the crystal, heat of fusion is released and must be conducted away; otherwise, it will raise the temperature of the crystal and remelt it. The area of the melt–crystal interface determines the rate at which this heat can be conducted away through the crystal, whereas the rate of pull determines the rate at which latent heat is released. Although the analysis is not a simple one, it is clear that to obtain an ingot with a large cross-sectional area, the pull speed must be slow. Typical growth rates are a few millimeters per minute.

The sizes and diameters of crystals grown by the Czochralski method are obviously limited by the equipment, though crystals 20–30 cm in diameter and 1–2 m in length are routinely grown for the IC fabrication industry. Also, the crystal orientation of the seed and its flatness with melt surface are important engineering requirements. For example, for very large scale integration (VLSI), the seed is placed with its (100) plane flat to the melt, so that the axis of the cylindrical ingot is along the [100] direction.

Following growth, the Si ingot is usually ground to a specified diameter. Using X-ray diffraction, the crystal orientation is identified and either a flat or an edge is ground along the ingot, as shown in Figure 1.56b. Subsequently, the ingot is cut into thin wafers by a rotating annular diamond saw. To remove any damage to the wafer surfaces caused by sawing and obtain flat, parallel surfaces, the wafers are lapped (ground flat with alumina powder and glycerine), chemically etched, and then polished. The wafers are then used in IC fabrication, usually as a substrate for the growth of a thin layer of crystal from the vapor phase.

The Czochralski technique is also used for growing Ge, GaAs, and InP single crystals, though each case has its own particular requirements. The main drawback of the Czochralski technique is that the final Si crystal inevitably contains oxygen impurities dissolved from the quartz crucible.

[15] The induced eddy currents in the graphite give rise to I^2R heating of the graphite susceptor.

1.11 GLASSES AND AMORPHOUS SEMICONDUCTORS

1.11.1 GLASSES AND AMORPHOUS SOLIDS

A characteristic property of the crystal structure is its periodicity and degree of symmetry. For each atom, the number of neighbors and their exact orientations are well defined; otherwise, the periodicity would be lost. There is therefore a **long-range order** resulting from strict adherence to a well-defined bond length and **relative bond angle** (or exact orientation of neighbors). Figure 1.57a schematically illustrates the presence of a clear, long-range order in a hypothetical two-dimensional crystal. Taking an arbitrary origin, we can predict the position of each atom anywhere in the crystal. We can perhaps use this to represent crystalline SiO_2 (silicon dioxide), for example, in two dimensions. In reality, a Si atom bonds with four oxygen atoms to form a tetrahedron, and the tetrahedra are linked at the corners to create a three-dimensional crystal structure.

Not all solids exhibit crystallinity. Many substances exist in a noncrystalline or amorphous form, due to their method of formation. For example, SiO_2 can have an amorphous structure, as illustrated schematically in two dimensions in Figure 1.57b. In the amorphous phase, SiO_2 is called **vitreous silica,** a form of glass, which has wide engineering applications, including optical fibers. The structure shown in the figure for vitreous silica is essentially that of a frozen liquid, or a **supercooled liquid.** Vitreous silica is indeed readily obtained by cooling the melt.

Many amorphous solids are formed by rapidly cooling or quenching the liquid to temperatures where the atomic motions are so sluggish that crystallization is virtually halted. (The cooling rate is measured relative to the crystallization rate, which depends on atomic diffusion.) We refer to these solids as **glasses.** In the liquid state, the atoms

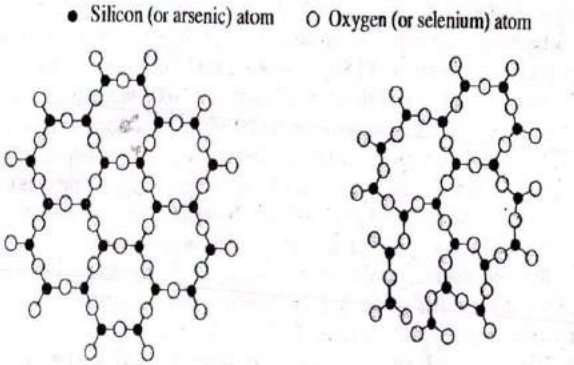


Figure 1.57 Crystalline and amorphous structures illustrated schematically in two dimensions.

have sufficient kinetic energy to break and make bonds frequently and to bend and twist their bonds. There are bond angle variations, as well as rotations of various atoms around bonds **(bond twisting)**. Thus, the bonding geometry around each atom is not necessarily identical to that of other atoms, which leads to the loss of long-range order and the formation of an amorphous structure, as illustrated in Figure 1.57b for the same material in Figure 1.57a. We may view Figure 1.57b as a snapshot of the structure of a liquid. As we move away from a reference atom, after the first and perhaps the second neighbors, random bending and twisting of the bonds is sufficient to destroy long-range order. The amorphous structure therefore lacks the long-range order of the crystalline state.

To reach the glassy state, the temperature is rapidly dropped well below the melting temperature where the atomic diffusion processes needed for arranging the atoms into a crystalline structure are infinitely slow on the time scale of the observation. The liquid structure thus becomes frozen. Figure 1.57b shows that for an amorphous structure, the coordination of each atom is well defined, because each atom must satisfy its chemical bonding requirement, but the whole structure lacks long-range order. Therefore, there is only a **short-range order** in an amorphous solid. The structure is a **continuous random network** of atoms (often called a CRN model of an amorphous solid). As a consequence of the lack of long-range order, amorphous materials do not possess such crystalline imperfections as grain boundaries and dislocations, which is a distinct advantage in certain engineering applications.

Whether a liquid forms a glass or a crystal structure on cooling depends on a combination of factors, such as the nature of the chemical bond between the atoms or molecules, the viscosity of the liquid (which determines how easily the atoms move), the rate of cooling, and the temperature relative to the melting temperature. For example, the oxides SiO_2, B_2O_3, GeO_2, and P_2O_5 have directional bonds that are a mixture of covalent and ionic bonds and the liquid is highly viscous. These oxides readily form glasses on cooling from the melt. On the other hand, it is virtually impossible to quench a pure metal, such as copper, from the melt, bypass crystallization, and form a glass. The metallic bonding is due to an electron gas permeating the space between the copper ions, and that bonding is nondirectional, which means that on cooling, copper ions are readily (and hence, quickly) shifted with respect to each other to form the crystal. There are, however, a number of metal–metal ($Cu_{66}Zr_{33}$) and metal–metalloid alloys ($Fe_{80}B_{20}$, $Pd_{80}Si_{20}$) that form glasses if quenched at ultrahigh cooling rates of 10^6–10^8 °C s^{-1}. In practice, such cooling rates are achieved by squirting a thin jet of the molten metal against a fast-rotating, cooled copper cylinder. On impact, the melt is frozen within a few milliseconds, producing a long ribbon of metallic glass. The process is known as **melt spinning** and is depicted in Figure 1.58.

Many solids used in various applications have an amorphous structure. The ordinary window glass $(SiO_2)_{0.8}(Na_2O)_{0.2}$ and the majority of glassware are common examples. Vitreous silica (SiO_2) mixed with germania (GeO_2) is used extensively in optical fibers. The insulating oxide layer grown on the Si wafer during IC fabrication is the amorphous form of SiO_2. Some intermetallic alloys, such as $Fe_{0.8}B_{0.2}$, can be rapidly quenched from the liquid (as shown in Figure 1.58) to obtain a glassy metal used in low-loss transformer cores. Arsenic triselenide, As_2Se_3, has a crystal structure that resembles the two-dimensional sketch in Figure 1.57a, where an As atom (valency III) bonds with three Se atoms, and a Se atom (valency VI) bonds with two As atoms. In the amorphous

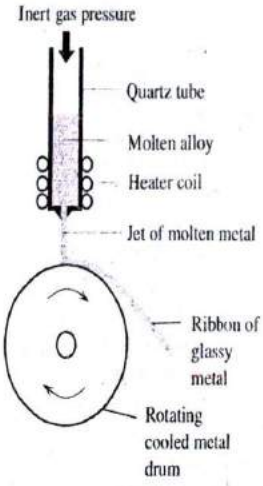




Figure 1.58 It is possible to rapidly quench a molten metallic alloy, thereby bypassing crystallization, and forming a glassy metal commonly called a metallic glass.

The process is called *melt spinning*.

Melt spinning involves squirting a jet of molten metal unto a rotating cool metal drum. The molten jet is instantly solidified into a glassy metal ribbon which is a few microns in thickness. The process produces roughly 1 to 2 kilometers of ribbon per minute.

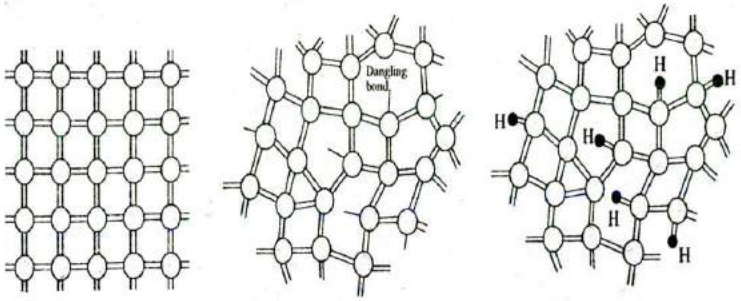
| SOURCE: Photo courtesy of the Estate of Fritz Goro.

phase, this crystal structure looks like the sketch in Figure 1.57b, in which the bonding requirements are only locally satisfied. The crystal can be prepared by condensation from the vapor phase, or by cooling the melt. The vapor-grown films of amorphous As_2Se_3 are used in some photoconductor drums in the photocopying industry.

1.11.2 CRYSTALLINE AND AMORPHOUS SILICON

A silicon atom in the silicon crystal forms four tetrahedrally oriented, covalent bonds with four neighbors, and the repetition of this exact bonding geometry with a well-defined bond length and angle leads to the diamond structure shown in Figure 1.6. A simplified two-dimensional sketch of the Si crystal is shown in Figure 1.59. The crystal has a clear long-range order. Single crystals of Si are commercially grown by the Czochralski crystal pulling technique.

It is also possible to grow amorphous silicon, denoted by a-Si, by the condensation of Si vapor onto a solid surface, called a substrate. For example, an electron beam is used to vaporize a silicon target in a vacuum; the Si vapor then condenses on a metallic substrate to form a thin layer of solid noncrystalline silicon. The technique, which is schematically depicted in Figure 1.60, is referred to as **electron beam deposition.** The structure of amorphous Si (a-Si) lacks the long-range order of crystalline Si (c-Si), even though each Si atom in a-Si, on average, prefers to bond



(a) Two-dimensional schematic representation of a silicon crystal.

(b) Two-dimensional schematic representation of the structure of amorphous silicon. The structure has voids and dangling bonds and there is no long-range order.

(c) Two-dimensional schematic representation of the structure of hydrogenated amorphous silicon. The number of hydrogen atoms shown is exaggerated.

Figure 1.59 Silicon can be grown as a semiconductor crystal or as an amorphous semiconductor film. Each line represents an electron in a bond. A full covalent bond has two lines, and a broken bond has one line.

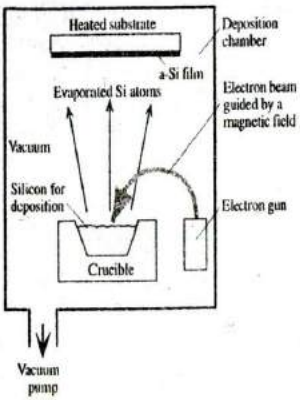


Figure 1.60 Amorphous silicon, a-Si, can be prepared by an electron beam evaporation of silicon.

Silicon has a high melting temperature, so an energetic electron beam is used to melt the crystal in the crucible locally and thereby vaporize Si atoms. Si atoms condense on a substrate placed above the crucible, to form a film of a-Si.

with four neighbors. The difference is that the relative angles between the Si–Si bonds in a-Si deviate considerably from those in the crystal, which obey a strict geometry. Therefore, as we move away from a reference atom in a-Si, eventually the periodicity for generating the crystalline structure is totally lost, as illustrated schematically in Figure 1.59. Furthermore, because the Si–Si bonds do not follow the equilibrium geometry, the bonds are strained and some are even missing, simply because the formation of a bond causes substantial bond bending. Consequently, the a-Si structure has many voids and incomplete bonds, or **dangling bonds,** as schematically depicted in Figure 1.59.

One way to reduce the density of dangling bonds is simply to terminate a dangling bond using hydrogen. Since hydrogen only has one electron, it can attach itself to a

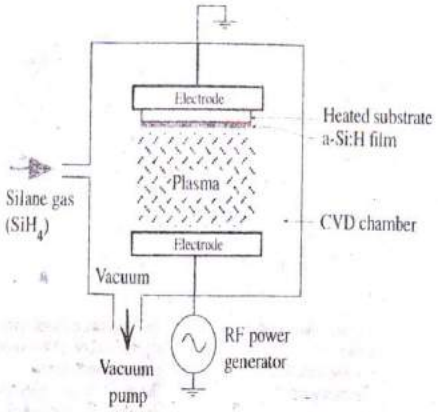


Figure 1.61 Hydrogenated amorphous silicon, a-Si:H, is generally prepared by the decomposition of silane molecules in a radio frequency (RF) plasma discharge. Si and H atoms condense on a substrate to form a film of a-Si:H.

dangling bond, that is, passivate the dangling bond. The structure resulting from hydrogen in amorphous silicon is called **hydrogenated amorphous Si (a-Si:H).**

Many electronic devices, such as a-Si:H solar cells, are based on a-Si being deposited with H to obtain a-Si:H, in which the hydrogen concentration is typically 10 at.% (atomic %). The process involves the decomposition of silane gas, SiH_4, in an electrical plasma in a vacuum chamber. Called **plasma-enhanced chemical vapor deposition (PECVD)**, the process is illustrated schematically in Figure 1.61. The silane gas molecules are dissociated in the plasma, and the Si and H atoms then condense onto a substrate to form a film of a-Si:H. If the substrate temperature is too hot, the atoms on the substrate surface will have sufficient kinetic energy, and hence the atomic mobility, to orient themselves to form a polycrystalline structure. Typically, the substrate temperature is ~250 °C. The advantage of a-Si:H is that it can be grown on large areas, for such applications as photovoltaic cells, flat panel thin-film transistor (TFT) displays, and the photoconductor drums used in some photocopying machines. Table 1.5 summarizes the properties of crystalline and amorphous silicon, in terms of structure and applications.

Table 1.5 Crystalline and amorphous silicon

	Crystalline Si (c-Si)	Amorphous Si (a-Si)	Hydrogenated a-Si (a-Si:H)
Structure	Diamond cubic.	Short-range order only. On average, each Si covalently bonds with four Si atoms. Has microvoids and dangling bonds.	Short-range order only. Structure typically contains 10% H. Hydrogen atoms passivate dangling bonds and relieve strain from bonds.
Typical preparation	Czochralski technique.	Electron beam evaporation of Si.	Chemical vapor deposition of silane gas by RF plasma.
Density (g cm^{-3})	2.33	About 3–10% less dense.	About 1–3% less dense.
Electronic applications	Discrete and integrated electronic devices.	None	Large-area electronic devices such as solar cells, flat panel displays, and some photoconductor drums used in photocopying.

1.12 SOLID SOLUTIONS AND TWO-PHASE SOLIDS

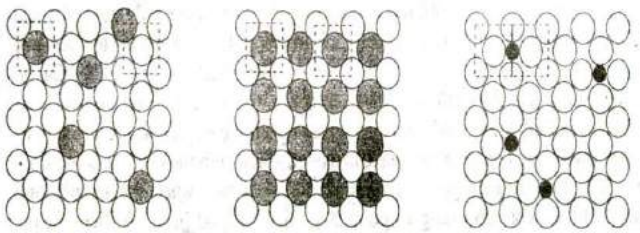
1.12.1 ISOMORPHOUS SOLID SOLUTIONS: ISOMORPHOUS ALLOYS

A **phase** of a material has the same composition, structure, and properties everywhere, so it is a homogeneous portion of the chemical system under consideration. In a given chemical system, one phase may be in contact with another phase. For example, at 0 °C, iced water will have solid and liquid phases in contact. Each phase, ice and water, has a distinct structure.

A bartender knows that alcohol and water are totally miscible; she can dilute whisky with as much water as she likes. When the two liquids are mixed, the molecules are randomly mixed with each other and the whole liquid is a homogenous mixture of the molecules. The liquid therefore has one phase; the properties of the liquid are the same everywhere. The same is not true when we try to mix water and oil. The mixture consists of two distinctly separate phases, oil and water, in contact. Each phase has a different composition, even though both are liquids.

Many solids are a homogeneous mixture of two types of separate atoms. For example, when nickel atoms are added to copper, Ni atoms substitute directly for the Cu atoms, and the resulting solid is a **solid solution,** as depicted in Figure 1.62a. The structure remains an FCC crystal whatever the amount of Ni we add, from 100% Cu to 100% Ni. The solid is a homogenous mixture of Cu and Ni atoms, with the same structure everywhere in the solid solution, which is called an **isomorphous solid solution.** The atoms in the majority make up **the solvent,** whereas the atoms in the minority are the **solute,** which is dissolved in the solvent. For a Cu–Ni alloy with a Ni content of less than 50 at.%, copper is the solvent and nickel is the solute.

The substitution of solute atoms for solvent atoms at various lattice sites of the solvent can be either random (disordered) or ordered. The two cases are schematically illustrated in Figure 1.62a and b, respectively. In many solid solutions, the substitution is random, but for certain compositions, the substitution becomes ordered. There is a



(a) Disordered substitutional solid solution. Example: Cu-Ni alloys ({100} planes)

(b) Ordered substitutional solid solution. Example: Cu–Zn alloy of composition 50% Cu–50% Zn. ({110} planes)

(c) Interstitial solid solution. Example: Small number of C atoms in FCC Fe (austenite). ({100} planes)

Figure 1.62 Solid solutions can be disordered substitutional, ordered substitutional, and interstitial substitutional.

Only one phase within the alloy has the same composition, structure, and properties everywhere.

distinct ordering of atoms around each solute atom such that the crystal structure resembles that of a compound. For example, β' brass has the composition 50 at.% Cu–50 at.% Zn. Each Zn atom is surrounded by eight Cu atoms and vice versa, as depicted in two dimensions in Figure 1.62b. The structure is that of a metallic compound between Cu and Zn.

Another type of solid solution is the **interstitial solid solution,** in which solute atoms occupy interstitial sites, or voids between atoms, in the crystal. Figure 1.62c shows an example in which a small number of carbon atoms have been dissolved in a γ-iron crystal (FCC) at high temperatures.

1.12.2 PHASE DIAGRAMS: Cu–Ni AND OTHER ISOMORPHOUS ALLOYS

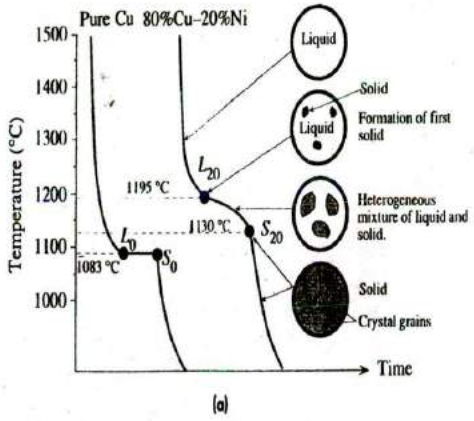
The Cu–Ni alloy is isomorphous. Unlike pure copper or pure nickel, when a Cu–Ni alloy melts, its melting temperature is not well defined. The alloy melts over a range of temperatures in which both the liquid and the solid coexist as a heterogeneous mixture. It is therefore instructive to know the phases that exist in a chemical system at various temperatures as a function of composition, and this need leads to the use of phase diagrams.

Suppose we take a crucible of molten copper and allow it to cool. Above its melting temperature (1083 °C), there is only the liquid phase. The temperature drops with time, as shown in Figure 1.63a, until at the melting or fusion temperature at point L_0 when copper crystals begin to **nucleate** (solidify) in the crucible. During solidification, the temperature remains constant. As long as we have both the liquid and solid phases coexisting, the temperature remains constant at 1083 °C. During this time, heat is given off as the Cu atoms in the melt attach themselves to the Cu crystals. This heat is called the **heat of fusion.** Once all the liquid has solidified (point S_0), the temperature begins to drop as the solid cools. There is therefore a sharp melting temperature for copper, at 1083 °C.

If we were to cool pure nickel from its melt, we would observe a behavior similar to that of pure copper, with a well-defined melting temperature at 1453 °C.

Now suppose we cool the melt of a Cu–Ni alloy with a composition[16] of 80 wt.% Cu and 20 wt.% Ni. In the melt, the two species of atoms are totally miscible, and there is only a single liquid phase. As the cooling proceeds, we reach the temperature 1195 °C, identified as point L_{20} in Figure 1.63a, where the first crystals of Cu–Ni alloy begin to appear. In this case, however, the temperature does not remain constant until the liquid is solidified, but continues to drop. Thus, there is no single melting temperature, but a range of temperatures over which both the liquid and the solid phases coexist in a heterogeneous mixture. We find that when the temperature reaches 1130 °C, corresponding to point S_{20}, all the liquid has solidified. Below 1130 °C, we have a single-phase solid that is an isomorphous solid solution of Cu and Ni. If we repeat these experiments for other compositions, we find a similar behavior; that is, freezing occurs over a transition temperature range. The beginning and end

[16] In materials science, we generally prefer to give alloy composition in wt.%, which henceforth will simply be %.



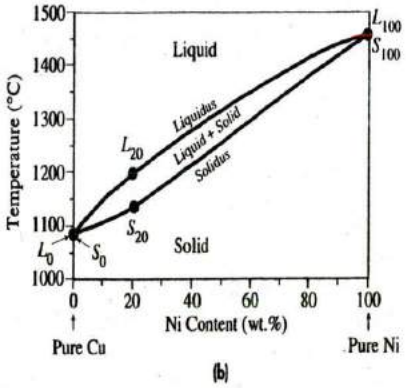


Figure 1.63 Solidification of an isomorphous alloy such as Cu–Ni.
(a) Typical cooling curves.
(b) The phase diagram marking the regions of existence for the phases.

of solidification, at points L and S, respectively, depend on the specific composition of the alloy.

To characterize the freezing or melting behavior of other compositions of Cu–Ni alloys, we can plot the temperatures for the beginning and end of solidification versus the composition and identify those temperature regions where various phases exist, as shown in Figure 1.63b. When we join all the points corresponding to the beginning of freezing, that is, all the L points, we obtain what is called the **liquidus curve.** For any given composition, only the liquid phase can exist above the liquidus curve. If we join all the points where the liquid has totally solidified, that is, all the S points, we have a curve called the **solidus curve.** At any temperature and composition below the solidus curve, we can only have the solid phase. The region between

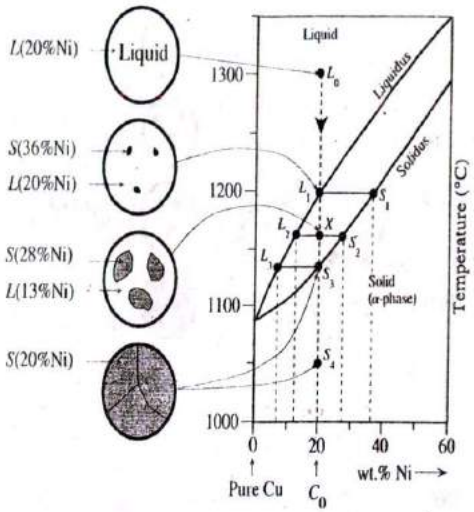


Figure 1.64 Cooling of a 80% Cu–20% Ni alloy from the melt to the solid state.

the liquidus and solidus curves marks where a heterogeneous mixture of liquid and solid phases exists.

Let's follow the cooling behavior of the 80% Cu–20% Ni alloy from the melt at 1300 °C down to the solid state at 1000 °C, as shown in Figure 1.64. The vertical dashed line at 20% Ni represents the overall composition of the alloy (the whole chemical system) and the cooling process corresponds to movement down this dashed line, starting from the liquid phase at L_0.

When the Cu–Ni alloy begins to solidify at 1195 °C, at point L_1, the first solid that forms is richer in Ni content. The only solid that can exist at this temperature has a composition S_1, which has a greater Ni content than the liquid, as shown in Figure 1.64. Intuitively, we can see this by noting that Cu, the component with the lower melting temperature, prefers to remain in the liquid, whereas Ni, which has a higher melting temperature, prefers to remain in the solid. When the temperature drops further, say to 1160 °C (indicated by X in the figure), the alloy is a heterogeneous mixture of liquid and solid. At this temperature, the only solid that can coexist with the liquid has a composition S_2. The liquid has the composition L_2. Since the liquid has lost some of its Ni atoms, the liquid composition is less than that at L_1. The liquidus and solidus curves therefore give the compositions of the liquid and solid phases coexisting in the heterogeneous mixture during melting.

At 1160 °C, the overall composition of the alloy (the whole chemical system) is still 20% Ni and is represented by point X in the phase diagram. When the temperature reaches 1130 °C, nearly all the liquid has been solidified. The solid has the composition S_3, which is 20% Ni, as we expect since the whole alloy is almost all solid. The last drops of the liquid in the alloy have the composition L_3, since at this temperature,

Table 1.6 Phase in the 80% Cu–20% Ni isomorphous alloy

Temperature, °C	Phases	Composition	Amount
1300	Liquid only	$L_0 = 20\%$ Ni	100%
1195	Liquid and solid	$L_1 = 20\%$ Ni	100%
		$S_1 = 36\%$ Ni	First solid appears
1160	Liquid and solid	$L_2 = 13\%$ Ni	53.3%
		$S_2 = 28\%$ Ni	46.7%
1130	Liquid and solid	$L_3 = 7\%$ Ni	The last liquid drop
		$S_3 = 20\%$ Ni	100%
1050	Solid only	$S_4 = 20\%$ Ni	100%

only the liquid with this composition can coexist with the solid at S_3. Table 1.6 summarizes the phases and their compositions, as observed during the cooling process depicted in Figure 1.64. By convention, all solid phases that can exist are labeled by different Greek letters. Since we can only have one solid phase, this is labeled the α-phase.

During the solidification process depicted in Figure 1.64, the solid composition changes from S_1 to S_2 to S_3. We tacitly assume that the cooling is sufficiently slow to allow time for atomic diffusion to change the composition of the whole solid. Therefore, the phase diagram in Figure 1.63b, which assumes near equilibrium conditions during cooling, is termed an **equilibrium phase diagram.** If the cooling is fast, there will be limited time for atomic diffusion in the solid phase, and the resulting solid will have a composition variation. The inner core will correspond to the solidification at S_1 and will be Ni rich. Since the solidification occurs quickly, the Ni atoms do not have time to diffuse out from the inner core to allow the composition in the solid to change from S_1 to S_2 to S_3. Thus, the outer region, the final solidification, will be Ni deficient (or Cu rich); its composition is not S_3 but less, because S_3 is the average composition in the whole solid. The solid structure will be **cored,** as depicted in Figure 1.65. The cooling process is then said to have occurred under nonequilibrium conditions, which leads to a segregation of the elements in the grains. Under nonequilibrium cooling conditions we cannot quantitatively use the equilibrium phase diagram in Figure 1.63b. The diagram can only serve as a qualitative guide.

The amounts of liquid and solid in the mixture can be determined from the phase diagram using the **lever rule,** which is based on the fact that the total mass of the alloy

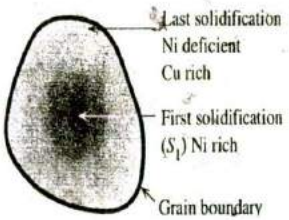


Figure 1.65 Segregation in a grain due to rapid cooling (nonequilibrium cooling).

remains the same throughout the entire cooling process. Let W_L and W_S be the **weight (or mass) fraction** of the liquid and solid phases in the alloy mixture. The compositions of the liquid and solid are denoted as C_L and C_S, respectively. The overall composition of the alloy is denoted C_O, which is the overall weight fraction of Ni in the alloy.

If we take the alloy to have a weight of unity, then the conservation of mass means that

$$W_L + W_S = 1$$

Further, the weight fraction of Ni in both the liquid and solid must add up to the composition C_O of Ni in the whole alloy, or

$$C_L W_L + C_S W_S = C_O$$

We can substitute for W_S in the above equation to find the weight fraction of the liquid and then that of the solid phase, as follows:

Lever rules

$$W_L = \frac{C_S - C_O}{C_S - C_L} \quad \text{and} \quad W_S = \frac{C_O - C_L}{C_S - C_L} \qquad [1.36]$$

To apply Equation 1.36, we first draw a line (called a **tie line**) from L_2 to S_2 corresponding to C_L and C_S, as shown in Figure 1.64. The line represents a "horizontal lever" and point X at C_O at this temperature is the lever's fulcrum. The lengths of the lever arms from the fulcrum to the liquidus and solidus curves are $(C_O - C_L)$ and $(C_S - C_O)$, respectively. The lever must be balanced by the weights W_L and W_S attached to the ends. The total length of the lever is $(C_S - C_L)$. At 1160 °C, $C_L = 0.13$ (13% Ni) and $C_S = 0.28$ (28% Ni), so the weight fraction of the liquid phase is

$$W_L = \frac{C_S - C_O}{C_S - C_L} = \frac{0.28 - 0.20}{0.28 - 0.13} = 0.533 \quad \text{or} \quad 53.3\%$$

Similarly, the weight fraction of the solid phase is 1 – 0.533 or 0.467.

1.12.3 ZONE REFINING AND PURE SILICON CRYSTALS

Zone refining is used for the production of high-purity crystals. Silicon, for example, has a high melting temperature, so any impurities present in the crystal decrease the melting temperature. This is similar to the depression of the melting temperature of pure Ni by the addition of Cu, as shown by the right-hand side of Figure 1.63b. We can represent the phase diagram of Si with small impurities as shown in Figure 1.66. Consider what happens if we have a rod of the solid and we melt only the left end by applying heat locally (using RF heating, for example). At the same time, we move the melted zone toward the right by moving the heater. We therefore melt the solid at A and refreeze it at B, as shown in Figure 1.67a.

The solid has an impurity concentration of C_O; when it melts at A, the melt initially also has the same concentration $C_L = C_O$. However, at temperature T_B, the melt

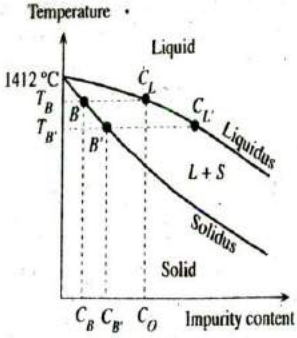
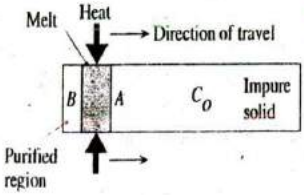
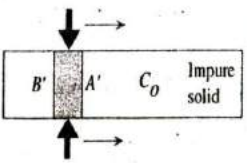


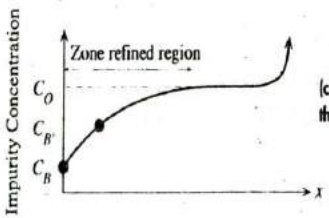
Figure 1.66 The phase diagram of Si with impurities near the low-concentration region.



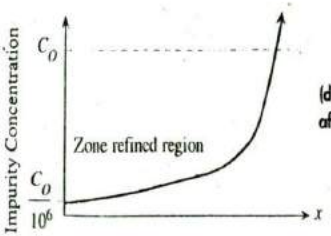
(a) Heat is applied locally starting at one end. The impurity concentration in the refrozen solid at B is $C_B < C_O$. The impurity concentration in the melt is $C_{L'} > C_O$.



(b) As the torch travels toward the right, the refrozen solid at B' has $C_{B'}$ where $C_B < C_{B'} < C_O$. The impurity concentration in the melt is now even greater than $C_{L'}$.



(c) The impurity concentration profile in the refrozen solid after one pass.



(d) Typical impurity concentration profile after many passes.

Figure 1.67 The principle of zone refining.

begins to solidify. At the start of solidification the solid that freezes has a composition C_B, which is considerably less than C_O, as is apparent in Figure 1.66. The cooling at B occurs rapidly, so the concentration C_B cannot adjust to the equilibrium value at the end of freezing. Thus, the solid that freezes at B has a lower concentration of impurities. The impurities have been pushed out of the solid at B and into the melt, whose impurity concentration increases from C_L to $C_{L'}$.

Next, refreezing at B', shown in Figure 1.67b, occurs at a lower temperature $T_{B'}$, because the melt concentration $C_{L'}$ is now greater than C_O. The solid that freezes at B' has the concentration $C_{B'}$, shown in Figure 1.66, which is greater than C_B but less than C_O. As the melted zone is floated toward the right, the melt that is solidified at B, B', etc., has a higher and higher impurity concentration, until its impurity content reaches that of the impure solid, at which point the concentration remains at C_O. When the melted zone approaches the far right where the freezing is halted, the impurities in the final melt appear in the last frozen region at the far right. The resulting impurity concentration profile is schematically depicted in Figure 1.67c. The region of impurity concentration below C_O is the **zone refined** section of the rod. The zone refining procedure can be repeated again, starting from the left toward the right, to reduce the impurity concentration even lower. The impurity concentration profile after many passes is sketched in Figure 1.67d. Although the profile is nonuniform, due to the segregation effect, the impurity concentrations in the zone refined section may be as low as a factor of 10^{-6}.

1.12.4 BINARY EUTECTIC PHASE DIAGRAMS AND Pb–Sn SOLDERS

When we dissolve salt in water, we obtain a brine solution. If we continue to add more salt, we eventually reach the solubility limit of salt in the solution, and the excess salt remains as a solid at the bottom of the container. We then have two coexisting phases: brine (liquid solution) and salt (solid), as shown in Figure 1.68. The solubility limit of one component in another in a mixture is represented by a **solvus curve** shown schematically in Figure 1.68 for salt in brine. In the solid state, there are many elements that can only be dissolved in small amounts in another solid.

Lead in the solid phase has an FCC crystal structure, and tin has a BCT (body-centered tetragonal) structure. Although the two elements are totally miscible in any

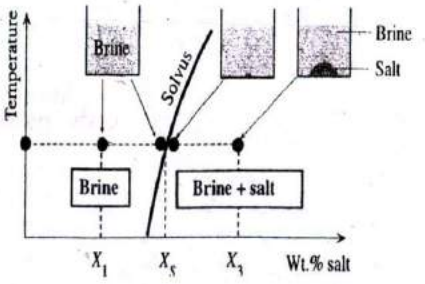


Figure 1.68 We can only dissolve so much salt in brine (solution of salt in water). Eventually we reach the solubility limit at X_S, which depends on the temperature. If we add more salt, then the excess salt does not dissolve and coexists with the brine. Past X_S we have two phases, brine (solution) and salt (solid).

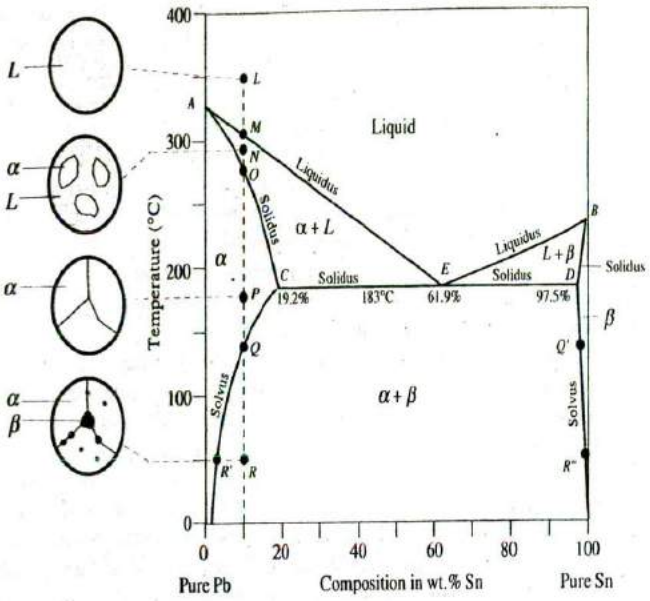


Figure 1.69 The equilibrium phase diagram of the Pb–Sn alloy.
The microstructures on the left show the observations at various points during the cooling of a 90% Pb–10% Sn from the melt along the dashed line (the *overall alloy composition* remains constant at 10% Sn).

proportion when melted, this is not so in the solid state. We can only dissolve so much Sn in solid Pb, and vice versa. We quickly reach the solubility limit, and the resulting solid is a mixture of two distinctly different solid phases. One solid phase, labeled α, is Pb rich and has the FCC structure with some Sn atoms dissolved in the crystal. The amount of Sn dissolved in α is given by the solvus curve of Sn in α at that temperature. The other phase, labeled β, is Sn rich and has the BCT structure with some Pb atoms dissolved in it. The amount of Pb dissolved in β is given by the solvus curve of Pb in β at that temperature.

The existence of various phases and their compositions as a function of temperature are given by the equilibrium phase diagram for the Pb–Sn alloy, shown in Figure 1.69. This is called an equilibrium **eutectic phase diagram.** The liquidus and solidus curves, as usual, mark the borders for the liquid and solid phases. Between the liquidus and solidus curves, we have a heterogeneous mixture of melt and solid. Unlike the Cu–Ni case, the melting temperature of both elements here is depressed with alloying. The liquidus and solidus curves thus decrease from both ends, starting at A and B. They meet at a point E, called the **eutectic point,** at 61.9% Sn and 183 °C. This point has a special significance: No liquid can exist below this temperature, so 183 °C is the lowest melting temperature of the alloy.

In addition, we must insert the solvus curves at both the Pb and Sn ends to mark the extent of solid-state solubility and hence identify the two-phase solid region. The solvus curve for the solubility limit of Sn in Pb meets the solidus curve at point C, 19.2% Sn. Similarly, the solubility limit of Pb in Sn meets the solvus curve at D. A characteristic feature of this phase diagram is that CD is a straight line through E at 183 °C. Below 183 °C, between the two solvus curves, we have a solid with two phases, α and β. This is identified as $\alpha + \beta$ in the diagram.

The usefulness of such a phase diagram is best understood by examining the phase transformations and microstructures during the cooling of a melt of a given composition alloy. Consider a 90% Pb–10% Sn alloy being cooled from the melt at 350 °C (point L) where there is only one phase, the liquid phase. At point M, 315 °C, few nuclei of the α-phase appear in the liquid. The composition of the α-phase is given by the solidus curve at 315 °C and is about 5% Sn. At point N, 290 °C, there is more α-phase in the mixture. The compositions of the liquid and α-phases are given respectively by the liquidus and solidus curves at 290 °C. At point O, 275 °C, all liquid has been solidified into the α-phase, which then has the composition 10% Sn.

Between M and O, the alloy is a coexistent mixture of the liquid phase (melt) and the solid α-phase. At point P, 175 °C, we still have only the α-phase. When we reach the solvus curve at point Q, 140 °C, we can no longer keep all the Sn dissolved in the α-phase, as we have reached the solubility limit of Sn in α. Some of the Sn atoms must diffuse out from the α-phase; they do so by forming a second solid phase, which is the β-phase. The β-phase nucleates within the α-phase (usually at the grain boundaries, where atomic diffusion occurs readily). The β-phase will contain as much dissolved Pb as is allowed by the solubility of Pb in the β-phase, which is given by the solvus curve on the Sn side and marked as point Q', about 98% Sn. Thus, the microstructure is now a mixture of the α and β phases.

As cooling proceeds, the two phases continue to coexist, but their relative proportions change. At R, 50 °C, the alloy is a mixture of the α-phase given by R'(4% Sn) and the β-phase given by R''(99% Sn). The relative amounts of α and β phases are given by the lever rule. Figure 1.69 illustrates the microstructure of the 90% Pb–10% Sn alloy as it is cooled.

An interesting phenomenon can be observed when we cool an alloy of the eutectic composition 38.1% Pb–61.9% Sn from the melt. The cooling process and the observed microstructures are illustrated in Figure 1.70; the microstructures are on the right. The temperature–time profile is also depicted in Figure 1.70. At point L, 350 °C, the alloy is all liquid; as it cools, its temperature drops until point E at 183 °C. At E, the temperature remains constant and a solid phase nucleates within the melt. With time, the amount of solid grows until all the liquid is solidified and the temperature begins to drop again. This behavior is much like that of a pure element, for which melting occurs at a well-defined temperature. This behavior only occurs for the eutectic composition (61.9% Sn), because this is the composition at which the liquidus and solidus curves meet at one temperature. Generally, the liquid with the eutectic composition will solidify through the **eutectic transformation** at the eutectic temperature, or

Eutectic transformation

$$L_{61.9\%\ Sn} \rightarrow \alpha_{19.2\%\ Sn} + \beta_{97.5\%\ Sn} \quad (183\ °C) \qquad [1.37]$$

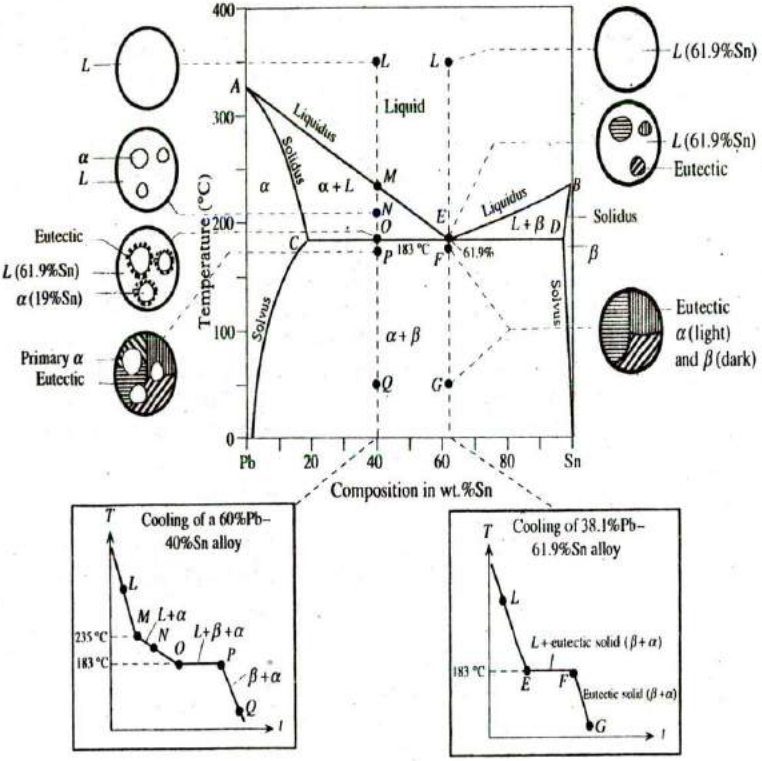


Figure 1.70 The alloy with the eutectic composition cools like a pure element, exhibiting a single solidification temperature at 183 °C.

The solid has the special eutectic structure. The alloy with the composition 60% Pb–40% Sn when solidified is a mixture of primary α and eutectic solid.

The solid that forms from the eutectic solidification has a special microstructure, consisting of alternating plates, or **lamellae,** of α and β phases, as shown in Figure 1.70. This is called the **eutectic microstructure** (or **eutectic solid**). The formation of a Pb-rich α-phase and an Sn-rich β-phase from the 61.9% Sn liquid requires the redistribution of the two types of atoms by atomic diffusion. Atomic diffusions are easier in the liquid than in the solid. The formation of a solid with alternating α and β layers allows the Pb and Sn atoms to diffuse in the liquid without having to move over long distances. The eutectic structure is not a phase itself, but a mixture of the two phases, α and β.

When cooled from the melt, an alloy with a composition between 19.2% Sn and 61.9% Sn solidifies into a mixture of α-phase and a eutectic solid (a mixture of α and β phases). Consider the cooling of an alloy with a composition of 40% Sn, starting from the liquid phase L at 350 °C, as shown in Figure 1.70. At point M (235 °C), the

first solid, the α-phase, nucleates. Its composition is about 15% Sn. At N, 210 °C, the alloy is a mixture of liquid, composition 50% Sn, and α-phase, composition 18% Sn. The composition of the liquid thus moves along the liquidus line from M toward E. At 183 °C, the liquid has the composition 61.9% Sn, or the eutectic composition, and therefore undergoes the eutectic transformation indicated in Equation 1.37. There is still α-phase in the alloy, but its composition is now 19.2% Sn; it does not take part in the eutectic transformation of the liquid. During the eutectic transformation, the temperature remains constant. When all the liquid has been solidified, we have a mixture of the preexisting α-phase, called **primary** α (or **proeutectic** α), and the newly formed eutectic solid. The final microstructure is shown in Figure 1.70 and consists of a primary α and a eutectic solid; therefore, two solid phases, α and β, coexist.

During cooling between points M and O, the alloy 60% Pb–40% Sn is a mixture of melt and α-phase, and it exhibits plastic-like characteristics while solidifying. Further, the temperature range for the solidification is about 183 °C to 235 °C, or about 50 °C. Such an alloy is preferable for such uses as soldering wiped joints to join pipes together, giving the plumber sufficient play for adjusting and wiping the joint. On the other hand, a solder with the eutectic composition (commercially, this is 40% Pb–60% Sn solder, which is close to the eutectic) has the lowest melting temperature and solidifies quickly. The liquid also has good wetting properties. Therefore, 40% Pb–60% Sn is widely used for soldering semiconductor devices, where good wetting and minimal exposure to high temperature are required.

EXAMPLE 1.17 **THE 60% Pb–40% Sn ALLOY** Consider the solidification of the 60% Pb–40% Sn alloy. What are the phases, compositions, and weight fractions of various phases existing in the alloy at 250 °C, 210 °C, 183. 5 °C (just above 183 °C), and 182.5 °C (just below 183 °C)?

SOLUTION

We again refer to the phase diagram in Figure 1.70 to identify which phases exist at what temperatures. At 250 °C, we only have the liquid phase. At 210 °C, point N, the liquid and the α-phase are in equilibrium. The composition of the α-phase is given by the solidus line; at 210 °C, C_α = 18% Sn. The composition of the liquid is given by the liquidus line; at 210 °C, C_L = 50% Sn. To find the weight fraction of α the alloy, we use the lever rule,

$$W_\alpha = \frac{C_L - C_O}{C_L - C_\alpha} = \frac{50 - 40}{50 - 18} = 0.313$$

From $W_\alpha + W_L = 1$, we obtain the weight fraction of the liquid phase, $W_L = 1 - 0.313 = 0.687$.

At 183.5 °C, point O, the composition of the α-phase is 19.2% Sn corresponding to C and that of the liquid is 61.9% Sn corresponding to E. The liquid therefore has the eutectic composition. The weight fractions are

$$W_\alpha = \frac{C_L - C_O}{C_L - C_\alpha} = \frac{61.9 - 40}{61.9 - 19.2} = 0.513$$

$$W_L = 1 - 0.513 = 0.487$$

As expected, the amount of α-phase increases during solidification; at the same time, its composition changes along the solidus curve. Just above 183 °C, about half the alloy is the solid α-phase and the other half is liquid with the eutectic composition. Thus, on solidification, the liquid

Table 1.7 The 60% Pb–40% Sn alloy

Temperature (°C)	Phases	Composition	Mass (g)	Microstructure and Comment
250	L	40% Sn	100	
235	L	40% Sn	100	The first solid (α-phase) nucleates in the
	α	15% Sn	0	liquid.
210	L	50% Sn	68.7	Mixture of liquid and α phases. More solid
	α	18% Sn	31.3	forms. Compositions change.
183.5	L	61.9% Sn	48.7	Liquid has the eutectic composition.
	α	19.2% Sn	51.3	
182.5	α	19.2% Sn	73.4	Eutectic (α and β phases) and primary
	β	97.5% Sn	26.6	α-phase.

[1] Assume mass of the alloy is 100 g.

undergoes the eutectic transformation and forms the eutectic solid. Just below 183 °C, therefore, the microstructure is the primary α-phase and the eutectic solid. Stated differently, below 183 °C, the α and β phases coexist, and β is in the eutectic structure. The weight fraction of the eutectic phase is the same as that of the liquid just above 183 °C, from which it was formed. The weight fractions of α and β in the whole alloy are given by the lever rule applied at point P, or

$$W_\alpha = \frac{C_\beta - C_O}{C_\beta - C_\alpha} = \frac{97.5 - 40}{97.5 - 19.2} = 0.734$$

$$W_\beta = \frac{C_O - C_\alpha}{C_\beta - C_\alpha} = \frac{40 - 19.2}{97.5 - 19.2} = 0.266$$

The microstructure at room temperature will be much like that just below 183 °C, at which the alloy is a two phase solid because atomic diffusions in the solid will not be sufficiently fast to allow the compositions to change. Table 1.7 summarizes the phases that exist in this alloy at various temperatures.

ADDITIONAL TOPICS

1.13 BRAVAIS LATTICES

An infinite periodic array of geometric points in space defines a **space lattice** or simply a **lattice.** Strictly, a lattice does not contain any atoms or molecules because it is simply an imaginary array of geometric points. A two-dimensional *simple square* lattice is shown in Figure 1.71a. In three dimensions, Figure 1.71a would correspond to the simple cubic (SC) lattice. The actual crystal is obtained from the lattice by placing an identical group of atoms (or molecules) at each lattice point. The identical group of atoms is called the **basis** of the crystal structure. Thus, conceptually, as illustrated in Figure 1.71a to c,

$$\textbf{Crystal} = \textbf{Lattice} + \textbf{Basis}$$

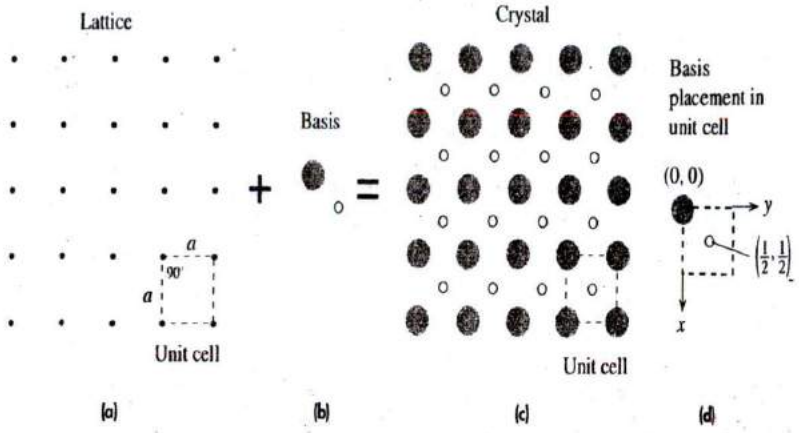


Figure 1.71

(a) A simple square lattice. The unit cell is a square with a side a.

(b) Basis has two atoms.

(c) Crystal = Lattice + Basis. The unit cell is a simple square with two atoms.

(d) Placement of basis atoms in the crystal unit cell.

The unit cell of the two-dimensional lattice in Figure 1.71a is a square which is characterized by the length a of one of the sides; a is called a **lattice parameter.** A given lattice can generate different *patterns* of atoms depending on the basis. The lattice in Figure 1.71a with the two-atom basis in Figure 1.71b produces the crystal in Figure 1.71c. Although the latter crystal appears as a body-centered square (similar to BCC in three dimensions), it is nonetheless a *simple square lattice* with two atoms comprising the basis. Suppose that the basis had only one atom; then the crystal would appear as the simple square lattice in Figure 1.71a (with each point now being an atom). The *patterns* in Figure 1.71a and c are different but the underlying lattice is the same. Because they have the same lattice, the two crystals would have certain identical symmetries. For example, for both crystals, a rotation by 90° about a lattice point would produce the same crystal structure.

To fully characterize the crystal, we also have to specify the locations of the basis atoms in the unit cell as in Figure 1.71d. By convention, we place a Cartesian coordinate system at the rear-left corner of the unit cell with the x and y axes along the square edges. We indicate the coordinates (x_i, y_i) of each ith atom in terms of the lattice parameters along x and y. Thus, the atoms in the unit cell in Figure 1.71d are at (0, 0) and at $(\frac{1}{2}, \frac{1}{2})$. The CsCl unit cell in Figure 1.38 appears as BCC, but it can be described by a SC lattice and a basis that has one Cl^- ion and one Cs^+ ion. The ions in the SC unit cell are located at (0, 0, 0) and at the cell center at $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. Similarly, the NaCl crystal in Figure 1.37 is an FCC lattice with a basis of Na^+ and Cl^- ions.

The diamond unit cell of silicon is an FCC lattice with two Si atoms constituting the basis. The two Si atoms are placed at (0, 0, 0) and $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Most of the important

III–V compound semiconductors such as GaAs, AlAs, InAs, InP, etc., which are widely used in numerous optoelectronic devices, have the zinc blende (ZnS) unit cell. The zinc blende unit cell consists of an FCC lattice and a basis that has the Zn and S atoms placed at (0, 0, 0) and $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, respectively.

We generally represent the *geometry* of the unit cell of a lattice as a parallelepiped with sides a, b, c and angles α, β, γ as depicted in Figure 1.40*a*. In the case of copper and iron, the geometry of the unit cell has $a = b = c$, $\alpha = \beta = \gamma = 90°$, and cubic

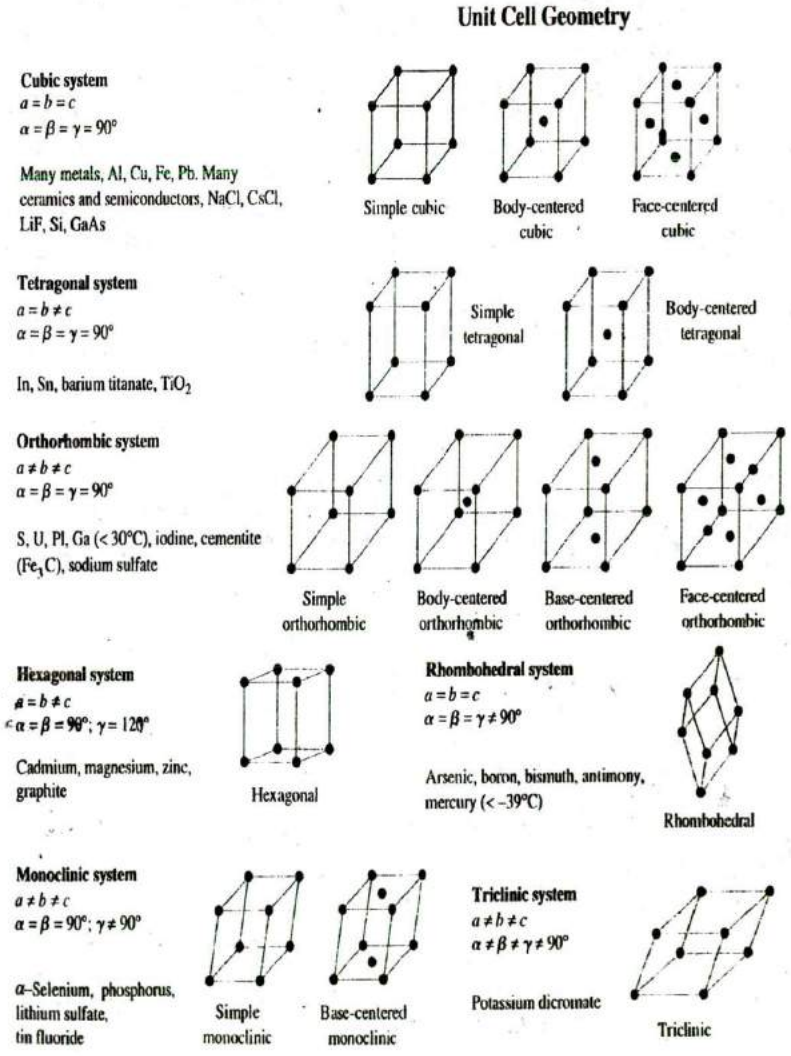


Figure 1.72 The seven crystal systems (unit-cell geometries) and fourteen Bravais lattices.

symmetry. For Zn, the unit cell has hexagonal geometry with $a = b \neq c$, $\alpha = \beta = 90°$, and $\gamma = 120°$ as shown in Figure 1.33d. Based on different lattice parameters, there are *seven* possible distinct unit-cell geometries, which we call **crystal systems** each with a particular distinct symmetry. The seven crystal systems are depicted in Figure 1.72 with typical examples. We are already familiar with the cubic and hexagonal systems. The seven crystal systems only categorize the unit cells based on the geometry of the unit cell and not in terms of the symmetry and periodicity of the lattice points. (One should not confuse the unit-cell geometry with the lattice, which is a periodic array of points.) In the cubic system, for example, there are three possible distinct lattices corresponding to SC, BCC, and FCC which are shown in Figure 1.72. All three have the same cubic geometry: $a = b = c$ and $\alpha = \beta = \gamma = 90°$.

Many distinctly different lattices, or distinct patterns of points, exist in three dimensions. There are 14 distinct lattices whose unit cells have one of the seven geometries as indicated in Figure 1.72. Each of these is called a **Bravais lattice.** The copper crystal, for example, has the FCC Bravais lattice, but arsenic, antimony, and bismuth crystals have the rhombohedral Bravais lattice. Tin's unit cell belongs to the **tetragonal** crystal system, and its crystal lattice is a **body-centered tetragonal (BCT).**

CD Selected Topics and Solved Problems

Selected Topics

Units and Conversions
Bonding: Bond Energies and Elastic Moduli
Secondary Bonding
Cohesive Energy: Ionic Bonding and Madelung Constant
Elementary Crystals
X-Ray Diffraction and Crystal Structures
Essential Mechanical Properties
Diffusion
Diffusion and Oxidation
Thermal Expansion
Surface Tension of Crystals

Solved Problems

van der Waals Bonding: Secondary Bonding and Bulk Modulus
Elementary Concepts in Material Science: Mean Atomic Separation, Bulk and Surface Atomic Concentrations, and Density
Elementary Crystals
Ionic Crystals

DEFINING TERMS

Activated state is the state that occurs temporarily during a transformation or reaction when the reactant atoms or molecules come together to form a particular arrangement (intermediate between reactants and products) that has a higher potential energy than the reactants. The potential energy barrier between the activated state and the reactants is the activation energy.

Activation energy is the potential energy barrier against the formation of a product. In other words, it is the minimum energy that the reactant atom or molecule must have to be able to reach the activated state and hence form a product.

Amorphous solid is a solid that exhibits no crystalline structure or long-range order. It only possesses a

short-range order in the sense that the nearest neighbors of an atom are well defined by virtue of chemical bonding requirements.

Anion is an atom that has gained negative charge by virtue of accepting one or more electrons. Usually, atoms of nonmetallic elements can gain electrons easily to become anions. Anions become attracted to the anode (positive terminal) in ionic conduction. Typical anions are the halogen ions F^-, Cl^-, Br^-, and I^-.

Atomic mass (or **relative atomic mass** or **atomic weight**) M_{at} of an element is the average atomic mass, in atomic mass units (amu), of all the naturally occurring isotopes of the element. Atomic masses are listed in the Periodic Table. The amount of an element that has 6.022×10^{23} atoms (the Avogadro number of atoms) has a mass in grams equal to the atomic mass.

Atomic mass unit (amu) is a convenient mass measurement equal to one-twelfth of the mass of a neutral carbon atom that has a mass number of $A = 12$ (6 protons and 6 neutrons). It has been found that amu = 1.66054×10^{27} kg, which is equivalent to $10^{-3}/N_A$, where N_A is Avogadro's number.

Atomic packing factor (APF) is the fraction of volume actually occupied by atoms in a crystal.

Avogadro's number (N_A) is the number of atoms in exactly 12 g of carbon-12. It is 6.022×10^{23}. Since atomic mass is defined as one-twelfth of the mass of the carbon-12 atom, the N_A number of atoms of any substance has a mass equal to the atomic mass M_{at}, in grams.

Basis represents an atom, a molecule, or a collection of atoms, that is placed at each lattice point to generate the true crystal structure of a substance. All crystals are thought of as a lattice with each point occupied by a basis.

Bond energy or **binding energy** is the work (or energy) needed to separate two atoms infinitely from their equilibrium separation in the molecule or solid.

Bulk modulus K is volume stress (pressure) needed per unit elastic volume strain and is defined by $p = -K\Delta$, where p is the applied volume stress (pressure) and Δ is the volume strain. K indicates the extent to which a body can be reversibly (and hence elastically) deformed in volume by an applied pressure.

Cation is an atom that has gained positive charge by virtue of losing one or more electrons. Usually, metal atoms can lose electrons easily to become cations. Cations become attracted to the cathode (negative terminal) in ionic conduction, as in gaseous discharge. The alkali metals, Li, Na, K, ..., easily lose their valence electron to become cations, Li^+, Na^+, K^+, ...

Coordination number is the number of nearest neighbors around a given atom in the crystal.

Covalent bond is the sharing of a pair of valence electrons between two atoms. For example, in H_2, the two hydrogen atoms share their electrons, so that each has a closed shell.

Crystal is a three-dimensional periodic arrangement of atoms, molecules, or ions. A characteristic property of the crystal structure is its periodicity and a degree of symmetry. For each atom, the number of neighbors and their exact orientations are well defined; otherwise the periodicity will be lost. Therefore, a long-range order results from strict adherence to a well-defined bond length and relative bond angle (that is, exact orientation of neighbors).

Crystallization is a process by which crystals of a substance are formed from another phase of that substance. Examples are solidification just below the fusion temperature from the melt, or condensation of the molecules from the vapor phase onto a substrate. The crystallization process initially requires the formation of small crystal nuclei, which contain a limited number (perhaps 10^3–10^4) of atoms or molecules of the substance. Following nucleation, the nuclei grow by atomic diffusion from the melt or vapor.

Diffusion is the migration of atoms by virtue of their random thermal motions.

Diffusion coefficient is a measure of the rate at which atoms diffuse. The rate depends on the nature of the diffusion process and is typically temperature dependent. The diffusion coefficient is defined as the magnitude of diffusion flux per unit concentration gradient.

Dislocation is a line imperfection within a crystal that extends over many atomic distances.

Edge dislocation is a line imperfection within a crystal that occurs when an additional, short plane of atoms

does not extend as far as its neighbors. The edge of this short plane constitutes a line of atoms where the bonding is irregular, that is, a line of imperfection called an edge dislocation.

Elastic modulus or **Young's modulus** (Y) is a measure of the ease with which a solid can be elastically deformed. The greater Y is, the more difficult it is to deform the solid elastically. When a solid of length ℓ is subjected to a tensile stress σ (force per unit area), the solid will extend elastically by an amount $\delta\ell$ where $\delta\ell/\ell$ is the strain ε. Stress and strain are related by $\sigma = Y\varepsilon$, so Y is the stress needed per unit elastic strain.

Electric dipole moment is formed when a positive charge $+Q$ is separated from a negative charge $-Q$ of equal magnitude. Even though the net charge is zero, there is nonetheless an electric dipole moment formed by the two charges $-Q$ and $+Q$ being separated by a finite distance. Just as two charges exert a Coulombic force on each other, two dipoles also exert an electrostatic force on each other that depends on the separation of dipoles and their relative orientation.

Electron affinity represents the energy that is needed to add an electron to a neutral atom to create a negative ion (*anion*). When an electron is added to Cl to form Cl^-, energy is actually released.

Electronegativity is a relative measure of the ability of an atom to attract the electrons in a bond it forms with another atom. The *Pauling scale of electronegativity* assigns an electronegativity value (a pure number) X to various elements, the highest being 4 for F, and the lowest values being for the alkali metal atoms, for which X are less than 1. The difference $X_A - X_B$ in the electronegativities of two atoms A and B is a measure of the polar or ionic character of the bond A–B between A and B. A molecule A–B would be polar, that is, possess a dipole moment, if X_A and X_B are different.

Equilibrium between two systems requires mechanical, thermal, and chemical equilibrium. Mechanical equilibrium means that the pressure should be the same in the two systems, so that one does not expand at the expense of the other. Thermal equilibrium implies that both have the same temperature. Equilibrium within a single-phase substance (*e.g.*, steam only or hydrogen gas only) implies uniform pressure and temperature within the system.

Equilibrium state of a system is the state in which the pressure and temperature in the system are uniform throughout. We say that the system possesses mechanical and thermal equilibrium.

Eutectic composition is an alloy composition of two elements that results in the lowest melting temperature compared to any other composition. A eutectic solid has a structure that is a mixture of two phases. The eutectic structure is usually special, such as alternating lamellae.

Face-centered cubic (FCC) lattice is a cubic lattice that has one lattice point at each corner of a cube and one at the center of each face. If there is a chemical species (atom or a molecule) at each lattice point, then the structure is an FCC crystal structure.

Frenkel defect is an ionic crystal imperfection that occurs when an ion moves into an interstitial site, thereby creating a vacancy in its original site. The imperfection is therefore a pair of point defects.

Grain is an individual crystal within a polycrystalline material. Within a grain, the crystal structure and orientation are the same everywhere and the crystal is oriented in one direction only.

Grain boundary is a surface region between differently oriented, adjacent grain crystals. The grain boundary contains a lattice mismatch between adjacent grains.

Heat is the amount of energy transferred from one system to another (or between the system and its surroundings) as a result of a temperature difference. Heat is not a new form of energy, but rather the transfer of energy from one body to another by virtue of the random motions of their molecules. When a hot body is in contact with a cold body, energy is transferred from the hot body to the cold one. The energy that is transferred is the excess mean kinetic energy of the molecules in the hot body. Molecules in the hot body have a higher mean kinetic energy and vibrate more violently. As a result of the collisions between the molecules, there is a net transfer of energy (heat) from the hot body to the cold one, until the molecules in both bodies have the same mean kinetic energy, that is, until their temperatures become equal.

Heat capacity at constant volume is the increase in the total energy E of the system per degree increase in the

temperature of the system with the volume remaining constant: $C = (\partial E/\partial T)_V$. Thus, the heat added to the system does no mechanical work due to a volume change but increases the internal energy. **Molar heat capacity** is the heat capacity for 1 mole of a substance. **Specific heat capacity** is the heat capacity per unit mass.

Interstitial site (interstice) is an unoccupied space between the atoms (or ions, or molecules) in a crystal.

Ionization energy is the energy required to remove an electron from a neutral atom; normally the most outer electron that has the least binding energy to the nucleus is removed to ionize an atom.

Isomorphous describes a structure that is the same everywhere (from *iso*, uniform, and *morphology*, structure).

Isotropic substance is a material that has the same property in all directions.

Kinetic molecular theory assumes that the atoms and molecules of all substances (gases, liquids, and solids) above absolute zero of temperature are in constant motion. Monatomic molecules (*e.g.*, He, Ne) in a gas exhibit constant and random translational motion, whereas the atoms in a solid exhibit constant vibrational motion.

Lattice is a regular array of points in space with a discernible periodicity. There are 14 distinct lattices possible in three-dimensional space. When an atom or molecule is placed at each lattice point, the resulting regular structure is a crystal structure.

Lattice parameters are (a) the lengths of the sides of the unit cell, and (b) the angles between the sides.

Mechanical work is qualitatively defined as the energy expended in displacing a constant force through a distance. When a force **F** is moved a distance $d\mathbf{x}$, work done $dW = \mathbf{F} \cdot d\mathbf{x}$. When we lift a body such as an apple of mass m (100 g) by a distance h (1 m), we do work by an amount $F\ \Delta x = mgh$ (1 J), which is then stored as the gravitational potential energy of the body. We have transferred energy from ourselves to the potential energy of the body by exchanging energy with it in the form of work. Further, in lifting the apple, the molecules have been displaced in orderly fashion, all upwards. Work therefore involves an orderly displacement of atoms and molecules of a substance in complete contrast to heat. When the volume V of a substance changes by dV when the pressure is P, the mechanical work involved is $P\,dV$ and is called the ***PV* work.**

Metallic bonding is the binding of metal atoms in a crystal through the attraction between the positive metal ions and the mobile valence electrons in the crystal. The valence electrons permeate the space between the ions.

Miller indices $(hk\ell)$ are indices that conveniently identify parallel planes in a crystal. Consider a plane with the intercepts, x_1, y_1, and z_1, in terms of lattice parameters a, b, and c. (For a plane passing through the origin, we shift the origin or use a parallel plane.) Then, $(hk\ell)$ are obtained by taking the reciprocals of x_1, y_1, and z_1 and clearing all fractions.

Miscibility of two substances is a measure of the mutual solubility of those two substances when they are in the same phase, such as liquid.

Mole of a substance is that amount of the substance that contains N_A number of atoms (or molecules), where N_A is Avogadro's number (6.023×10^{23}). One mole of a substance has a mass equal to its atomic (molecular) mass, in grams. For example, 1 mole of copper contains 6.023×10^{23} atoms and has a mass of 63.55 g.

Phase of a system is a homogeneous portion of the chemical system that has the same composition, structure, and properties everywhere. In a given chemical system, one phase may be in contact with another phase of the system. For example, iced water at 0 °C will have solid and liquid phases in contact. Each phase, solid ice and liquid water, has a distinct structure.

Phase diagram is a temperature versus composition diagram in which the existence and coexistence of various phases are identified by regions and lines. Between the liquidus and solidus lines, for example, the material is a heterogeneous mixture of the liquid and solid phases.

Planar concentration of atoms is the number of atoms per unit area on a given $(hk\ell)$ plane in a crystal.

Polarization is the separation of positive and negative charges in a system, which results in a net electric dipole moment.

Polymorphism or **allotropy** is a material attribute that allows the material to possess more than one crystal structure. Each possible crystal structure is called a polymorph. Generally, the structure of the polymorph depends on the temperature and pressure, as well as on the method of preparation of the solid. (For example, diamond can be prepared from graphite by the application of very high pressures.)

Primary bond is a strong interatomic bond, typically greater then 1 eV/atom, that involves ionic, covalent, or metallic bonding.

Property is a system characteristic or an attribute that we can measure. Pressure, volume, temperature, mass, energy, electrical resistivity, magnetization, polarization, and color are all properties of matter. Properties such as pressure, volume, and temperature can only be attributed to a system of many particles (which we treat as a continuum). Note that heat and work are not properties of a substance; instead, they represent energy transfers involved in producing changes in the properties.

Saturated solution is a solution that has the maximum possible amount of solute dissolved in a given amount of solvent at a specified temperature and pressure.

Schottky defect is an ionic crystal imperfection that occurs when a pair of ions is missing, that is, when there is a cation and anion pair vacancy.

Screw dislocation is a crystal defect that occurs when one portion of a perfect crystal is twisted or skewed with respect to another portion on only one side of a line.

Secondary bond is a weak bond, typically less than 0.1 eV/atom, which is due to dipole–dipole interactions between the atoms or molecules.

Solid solution is a homogeneous crystalline phase that contains two or more chemical components.

Solute is the minor chemical component of a solution; the component that is usually added in small amounts to a solvent to form a solution.

Solvent is the major chemical component of a solution.

Stoichiometric compounds are compounds with an integer ratio of atoms, as in CaF_2, in which two fluorine atoms bond with one calcium atom.

Strain is a relative measure of the deformation a material exhibits under an applied stress. Under an applied tensile (or compressive) stress, strain ε is the change in the length per unit original length. When a shear stress is applied, the deformation involves a shear angle. **Shear strain** is the tangent of the shear angle that is developed by the application of the shearing stress. **Volume strain** Δ is the change in the volume per unit original volume.

Stress is force per unit area. When the applied force F is perpendicular to the area A, stress $\sigma = F/A$ is either tensile or compressive. If the applied force is tangential to the area, then stress is **shear stress,** $\tau = F/A$.

Thermal expansion is the change in the length or volume of a substance due to a change in the temperature. **Linear coefficient of thermal expansion** λ is the fractional change in the length per unit temperature change or $\Delta L/L_o = \lambda\, \Delta T$. Volume coefficient of expansion α_V is the fractional change in the volume per unit temperature change; $\alpha_V \approx 3\lambda$.

Unit cell is the most convenient small cell in a crystal structure that carries the characteristics of the crystal. The repetition of the unit cell in three dimensions generates the whole crystal structure.

Vacancy is a point defect in a crystal, where a normally occupied lattice site is missing an atom.

Valence electrons are the electrons in the outer shell of an atom. Since they are the farthest away from the nucleus, they are the first electrons involved in atom-to-atom interactions.

Young's modulus see **elastic modulus.**

QUESTIONS AND PROBLEMS

1.1 **Virial theorem** The Li atom has a nucleus with a $+3e$ positive charge, which is surrounded by a full $1s$ shell with two electrons, and a single valence electron in the outer $2s$ subshell. The atomic radius of the Li atom is about 0.17 nm. Using the Virial theorem, and assuming that the valence electron sees the nuclear $+3e$ shielded by the two $1s$ electrons, that is, a net charge of $+e$, estimate the ionization energy

of Li (the energy required to free the 2*s* electron). Compare this value with the experimental value of 5.39 eV. Suppose that the actual nuclear charge seen by the valence electron is not $+e$ but a little higher, say $+1.25e$, due to the imperfect shielding provided by the closed 1*s* shell. What would be the new ionization energy? What is your conclusion?

1.2 Atomic mass and molar fractions

a. Consider a multicomponent alloy containing N elements. If $w_1, w_2, \ldots, w_N$ are the weight fractions of components $1, 2, \ldots, N$ in the alloy and $M_1, M_2, \ldots, M_N$ are the respective atomic masses of the elements, show that the atomic fraction of the ith component is given by

$$n_i = \frac{w_i/M_i}{\dfrac{w_1}{M_1} + \dfrac{w_2}{M_2} + \cdots + \dfrac{w_N}{M_N}}.$$

Weight to atomic percentage

b. Suppose that a substance (compound or an alloy) is composed of N elements, $A, B, C, \ldots$ and that we know their atomic (or molar) fractions $n_A, n_B, n_C, \ldots$. Show that the weight fractions $w_A, w_B, w_C, \ldots$ are given by

$$w_A = \frac{n_A M_A}{n_A M_A + n_B M_B + n_C M_C + \cdots}$$

$$w_B = \frac{n_B M_B}{n_A M_A + n_B M_B + n_C M_C + \cdots}$$

Atomic to weight percentage

c. Consider the semiconducting II–VI compound cadmium selenide, CdSe. Given the atomic masses of Cd and Se, find the weight fractions of Cd and Se in the compound and grams of Cd and Se needed to make 100 grams of CdSe.

d. A Se–Te–P glass alloy has the composition 77 wt.% Se, 20 wt.% Te, and 3 wt.% P. Given their atomic masses, what are the atomic fractions of these constituents?

1.3 The covalent bond Consider the H_2 molecule in a simple way as two touching H atoms, as depicted in Figure 1.73. Does this arrangement have a lower energy than two separated H atoms? Suppose that electrons totally correlate their motions so that they move to avoid each other as in the snapshot in Figure 1.73. The radius r_o of the hydrogen atom is 0.0529 nm. The electrostatic potential energy of two charges Q_1 and Q_2 separated by a distance r is given by $Q_1 Q_2/(4\pi\varepsilon_o r)$. Using the virial theorem as in Example 1.1 consider the following:

a. Calculate the total electrostatic potential energy PE of all the charges when they are arranged as shown in Figure 1.73. In evaluating the PE of the whole collection of charges you must consider all pairs of charges and, at the same time, avoid double counting of interactions between the same pair of charges. The total PE is the sum of the following: electron 1 interacting with the proton at a distance r_o on the left, proton at r_o on the right, and electron 2 at a distance $2r_o$ + electron 2 interacting with a proton at r_o and another proton at $3r_o$ + two protons, separated by $2r_o$, interacting with each other. Is this configuration energetically favorable?

b. Given that in the isolated H atom the PE is $2 \times (-13.6\ \text{eV})$, calculate the change in PE in going from two isolated H atoms to the H_2 molecule. Using the virial theorem, find the change in the total energy and hence the covalent bond energy. How does this compare with the experimental value of 4.51 eV?

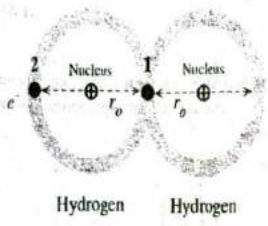


Figure 1.73 A simplified view of the covalent bond in H_2. A snapshot at one instant.

1.4 **Ionic bonding and CsCl** The potential energy E per Cs^+–Cl^- pair within the CsCl crystal depends on the interionic separation r in the same fashion as in the NaCl crystal,

Energy per ion pair in ionic crystals

$$E(r) = -\frac{e^2 M}{4\pi\varepsilon_o r} + \frac{B}{r^m} \qquad [1.38]$$

where for CsCl, $M = 1.763$, $B = 1.192 \times 10^{-104}$ J m^9 or 7.442×10^{-5} eV (nm)^9, and $m = 9$. Find the equilibrium separation (r_o) of the ions in the crystal and the ionic bonding energy, that is, the ionic cohesive energy, and compare the latter value to the experimental value of 657 kJ mol^{-1}. Given that the *ionization energy* of Cs is 3.89 eV and the *electron affinity* of Cl (energy released when an electron is added) is 3.61 eV, calculate the atomic cohesive energy of the CsCl crystal as joules per mole.

1.5 **Madelung constant** If we were to examine the NaCl crystal in three dimensions, we would find that each Na^+ ion has

6 Cl^- ions as *nearest* neighbors at a distance r

12 Na^+ ions as *second* nearest neighbors at a distance $r\sqrt{2}$

8 Cl^- ions as *third* nearest neighbors at a distance $r\sqrt{3}$

and so on. Show that the electrostatic potential energy of the Na^+ atom can be written as

Madelung constant M for NaCl

$$E(r) = -\frac{e^2}{4\pi\varepsilon_o r}\left[6 - \frac{12}{\sqrt{2}} + \frac{8}{\sqrt{3}} - \cdots\right] = -\frac{e^2 M}{4\pi\varepsilon_o r}$$

where M, called the **Madelung constant,** is given by the summation in the square brackets for this particular ionic crystal structure (NaCl). Calculate M for the first three terms and compare it with $M = 1.7476$, its value had we included the higher terms. What is your conclusion?

***1.6** **Bonding and bulk modulus** In general, the potential energy E per atom, or per ion pair, in a crystal as a function of interatomic (interionic) separation r can be written as the sum of an attractive PE and a repulsive PE,

General PE curve for bonding

$$E(r) = -\frac{A}{r^n} + \frac{B}{r^m} \qquad [1.39]$$

where A and n are constants characterizing the attractive PE and B and m are constants characterizing the repulsive PE. This energy is minimum when the crystal is in equilibrium. The magnitude of the minimum energy and its location r_o define the bonding energy and the equilibrium interatomic (or interionic) separation, respectively.

When a pressure P is applied to a solid, its original volume V_o shrinks to V by an amount $\Delta V = V - V_o$. The bulk modulus K relates the volume strain $\Delta V/V$ to the applied pressure P by

Bulk modulus definition

$$P = -K\frac{\Delta V}{V_o} \qquad [1.40]$$

The bulk modulus K is related to the energy curve. In its simplest form (assuming a simple cubic unit cell) K can be estimated from Equation 1.39 by

Bulk modulus

$$K = \frac{1}{9cr_o}\left[\frac{d^2E}{dr^2}\right]_{r=r_o} \qquad [1.41]$$

where c is a numerical factor, of the order of unity, given by b/p where p is the number of atoms or ion pairs in the unit cell and b is a numerical factor that relates the cubic unit cell lattice parameter a_o to the equilibrium interatomic (interionic) separation r_o by $b = a_o^3/r_o^3$.

a. Show that the bond energy and equilibrium separation are given by

$$E_{\text{bond}} = \frac{A}{r_o^n}\left(1 - \frac{n}{m}\right) \quad \text{and} \quad r_o = \left(\frac{Bn}{An}\right)^{1/(m-n)}$$

b. Show that the bulk modulus is given by

$$K = \frac{An}{9cr_o^{n+3}}(m-n) \quad \text{or} \quad K = \frac{mnE_{bond}}{9cr_o^3}$$

c. For a NaCl-type crystal, Na^+ and Cl^- ions touch along the cube edge so that $r_o = (a_o/2)$. Thus, $a^3 = 2^3 r_o^3$ and $b = 2^3 = 8$. There are four ion pairs in the unit cell, $p = 4$. Thus, $c = b/p = 8/4 = 2$. Using the values from Example 1.2, calculate the bulk modulus of NaCl.

***1.7 Van der Waals bonding** Below 24.5 K, Ne is a crystalline solid with an FCC structure. The interatomic interaction energy per atom can be written as

$$E(r) = -2\varepsilon\left[14.45\left(\frac{\sigma}{r}\right)^6 - 12.13\left(\frac{\sigma}{r}\right)^{12}\right] \quad \text{(eV/atom)}$$

where ε and σ are constants that depend on the polarizability, the mean dipole moment, and the extent of overlap of core electrons. For crystalline Ne, $\varepsilon = 3.121 \times 10^{-3}$ eV and $\sigma = 0.274$ nm.

a. Show that the equilibrium separation between the atoms in an inert gas crystal is given by $r_o = (1.090)\sigma$. What is the equilibrium interatomic separation in the Ne crystal?

b. Find the bonding energy per atom in solid Ne.

c. Calculate the density of solid Ne (atomic mass = 20.18).

1.8 Kinetic molecular theory

a. In a particular Ar-ion laser tube the gas pressure due to Ar atoms is about 0.1 torr at 25 °C when the laser is off. What is the concentration of Ar atoms per cm^3 at 25 °C in this laser? (760 torr = 1 atm = 1.013×10^5 Pa.)

b. In the He–Ne laser tube He and Ne gases are mixed and sealed. The total pressure P in the gas is given by contributions arising from He and Ne atoms:

$$P = P_{He} + P_{Ne}$$

where P_{He} and P_{Ne} are the *partial pressures* of He and Ne in the gas mixture, that is, pressures due to He and Ne gases alone,

$$P_{He} = \frac{N_{He}}{N_A}\left(\frac{RT}{V}\right) \quad \text{and} \quad P_{Ne} = \frac{N_{Ne}}{N_A}\left(\frac{RT}{V}\right)$$

In a particular He–Ne laser tube the ratio of He and Ne atoms is 7:1, and the total pressure is about 1 torr at 22 °C. Calculate the concentrations of He and Ne atoms in the gas at 22 °C. What is the pressure at an operating temperature of 130 °C?

1.9 Kinetic molecular theory Calculate the effective (rms) speeds of the He and Ne atoms in the He–Ne gas laser tube at room temperature (300 K).

***1.10 Kinetic molecular theory and the Ar-ion laser** An argon-ion laser has a laser tube that contains Ar atoms that produce the laser emission when properly excited by an electrical discharge. Suppose that the gas temperature inside the tube is 1300 °C (very hot).

a. Calculate the mean speed (v_{av}), rms velocity ($v_{rms} = \sqrt{\overline{v^2}}$), and the rms speed ($v_{rms,x} = \sqrt{\overline{v_x^2}}$) in one particular direction of the Ar atoms in the laser tube, assuming 1300 °C . (See Example 1.10.)

b. Consider a light source that is emitting waves and is moving toward an observer, somewhat like a whistling train moving toward a passenger. If f_o is the frequency of the light waves emitted at the source, then, due to the *Doppler effect*, the observer measures a higher frequency f that depends on the velocity v_{Ar} of the source moving toward the observer and the speed c of light,

$$f = f_o\left(1 + \frac{v_{Ar}}{c}\right)$$

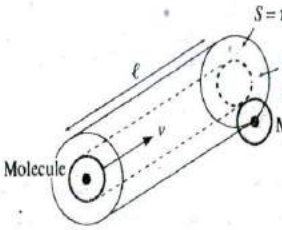
It is the Ar ions that emit the laser output light in the Ar-ion laser. The emission wavelength $\lambda_o = c/f_o$ is 514.5 nm. Calculate the wavelength λ registered by an observer for those atoms that are moving with a mean speed v_{Ar} toward the observer. Those atoms that are moving away from the observer will result in a lower observed frequency because v_{Ar} will be negative. Estimate the width of the wavelengths (the difference between the longest and shortest wavelengths) emitted by the Ar-ion laser.

*1.11 **Vacuum deposition** Consider air as composed of nitrogen molecules N_2.

a. What is the concentration n (number of molecules per unit volume) of N_2 molecules at 1 atm and 27°C?

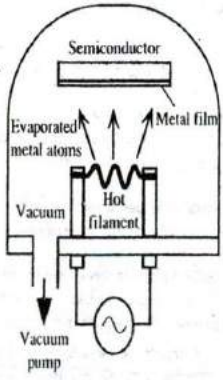
b. Estimate the mean separation between the N_2 molecules.

c. Assume each molecule has a finite size that can be represented by a sphere of radius r. Also assume that ℓ is the **mean free path,** defined as the mean distance a molecule travels before colliding with another molecule, as illustrated in Figure 1.74a. If we consider the motion of one N_2 molecule, with all the others stationary, it is apparent that if the path of the traveling molecule crosses the cross-sectional area $S = \pi(2r)^2$, there will be a collision. Since ℓ is the mean distance between collisions, there must be at least one stationary molecule within the volume $S\ell$,



(a) A molecule moving with a velocity v travels a mean distance ℓ between collisions. Since the collision cross-sectional area is S, in the volume $S\ell$ there must be at least one molecule. Consequently, $n(S\ell) = 1$.





(b) Vacuum deposition of metal electrodes by thermal evaporation.

Figure 1.74

Walter Houser Brattain (1902–1987), experimenting with metal contacts on copper oxide (1935) at Bell Telephone Labs. A vacuum evaporation chamber is used to deposit the metal electrode.

| SOURCE: Bell Telephone Laboratories, courtesy AIP Emilio Segrè Visual Archives.

as shown in Figure 1.74a. Since n is the concentration, we must have $n(S\ell) = 1$ or $\ell = 1/(\pi 4r^2 n)$. However, this must be corrected for the fact that all the molecules are in motion, which only introduces a numerical factor, so that

$$\ell = \frac{1}{2^{1/2} 4\pi r^2 n}$$

Assuming a radius r of 0.1 nm, calculate the mean free path of N_2 molecules between collisions at 27 °C and 1 atm.

d. Assume that an Au film is to be deposited onto the surface of a Si chip to form metallic interconnections between various devices. The deposition process is generally carried out in a vacuum chamber and involves the condensation of Au atoms from the vapor phase onto the chip surface. In one procedure, a gold wire is wrapped around a tungsten filament, which is heated by passing a large current through the filament (analogous to the heating of the filament in a light bulb) as depicted in Figure 1.74b. The Au wire melts and wets the filament, but as the temperature of the filament increases, the gold evaporates to form a vapor. Au atoms from this vapor then condense onto the chip surface, to solidify and form the metallic connections. Suppose that the source (filament)-to-substrate (chip) distance L is 10 cm. Unless the mean free path of air molecules is much longer than L, collisions between the metal atoms and air molecules will prevent the deposition of the Au onto the chip surface. Taking the mean free path ℓ to be $100L$, what should be the pressure inside the vacuum system? (Assume the same r for Au atoms.)

1.12 Heat capacity

a. Calculate the heat capacity per mole and per gram of N_2 gas, neglecting the vibrations of the molecule. How does this compare with the experimental value of 0.743 J g^{-1} K^{-1}?

b. Calculate the heat capacity per mole and per gram of CO_2 gas, neglecting the vibrations of the molecule. How does this compare with the experimental value of 0.648 J K^{-1} g^{-1}? Assume that the CO_2 molecule is linear (O–C–O) so that it has two rotational degrees of freedom.

c. Based on the Dulong–Petit rule, calculate the heat capacity per mole and per gram of solid silver. How does this compare with the experimental value of 0.235 J K^{-1} g^{-1}?

d. Based on the Dulong–Petit rule, calculate the heat capacity per mole and per gram of the silicon crystal. How does this compare with the experimental value of 0.71 J K^{-1} g^{-1}?

1.13 Dulong–Petit atomic heat capacity

Express the Dulong–Petit rule for the molar heat capacity as heat capacity per atom and in the units of eV K^{-1} per atom, called the **atomic heat capacity.** CsI is an ionic crystal used in optical applications that require excellent infrared transmission at very long wavelengths (up to 55 μm). It has the CsCl crystal structure with one Cs^+ and one I^- ion in the unit cell. Given the density of CsI as 4.51 g cm^{-3}, calculate the specific heat capacity of CsI and compare it with the experimental value of 0.2 J K^{-1} g^{-1}. What is your conclusion?

1.14 Dulong–Petit specific heat capacity of alloys and compounds

a. Consider an alloy AB, such as solder, or a compound material such as MgO, composed of n_A, atomic fractions of A, and n_B, atomic fractions of B. (The atomic fraction of A is the same as its molar fraction.) Let M_A and M_B be the atomic weights of A and B, in g mol^{-1}. The mean atomic weight per atom in the alloy or compound is then

$$\overline{M} = n_A M_A + n_B M_B$$

Average atomic weight

Show that the *Dulong–Petit rule* for the specific heat capacity c_s leads to

$$c_s = \frac{C_m}{\overline{M}} = \frac{25}{n_A M_A + n_B M_B} \quad \text{J K}^{-1}\,\text{g}^{-1}$$

Specific heat capacity

b. Calculate the specific heat capacity of Pb–Sn solder assuming that its composition is 38 wt.% Pb and 62 wt.% Sn.

c. Calculate the specific heat capacities of Pb and Sn individually as c_{vA} and c_{vB}, respectively, and then calculate the c_v for the alloy using

Alloy specific heat capacity

$$c_v = c_{vA} w_A + c_{vB} w_B$$

where w_A and w_B are the weight fractions of *A* (Pb) and *B* (Sn) in the alloy (solder). Compare your result with part (*a*). What is your conclusion?

d. ZnSe is an important optical material (used in infrared windows and lenses and high-power CO_2 laser optics) and also an important II–VI semiconductor that can be used to fabricate blue-green laser diodes. Calculate the specific heat capacity of ZnSe, and compare the calculation to the experimental value of 0.345 $J\,K^{-1}\,g^{-1}$.

1.15 Thermal expansion

a. If λ is the thermal expansion coefficient, show that the thermal expansion coefficient for an area is 2λ. Consider an aluminum square sheet of area 1 cm^2. If the thermal expansion coefficient of Al at room temperature (25 °C) is about $24 \times 10^{-6}\ K^{-1}$, at what temperature is the percentage change in the area +1%?

b. A particular incandescent light bulb (100 W, 120 V) has a tungsten (W) filament of length 57.9 cm and a diameter of 63.5 μm. Calculate the length of the filament at 2300 °C, the approximate operating temperature of the filament inside the bulb. The linear expansion coefficient λ of W is approximately $4.50 \times 10^{-6}\ K^{-1}$ at 300 K. How would you improve your calculation?

1.16 Thermal expansion of Si The expansion coefficient of silicon over the temperature range 120–1500 K is given by Okada and Tokumaru (1984) as

Silicon linear expansion coefficient

$$\lambda = 3.725 \times 10^{-6}[1 - e^{-5.88 \times 10^{-3}(T-124)}] + 5.548 \times 10^{-10} T$$

where λ is in K^{-1} (or $°C^{-1}$) and T is in kelvins.

a. By expanding the above function around 20 °C (293 K) show that,

Silicon linear expansion coefficient

$$\lambda = 2.5086 \times 10^{-6} + (8.663 \times 10^{-9})(T - 293) - (2.3839 \times 10^{-11})(T - 293)^2$$

b. The change $\delta\rho$ in the density due to a change δT in the temperature, from Example 1.5, is given by

$$\delta\rho = -\rho_o \alpha_V\, \delta T = -3\rho_o \lambda\, \delta T$$

Given the density of Si as 2.329 $g\,cm^{-3}$ at 20 °C, calculate the density at 1000 °C by using the full expression and by using the polynomials expansion of λ. What is your conclusion?

1.17 Thermal expansion of GaP and GaAs

a. GaP has the zinc blende structure. The linear expansion coefficient in GaP has been measured as follows: $\lambda = 4.65 \times 10^{-6}\ K^{-1}$ at 300 K; $5.27 \times 10^{-6}\ K^{-1}$ at 500 K; $5.97 \times 10^{-6}\ K^{-1}$ at 800 K. Calculate the coefficients, *A*, *B*, and *C* in

$$\frac{dL}{L_o\, dT} = \lambda(T) = A + B(T - T_o) + C(T - T_o)^2 + \cdots$$

where $T_o = 300$ K. The lattice constant of GaP, *a*, at 27 °C is 0.5451 nm. Calculate the lattice constant at 300 °C.

b. The linear expansion coefficient of GaAs over 200–1000 K is given by

GaAs linear expansion coefficient

$$\lambda = 4.25 \times 10^{-6} + (5.82 \times 10^{-9})T - (2.82 \times 10^{-12})T^2$$

where T is in kelvins. The lattice constant *a* at 300 K is 0.56533 nm. Calculate the lattice constant and the density at −40°C.

1.18 Electrical noise Consider an amplifier with a bandwidth *B* of 5 kHz, corresponding to a typical speech bandwidth. Assume the input resistance of the amplifier is 1 MΩ. What is the rms noise voltage at the input? What will happen if the bandwidth is doubled to 10 kHz? What is your conclusion?

1.19 **Thermal activation** A certain chemical oxidation process (*e.g.*, SiO_2) has an activation energy of 2 eV atom^{-1}.

a. Consider the material exposed to pure oxygen gas at a pressure of 1 atm at 27 °C. Estimate how many oxygen molecules per unit volume will have energies in excess of 2 eV? (Consider the numerical integration of Equation 1.24.)

b. If the temperature is 900 °C, estimate the number of oxygen molecules with energies more than 2 eV. What happens to this concentration if the pressure is doubled?

1.20 **Diffusion in Si** The diffusion coefficient of boron (B) atoms in a single crystal of Si has been measured to be 1.5×10^{-18} m^2 s^{-1} at 1000 °C and 1.1×10^{-16} m^2 s^{-1} at 1200 °C.

a. What is the activation energy for the diffusion of B, in eV/atom?

b. What is the preexponential constant D_o?

c. What is the rms distance (in micrometers) diffused in 1 hour by the B atom in the Si crystal at 1200 °C and 1000 °C?

d. The diffusion coefficient of B in polycrystalline Si has an activation energy of 2.4–2.5 eV/atom and $D_o = (1.5 - 6) \times 10^{-7}$ m^2 s^{-1}. What constitutes the diffusion difference between the single crystal sample and the polycrystalline sample?

1.21 **Diffusion in SiO_2** The diffusion coefficient of P atoms in SiO_2 has an activation energy of 2.30 eV/atom and $D_o = 5.73 \times 10^{-9}$ m^2 s^{-1}. What is the rms distance diffused in 1 hour by P atoms in SiO_2 at 1200 °C?

1.22 **BCC and FCC crystals**

a. Molybdenum has the BCC crystal structure, a density of 10.22 g cm^{-3}, and an atomic mass of 95.94 g mol^{-1}. What is the atomic concentration, lattice parameter a, and atomic radius of molybdenum?

b. Gold has the FCC crystal structure, a density of 19.3 g cm^{-3}, and an atomic mass of 196.97 g mol^{-1}. What is the atomic concentration, lattice parameter a, and atomic radius of gold?

1.23 **BCC and FCC crystals**

a. Tungsten (W) has the BCC crystal structure. The radius of the W atom is 0.1371 nm. The atomic mass of W is 183.8 amu (g mol^{-1}). Calculate the number of W atoms per unit volume and density of W.

b. Platinum (Pt) has the FCC crystal structure. The radius of the Pt atom is 0.1386 nm. The atomic mass of Pt is 195.09 amu (g mol^{-1}). Calculate the number of Pt atoms per unit volume and density of Pt.

1.24 **Planar and surface concentrations** Niobium (Nb) has the BCC crystal with a lattice parameter $a = 0.3294$ nm. Find the planar concentrations as the number of atoms per nm^2 of the (100), (110), and (111) planes. Which plane has the most concentration of atoms per unit area? Sometimes the number of atoms per unit area $n_{surface}$ on the surface of a crystal is estimated by using the relation $n_{surface} = n_{bulk}^{2/3}$, where n_{bulk} is the concentration of atoms in the bulk. Compare $n_{surface}$ values with the planar concentrations that you calculated and comment on the difference. [*Note:* The BCC (111) plane does not cut through the center atom and the (111) has one-sixth of an atom at each corner.]

1.25 **Diamond and zinc blende** Si has the diamond and GaAs has the zinc blende crystal structure. Given the lattice parameters of Si and GaAs, $a = 0.543$ nm and $a = 0.565$ nm, respectively, and the atomic masses of Si, Ga, and As as 28.08, 69.73, and 74.92, respectively, calculate the density of Si and GaAs. What is the atomic concentration (atoms per unit volume) in each crystal?

1.26 **Zinc blende, NaCl, and CsCl**

a. InAs is a III-V semiconductor that has the zinc blende structure with a lattice parameter of 0.606 nm. Given the atomic masses of In (114.82 g mol^{-1}) and As (74.92 g mol^{-1}), find the density.

b. CdO has the NaCl crystal structure with a lattice parameter of 0.4695 nm. Given the atomic masses of Cd (112.41 g mol^{-1}) and O (16.00 g mol^{-1}), find the density.

c. KCl has the same crystal structure as NaCl. The lattice parameter a of KCl is 0.629 nm. The atomic masses of K and Cl are 39.10 g mol^{-1} and 35.45 g mol^{-1}, respectively. Calculate the density of KCl.

1.27 Crystallographic directions and planes Consider the cubic crystal system.

a. Show that the line $[hk\ell]$ is perpendicular to the $(hk\ell)$ plane.

b. Show that the spacing between adjacent $(hk\ell)$ planes is given by

$$d = \frac{a}{\sqrt{h^2 + k^2 + \ell^2}}.$$

1.28 Si and SiO_2

a. Given the Si lattice parameter $a = 0.543$ nm, calculate the number of Si atoms per unit volume, in nm^{-3}.

b. Calculate the number of atoms per m^2 and per nm^2 on the (100), (110), and (111) planes in the Si crystal as shown in Figure 1.75. Which plane has the most number of atoms per unit area?

c. The density of SiO_2 is 2.27 g cm^{-3}. Given that its structure is amorphous, calculate the number of molecules per unit volume, in nm^{-3}. Compare your result with (a) and comment on what happens when the surface of an Si crystal oxidizes. The atomic masses of Si and O are 28.09 and 16, respectively.

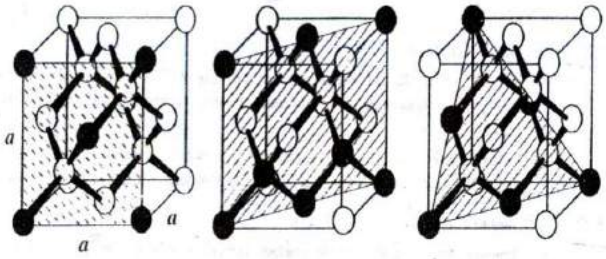


Figure 1.75 Diamond cubic crystal structure and planes.
Determine what portion of a black-colored atom belongs to the plane that is hatched.

1.29 Vacancies in metals

a. The energy of formation of a vacancy in the copper crystal is about 1 eV. Calculate the concentration of vacancies at room temperature (300 K) and just below the melting temperature, 1084 °C. Neglect the change in the density which is small.

b. The following table shows the energies of vacancy formation in various metals with *close-packed* crystal structures and the melting temperature T_m. Plot E_v in eV versus. T_m in kelvins, and explore if there is a correlation between a and T_m. Some materials engineers take E_v to be very roughly $10kT_m$. Do you think that they are correct? (Justify.)

	Metal								
	Al	Ag	Au	Cu	Mg	Pt	Pb	Ni	Pd
Crystal	FCC	FCC	FCC	FCC	HCP	FCC	FCC	FCC	FCC
E_v (eV)	0.70–0.76	1.0–1.1	0.90–0.98	1–1.28	0.89	1.3–1.5	0.50	1.63–1.79	1.54–1.85
T_m (°C)	660	962	1064	1085	650	1768	328	1455	1555

1.30 **Vacancies in silicon** In device fabrication, Si is frequently doped by the diffusion of impurities (dopants) at high temperatures, typically 950–1200°C. The energy of vacancy formation in the Si crystal is about 3.6 eV. What is the equilibrium concentration of vacancies in a Si crystal at 1000 °C? Neglect the change in the density with temperature which is less than 1 percent in this case.

1.31 **Pb–Sn solder** Consider the soldering of two copper components. When the solder melts, it wets both metal surfaces. If the surfaces are not clean or have an oxide layer, the molten solder cannot wet the surfaces and the soldering fails. Assume that soldering takes place at 250 °C, and consider the diffusion of Sn atoms into the copper (the Sn atom is smaller than the Pb atom and hence diffuses more easily).

a. The diffusion coefficient of Sn in Cu at two temperatures is $D = 1.69 \times 10^{-9}$ cm^2 hr^{-1} at 400 °C and $D = 2.48 \times 10^{-7}$ cm^2 hr^{-1} at 650 °C. Calculate the rms distance diffused by an Sn atom into the copper, assuming the cooling process takes 10 seconds.

b. What should be the composition of the solder if it is to begin freezing at 250 °C?

c. What are the components (phases) in this alloy at 200 °C? What are the compositions of the phases and their relative weights in the alloy?

d. What is the microstructure of this alloy at 25 °C? What are weight fractions of the α and β phases assuming near equilibrium cooling?

1.32 **Pb–Sn solder** Consider 50% Pb–50% Sn solder.

a. Sketch the temperature-time profile and the microstructure of the alloy at various stages as it is cooled from the melt.

b. At what temperature does the solid melt?

c. What is the temperature range over which the alloy is a mixture of melt and solid? What is the structure of the solid?

d. Consider the solder at room temperature following cooling from 182 °C. Assume that the rate of cooling from 182 °C to room temperature is faster than the atomic diffusion rates needed to change the compositions of the α and β phases in the solid. Assuming the alloy is 1 kg, calculate the masses of the following components in the solid:

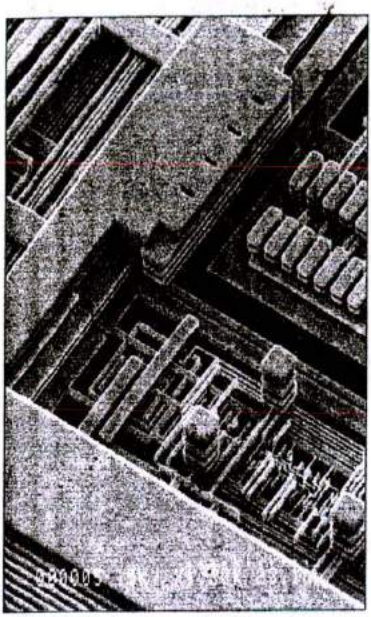
1. The primary α.
2. α in the whole alloy.
3. α in the eutectic solid.
4. β in the alloy. (Where is the β-phase?)

e. Calculate the specific heat of the solder given the atomic masses of Pb (207.2) and Sn (118.71).



Walter Houser Brattain (1902–1987), one of the inventors of the transistor, looking at a vacuum evaporator used for depositing metal film electrodes on semiconductors (1937).

| SOURCE: AIP Emilio Segrè Visual Archives, Brattain Collection.



Highly magnified scanning electron microscope (SEM) view of IBM's six-level copper interconnect technology in an integrated circuit chip. The aluminum in transistor interconnections in a silicon chip has been replaced by copper that has a higher conductivity (by nearly 40%) and also a better ability to carry higher current densities without electromigration. Lower copper interconnect resistance means higher speeds and lower RC constants (1997).

| SOURCE: Courtesy of IBM Corporation.



SEM view of three levels of copper interconnect metallization in IBM's new faster CMOS integrated circuits (1997).

| SOURCE: Courtesy of IBM Corporation.

CHAPTER

2

Electrical and Thermal Conduction in Solids

Electrical conduction involves the motion of charges in a material under the influence of an applied electric field. A material can generally be classified as a conductor if it contains a large number of "free" or mobile charge carriers. In metals, due to the nature of metallic bonding, the valence electrons from the atoms form a sea of electrons that are free to move within the metal and are therefore called conduction electrons. In this chapter, we will treat the conduction electrons in metal as "free charges" that can be accelerated by an applied electric field. In the presence of an electric field, the conduction electrons attain an average velocity, called the drift velocity, that depends on the field. By applying Newton's second law to electron motion and using such concepts as mean free time between electron collisions with lattice vibrations, crystal defects, impurities, etc., we will derive the fundamental equations that govern electrical conduction in solids. A key concept will be the drift mobility, which is a measure of the ease with which charge carriers in the solid drift under the influence of an external electric field.

Good electrical conductors, such as metals, are also known to be good thermal conductors. The conduction of thermal energy from higher to lower temperature regions in a metal involves the conduction electrons carrying the energy. Consequently, there is an innate relationship between the electrical and thermal conductivities, which is supported by theory and experiments.

2.1 CLASSICAL THEORY: THE DRUDE MODEL

2.1.1 METALS AND CONDUCTION BY ELECTRONS

The electric current density J is defined as the net amount of charge flowing across a unit area per unit time, that is,

Current density definition

$$J = \frac{\Delta q}{A\,\Delta t}$$

where Δq is the net quantity of charge flowing through an area A in time Δt. Figure 2.1 shows the net flow of electrons in a conductor section of cross-sectional area A in the presence of an applied field $\mathcal{E}_x$. Notice that the direction of electron motion is opposite to that of the electric field $\mathcal{E}_x$ and of conventional current, because the electrons experience a Coulombic force $e\mathcal{E}_x$ in the x direction, due to their negative charge.

We know that the conduction electrons are actually moving around randomly[1] in the metal, but we will assume that as a result of the application of the electric field $\mathcal{E}_x$, they all acquire a net velocity in the x direction. Otherwise, there would be no net flow of charge through area A.

The average velocity of the electrons in the x direction at time t is denoted $v_{dx}(t)$. This is called the **drift velocity**, which is the instantaneous velocity v_x in the x direction averaged over many electrons (perhaps, $\sim 10^{28}$ m^{-3}); that is

Definition of drift velocity

$$v_{dx} = \frac{1}{N}[v_{x1} + v_{x2} + v_{x3} + \cdots + v_{xN}] \qquad [2.1]$$

where v_{xi} is the x direction velocity of the ith electron, and N is the number of conduction electrons in the metal. Suppose that n is the number of electrons per unit volume in the conductor ($n = N/V$). In time Δt, electrons move a distance $\Delta x = v_{dx}\,\Delta t$, so the total charge Δq crossing the area A is $enA\,\Delta x$. This is valid because all the electrons within distance Δx pass through A; thus, $n(A\,\Delta x)$ is the total number of electrons crossing A in time Δt.

The current density in the x direction is

$$J_x = \frac{\Delta q}{A\,\Delta t} = \frac{enAv_{dx}\,\Delta t}{A\,\Delta t} = env_{dx}$$

This general equation relates J_x to the average velocity v_{dx} of the electrons. It must be appreciated that the average velocity at one time may not be the same as at another time, because the applied field, for example, may be changing: $\mathcal{E}_x = \mathcal{E}_x(t)$. We therefore allow for a time-dependent current by writing

Current density and drift velocity

$$J_x(t) = env_{dx}(t) \qquad [2.2]$$

To relate the current density J_x to the electric field $\mathcal{E}_x$, we must examine the effect of the electric field on the motion of the electrons in the conductor. To do so, we will consider the copper crystal.

[1] All the conduction electrons are "free" within the metal and move around randomly, being scattered from vibrating metal ions, as we discuss in this chapter.

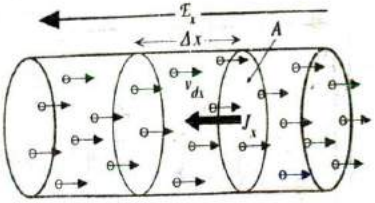


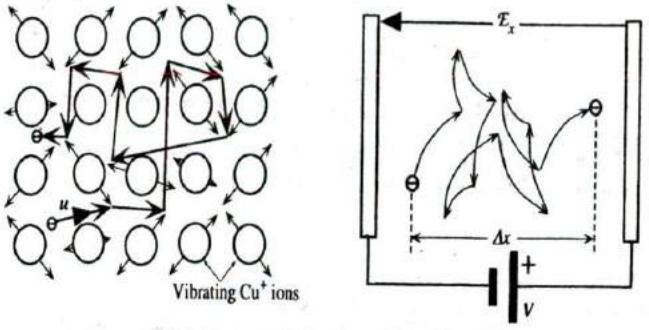
Figure 2.1 Drift of electrons in a conductor in the presence of an applied electric field.
Electrons drift with an average velocity v_{dx} in the x direction.

The copper atom has a single valence electron in its 4*s* subshell, and this electron is loosely bound. The solid metal consists of positive ion cores, Cu^+, at regular sites, in the face-centered cubic (FCC) crystal structure. The valence electrons detach themselves from their parents and wander around freely in the solid, forming a kind of electron cloud or gas. These mobile electrons are free to respond to an applied field, creating a current density J_x. The valence electrons in the electron gas are therefore **conduction electrons.**

The attractive forces between the negative electron cloud and the Cu^+ ions are responsible for metallic bonding and the existence of the solid metal. (This simplistic view of metal was depicted in Figure 1.7 for copper.) The electrostatic attraction between the conduction electrons and the positive metal ions, like the electrostatic attraction between the electron and the proton in the hydrogen atom, results in the conduction electron having both potential energy *PE* and kinetic energy *KE*. The conduction electrons move about the crystal lattice in the same way that gas atoms move randomly in a cylinder. Although the average *KE* for gas atoms is $\frac{3}{2}kT$, this is not the case for electrons in a metal, because these electrons strongly interact with the metal ions and with each other as a result of electrostatic interactions.

The mean *KE* of the conduction electrons in a metal is primarily determined by the electrostatic interaction of these electrons with the positive metal ions and also with each other. For most practical purposes, we will therefore neglect the temperature dependence of the mean *KE* compared with other factors that control the behavior of the conduction electrons in the metal crystal. We can speculate from Example 1.1, that the magnitude of mean *KE* must be comparable to the magnitude of the mean *PE* of electrostatic interaction[2] or, stated differently, to the metal bond energy which is several electron volts per atom. If u is the **mean speed** of the conduction electrons, then, from electrostatic interactions alone, we expect $\frac{1}{2}m_e u^2$ to be several electron volts which means that u is typically $\sim 10^6$ m s^{-1}. This purely classical and intuitive reasoning is not sufficient, however, to show that the mean speed u is relatively temperature insensitive and much greater than that expected from kinetic molecular theory. The true reasons are quantum mechanical and are discussed in Chapter 4. (They arise from what is called the Pauli exclusion principle.)

[2] There is a theorem in classical mechanics called **the virial theorem**, which states that for a collection of particles, the mean *KE* has half the magnitude of the mean *PE* if the only forces acting on the particles are such that they follow an inverse square law dependence on the particle-particle separation (as in Coulombic and gravitational forces).



(a) A conduction electron in the electron gas moves about randomly in a metal (with a mean speed u) being frequently and randomly scattered by thermal vibrations of the atoms. In the absence of an applied field there is no net drift in any direction.

(b) In the presence of an applied field, $\mathcal{E}_x$, there is a net drift along the x direction. This net drift along the force of the field is superimposed on the random motion of the electron. After many scattering events the electron has been displaced by a net distance, Δx, from its initial position toward the positive terminal.

Figure 2.2 Motion of a conduction electron in a metal.

In general, the copper crystal will not be perfect and the atoms will not be stationary. There will be crystal defects, vacancies, dislocations, impurities, etc., which will scatter the conduction electrons. More importantly, due to their thermal energy, the atoms will vibrate about their lattice sites (equilibrium positions), as depicted in Figure 2.2a. An electron will not be able to avoid collisions with vibrating atoms; consequently, it will be "scattered" from one atom to another. In the absence of an applied field, the path of an electron may be visualized as illustrated in Figure 2.2a, where scattering from lattice vibrations causes the electron to move randomly in the lattice. On those occasions when the electron reaches a crystal surface, it becomes "deflected" (or "bounced") back into the crystal. Therefore, in the absence of a field, after some duration of time, the electron crosses its initial x plane position again. Over a long time, the electrons therefore show no net displacement in any one direction.

When the conductor is connected to a battery and an electric field is applied to the crystal, as shown in Figure 2.2b, the electron experiences an acceleration in the x direction in addition to its random motion, so after some time, it will drift a finite distance in the x direction. The electron accelerates along the x direction under the action of the force $e\mathcal{E}_x$, and then it suddenly collides with a vibrating atom and loses the gained velocity. Therefore, there is an average velocity in the x direction, which, if calculated, determines the current via Equation 2.2. Note that since the electron experiences an acceleration in the x direction, its trajectory between collisions is a parabola, like the trajectory of a golf ball experiencing acceleration due to gravity.

To calculate the drift velocity v_{dx} of the electrons due to applied field $\mathcal{E}_x$, we first consider the velocity v_{xi} of the ith electron in the x direction at time t. Suppose its last collision was at time t_i; therefore, for time $(t - t_i)$, it accelerated *free of collisions*, as

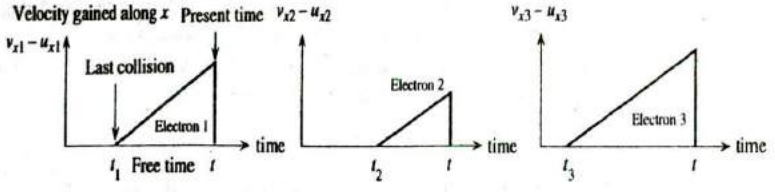


Figure 2.3 Velocity gained in the x direction at time t from the electric field ($\mathcal{E}_x$) for three electrons. There will be N electrons to consider in the metal.

indicated in Figure 2.3. Let u_{xi} be the velocity of electron i in the x direction just after the collision. We will call this the initial velocity. Since $e\mathcal{E}_x/m_e$ is the acceleration of the electron, the velocity v_{xi} in the x direction at time t will be

$$v_{xi} = u_{xi} + \frac{e\mathcal{E}_x}{m_e}(t - t_i)$$

However, this is only for the ith electron. We need the average velocity v_{dx} for all such electrons along x. We average the expression for $i = 1$ to N electrons, as in Equation 2.1. We assume that immediately after a collision with a vibrating ion, the electron may move in any random direction; that is, it can just as likely move along the negative or positive x, so that u_{xi} averaged over many electrons is zero. Thus,

$$v_{dx} = \frac{1}{N}[v_{x1} + v_{x2} + \cdots + v_{xN}] = \frac{e\mathcal{E}_x}{m_e}\overline{(t - t_i)} \qquad \textit{Drift velocity}$$

where $\overline{(t - t_i)}$ is the **average free time** for N electrons between collisions.

Suppose that τ is the mean free time, or **the mean time between collisions** (also known as the **mean scattering time**). For some electrons, $(t - t_i)$ will be greater than τ, and for others, it will be shorter, as shown in Figure 2.3. Averaging $(t - t_i)$ for N electrons will be the same as τ. Thus, we can substitute τ for $\overline{(t - t_i)}$ in the previous expression to obtain

$$v_{dx} = \frac{e\tau}{m_e}\mathcal{E}_x \qquad [2.3]$$

Equation 2.3 shows that the drift velocity increases linearly with the applied field. The constant of proportionality $e\tau/m_e$ has been given a special name and symbol. It is called the **drift mobility** μ_d, which is defined as

$$v_{dx} = \mu_d \mathcal{E}_x \qquad [2.4] \qquad \textit{Definition of drift mobility}$$

where

$$\mu_d = \frac{e\tau}{m_e} \qquad [2.5] \qquad \textit{Drift mobility and mean free time}$$

Equation 2.5 relates the drift mobility of the electrons to their mean scattering time τ. To reiterate, τ, which is also called the **relaxation time,** is directly related to

the microscopic processes that cause the scattering of the electrons in the metal; that is, lattice vibrations, crystal imperfections, and impurities, to name a few.

From the expression for the drift velocity v_{dx}, the current density J_x follows immediately by substituting Equation 2.4 into 2.2, that is,

Ohm's law

$$J_x = en\mu_d\mathcal{E}_x \tag{2.6}$$

Therefore, the current density is proportional to the electric field and the conductivity σ is the term multiplying $\mathcal{E}_x$, that is,

Unipolar conductivity

$$\sigma = en\mu_d \tag{2.7}$$

It is gratifying that by treating the electron as a particle and applying classical mechanics ($F = ma$), we are able to derive Ohm's law. We should note, however, that we assumed τ to be independent of the field.

Drift mobility is important because it is a widely used electronic parameter in semiconductor device physics. The drift mobility gauges how fast electrons will drift when driven by an applied field. If the electron is not highly scattered, then the mean free time between collisions will be long, τ will be large, and by Equation 2.5, the drift mobility will also be large; the electrons will therefore be highly mobile and be able to "respond" to the field. However, a large drift mobility does not necessarily imply high conductivity, because σ also depends on the concentration of conduction electrons n.

The mean time between collisions τ has further significance. Its reciprocal $1/\tau$ represents the **mean frequency of collisions** or **scattering events**; that is, $1/\tau$ is the mean probability per unit time that the electron will be scattered (see Example 2.1). Therefore, during a small time interval δt, the probability of scattering will be $\delta t/\tau$. The probability of scattering per unit time $1/\tau$ is time independent and depends only on the nature of the electron scattering mechanism.

There is one important assumption in the derivation of the drift velocity v_{dx} in Equation 2.3. We obtained v_{dx} by averaging the velocities v_{xi} of N electrons along x at one instant, as defined in Equation 2.1. The drift velocity therefore represents the average velocity of *all* the electrons along x at one instant; that is, v_{dx} is a number average at one instant. Figure 2.2b shows that after many collisions, after a time interval $\Delta t \gg \tau$, an electron would have been displaced by a net distance Δx along x. The term $\Delta x/\Delta t$ represents the effective velocity with which the electron drifts along x. It is an average velocity for one electron over many collisions, that is, over a long time (hence, $\Delta t \gg \tau$), so $\Delta x/\Delta t$ is a time average. Provided that Δt contains many collisions, it is reasonable to expect that the drift velocity $\Delta x/\Delta t$ from the time average for one electron is the same as the drift velocity v_{dx} per electron from averaging for all electrons at one instant, as in Equation 2.1, or

Drift velocity

$$\frac{\Delta x}{\Delta t} = v_{dx}$$

The two velocities are the same only under steady-state conditions ($\Delta t \gg \tau$). The proof may be found in more advanced texts.

EXAMPLE 2.1

PROBABILITY OF SCATTERING PER UNIT TIME AND THE MEAN FREE TIME If $1/\tau$ is defined as the mean probability per unit time that an electron is scattered, show that the mean time between collisions is τ.

SOLUTION

Consider an infinitesimally small time interval dt at time t. Let N be the number of unscattered electrons at time t. The probability of scattering during dt is $(1/\tau)\,dt$, and the number of scattered electrons during dt is $N(1/\tau)\,dt$. The change dN in N is thus

$$dN = -N\left(\frac{1}{\tau}\right)dt$$

The negative sign indicates a reduction in N because, as electrons become scattered, N decreases. Integrating this equation, we can find N at any time t, given that at time $t = 0$, N_0 is the total number of unscattered electrons. Therefore,

Unscattered electron concentration

$$N = N_0 \exp\left(-\frac{t}{\tau}\right)$$

This equation represents the number of unscattered electrons at time t. It reflects an exponential decay law for the number of unscattered electrons. The **mean free time** $\bar{t}$ can be calculated from the mathematical definition of $\bar{t}$,

Mean free time

$$\bar{t} = \frac{\int_0^\infty tN\,dt}{\int_0^\infty N\,dt} = \tau$$

where we have used $N = N_0 \exp(-t/\tau)$. Clearly, $1/\tau$ is the mean probability of scattering per unit time.

EXAMPLE 2.2

ELECTRON DRIFT MOBILITY IN METALS Calculate the drift mobility and the mean scattering time of conduction electrons in copper at room temperature, given that the conductivity of copper is $5.9 \times 10^5\ \Omega^{-1}\,\text{cm}^{-1}$. The density of copper is $8.96\ \text{g cm}^{-3}$ and its atomic mass is $63.5\ \text{g mol}^{-1}$.

SOLUTION

We can calculate μ_d from $\sigma = en\mu_d$ because we already know the conductivity σ. The number of free electrons n per unit volume can be taken as equal to the number of Cu atoms per unit volume, if we assume that each Cu atom donates one electron to the conduction electron gas in the metal. One mole of copper has N_A (6.02×10^{23}) atoms and a mass of 63.5 g. Therefore, the number of copper atoms per unit volume is

$$n = \frac{dN_A}{M_{at}}$$

where d = density = $8.96\ \text{g cm}^{-3}$, and M_{at} = atomic mass = $63.5\ (\text{g mol}^{-1})$. Substituting for d, N_A, and M_{at}, we find $n = 8.5 \times 10^{22}$ electrons cm^{-3}.

The electron drift mobility is therefore

$$\mu_d = \frac{\sigma}{en} = \frac{5.9 \times 10^5\ \Omega^{-1}\,\text{cm}^{-1}}{[(1.6 \times 10^{-19}\ \text{C})(8.5 \times 10^{22}\ \text{cm}^{-3})]}$$

$$= 43.4\ \text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$$

From the drift mobility we can calculate the mean free time τ between collisions by using Equation 2.5,

$$\tau = \frac{\mu_d m_e}{e} = \frac{(43.4 \times 10^{-4}\ \text{m}^2\ \text{V}^{-1}\ \text{s}^{-1})(9.1 \times 10^{-31}\ \text{kg})}{1.6 \times 10^{-19}\ \text{C}} = 2.5 \times 10^{-14}\ \text{s}$$

Note that the mean speed u of the conduction electrons is about $1.5 \times 10^6\ \text{m s}^{-1}$, so that their mean free path is about 37 nm.

EXAMPLE 2.3 **DRIFT VELOCITY AND MEAN SPEED** What is the applied electric field that will impose a drift velocity equal to 0.1 percent of the mean speed $u(\sim 10^6\ \text{m s}^{-1})$ of conduction electrons in copper? What is the corresponding current density and current through a Cu wire of diameter 1 mm?

SOLUTION

The drift velocity of the conduction electrons is $v_{dx} = \mu_d \mathcal{E}_x$, where μ_d is the drift mobility, which for copper is $43.4\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$ (see Example 2.2). With $v_{dx} = 0.001u = 10^3\ \text{m s}^{-1}$, we have

$$\mathcal{E}_x = \frac{v_{dx}}{\mu_d} = \frac{10^3\ \text{m s}^{-1}}{43.4 \times 10^{-4}\ \text{m}^2\ \text{V}^{-1}\ \text{s}^{-1}} = 2.3 \times 10^5\ \text{V m}^{-1} \quad \text{or} \quad 230\ \text{kV m}^{-1}$$

This is an unattainably large electric field in a metal. Given the conductivity σ of copper, the equivalent current density is

$$J_x = \sigma \mathcal{E}_x = (5.9 \times 10^7\ \Omega^{-1}\ \text{m}^{-1})(2.3 \times 10^5\ \text{V m}^{-1})$$
$$= 1.4 \times 10^{13}\ \text{A m}^{-2} \quad \text{or} \quad 1.4 \times 10^7\ \text{A mm}^{-2}$$

This means a current of 1.1×10^7 A through a 1 mm diameter wire! It is clear from this example that for all practical purposes, even under the highest working currents and voltages, the drift velocity is much smaller than the mean speed of the electrons. Consequently, when an electric field is applied to a conductor, for all practical purposes, the mean speed is unaffected.

EXAMPLE 2.4 **DRIFT VELOCITY IN A FIELD: A CLOSER LOOK** There is another way to explain the observed dependence of the drift velocity on the field, and Equation 2.3. Consider the path of a conduction electron in an applied field $\mathcal{E}$ as shown in Figure 2.4. Suppose that at time $t = 0$ the electron has just been scattered from a lattice vibration. Let u_{x1} be the initial velocity in the x direction just after this initial collision (to which we assign a collision number of zero). We will assume that immediately after a collision, the velocity of the electron is in a random *direction*. Suppose that the first collision occurs at time t_1. Since $e\mathcal{E}_x/m_e$ is the acceleration, the distance s_1 covered in the x direction during the free time t_1 will be



$$s_1 = u_{x1}t_1 + \frac{1}{2}\left(\frac{e\mathcal{E}_x}{m_e}\right)t_1^2$$

At time t_1, the electron collides with a lattice vibration (its first collision), and the velocity is randomized again to become u_{x2}. The whole process is then repeated during the next interval which lasts for a free time t_2, and the electron traverses a distance s_2 along x, and so on. To find the overall distance traversed by the electron after p such scattering events, we sum all the

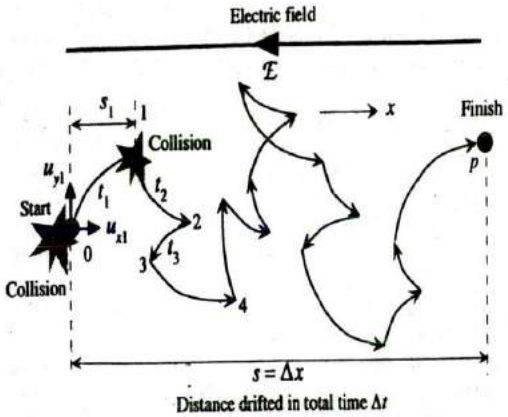


Figure 2.4 The motion of a single electron in the presence of an electric field $\mathcal{E}$. During a time interval t_i, the electron traverses a distance s_i along x. After p collisions, it has drifted a distance $s = \Delta x$.

above distances $s_1, s_2, \ldots$ for p free time intervals,

$$s = s_1 + s_2 + \cdots + s_p = [u_{x1}t_1 + u_{x2}t_2 + \cdots + u_{xp}t_p] + \frac{1}{2}\left(\frac{e\mathcal{E}_x}{m_e}\right)[t_1^2 + t_1^2 + \cdots + t_p^2] \quad [2.8]$$

Since after a collision the "initial" velocity u_x is always random, the first term has u_x values that are randomly negative and positive, so for many collisions (large p) the first term on the right-hand side of Equation 2.8 is nearly zero and can certainly be neglected compared with the second term. Thus, after many collisions, the net distance $s = \Delta x$ traversed in the x direction is given by the second term in Equation 2.8, which is the electric field induced displacement term. If $\overline{t^2}$ is the **mean square free time,** then

$$s = \frac{1}{2}\left(\frac{e\mathcal{E}_x}{m_e}\right)p\overline{t^2}$$

where

$$\overline{t^2} = \frac{1}{p}[t_1^2 + t_1^2 + \cdots + t_p^2]$$

Suppose that τ is the **mean free time between collisions,** where $\tau = (t_1 + t_2 + \cdots + t_p)/p$. Then from straightforward elementary statistics it can be shown that $\overline{t^2} = 2(\overline{t})^2 = 2\tau^2$. So in terms of the mean free time τ between collisions, the overall distance $s = \Delta x$ drifted in the x direction after p collisions is

$$s = \frac{e\mathcal{E}_x}{m_e}(p\tau^2)$$

Further, since the total time Δt taken for these p scattering events is simply $p\tau$, the drift velocity v_{dx} is given by $\Delta x/\Delta t$ or $s/(p\tau)$, that is,

$$v_{dx} = \frac{e\tau}{m_e}\mathcal{E}_x \quad [2.9]$$

This is the same expression as Equation 2.3, except that τ is defined here as the average free time for a single electron over a long time, that is, over many collisions, whereas previously it was the mean free time averaged over many electrons. Further, in Equation 2.9 v_{dx} is an average drift for an electron over a long time, over many collisions. In Equation 2.1 v_{dx} is the

average velocity averaged over all electrons at one instant. For all practical purposes, the two are equivalent. (The equivalence breaks down when we are interested in events over a time scale that is comparable to one scattering, $\sim 10^{-14}$ second.)

The drift mobility μ_d from Equation 2.9 is identical to that of Equation 2.5, $\mu_d = e\tau/m_e$. Suppose that the mean speed of the electrons (not the drift velocity) is u. Then an electron moves a distance $\ell = u\tau$ in mean free time τ, which is called the **mean free path.** The drift mobility and conductivity become,

Drift mobility and conductivity and mean free path

$$\mu_d = \frac{e\ell}{m_e u} \quad \text{and} \quad \sigma = en\mu_d = \frac{e^2 n\ell}{m_e u} \qquad [2.10]$$

Equations 2.3 and 2.10 both assume that after each collision the velocity is randomized. The scattering process, lattice scattering, is able to randomize the velocity in one single scattering. In general not all electron scattering processes can randomize the velocity in one scattering process. If it takes more than one collision to randomize the velocity, then the electron is able to carry with it some velocity gained from a previous collision and hence possesses a higher drift mobility. In such cases one needs to consider the effective mean free path a carrier has to move to eventually randomize the velocity gained; this is a point considered in Chapter 4 when we calculate the resistivity at low temperatures.

2.2 TEMPERATURE DEPENDENCE OF RESISTIVITY: IDEAL PURE METALS

When the conduction electrons are only scattered by thermal vibrations of the metal ions, then τ in the mobility expression $\mu_d = e\tau/m_e$ refers to the mean time between scattering events by this process. The resulting conductivity and resistivity are denoted by σ_T and ρ_T, where the subscript T represents "thermal vibration scattering."

To find the temperature dependence of σ, we first consider the temperature dependence of the mean free time τ, since this determines the drift mobility. An electron moving with a mean speed u is scattered when its path crosses the cross-sectional area S of a scattering center, as depicted in Figure 2.5. The scattering center

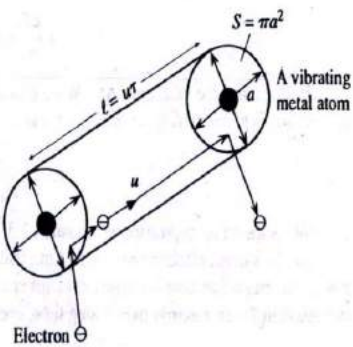


Figure 2.5 Scattering of an electron from the thermal vibrations of the atoms. The electron travels a mean distance $\ell = u\tau$ between collisions. Since the scattering cross-sectional area is S, in the volume St there must be at least one scatterer, $N_s(Su\tau) = 1$.

may be a vibrating atom, impurity, vacancy, or some other crystal defect. Since τ is the mean time taken for one scattering process, the **mean free path** ℓ of the electron between scattering processes is $u\tau$. If N_s is the concentration of scattering centers, then in the volume $S\ell$, there is one scattering center, that is, $(Su\tau)N_s = 1$. Thus, the mean free time is given by

Mean free time between collisions

$$\tau = \frac{1}{SuN_s} \qquad [2.11]$$

The mean speed u of conduction electrons in a metal can be shown to be only slightly temperature dependent.[3] In fact, electrons wander randomly around in the metal crystal with an almost constant mean speed that depends largely on their concentration and hence on the crystal material. Taking the number of scattering centers per unit volume to be the atomic concentration, the temperature dependence of τ then arises essentially from that of the cross-sectional area S. Consider what a free electron "sees" as it approaches a vibrating crystal atom as in Figure 2.5. Because the atomic vibrations are random, the atom covers a cross-sectional area πa^2, where a is the amplitude of the vibrations. If the electron's path crosses πa^2, it gets scattered. Therefore, the mean time between scattering events τ is inversely proportional to the area πa^2 that scatters the electron, that is, $\tau \propto 1/\pi a^2$.

The thermal vibrations of the atom can be considered to be simple harmonic motion, much the same way as that of a mass M attached to a spring. The average kinetic energy of the oscillations is $\frac{1}{4}Ma^2\omega^2$, where ω is the oscillation frequency. From the kinetic theory of matter, this average kinetic energy must be on the order of $\frac{1}{2}kT$. Therefore,

$$\tfrac{1}{4}Ma^2\omega^2 \approx \tfrac{1}{2}kT$$

so $a^2 \propto T$. Intuitively, this is correct because raising the temperature increases the amplitude of the atomic vibrations. Thus,

$$\tau \propto \frac{1}{\pi a^2} \propto \frac{1}{T} \quad \text{or} \quad \tau = \frac{C}{T}$$

where C is a temperature-independent constant. Substituting for τ in $\mu_d = e\tau/m_e$, we obtain

$$\mu_d = \frac{eC}{m_e T}$$

So, the resistivity of a metal is

$$\rho_T = \frac{1}{\sigma_T} = \frac{1}{en\mu_d} = \frac{m_e T}{e^2 nC}$$

[3] The fact that the mean speed of electrons in a metal is only weakly temperature dependent can be proved from what it called the Fermi–Dirac statistics for the collection of electrons in a metal (see Chapter 4). This result contrasts sharply with the kinetic molecular theory of gases (Chapter 1), which predicts that the mean speed of molecules is proportional to $\sqrt{T}$. For the time being, we simply use a constant mean speed u for the conduction electrons in a metal.

Pure metal resistivity due to thermal vibrations of the crystal

that is,

$$\rho_T = AT \qquad [2.12]$$

where A is a temperature-independent constant. This shows that the resistivity of a pure metal wire increases linearly with the temperature, and that the resistivity is due simply to the scattering of conduction electrons by the thermal vibrations of the atoms. We term this conductivity **lattice-scattering-limited conductivity.**

EXAMPLE 2.5 **TEMPERATURE DEPENDENCE OF RESISTIVITY** What is the percentage change in the resistance of a pure metal wire from Saskatchewan's summer to winter, neglecting the changes in the dimensions of the wire?

SOLUTION

Assuming 20 °C for the summer and perhaps −30 °C for the winter, from $R \propto \rho = AT$, we have

$$\frac{R_{\text{summer}} - R_{\text{winter}}}{R_{\text{summer}}} = \frac{T_{\text{summer}} - T_{\text{winter}}}{T_{\text{summer}}} = \frac{(20 + 273) - (-30 + 273)}{(20 + 273)}$$

$$= 0.171 \quad \text{or} \quad 17\%$$

Notice that we have used the absolute temperature for T. How will the outdoor cable power losses be affected?

EXAMPLE 2.6 **DRIFT MOBILITY AND RESISTIVITY DUE TO LATTICE VIBRATIONS** Given that the mean speed of conduction electrons in copper is 1.5×10^6 m s^{-1} and the frequency of vibration of the copper atoms at room temperature is about 4×10^{12} s^{-1}, estimate the drift mobility of electrons and the conductivity of copper. The density d of copper is 8.96 g cm^{-3} and the atomic mass M_{at} is 63.56 g mol^{-1}.

SOLUTION

The method for calculating the drift mobility and hence the conductivity is based on evaluating the mean free time τ via Equation 2.11, that is, $\tau = 1/SuN_s$. Since τ is due to scattering from atomic vibrations, N_s is the atomic concentration,

$$N_s = \frac{dN_A}{M_{at}} = \frac{(8.96 \times 10^3 \text{ kg m}^{-3})(6.02 \times 10^{23} \text{ mol}^{-1})}{63.56 \times 10^{-3} \text{ kg mol}^{-1}}$$

$$= 8.5 \times 10^{28} \text{ m}^{-3}$$

The cross-sectional area $S = \pi a^2$ depends on the amplitude a of the thermal vibrations as shown in Figure 2.5. The average kinetic energy KE_{av} associated with a vibrating mass M attached to a spring is given by $KE_{av} = \frac{1}{4}Ma^2\omega^2$, where ω is the angular frequency of the vibration ($\omega = 2\pi 4 \times 10^{12}$ rad s^{-1}). Applying this equation to the vibrating atom and equating the average kinetic energy KE_{av} to $\frac{1}{2}kT$, by virtue of equipartition of energy theorem, we have $a^2 = 2kT/M\omega^2$ and thus

$$S = \pi a^2 = \frac{2\pi kT}{M\omega^2} = \frac{2\pi(1.38 \times 10^{-23} \text{ J K}^{-1})(300 \text{ K})}{\left(\dfrac{63.56 \times 10^{-3} \text{ kg mol}^{-1}}{6.022 \times 10^{23} \text{ mol}^{-1}}\right)(2\pi \times 4 \times 10^{12} \text{ rad s}^{-1})^2}$$

$$= 3.9 \times 10^{-22} \text{ m}^2$$

Therefore,

$$\tau = \frac{1}{SuN_s} = \frac{1}{(3.9 \times 10^{-22}\ \text{m}^2)(1.5 \times 10^6\ \text{m s}^{-1})(8.5 \times 10^{28}\ \text{m}^{-3})}$$
$$= 2.0 \times 10^{-14}\ \text{s}$$

The drift mobility is

$$\mu_d = \frac{e\tau}{m_e} = \frac{(1.6 \times 10^{-19}\ \text{C})(2.0 \times 10^{-14}\ \text{s})}{(9.1 \times 10^{-31}\ \text{kg})}$$
$$= 3.5 \times 10^{-3}\ \text{m}^2\ \text{V}^{-1}\ \text{s}^{-1} = 35\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$$

The conductivity is then

$$\sigma = en\mu_d = (1.6 \times 10^{-19}\ \text{C})(8.5 \times 10^{22}\ \text{cm}^{-3})(35\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1})$$
$$= 4.8 \times 10^5\ \Omega^{-1}\ \text{cm}^{-1}$$

The experimentally measured value for the conductivity is $5.9 \times 10^5\ \Omega^{-1}\ \text{cm}^{-1}$, so our crude calculation based on Equation 2.11 is actually only 18 percent lower, which is not bad for an estimate. (As we might have surmised, the agreement is brought about by using reasonable values for the mean speed u and the atomic vibrational frequency ω. These values were taken from quantum mechanical calculations, so our evaluation for τ was not truly based on classical concepts.)

2.3 MATTHIESSEN'S AND NORDHEIM'S RULES

2.3.1 MATTHIESSEN'S RULE AND THE TEMPERATURE COEFFICIENT OF RESISTIVITY (α)

The theory of conduction that considers scattering from lattice vibrations only works well with pure metals; unfortunately, it fails for metallic alloys. Their resistivities are only weakly temperature dependent. We must therefore search for a different type of scattering mechanism.

Consider a metal alloy that has randomly distributed impurity atoms. An electron can now be scattered by the impurity atoms because they are not identical to the host atoms, as illustrated in Figure 2.6. The impurity atom need not be larger than the host atom; it can be smaller. As long as the impurity atom results in a local distortion of the crystal lattice, it will be effective in scattering. One way of looking at the scattering process from an impurity is to consider the scattering cross section. What actually scatters the electron is a local, unexpected change in the potential energy PE of the electron as it approaches the impurity, because the force experienced by the electron is given by

$$F = -\frac{d(PE)}{dx}$$

For example, when an impurity atom of a different size compared to the host atom is placed into the crystal lattice, the impurity atom distorts the region around it, either by

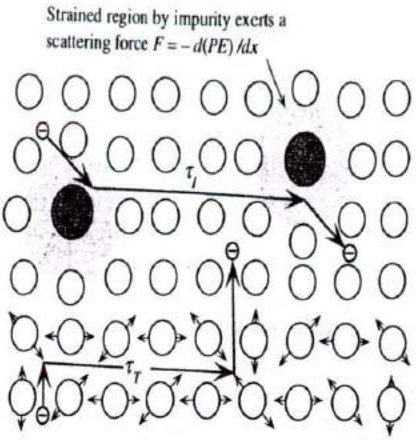


Figure 2.6 Two different types of scattering processes involving scattering from impurities alone and from thermal vibrations alone.

pushing the host atoms farther away, or by pulling them in, as depicted in Figure 2.6. The cross section that scatters the electron is the lattice region that has been elastically distorted by the impurity (the impurity atom itself and its neighboring host atoms), so that in this zone, the electron suddenly experiences a force $F = -d(PE)/dx$ due to a sudden change in the PE. This region has a large scattering cross section, since the distortion induced by the impurity may extend a number of atomic distances. These impurity atoms will therefore hinder the motion of the electrons, thereby increasing the resistance.

We now effectively have two types of mean free times between collisions: one, τ_T, for scattering from thermal vibrations only, and the other, τ_I, for scattering from impurities only. We define τ_T as the mean time between scattering events arising from thermal vibrations alone and τ_I as the mean time between scattering events arising from collisions with impurities alone. Both are illustrated in Figure 2.6.

In general, an electron may be scattered by both processes, so the effective mean free time τ between any two scattering events will be less than the individual scattering times τ_T and τ_I. The electron will therefore be scattered when it collides with either an atomic vibration or an impurity atom. Since in unit time, $1/\tau$ is the net probability of scattering, $1/\tau_T$ is the probability of scattering from lattice vibrations alone, and $1/\tau_I$ is the probability of scattering from impurities alone, then within the realm of elementary probability theory for independent events, we have

Overall frequency of scattering

$$\frac{1}{\tau} = \frac{1}{\tau_T} + \frac{1}{\tau_I} \qquad [2.13]$$

In writing Equation 2.13 for the various probabilities, we make the reasonable assumption that, to a greater extent, the two scattering mechanisms are essentially independent. Here, the effective mean scattering time τ is clearly smaller than both τ_T and τ_I. We can also interpret Equation 2.13 as follows: In unit time, the overall number of

collisions ($1/\tau$) is the sum of the number of collisions with thermal vibrations alone ($1/\tau_T$) and the number of collisions with impurities alone ($1/\tau_I$).

The drift mobility μ_d depends on the effective scattering time τ via $\mu_d = e\tau/m_e$, so Equation 2.13 can also be written in terms of the drift mobilities determined by the various scattering mechanisms. In other words,

$$\frac{1}{\mu_d} = \frac{1}{\mu_L} + \frac{1}{\mu_I} \qquad [2.14]$$

Effective drift mobility

where μ_L is the **lattice-scattering-limited drift mobility,** and μ_I is the **impurity-scattering-limited drift mobility.** By definition, $\mu_L = e\tau_T/m_e$ and $\mu_I = e\tau_I/m_e$. The effective (or overall) resistivity ρ of the material is simply $1/en\mu_d$, or

$$\rho = \frac{1}{en\mu_d} = \frac{1}{en\mu_L} + \frac{1}{en\mu_I}$$

which can be written

$$\rho = \rho_T + \rho_I \qquad [2.15]$$

Matthiessen's rule

where $1/en\mu_L$ is defined as the resistivity due to scattering from thermal vibrations, and $1/en\mu_I$ is the resistivity due to scattering from impurities, or

$$\rho_T = \frac{1}{en\mu_L} \quad \text{and} \quad \rho_I = \frac{1}{en\mu_I}$$

Resistivities due to lattice and impurity scattering

The final result in Equation 2.15 simply states that the effective resistivity ρ is the sum of two contributions. First, $\rho_T = 1/en\mu_L$ is the resistivity due to scattering by thermal vibrations of the host atoms. For those near-perfect pure metal crystals, this is the dominating contribution. As soon as we add impurities, however, there is an additional resistivity, $\rho_I = 1/en\mu_I$, which arises from the scattering of the electrons from the impurities. The first term is temperature dependent because $\tau_T \propto T^{-1}$ (see Section 2.2), but the second term is not.

The mean time τ_I between scattering events involving electron collisions with impurity atoms depends on the separation between the impurity atoms and therefore on the concentration of those atoms (see Figure 2.6). If ℓ_I is the mean separation between the impurities, then the mean free time between collisions with impurities alone will be ℓ_I/u, which is temperature independent because ℓ_I is determined by the impurity concentration N_I (*i.e.*, $\ell_I = N_I^{-1/3}$), and the mean speed of the electrons u is nearly constant in a metal. In the absence of impurities, τ_I is infinitely long, and thus $\rho_I = 0$. The summation rule of resistivities from different scattering mechanisms, as shown by Equation 2.15, is called **Matthiessen's rule.**

There may also be electrons scattering from dislocations and other crystal defects, as well as from grain boundaries. All of these scattering processes add to the resistivity of a metal, just as the scattering process from impurities. We can therefore write the effective resistivity of a metal as

$$\rho = \rho_T + \rho_R \qquad [2.16]$$

Matthiessen's rule

where ρ_R is called the **residual resistivity** and is due to the scattering of electrons by impurities, dislocations, interstitial atoms, vacancies, grain boundaries, etc. (which means that ρ_R also includes ρ_I). The residual resistivity shows very little temperature dependence, whereas $\rho_T = AT$, so the effective resistivity ρ is given by

$$\rho \approx AT + B \qquad [2.17]$$

where A and B are temperature-independent constants.

Equation 2.17 indicates that the resistivity of a metal varies almost linearly with the temperature, with A and B depending on the material. Instead of listing A and B in resistivity tables, we prefer to use a temperature coefficient that refers to small, normalized changes around a reference temperature. The **temperature coefficient of resistivity (TCR)** α_0 is defined as the fractional change in the resistivity per unit temperature increase at the reference temperature T_0, that is,

Definition of temperature coefficient of resistivity

$$\alpha_0 = \frac{1}{\rho_0}\left[\frac{\delta\rho}{\delta T}\right]_{T=T_0} \qquad [2.18]$$

where ρ_0 is the resistivity at the reference temperature T_0, usually 273 K (0 °C) or 293 K (20 °C), and $\delta\rho = \rho - \rho_0$ is the change in the resistivity due to a small increase in temperature, $\delta T = T - T_0$.

When the resistivity follows the behavior $\rho \approx AT + B$ in Equation 2.17, then according to Equation 2.18, α_0 is constant over a temperature range T_0 to T, and Equation 2.18 leads to the well-known equation,

Temperature dependence of resistivity

$$\rho = \rho_0[1 + \alpha_0(T - T_0)] \qquad [2.19]$$

Equation 2.19 is actually only valid when α_0 is constant over the temperature range of interest, which requires Equation 2.17 to hold. Over a limited temperature range, this will usually be the case. Although it is not obvious from Equation 2.19, we should note that α_0 depends on the reference temperature T_0, by virtue of ρ_0 depending on T_0.

The equation $\rho = AT$, which we used for pure-metal crystals to find the change in the resistance with temperature, is only approximate; nonetheless, for pure metals, it is useful to recall in the absence of tabulated data. To determine how good the formula $\rho = AT$ is, put it in Equation 2.19, which leads to $\alpha_0 = T_0^{-1}$. If we take the reference temperature T_0 as 273 K (0 °C), then α_0 is simply 1/273 K; stated differently, Equation 2.19 is then equivalent to $\rho = AT$.

Table 2.1 shows that $\rho \propto T$ is not a bad approximation for some of the familiar pure metals used as conductors (Cu, Al, Au, etc.), but it fails badly for others, such as indium, antimony, and, in particular, the magnetic metals, iron and nickel.

The temperature dependence of the resistivity of various metals is shown in Figure 2.7, where it is apparent that except for the magnetic materials, such as iron and nickel, the linear relationship $\rho \propto T$ seems to be approximately obeyed almost all the way to the melting temperature for many pure metals. It should also be noted that for the alloys, such as nichrome (Ni–Cr), the resistivity is essentially dominated by the residual resistivity, so the resistivity is relatively temperature insensitive, with a very small TCR.

Table 2.1 Resistivity, thermal coefficient of resistivity α_0 at 273 K (0 °C) for various metals. The resistivity index n in $\rho \propto T^n$ for some of the metals is also shown.

Metal	ρ_0 (nΩ m)	$\alpha_0\left(\frac{1}{K}\right)$	n	Comment
Aluminum, Al	25.0	$\frac{1}{233}$	1.20	
Antimony, Sb	38	$\frac{1}{196}$	1.40	
Copper, Cu	15.7	$\frac{1}{232}$	1.15	
Gold, Au	22.8	$\frac{1}{251}$	1.11	
Indium, In	78.0	$\frac{1}{196}$	1.40	
Platinum, Pt	98	$\frac{1}{255}$	0.94	
Silver, Ag	14.6	$\frac{1}{244}$	1.11	
Tantalum, Ta	117	$\frac{1}{294}$	0.93	
Tin, Sn	110	$\frac{1}{217}$	1.11	
Tungsten, W	50	$\frac{1}{220}$	1.20	
Iron, Fe	84.0	$\frac{1}{152}$	1.80	Magnetic metal: $273 < T < 1043$ K
Nickel, Ni	59.0	$\frac{1}{125}$	1.72	Magnetic metal: $273 < T < 627$ K

SOURCE: Data were extracted and combined from several sources. Typical values.

Frequently, the resistivity versus temperature behavior of pure metals can be empirically represented by a power law of the form

$$\rho = \rho_0 \left[\frac{T}{T_0}\right]^n \qquad [2.20]$$

Resistivity of pure metals

where ρ_0 is the resistivity at the reference temperature T_0, and n is a characteristic index that best fits the data. Table 2.1 lists some typical n values for various pure metals above 0 °C. It is apparent that for the nonmagnetic metals, n is close to unity, whereas it is closer to 2 than 1 for the magnetic metals Fe and Ni. In iron, for example, the conduction electron is not scattered simply by atomic vibrations, as in copper, but is affected by its magnetic interaction with the Fe ions in the lattice. This leads to a complicated temperature dependence.

Although our oversimplified theoretical analysis predicts a linear $\rho = AT + B$ behavior for the resistivity down to the lowest temperatures, this is not true in reality,

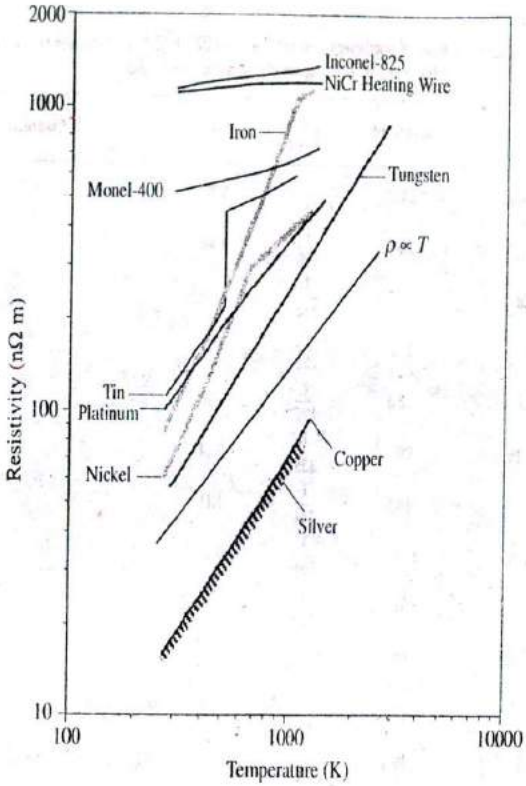


Figure 2.7 The resistivity of various metals as a function of temperature above 0 °C.

Tin melts at 505 K, whereas nickel and iron go through a magnetic-to-nonmagnetic (Curie) transformation at about 627 K and 1043 K, respectively. The theoretical behavior ($\rho \sim T$) is shown for reference.

SOURCE: Data selectively extracted from various sources, including sections in *Metals Handbook*, 10th ed., 2 and 3. Metals Park, Ohio: ASM, 1991.

as depicted for copper in Figure 2.8. As the temperature decreases, typically below ~100 K for many metals, our simple and gross assumption that all the atoms are vibrating with a constant frequency fails. Indeed, the number of atoms that are vibrating with sufficient energy to scatter the conduction electrons starts to decrease rapidly with decreasing temperature, so the resistivity due to scattering from thermal vibrations becomes more strongly temperature dependent. The mean free time $\tau = 1/SuN_s$ becomes longer and strongly temperature dependent, leading to a smaller resistivity than the $\rho \propto T$ behavior. A full theoretical analysis, which is beyond the scope of this chapter, shows that $\rho \propto T^5$. Thus, at the lowest temperature, from Matthiessen's rule, the resistivity becomes $\rho = DT^5 + \rho_R$, where D is a constant. Since the slope of ρ versus T is $d\rho/dT = 5DT^4$, which tends to zero as T becomes small, we have ρ curving

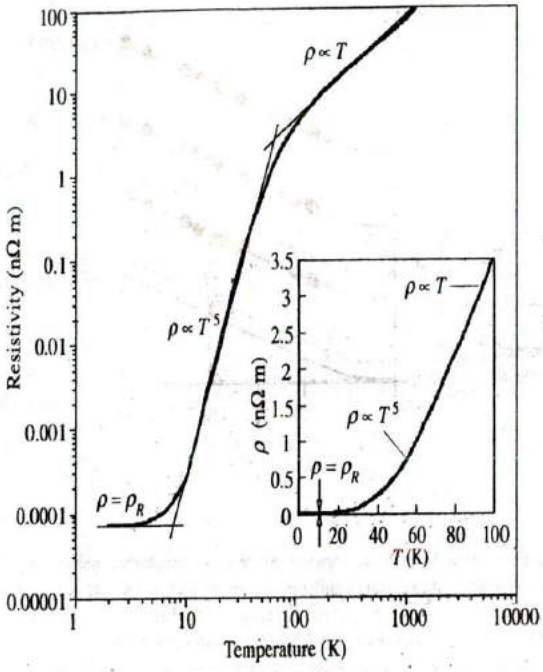


Figure 2.8 The resistivity of copper from lowest to highest temperatures (near melting temperature, 1358 K) on a log-log plot.
Above about 100 K, $\rho \propto T$, whereas at low temperatures, $\rho \propto T^5$, and at the lowest temperatures ρ approaches the residual resistivity ρ_R. The inset shows the ρ vs. T behavior below 100 K on a linear plot. (ρ_R is too small on this scale.)

toward ρ_R as T decreases toward 0 K. This is borne out by experiments, as shown in Figure 2.8 for copper. Therefore, at the lowest temperatures of interest, the resistivity is limited by scattering from impurities and crystal defects.[4]

EXAMPLE 2.7

MATTHIESSEN'S RULE Explain the typical resistivity versus temperature behavior of annealed and cold-worked (deformed) copper containing various amounts of Ni as shown in Figure 2.9.

SOLUTION

When small amounts of nickel are added to copper, the resistivity increases by virtue of Matthiessen's rule, $\rho = \rho_T + \rho_R + \rho_I$, where ρ_T is the resistivity due to scattering from thermal vibrations; ρ_R is the residual resistivity of the copper crystal due to scattering from crystal defects, dislocations, trace impurities, etc.; and ρ_I is the resistivity arising from Ni addition

[4] At sufficiently low temperatures (typically, below 10–20 K for many metals and below ~135 K for certain ceramics) certain materials exhibit superconductivity in which the resistivity vanishes ($\rho = 0$), even in the presence of impurities and crystal defects. Superconductivity and its quantum mechanical origin will be explained in Chapter 8.

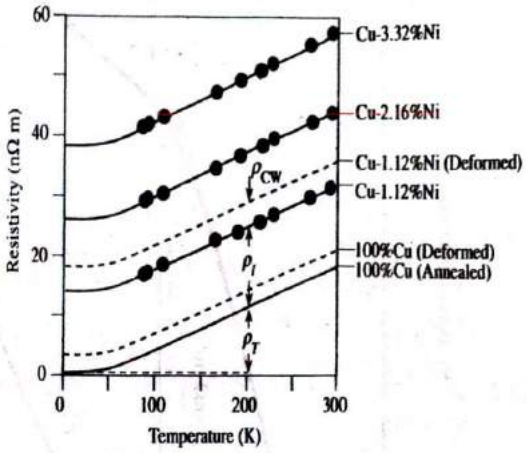


Figure 2.9 Typical temperature dependence of the resistivity of annealed and cold-worked (deformed) copper containing various amounts of Ni in atomic percentage.
SOURCE: Data adapted from J.O. Linde, *Ann Physik*, 5, 219 (Germany, 1932).

alone (scattering from Ni impurity regions). Since ρ_I is temperature independent, for small amounts of Ni addition, ρ_I will simply shift up the ρ versus T curve for copper, by an amount proportional to the Ni content, $\rho_I \propto N_{Ni}$, where N_{Ni} is the Ni impurity concentration. This is apparent in Figure 2.9, where the resistivity of Cu–2.16% Ni is almost twice that of Cu–1.12% Ni. Cold working (CW) or deforming a metal results in a higher concentration of dislocations and therefore increases the residual resistivity ρ_R by ρ_{CW}. Thus, cold-worked samples have a resistivity curve that is shifted up by an additional amount ρ_{CW} that depends on the extent of cold working.

EXAMPLE 2.8 **TEMPERATURE COEFFICIENT OF RESISTIVITY α AND RESISTIVITY INDEX n** If α_0 is the temperature coefficient of resistivity (TCR) at temperature T_0 and the resistivity obeys the equation

$$\rho = \rho_0 \left[\frac{T}{T_0}\right]^n$$

show that

$$\alpha_0 = \frac{n}{T_0}\left[\frac{T}{T_0}\right]^{n-1}$$

What is your conclusion?

Experiments indicate that $n = 1.2$ for W. What is its α_0 at 20 °C? Given that, experimentally, $\alpha_0 = 0.00393$ for Cu at 20 °C, what is n?

SOLUTION

Since the resistivity obeys $\rho = \rho_0(T/T_0)^n$, we substitute this equation into the definition of TCR,

$$\alpha_0 = \frac{1}{\rho_0}\left[\frac{d\rho}{dT}\right] = \frac{n}{T_0}\left[\frac{T}{T_0}\right]^{n-1}$$

It is clear that, in general, α_0 depends on the temperature T, as well as on the reference temperature T_0. The TCR is only independent of T when $n = 1$.

At $T = T_0$, we have

$$\frac{\alpha_0 T_0}{n} = 1 \quad \text{or} \quad n = \alpha_0 T_0$$

For W, $n = 1.2$, so at $T = T_0 = 293$ K, we have $\alpha_{293\,K} = 0.0041$, which agrees reasonably well with $\alpha_{293\,K} = 0.0045$, frequently found in data books.

For Cu, $\alpha_{293\,K} = 0.00393$, so that $n = 1.15$, which agrees with the experimental value of n.

TCR AT DIFFERENT REFERENCE TEMPERATURES If α_1 is the temperature coefficient of resistivity (TCR) at temperature T_1 and α_0 is the TCR at T_0, show that **EXAMPLE 2.9**

$$\alpha_1 = \frac{\alpha_0}{1 + \alpha_0(T_1 - T_0)}$$

SOLUTION

Consider the resistivity at temperature T in terms of α_0 and α_1:

$$\rho = \rho_0[1 + \alpha_0(T - T_0)] \quad \text{and} \quad \rho = \rho_1[1 + \alpha_1(T - T_1)]$$

These equations are expected to hold at any temperature T, so the first and second equations at T_1 and T_0, respectively, give

$$\rho_1 = \rho_0[1 + \alpha_0(T_1 - T_0)] \quad \text{and} \quad \rho_0 = \rho_1[1 + \alpha_1(T_0 - T_1)]$$

These two equations can be readily solved to eliminate ρ_0 and ρ_1 to obtain

$$\alpha_1 = \frac{\alpha_0}{1 + \alpha_0(T_1 - T_0)}$$

TEMPERATURE OF THE FILAMENT OF A LIGHT BULB **EXAMPLE 2.10**

a. Consider a 40 W, 120 V incandescent light bulb. The tungsten filament is 0.381 m long and has a diameter of 33 μm. Its resistivity at room temperature is $5.51 \times 10^{-8}\ \Omega$ m. Given that the resistivity of the tungsten filament varies at $T^{1.2}$, estimate the temperature of the bulb when it is operated at the rated voltage, that is, when it is lit directly from a power outlet, as shown schematically in Figure 2.10. Note that the bulb dissipates 40 W at 120 V.

b. Assume that the electrical power dissipated in the tungsten wire is radiated from the surface of the filament. The radiated electromagnetic power at the absolute temperature T can

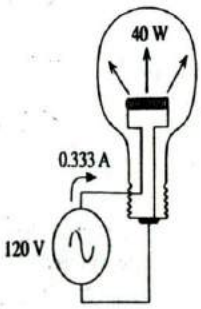


Figure 2.10 Power radiated from a light bulb is equal to the electrical power dissipated in the filament.

be described by **Stefan's law,** as follows:

$$P_{\text{radiated}} = \epsilon\sigma_S A(T^4 - T_0^4)$$

where σ_S is Stefan's constant (5.67×10^{-8} W m^{-2} K^{-4}), ϵ is the emissivity of the surface (0.35 for tungsten), A is the surface area of the tungsten filament, and T_0 is the room temperature (293 K). For $T^4 \gg T_0^4$, the equation becomes

$$P_{\text{radiated}} = \epsilon\sigma_S A T^4$$

Assuming that all the electrical power is radiated as electromagnetic waves from the surface, estimate the temperature of the filament and compare it with your answer in part (a).

SOLUTION

a. When the bulb is operating at 120 V, it is dissipating 40 W, which means that the current is

$$I = \frac{P}{V} = \frac{40\text{ W}}{120\text{ V}} = 0.333\text{ A}$$

The resistance of the filament at the operating temperature T must be

$$R = \frac{V}{I} = \frac{120}{0.333} = 360\,\Omega$$

Since $R = \rho L/A$, the resistivity of tungsten at the operating temperature T must be

$$\rho(T) = \frac{R(\pi D^2/4)}{L} = \frac{360\,\Omega\,\pi(33 \times 10^{-6}\text{ m})^2}{4(0.381\text{ m})} = 8.08 \times 10^{-7}\,\Omega\,\text{m}$$

But, $\rho(T) = \rho_0(T/T_0)^{1.2}$, so that

$$T = T_0\left(\frac{80.8 \times 10^{-8}}{5.51 \times 10^{-8}}\right)^{1/1.2}$$

$$= 2746\text{ K} \quad \text{or} \quad 2473\text{ °C} \quad \text{(melting temperature of W is about 3680, K)}$$

b. To calculate T from the radiation law, we note that $T = [P_{\text{radiated}}/\epsilon\sigma_S A]^{1/4}$.

The surface area is

$$A = L(\pi D) = (0.381)(\pi\,33 \times 10^{-6}) = 3.95 \times 10^{-5}\text{ m}^2$$

Then,

$$T = \left[\frac{P_{\text{radiated}}}{\epsilon\sigma_S A}\right]^{1/4} = \left[\frac{40\text{ W}}{(0.35)(5.67 \times 10^{-8}\text{ W m}^{-2}\text{ K}^{-4})(3.95 \times 10^{-5}\text{ m}^2)}\right]^{1/4}$$

$$= [5.103 \times 10^{13}]^{1/4} = 2673\text{ K} \quad \text{or} \quad 2400\text{ °C}$$

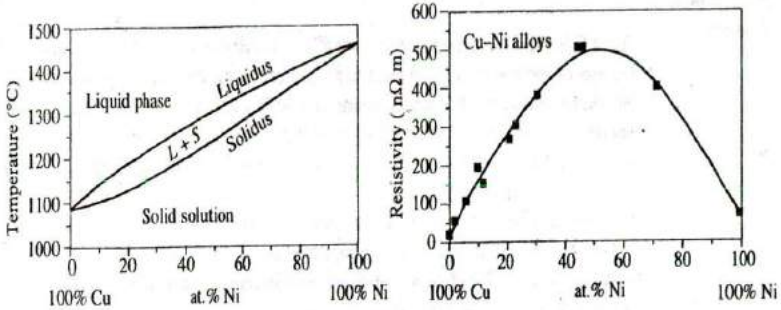
The difference between the two methods is less than 3 percent.

2.3.2 SOLID SOLUTIONS AND NORDHEIM'S RULE

In an isomorphous alloy of two metals, that is, a binary alloy that forms a solid solution, we would expect Equation 2.15 to apply, with the temperature-independent impurity contribution ρ_I increasing with the concentration of solute atoms. This means that as the alloy concentration increases, the resistivity ρ increases and becomes less temperature

Table 2.2 The effect of alloying on the resistivity

Material	Resistivity at 20 °C (nΩ m)	α at 20 °C (1/K)
Nickel	69	0.006
Chrome	129	0.003
Nichrome	1120	0.0003



(a) Phase diagram of the Cu–Ni alloy system. Above the liquidus line only the liquid phase exists. In the L + S region, the liquid (L) and solid (S) phases coexist whereas below the solidus line, only the solid phase (a solid solution) exists.

(b) The resistivity of the Cu–Ni alloy as a function of Ni content (at.%) at room temperature.

Figure 2.11 The Cu–Ni alloy system.
SOURCE: Data extracted from *Metals Handbook*, 10th ed., 2 and 3, Metals Park, Ohio: ASM, 1991, and M. Hansen and K. Anderko, *Constitution of Binary Alloys*, New York: McGraw-Hill, 1958.

dependent as ρ_I overwhelms ρ_T, leading to $\alpha \ll 1/273$. This is the advantage of alloys in resistive components. Table 2.2 shows that when 80% nickel is alloyed with 20% chromium, the resistivity of Ni increases almost 16 times. In fact, the alloy is called **nichrome** and is widely used as a heater wire in household appliances and industrial furnaces.

As a further example of the resistivity of a solid solution, consider the copper–nickel alloy. The phase diagram for this alloy system is shown in Figure 2.11a. It is clear that the alloy forms a one-phase solid solution for all compositions. Both Cu and Ni have the same FCC crystal structure, and since the Cu atom is only slightly larger than the Ni atom by about ~3 percent (easily checked on the Periodic Table), the Cu–Ni alloy will therefore still be FCC, but with Cu and Ni atoms randomly mixed, resulting in a solid solution. When Ni is added to copper, the impurity resistivity ρ_I in Equation 2.15 will increase with the Ni concentration. Experimental results for this alloy system are shown in Figure 2.11b. It should be apparent that when we reach 100% Ni, we again have a pure metal whose resistivity must be small. Therefore, ρ versus Ni concentration must pass through a maximum, which for the Cu–Ni alloy seems to be at around ~50% Ni.

There are other binary solid solutions that reflect similar behavior to that depicted in Figure 2.11, such as Cu–Au, Ag–Au, Pt–Pd, Cu–Pd, to name a few. Quite often, the use of an alloy for a particular application is necessitated by the mechanical properties, rather than the desired electrical resistivity alone. For example, brass, which is 70% Cu–30% Zn in solid solution, has a higher strength compared to pure copper; as such, it is a suitable metal for the prongs of an electrical plug.

Nordheim's rule for solid solutions

An important semiempirical equation that can be used to predict the resistivity of an alloy is **Nordheim's rule** which relates the impurity resistivity ρ_I to the atomic fraction X of solute atoms in a solid solution, as follows:

$$\rho_I = CX(1 - X) \qquad [2.21]$$

where C is the constant termed the **Nordheim coefficient,** which represents the effectiveness of the solute atom in increasing the resistivity. Nordheim's rule assumes that the solid solution has the solute atoms randomly distributed in the lattice, and these random distributions of impurities cause the electrons to become scattered as they whiz around the crystal. For sufficiently small amounts of impurity, experiments show that the increase in the resistivity ρ_I is nearly always simply proportional to the impurity concentration X, that is, $\rho_I \propto X$, which explains the initial approximately equal increments of rise in the resistivity of copper with 1.11% Ni and 2.16% Ni additions as shown in Figure 2.9. For dilute solutions, Nordheim's rule predicts the same linear behavior, that is, $\rho_I = CX$ for $X \ll 1$.

Table 2.3 lists some typical Nordheim coefficients for various additions to copper and gold. The value of the Nordheim coefficient depends on the type of solute and the solvent. A solute atom that is drastically different in size to the solvent atom will result in a bigger increase in ρ_I and will therefore lead to a larger C. An important assumption

Table 2.3 Nordheim coefficient C (at 20 °C) for dilute alloys obtained from $\rho_I = CX$ and $X < 1$ at.%*

Solute in Solvent (element in matrix)	C (nΩ m)	Maximum Solubility at 25 °C (at.%)
Au in Cu matrix	5500	100
Mn in Cu matrix	2900	24
Ni in Cu matrix	1200	100
Sn in Cu matrix	2900	0.6
Zn in Cu matrix	300	30
Cu in Au matrix	450	100
Mn in Au matrix	2410	25
Ni in Au matrix	790	100
Sn in Au matrix	3360	5
Zn in Au matrix	950	15

*NOTE: For many isomorphous alloys C may be different at higher concentrations; that is, it may depend on the composition of the alloy.

SOURCES: D.G. Fink and D. Christiansen, eds., *Electronics Engineers' Handbook*, 2nd ed., New York, McGraw-Hill, 1982. J. K. Stanley, *Electrical and Magnetic Properties of Metals*, Metals Park, OH, American Society for Metals, 1963. Solubility data from M. Hansen and K. Anderko, *Constitution of Binary Alloys*, 2nd ed., New York, McGraw-Hill, 1985.

in Nordheim's rule in Equation 2.21 is that the alloying does not significantly vary the number of conduction electrons per atom in the alloy. Although this will be true for alloys with the same valency, that is, from the same column in the Periodic Table (*e.g.*, Cu–Au, Ag–Au), it will not be true for alloys of different valency, such as Cu and Zn. In pure copper, there is just one conduction electron per atom, whereas each Zn atom can donate two conduction electrons. As the Zn content in brass is increased, more conduction electrons become available per atom. Consequently, the resistivity predicted by Equation 2.21 at high Zn contents is greater than the actual value because C refers to dilute alloys. To get the correct resistivity from Equation 2.21 we have to lower C, which is equivalent to using an effective Nordheim coefficient C_{eff} that decreases as the Zn content increases. In other cases, for example, in Cu–Ni alloys, we have to increase C at high Ni concentrations to account for additional electron scattering mechanisms that develop with Ni addition. Nonetheless, the Nordheim rule is still useful for predicting the resistivities of dilute alloys, particularly in the low-concentration region.

Combined Matthiessen and Nordheim rules

With Nordheim's rule in Equation 2.21, the resistivity of an alloy of composition X is

$$\rho = \rho_{\text{matrix}} + CX(1 - X) \qquad [2.22]$$

where $\rho_{\text{matrix}} = \rho_T + \rho_R$ is the resistivity of the matrix due to scattering from thermal vibrations and from other defects, in the absence of alloying elements. To reiterate, the value of C depends on the alloying element and the matrix. For example, C for gold in copper would be different than C for copper in gold, as shown in Table 2.3.

In solid solutions, at some concentrations of certain binary alloys, such as 75% Cu–25% Au and 50% Cu–50% Au, the annealed solid has an orderly structure; that is, the Cu and Au atoms are not randomly mixed, but occupy regular sites. In fact, these compositions can be viewed as pure compound—like the solids Cu_3Au and CuAu. The resistivities of Cu_3Au and CuAu will therefore be less than the same composition random alloy that has been quenched from the melt. As a consequence, the resistivity ρ versus composition X curve does not follow the dashed parabolic curve throughout; rather, it exhibits sharp falls at these special compositions, as illustrated in Figure 2.12.

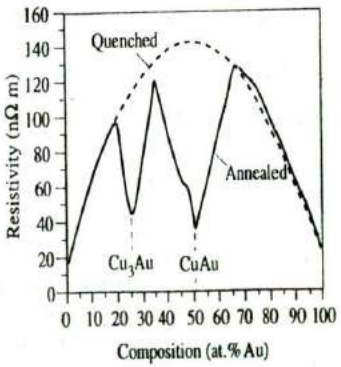


Figure 2.12 Electrical resistivity vs. composition at room temperature in Cu–Au alloys.

The quenched sample (dashed curve) is obtained by quenching the liquid, and the Cu and Au atoms are randomly mixed. The resistivity obeys the Nordheim rule. When the quenched sample is annealed or the liquid is slowly cooled (solid curve), certain compositions (Cu_3Au and CuAu) result in an ordered crystalline structure in which the Cu and Au atoms are positioned in an ordered fashion in the crystal and the scattering effect is reduced.

EXAMPLE 2.11 **NORDHEIM'S RULE** The alloy 90 wt.% Au–10 wt.% Cu is sometimes used in low-voltage dc electrical contacts, because pure gold is mechanically soft and the addition of copper increases the hardness of the metal without sacrificing the corrosion resistance. Predict the resistivity of the alloy and compare it with the experimental value of 108 nΩ m.

SOLUTION

We apply Equation 2.22, $\rho(X) = \rho_{Au} + CX(1 - X)$ but with 10 wt.% Cu converted to the atomic fraction for X. If w is the weight fraction of Cu, $w = 0.1$, and if M_{Au} and M_{Cu} are the atomic masses of Au and Cu, then the atomic fraction X of Cu is given by (see Example 1.2),

$$X = \frac{w/M_{Cu}}{w/M_{Cu} + (1 - w)/M_{Au}} = \frac{0.1/63.55}{(0.1/63.55) + (0.90/197)} = 0.256$$

Given that $\rho_{Au} = 22.8$ nΩ m and $C = 450$ nΩ m,

$$\rho = \rho_{Au} + CX(1 - X) = (22.8 \text{ n}\Omega \text{ m}) + (450 \text{ n}\Omega \text{ m})(0.256)(1 - 0.256)$$

$$= 108.5 \text{ n}\Omega \text{ m}$$

This value is only 0.5% different from the experimental value.

EXAMPLE 2.12 **RESISTIVITY DUE TO IMPURITIES** The mean speed of conduction electrons in copper is about 1.5×10^6 m s^{-1}. Its room temperature resistivity is 17 nΩ m, and the atomic concentration N_{at} in the crystal is 8.5×10^{22} cm^{-3}. Suppose that we add 1 at.% Au to form a solid solution. What is the resistivity of the alloy, the effective mean free path, and the mean free path due to collisions with Au atoms only?

SOLUTION

According to Table 2.3, the Nordheim coefficient C of Au in Cu is 5500 nΩ m. With $X = 0.01$ (1 at.%), the overall resistivity from Equation 2.22 is

$$\rho = \rho_{matrix} + CX(1 - X) = 17 \text{ n}\Omega \text{ m} + (5500 \text{ n}\Omega \text{ m})(0.01)(1 - 0.01)$$

$$= 17 \text{ n}\Omega \text{ m} + 54.45 \text{ n}\Omega \text{ m} = 71.45 \text{ n}\Omega \text{ m}$$

Suppose that ℓ is the overall or effective mean free path and τ is the effective mean free time between scattering events (includes both scattering from lattice vibrations and impurities). Since $\ell = u\tau$, and the effective drift mobility $\mu_d = e\tau/m_e$, the expression for the conductivity becomes

Conductivity and mean free path

$$\sigma = en\mu_d = \frac{e^2 n\tau}{m_e} = \frac{e^2 n\ell}{m_e u}$$

We can now calculate the effective mean free path ℓ in the alloy given that copper has a valency of 1 and the electron concentration $n = N_{at}$,

$$\frac{1}{71.5 \times 10^{-9} \ \Omega \text{ m}} = \frac{(1.6 \times 10^{-19}\text{C})^2(8.5 \times 10^{28} \text{ m}^{-3})\ell}{(9.1 \times 10^{-31}\text{kg})(1.5 \times 10^6 \text{ m s}^{-1})}$$

which gives $\ell = 8.8$ nm. We can repeat the calculation for pure copper using $\sigma = 1/\rho_{matrix} = 1/(17 \times 10^{-9} \ \Omega \text{ m})$ to find $\ell_{Cu} = 37$ nm. The mean free path is reduced approximately by 4 times by adding only 1 at.% Au. The mean free path ℓ_i due to scattering from impurities only can be found from Equation 2.13 multiplied through by $1/u$, or by using

Matthiessen's rule in Equation 2.14:

$$\frac{1}{\ell} = \frac{1}{\ell_{Cu}} + \frac{1}{\ell_I}$$

Substituting $\ell_{Cu} = 37$ nm and $\ell = 8.8$ nm, we find $\ell_I = 11.5$ nm.

We can take these calculations one step further. If N_I is the impurity concentration in the alloy, then $N_I = 0.01 N_{at} = 0.01(8.5 \times 10^{28}\ \text{m}^{-3}) = 8.5 \times 10^{26}\ \text{m}^{-3}$. The mean separation d_I between the impurities can be estimated roughly from $d_I \approx 1/N_I^{1/3}$, which gives $d_I \approx 1.0$ nm. It is clear that not all Au atoms can be involved in scattering the electrons since ℓ_I is much longer than d_I. (Another way to look at it is to say that it takes more than just one collision with an impurity to randomize the velocity of the electron.)

2.4 RESISTIVITY OF MIXTURES AND POROUS MATERIALS

2.4.1 HETEROGENEOUS MIXTURES

Nordheim's rule only applies to solid solutions that are single-phase solids. In other words, it is valid for homogeneous mixtures in which the atoms are mixed at the atomic level throughout the solid, as in the Cu–Ni alloy. The classic problem of determining the effective resistivity of a multiphase solid is closely related to the evaluation of the effective dielectric constant, effective thermal conductivity, effective elastic modulus, effective Poisson's ratio, etc., for a variety of mixtures, including such composite materials as fiberglass. Indeed, many of the mixture rules are identical.

Consider a material with two distinct phases α and β, which are stacked in layers as illustrated in Figure 2.13a. Let us evaluate the effective resistivity for current flow

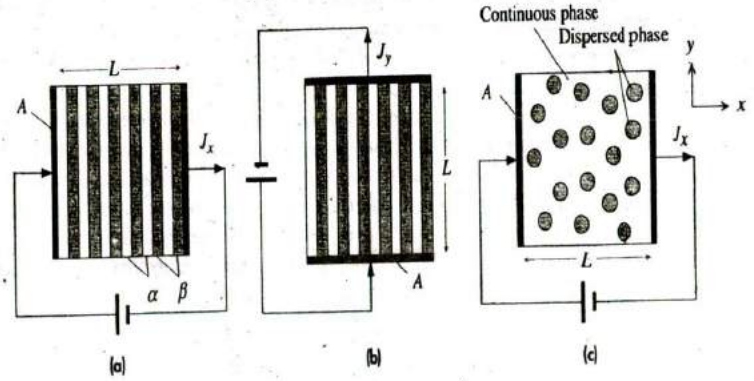


Figure 2.13 The effective resistivity of a material with a layered structure.
(a) Along a direction perpendicular to the layers.
(b) Along a direction parallel to the plane of the layers.
(c) Materials with a dispersed phase in a continuous matrix.

in the x direction. Since the layers are in series, the effective resistance R_{eff} for the whole material is

Effective resistance

$$R_{\text{eff}} = \frac{L_\alpha \rho_\alpha}{A} + \frac{L_\beta \rho_\beta}{A} \qquad [2.23]$$

where L_α is the total length (thickness) of the α-phase layers, and L_β is the total length of the β-phase layers, $L_\alpha + L_\beta = L$ is the length of the sample, and A is the cross-sectional area. Let χ_α and χ_β be the volume fractions of the α and β phases. The effective resistance is defined by

$$R_{\text{eff}} = \frac{L\rho_{\text{eff}}}{A}$$

where ρ_{eff} is the **effective resistivity.** Using $\chi_\alpha = L_\alpha/L$ and $\chi_\beta = L_\beta/L$ in Equation 2.23, we find

Resistivity–mixture rule

$$\rho_{\text{eff}} = \chi_\alpha \rho_\alpha + \chi_\beta \rho_\beta \qquad [2.24]$$

which is called the **resistivity–mixture rule** (or the **series rule of mixtures**).

If we are interested in the effective resistivity in the y direction, as shown in Figure 2.13b, obviously the α and β layers are in parallel, so an effective conductivity could be calculated in the same way as we did for the series case to find the **parallel rule of mixtures,** that is,

Conductivity–mixture rule

$$\sigma_{\text{eff}} = \chi_\alpha \sigma_\alpha + \chi_\beta \sigma_\beta \qquad [2.25]$$

where σ is the electrical conductivity of those phases identified by the subscript. Notice that the parallel rule uses the conductivity, and the series rule uses the resistivity. Equation 2.25 is often referred to as the **conductivity–mixture rule.**

Although these two rules refer to special cases, in general, for a random mixture of phase α and phase β, we would not expect either equation to apply rigorously. When the resistivities of two randomly mixed phases are not markedly different, the series mixture rule can be applied at least approximately, as we will show in Example 2.13.

However, if the resistivity of one phase is appreciably different than the other, there are two semiempirical rules that are quite useful in materials engineering.[5] Consider a heterogeneous material that has a dispersed phase (labeled d), in the form of particles, in a continuous phase (labeled c) that acts as a matrix, as depicted in Figure 2.13c. Assume that ρ_c and ρ_d are the resistivities of the continuous and dispersed phases, and χ_c and χ_d are their volume fractions. If the dispersed phase is much more resistive with respect to the matrix, that is, $\rho_d > 10\rho_c$, then

Mixture rule

$$\rho_{\text{eff}} = \rho_c \frac{\left(1 + \frac{1}{2}\chi_d\right)}{(1 - \chi_d)} \qquad (\rho_d > 10\rho_c) \qquad [2.26]$$

[5] Over the years, the task of predicting the resistivity of a mixture has challenged many theorists and experimentalists, including Lord Rayleigh who, in 1892, published an excellent exposition on the subject in the Philosophical Magazine. An extensive treatment of mixtures can be found in a paper by J. A. Reynolds and J. M. Hough published in 1957 (Proceedings of the Physical Society, 70, no. 769, London), which contains nearly all the mixture rules for the resistivity.

On the other hand, if $\rho_d < (\rho_c/10)$, then

$$\rho_{\text{eff}} = \rho_c \frac{(1 - \chi_d)}{(1 + 2\chi_d)} \qquad (\rho_d < 0.1\rho_c) \qquad [2.27]$$

Mixture rule

We therefore have at least four mixture rules at our disposal, the uses of which depend on the mixture geometry and the resistivities of the various phases. The problem is identifying which one to use for a given material, which in turn requires a knowledge of the microstructure and properties of the constituents. It should be emphasized that, at best, Equations 2.24 to 2.27 provide only a reasonable estimate of the effective resistivity of the mixture.[6]

Equations 2.26 and 2.27 are simplified special cases of a more general mixture rule due to Reynolds and Hough (1957). Consider a mixture that consists of a continuous conducting phase with a conductivity σ_c that has dispersed spheres of another phase of conductivity σ_d and of volume fraction χ, similar to Figure 2.13c. The effective conductivity of the mixture is given by

Reynolds and Hough rule for mixture of dispersed phases

$$\frac{\sigma - \sigma_c}{\sigma + 2\sigma_c} = \chi \frac{\sigma_d - \sigma_c}{\sigma_d + 2\sigma_c} \qquad [2.28]$$

It is assumed that the spheres are randomly dispersed in the material. It is left as an exercise to show that if $\sigma_d \ll \sigma_c$, then Equation 2.28 reduces to Equation 2.26. A good application would be the calculation of the effective resistivity of porous carbon electrodes, which can be 50–100 percent higher than the resistivity of bulk polycrystalline carbon (graphite). If, on the other hand, $\sigma_d \gg \sigma_c$, the dispersed phase is very conducting, for example, silver particles mixed into a graphite paste to increase the conductivity of the paste, then Equation 2.28 reduces to Equation 2.27. The usefulness of Equation 2.28 cannot be underestimated inasmuch as there are many types of materials in engineering that are mixtures of one type or another.

EXAMPLE 2.13

THE RESISTIVITY-MIXTURE RULE Consider a two-phase alloy consisting of phase α and phase β randomly mixed as shown in Figure 2.14a. The solid consists of a random mixture of two types of resistivities, ρ_α of α and ρ_β of β. We can divide the solid into a bundle of N parallel fibers of length L and cross-sectional area A/N, as shown in Figure 2.14b. In this fiber (infinitesimally thin), the α and β phases are in series, so if $\chi_\alpha = V_\alpha/V$ is the volume fraction of phase α and χ_β is that of β, then the total length of all α regions present in the fiber is $\chi_\alpha L$, and the total length of β regions is $\chi_\beta L$. The two resistances are in series, so the fiber resistance is

$$R_{\text{fiber}} = \frac{\rho_\alpha(\chi_\alpha L)}{(A/N)} + \frac{\rho_\beta(\chi_\beta L)}{(A/N)}$$

But the resistance of the solid is made up of N such fibers in parallel, that is,

$$R_{\text{solid}} = \frac{R_{\text{fiber}}}{N} = \frac{\rho_\alpha \chi_\alpha L}{A} + \frac{\rho_\beta \chi_\beta L}{A}$$

---

[6] More accurate mixture rules have been established for various types of mixtures with components possessing widely different properties, which the keen reader can find in P. L. Rossiter, *The Electrical Resistivity of Metals and Alloys* (Cambridge University Press, Cambridge, 1987).

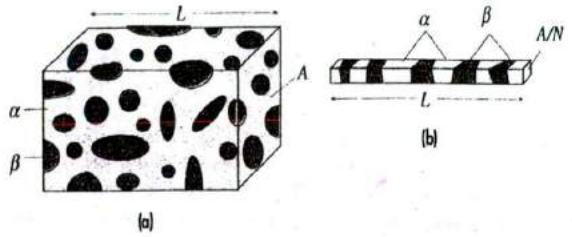


Figure 2.14
(a) A two-phase solid.
(b) A thin fiber cut out from the solid.

By definition, $R_{solid} = \rho_{eff} L/A$, where ρ_{eff} is the effective resistivity of the material, so

$$\frac{\rho_{eff} L}{A} = \frac{\rho_\alpha \chi_\alpha L}{A} + \frac{\rho_\beta \chi_\beta L}{A}$$

Thus, for a two-phase solid, the effective resistivity will be

Resistivity mixture rule

$$\rho_{eff} = \chi_\alpha \rho_\alpha + \chi_\beta \rho_\beta$$

If the densities of the two phases are not too different, we can use weight fractions instead of volume fractions. The series rule fails when the resistivities of the phases are vastly different. A major (and critical) tacit assumption here is that the current flow lines are all parallel, so that no current crosses from one fiber to another. Only then can we say that the effective resistance is R_{fiber}/N.

EXAMPLE 2.14 **A COMPONENT WITH DISPERSED AIR PORES** What is the effective resistivity of 95/5 (95% Cu–5% Sn) bronze, which is made from powdered metal containing dispersed pores at 15% (volume percent, vol.%). The resistivity of 95/5 bronze is 1×10^{-7} Ω m.

SOLUTION

Pores are infinitely more resistive ($\rho_d = \infty$) than the bronze matrix, so we use Equation 2.26,

$$\rho_{eff} = \rho_c \frac{1 + \frac{1}{2}\chi_d}{1 - \chi_d} = (1 \times 10^{-7}\ \Omega\ \text{m}) \frac{1 + \frac{1}{2}(0.15)}{1 - 0.15} = 1.27 \times 10^{-7}\ \Omega\ \text{m}$$

EXAMPLE 2.15 **COMBINED NORDHEIM AND MIXTURE RULES** Brass is an alloy composed of Cu and Zn. The alloy is a solid solution for Zn content less than 30 wt.%. Consider a brass component made from sintering 90 at.% Cu and 10 at.% Zn brass powder. The component contains dispersed air pores at 15% (vol.%). The Nordheim coefficient C of Zn in Cu is 300 nΩ m, under very dilute conditions. Each Zn atom donates two, whereas each Cu atom of the matrix donates one conduction electron, so that the Cu–Zn alloy has a higher electron concentration than in the Cu crystal itself. Predict the effective resistivity of this brass component.

SOLUTION

We first calculate the resistivity of the alloy without the pores, which forms the continuous phase in the powdered material. The simple Nordheim's rule predicts that

$$\rho_{brass} = \rho_{copper} + CX(1 - X) = 17\ \text{n}\Omega\ \text{m} + 300(0.1)(1 - 0.1) = 44\ \text{n}\Omega\ \text{m}$$

The experimental value, about 40 nΩ m, is actually less because Zn has a valency of 2, and when a Zn atom replaces a host Cu atom, it donates two electrons instead of one. We can very roughly adjust the calculated resistivity by noting that a 10 at.% Zn addition increases the

conduction electron concentration by 10% and hence reduces the resistivity ρ_{brass} by 10% to 40 nΩ m.

The powdered metal has $\chi_d = 0.15$, which is the volume fraction of the dispersed phase, that is, the air pores, and $\rho_c = \rho_{brass} = 40$ nΩ m is the resistivity of the continuous matrix. The effective resistivity of the powdered metal is given by

$$\rho_{eff} = \rho_c \frac{1 + \frac{1}{2}\chi_d}{1 - \chi_d} = (40 \text{ n}\Omega \text{ m}) \frac{1 + \frac{1}{2}(0.15)}{1 - (0.15)} = 50.6 \text{ n}\Omega \text{ m}$$

If we use the simple conductivity mixture rule, ρ_{eff} is 47.1 nΩ m, and it is underestimated.

The *effective* Nordheim coefficient C_{eff} at the composition of interest is about 255 nΩ m, which would give $\rho_{brass} = \rho_o + C_{eff} X(1 - X) = 40$ nΩ m. It is left as an exercise to show that the effective number of conduction electrons per atom in the alloy is $1 + X$ so that we must divide the ρ_{brass} calculated above by $(1 + X)$ to obtain the correct resistivity of brass if we use the listed value of C under dilute conditions. (See Question 2.8.)

2.4.2 TWO-PHASE ALLOY (Ag–Ni) RESISTIVITY AND ELECTRICAL CONTACTS

Certain binary alloys, such as Pb–Sn and Cu–Ag, only exhibit a single-phase alloy structure over very small composition ranges. For most compositions, these alloys form a two-phase heterogeneous mixture of phases α and β. A typical phase diagram for such a eutectic binary alloy system is shown in Figure 2.15a, which could be a

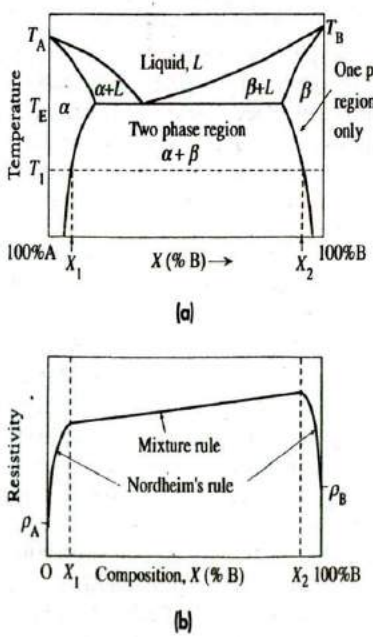


Figure 2.15 Eutectic-forming alloys, e.g., Cu–Ag.
(a) The phase diagram for a binary, eutectic-forming alloy.
(b) The resistivity versus composition for the binary alloy.

schematic scheme for the Cu–Ag system or the Pb–Sn system. The phase diagram identifies the phases existing in the alloy at a given temperature and composition. If the overall composition X is less than X_1, then at T_1, the alloy will consist of phase α only. This phase is Cu rich. When the composition X is between X_1 and X_2, then the alloy will consist of the two phases α and β randomly mixed. The phase α is Cu rich (that is, it has composition X_1) and the phase β is Ag rich (composition X_2). The relative amounts of each phase are determined by the well-known **lever rule,** which means that we can determine the volume fractions of α and β, χ_α and χ_β, as the alloy composition is changed from X_1 to X_2.

For this alloy system, the dependence of the resistivity on the alloy composition is shown in Figure 2.15b. Between O and X_1 (% Ag), the solid is one phase (isomorphous); therefore, in this region, ρ increases with the concentration of Ag by virtue of Nordheim's rule. At X_1, the solubility limit of Ag in Cu is reached, and after X_1, a second phase, which is β rich, is formed. Thus, in the composition range X_1 to X_2, we have a mixture of α and β phases, so ρ is given by Equation 2.24 for mixtures and is therefore less than that for a single-phase alloy of the same composition. Similarly, at the Ag end ($X_2 < X < 100\%$), as Cu is added to Ag, between 100% Ag and the solubility limit at X_2, the resistivity is determined by Nordheim's rule. The expected behavior of the resistivity of an eutectic binary alloy over the whole composition range is therefore as depicted in Figure 2.15b.

Electrical, thermal, and other physical properties make copper the most widely used metallic conductor. For many electrical applications, high-conductivity copper, having extremely low oxygen and other impurity contents, is produced. Although aluminum has a conductivity of only about half that of copper, it is also frequently used as an electrical conductor. On the other hand, silver has a higher conductivity than copper, but its cost prevents its use, except in specialized applications. Switches often have silver contact specifications, though it is likely that the contact metal is actually a silver alloy. In fact, silver has the highest electrical and thermal conductivity and is consequently the natural choice for use in electrical contacts. In the form of alloys with various other metals, it is used extensively in make-and-break switching applications for currents of up to about 600 A. The precious metals, gold, platinum, and palladium, are extremely resistant to corrosion; consequently, in the form of various alloys, particularly with Ag, they are widely used in electrical contacts. For example, Ag–Ni alloys are common electrical contact materials for the switches in many household appliances.

It is frequently necessary to improve the mechanical properties of a metal alloy without significantly impairing its electrical conductivity. Solid-solution alloying improves mechanical strength, but at the expense of conductivity. A compromise must often be found between electrical and mechanical properties. Most often, strength is enhanced by introducing a second phase that does not have such an adverse effect on the conductivity. For example, Ag–Pd alloys form a solid solution such that the resistivity increases appreciably due to Nordheim's rule. The resistivity of Ag–Pd is mainly controlled by the scattering of electrons from Pd atoms randomly mixed in the Ag matrix. In contrast, Ag and Ni form a two-phase alloy, a mixture of Ag-rich and Ni-rich phases. The Ag–Ni alloy is almost as strong as the Ag–Pd alloy, but it has a lower resistivity because the mixture rule volume averages the two resistivities.

2.5 THE HALL EFFECT AND HALL DEVICES

An important phenomenon that we can comfortably explain using the "electron as a particle" concept is the Hall effect, which is illustrated in Figure 2.16. When we apply a magnetic field in a perpendicular direction to the applied field (which is driving the current), we find there is a transverse field in the sample that is perpendicular to the direction of both the applied field $\mathcal{E}_x$ and the magnetic field B_z, that is, in the y direction. Putting a voltmeter across the sample, as in Figure 2.16, gives a voltage reading V_H. The applied field $\mathcal{E}_x$ drives a current J_x in the sample. The electrons move in the $-x$ direction, with a drift velocity v_{dx}. Because of the magnetic field, there is a force (called the **Lorentz force**) acting on each electron and given by $F_y = -ev_{dx}B_z$. The direction of this Lorentz force is the $-y$ direction, which we can show by applying the corkscrew rule, because, in vector notation, the force **F** acting on a charge q moving with a velocity **v** in a magnetic field **B** is given through the vector product

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B} \qquad [2.29]$$

Lorentz force

All moving charges experience the Lorentz force in Equation 2.29 as shown schematically in Figure 2.17. In our example of a metal in Figure 2.16, this Lorentz force is the $-y$ direction, so it pushes the electrons downward, as a result of which there is a negative charge accumulation near the bottom of the sample and a positive charge near the top of the sample, due to exposed metal ions (*e.g.*, Cu^+).

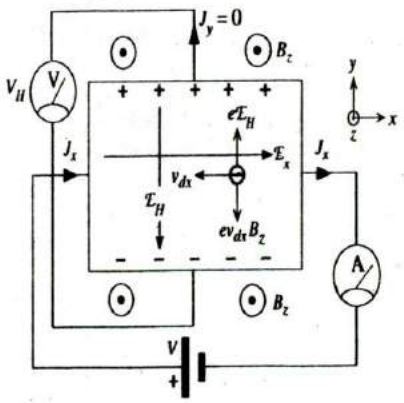


Figure 2.16 Illustration of the Hall effect. The z direction is out of the plane of the paper. The externally applied magnetic field is along the z direction.

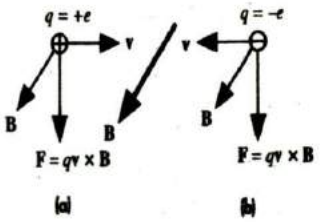


Figure 2.17 A moving charge experiences a Lorentz force in a magnetic field.
(a) A positive charge moving in the x direction experiences a force downward.
(b) A negative charge moving in the $-x$ direction also experiences a force downward.

The accumulation of electrons near the bottom results in an internal electric field $\mathcal{E}_H$ in the $-y$ direction. This is called the **Hall field** and gives rise to a Hall voltage V_H between the top and bottom of the sample. Electron accumulation continues until the increase in $\mathcal{E}_H$ is sufficient to stop the further accumulation of electrons. When this happens, the magnetic-field force $ev_{dx}B_z$ that pushes the electrons down just balances the force $e\mathcal{E}_H$ that prevents further accumulation. Therefore, in the steady state,

$$e\mathcal{E}_H = ev_{dx}B_z$$

However, $J_x = env_{dx}$. Therefore, we can substitute for v_{dx} to obtain $e\mathcal{E}_H = J_xB_z/n$ or

$$\mathcal{E}_H = \left(\frac{1}{en}\right)J_xB_z \qquad [2.30]$$

Definition of Hall coefficient

A useful parameter called the **Hall coefficient** R_H is defined as

$$R_H = \frac{\mathcal{E}_y}{J_xB_z} \qquad [2.31]$$

The quantity R_H measures the resulting Hall field, along y, per unit transverse applied current and magnetic field. The larger R_H, the greater $\mathcal{E}_y$ for a given J_x and B_z. Therefore, R_H is a gauge of the magnitude of the Hall effect. A comparison of Equations 2.30 and 2.31 shows that for metals,

Hall coefficient for electron conduction

$$R_H = -\frac{1}{en} \qquad [2.32]$$

The reason for the negative sign is that $\mathcal{E}_H = -\mathcal{E}_y$, which means that $\mathcal{E}_H$ is in the $-y$ direction.

Inasmuch as R_H depends inversely on the free electron concentration, its value in metals is much less than that in semiconductors. In fact, Hall-effect devices (such as magnetometers) always employ a semiconductor material, simply because the R_H is larger. Table 2.4 lists the Hall coefficients of various metals. Note that this is negative

Table 2.4 Hall coefficient and Hall mobility ($\mu_H = |\sigma R_H|$) of selected metals

Metal	n [m^{-3}] ($\times 10^{28}$)	R_H (Experimental) [$m^3A^{-1}s^{-1}$] ($\times 10^{-11}$)	$\mu_H = \lvert\sigma R_H\rvert$ [$m^2V^{-1}s^{-1}$] ($\times 10^{-4}$)
Ag	5.85	−9.0	57
Al	18.06	−3.5	13
Au	5.90	−7.2	31
Be	24.2	−3.4	?
Cu	8.45	−5.5	32
Ga	15.3	−6.3	3.6
In	11.49	−2.4	2.9
Mg	8.60	−9.4	22
Na	2.56	−25	53

SOURCES: Data from various sources, including C. Nording and J. Osterman, *Physics Handbook*, Bromley, England: Chartwell-Bratt Ltd., 1982.



Magnetically operated Hall-effect position sensor as available from Micro Switch.

for most metals, although a few metals exhibit a positive Hall coefficient (see Be in Table 2.4). The reasons for the latter involve the band theory of solids, which we will discuss in Chapter 4.

Since the Hall voltage depends on the product of two quantities, the current density J_x and the transverse applied magnetic field B_z, we see that the effect naturally multiplies two independently variable quantities. Therefore, it provides a means of carrying out a multiplication process. One obvious application is measuring the power dissipated in a load, where the load current and voltage are multiplied. There are many instances when it is necessary to measure magnetic fields, and the Hall effect is ideally suited to such applications. Commercial Hall-effect magnetometers can measure magnetic fields as low as 10 nT, which should be compared to the earth's magnetic field of ~50 μT. Depending on the application, manufacturers use different semiconductors to obtain the desired sensitivity. Hall-effect semiconductor devices are generally inexpensive, small, and reliable. Typical commercial, linear Hall-effect sensor devices are capable of providing a Hall voltage of ~10 mV per mT of applied magnetic field.

The Hall effect is also widely used in magnetically actuated electronic switches. The application of a magnetic field, say from a magnet, results in a Hall voltage that is amplified to trigger an electronic switch. The switches invariably use Si and are readily available from various companies. Hall-effect electronic switches are used as noncontacting keyboard and panel switches that last almost forever, as they have no mechanical contact assembly. Another advantage is that the electrical contact is "bounce" free. There are a variety of interesting applications for Hall-effect switches, ranging from ignition systems, to speed controls, position detectors, alignment controls, brushless dc motor commutators, etc.

EXAMPLE 2.16

HALL-EFFECT WATTMETER The Hall effect can be used to implement a wattmeter to measure electrical power dissipated in a load. The schematic sketch of the Hall-effect wattmeter is shown in Figure 2.18, where the Hall-effect sample is typically a semiconductor material (usually Si). The load current I_L passes through two coils, which are called current coils and are shown as C in Figure 2.18. These coils set up a magnetic field B_z such that $B_z \propto I_L$. The Hall-effect sample is positioned in this field between the coils. The voltage V_L across the load drives a current

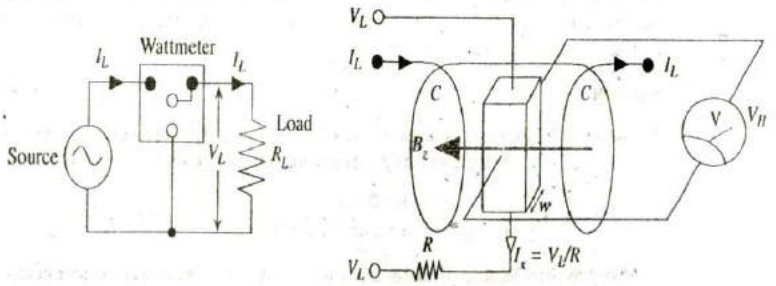


Figure 2.18 Wattmeter based on the Hall effect.

Load voltage and load current have L as subscript; C denotes the current coils for setting up a magnetic field through the Hall-effect sample (semiconductor).

$I_x = V_L/R$ through the sample, where R is a series resistance that is much larger than the resistance of the sample and that of the load. Normally, the current I_x is very small and negligible compared to the load current. If w is the width of the sample, then the measured Hall voltage is

$$V_H = wE_H = wR_H J_x B_z \propto I_x B_z \propto V_L I_L$$

which is the electrical power dissipated in the load. The voltmeter that measures V_H can now be calibrated to read directly the power dissipated in the load.

EXAMPLE 2.17 **HALL MOBILITY** Show that if R_H is the Hall coefficient and σ is the conductivity of a metal, then the drift mobility of the conduction electrons is given by

$$\mu_d = |\sigma R_H| \qquad [2.33]$$

The Hall coefficient and conductivity of copper at 300 K have been measured to be -0.55×10^{-10} m^3 A^{-1} s^{-1} and 5.9×10^7 Ω^{-1} m^{-1}, respectively. Calculate the drift mobility of electrons in copper.

SOLUTION

Consider the expression for

$$R_H = \frac{-1}{en}$$

Since the conductivity is given by $\sigma = en\mu_d$, we can substitute for en to obtain

$$R_H = \frac{-\mu_d}{\sigma} \quad \text{or} \quad \mu_d = -R_H\sigma$$

which is Equation 2.33. The drift mobility can thus be determined from R_H and σ.

The product of σ and R_H is called the **Hall mobility** μ_H. Some values for the Hall mobility of electrons in various metals are listed in Table 2.4. From the expression in Equation 2.33, we get

$$\mu_d = -(-0.55 \times 10^{-10} \text{ m}^3 \text{ A}^{-1} \text{ s}^{-1})(5.9 \times 10^7 \ \Omega^{-1} \text{ m}^{-1}) = 3.2 \times 10^{-3} \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$$

It should be mentioned that Equation 2.33 is an oversimplification. The actual relationship involves a numerical factor that multiplies the right term in Equation 2.33. The factor depends on the charge carrier scattering mechanism that controls the drift mobility.

EXAMPLE 2.18 **CONDUCTION ELECTRON CONCENTRATION FROM THE HALL EFFECT** Using the electron drift mobility from Hall-effect measurements (Table 2.4), calculate the concentration of conduction electrons in copper, and then determine the average number of electrons contributed to the free electron gas per copper atom in the solid.

SOLUTION

The number of conduction electrons is given by $n = \sigma/e\mu_d$. The conductivity of copper is 5.9×10^7 Ω^{-1} m^{-1}, whereas from Table 2.4, the electron drift mobility is 3.2×10^{-3} m^2 V^{-1} s^{-1}. So,

$$n = \frac{(5.9 \times 10^7 \ \Omega^{-1} \text{ m}^{-1})}{[(1.6 \times 10^{-19} \text{ C})(3.2 \times 10^{-3} \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1})]} = 1.15 \times 10^{29} \text{ m}^{-3}$$

Since the concentration of copper atoms is 8.5×10^{28} m^{-3}, the average number of electrons contributed per atom is $(1.15 \times 10^{29} \text{ m}^{-3})/(8.5 \times 10^{28} \text{ m}^{-3}) \approx 1.36$.

26 THERMAL CONDUCTION

26.1 THERMAL CONDUCTIVITY

Experience tells us that metals are both good electrical and good thermal conductors. We may therefore surmise that the free conduction electrons in a metal must also play a role in heat conduction. Our conjecture is correct for metals, but not for other materials. The transport of heat in a metal is accomplished by the electron gas (conduction electrons), whereas in nonmetals, the conduction is due to lattice vibrations.

When a metal piece is heated at one end, the amplitude of the atomic vibrations, and thus the average kinetic energy of the electrons, in this region increases, as depicted in Figure 2.19. Electrons gain energy from energetic atomic vibrations when the two collide. By virtue of their increased random motion, these energetic electrons then transfer the extra energy to the colder regions by colliding with the atomic vibrations there. Thus, electrons act as "energy carriers."

The thermal conductivity of a material, as its name implies, measures the ease with which heat, that is, thermal energy, can be transported through the medium. Consider the metal rod shown in Figure 2.20, which is heated at one end. Heat will flow from the hot end to the cold end. Experiments show that the rate of heat flow, $Q' = dQ/dt$, through a thin section of thickness δx is proportional to the temperature gradient $\delta T/\delta x$ and the cross-sectional area A, so

$$Q' = -A\kappa \frac{\delta T}{\delta x} \qquad [2.34]$$

Fourier's law of thermal conduction

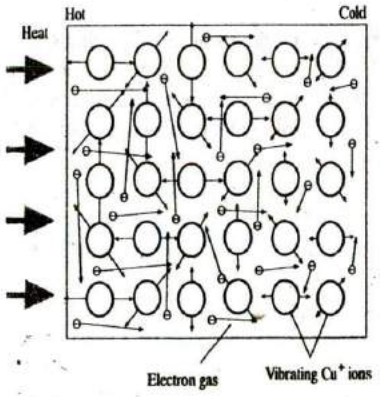


Figure 2.19 Thermal conduction in a metal involves transferring energy from the hot region to the cold region by conduction electrons.

More energetic electrons (shown with longer velocity vectors) from the hotter regions arrive at cooler regions, collide with lattice vibrations, and transfer their energy. Lengths of arrowed lines on atoms represent the magnitudes of atomic vibrations.

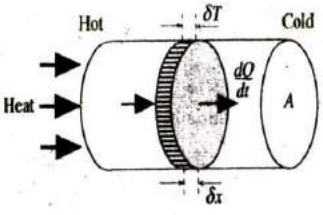


Figure 2.20 Heat flow in a metal rod heated at one end.

Consider the rate of heat flow, dQ/dt, across a thin section δx of the rod. The rate of heat flow is proportional to the temperature gradient $\delta T/\delta x$ and the cross-sectional area A.

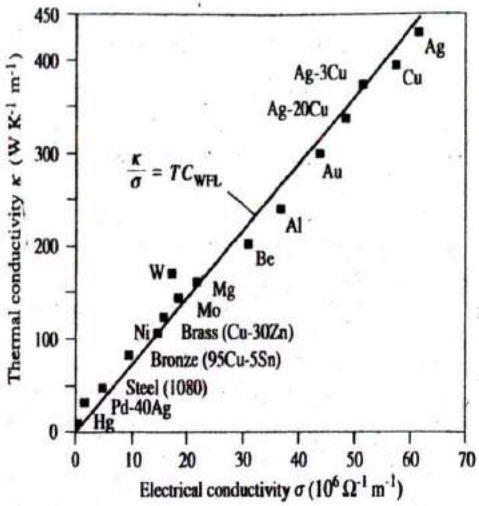


Figure 2.21 Thermal conductivity κ versus electrical conductivity σ for various metals (elements and alloys) at 20 °C.
The solid line represents the WFL law with $C_{WFL} \approx 2.44 \times 10^{-8}\ W\,\Omega\,K^{-2}$.

where κ is a material-dependent **constant of proportionality** that we call the **thermal conductivity.** The negative sign indicates that the heat flow direction is that of decreasing temperature. Equation 2.34 is often referred to as **Fourier's law** of heat conduction and is a defining equation for κ. The driving force for the heat flow is the temperature gradient $\delta T/\delta x$. If we compare Equation 2.34 with Ohm's law for the electric current I, we see that

Ohm's law of electrical conduction

$$I = -A\sigma\frac{\delta V}{\delta x} \qquad [2.35]$$

which shows that in this case, the driving force is the potential gradient, that is, the electric field.[7] In metals, electrons participate in the processes of charge and heat transport, which are characterized by σ and κ, respectively. Therefore, it is not surprising to find that the two coefficients are related by the **Wiedemann–Franz–Lorenz law,**[8] which is

Wiedemann–Franz–Lorenz law

$$\frac{\kappa}{\sigma T} = C_{WFL} \qquad [2.36]$$

where $C_{WFL} = \pi^2 k^2/3e^2 = 2.44 \times 10^{-8}\ W\,\Omega\,K^{-2}$ is a constant called the **Lorenz number** (or the Wiedemann–Franz–Lorenz coefficient).

Experiments on a wide variety of metals, ranging from pure metals to various alloys, show that Equation 2.36 is reasonably well obeyed at close to room temperature and above, as illustrated in Figure 2.21. Since the electrical conductivity of pure metals is inversely proportional to the temperature, we can immediately conclude that the thermal conductivity of these metals must be relatively temperature independent at room temperature and above.

[7] Recall that $J = \sigma\mathcal{E}$ which is equivalent to Equation 2.35.

[8] Historically, Wiedemann and Franz noted in 1853 that κ/σ is the same for all metals at the same temperature. Lorenz in 1881 showed that κ/σ is proportional to the temperature with a proportionality constant that is nearly the same for many metals. The law stated in Equation 2.36 reflects both observations. By the way, Lorenz, who was a Dane, should not be confused with Lorentz, who was Dutch.

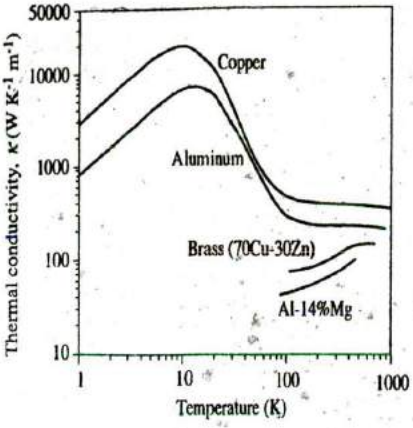


Figure 2.22 Thermal conductivity versus temperature for two pure metals (Cu and Al) and two alloys (brass and Al-14% Mg).

SOURCE: Data extracted from Y. S. Touloukian, *et al.*, *Thermophysical Properties of Matter*, vol. 1: "Thermal Conductivity, Metallic Elements and Alloys," New York: Plenum, 1970.

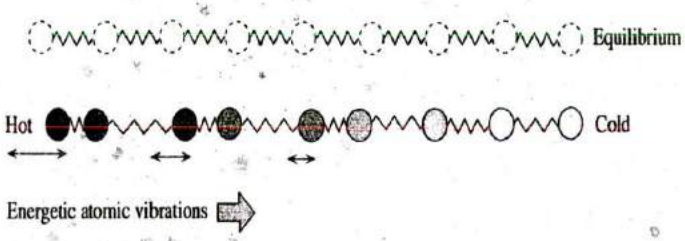


Figure 2.23 Conduction of heat in insulators involves the generation and propogation of atomic vibrations through the bonds that couple the atoms (an intuitive figure).

Figure 2.22 shows the temperature dependence of κ for copper and aluminum down to the lowest temperatures. It can be seen that for these two metals, above ~100 K, the thermal conductivity becomes temperature independent, in agreement with Equation 2.36. Qualitatively, above ~100 K, κ is constant, because heat conduction depends essentially on the rate at which the electron transfers energy from one atomic vibration to another as it collides with them (Figure 2.19). This rate of energy transfer depends on the mean speed of the electron u, which increases only fractionally with the temperature. In fact, the fractionally small inerease in u is more than sufficient to carry the energy from one collision to another and thereby excite more energetic lattice vibrations in the colder regions.

Nonmetals do not have any free conduction electrons inside the crystal to transfer thermal energy from hot to cold regions of the material. In nonmetals, the energy transfer involves lattice vibrations, that is, atomic vibrations of the crystal. We know that we can view the atoms and bonds in a crystal as balls connected together through springs as shown for one chain of atoms in Figure 2.23. As we know from the kinetic molecular theory, all the atoms would be vibrating and the average vibrational kinetic energy would be proportional to the temperature. Intuitively, as depicted in Figure 2.23, when we heat one end of a crystal, we set up large-amplitude atomic vibrations at this hot end. The springs *couple* the vibrations to neighboring atoms and thus allow the large-amplitude vibrations to propagate, as a **vibrational wave**, to the cooler regions of the crystal. If we were to grab the left-end atom in Figure 2.23 and vibrate it violently, we

would be sending vibrational waves down the ball-spring-ball chain. The efficiency of heat transfer depends not only on the efficiency of coupling between the atoms, and hence on the nature of interatomic bonding, but also on how the vibrational waves propogate in the crystal and how they are scattered by crystal imperfections and by their interactions with other vibrational waves; this topic is discussed in Chapter 4. The stronger the coupling, the greater will be the thermal conductivity, a trend that is intuitive but also borne out by experiments. Diamond has an exceptionally strong covalent bond and also has a very high thermal conductivity; $\kappa \approx 1000\ \mathrm{W\ m^{-1}\ K^{-1}}$. On the other hand, polymers have weak secondary bonding between the polymer chains and their thermal conductivities are very poor; $\kappa < 1\ \mathrm{W\ m^{-1}\ K^{-1}}$.

The thermal conductivity, in general, depends on the temperature. Different classes of materials exhibit different κ values and also different κ versus T behavior. Table 2.5

Table 2.5 Typical thermal conductivities of various classes of materials at 25 °C

Material	κ (W m^{-1} K^{-1})
Pure metal	
Nb	52
Fe	80
Zn	113
W	178
Al	250
Cu	390
Ag	420
Metal alloys	
Stainless steel	12–16
55% Cu–45% Ni	19.5
70% Ni–30% Cu	25
1080 steel	50
Bronze (95% Cu–5% Sn)	80
Brass (63% Cu–37% Zn)	125
Dural (95% Al–4% Cu–1% Mg)	147
Ceramics and glasses	
Glass-borosilicate	0.75
Silica-fused (SiO_2)	1.5
S_3N_4	20
Alumina (Al_2O_3)	30
Sapphire (Al_2O_3)	37
Beryllium (BeO)	260
Diamond	~1000
Polymers	
Polypropylene	0.12
PVC	0.17
Polycarbonate	0.22
Nylon 6,6	0.24
Teflon	0.25
Polyethylene, low density	0.3
Polyethylene, high density	0.5

summarizes κ at room temperature for various classes of materials. Notice how ceramics have a very large range of κ values.

THERMAL CONDUCTIVITY A 95/5 (95% Cu–5% Sn) bronze bearing made of powdered metal contains 15% (vol.%) porosity. Calculate its thermal conductivity at 300 K, given that the electrical conductivity of 95/5 bronze is $10^7\ \Omega^{-1}\ m^{-1}$. **EXAMPLE 2.19**

SOLUTION

Recall that in Example 2.14, we found the electrical resistivity of the same bronze by using the mixture rule in Equation 2.26 in Section 2.4. We can use the same mixture rule again here, but we need the thermal conductivity of 95/5 bronze. From $\kappa/\sigma T = C_{WFL}$, we have

$$\kappa = \sigma T C_{WFL} = (1 \times 10^7)(300)(2.44 \times 10^{-8}) = 73.2\ W\ m^{-1}\ K^{-1}$$

Thus, the effective thermal conductivity is

$$\frac{1}{\kappa_{eff}} = \frac{1}{\kappa_c}\left[\frac{1 + \frac{1}{2}\chi_d}{1 - \chi_d}\right] = \frac{1}{(73.2\ W\ m^{-1}\ K^{-1})}\left[\frac{1 + \frac{1}{2}(0.15)}{1 - 0.15}\right]$$

so that

$$\kappa_{eff} = 57.9\ W\ m^{-1}\ K^{-1}$$

2.6.2 THERMAL RESISTANCE

Consider a component of length L that has a temperature difference ΔT between its ends as in Figure 2.24a. The temperature gradient is $\Delta T/L$. Thus, the rate of heat flow, or the **heat current**, is

$$Q' = A\kappa\frac{\Delta T}{L} = \frac{\Delta T}{(L/\kappa A)} \qquad [2.37]$$ *Fourier's law*

This should be compared with Ohm's law in electric circuits,

$$I = \frac{\Delta V}{R} = \frac{\Delta V}{(L/\sigma A)} \qquad [2.38]$$ *Ohm's law*

where ΔV is the voltage difference across a conductor of resistance R, and I is the electric current.

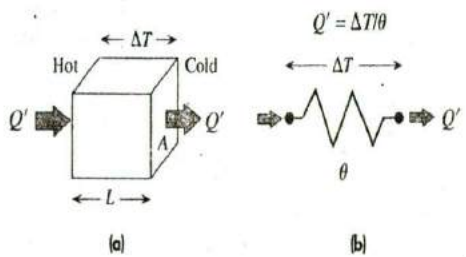


Figure 2.24 Conduction of heat through a component in (a) can be modeled as a thermal resistance θ shown in (b) where $Q' = \Delta T/\theta$.

In analogy with electrical resistance, we may define **thermal resistance** θ by

Definition of thermal resistance

$$Q' = \frac{\Delta T}{\theta} \qquad [2.39]$$

where, in terms of thermal conductivity,

Thermal resistance

$$\theta = \frac{L}{\kappa A} \qquad [2.40]$$

The rate of heat flow Q' and the temperature difference ΔT correspond to the electric current I and potential difference ΔV, respectively. Thermal resistance is the thermal analog of electrical resistance and its thermal circuit representation is shown in Figure 2.24b.

EXAMPLE 2.20 **THERMAL RESISTANCE** A brass disk of electrical resistivity 50 nΩ m conducts heat from a heat source to a heat sink at a rate of 10 W. If its diameter is 20 mm and its thickness is 30 mm, what is the temperature drop across the disk, neglecting the heat losses from the surface?

SOLUTION

We first determine the thermal conductivity:

$$\kappa = \sigma T C_{WFL} = (5 \times 10^{-8}\ \Omega\ \text{m})^{-1}(300\ \text{K})(2.44 \times 10^{-8}\ \text{W}\ \Omega\ \text{K}^{-2})$$
$$= 146\ \text{W m}^{-1}\ \text{K}^{-1}$$

The thermal resistance is

$$\theta = \frac{L}{\kappa A} = \frac{(30 \times 10^{-3}\ \text{m})}{\pi(10 \times 10^{-3}\ \text{m})^2(146\ \text{W m}^{-1}\ \text{K}^{-1})} = 0.65\ \text{K W}^{-1}$$

Therefore, the temperature drop is

$$\Delta T = \theta Q' = (0.65\ \text{K W}^{-1})(10\ \text{W}) = 6.5\ \text{K or °C}$$

2.7 ELECTRICAL CONDUCTIVITY OF NONMETALS

All metals are good conductors because they have a very large number of conduction electrons free inside the metal. We should therefore expect solids that do not have metallic bonding to be very poor conductors, indeed insulators. Figure 2.25 shows the range of conductivities exhibited by a variety of solids. Based on typical values of the conductivity, it is possible to empirically classify various materials into conductors, semiconductors, and insulators as in Figure 2.25. It is apparent that nonmetals are not perfect insulators with zero conductivity. There is no well-defined sharp boundary between what we call insulators and semiconductors. Conductors are intimately identified with metals. It is more appropriate to view insulators as **high resistivity** (or **low conductivity**) **materials.** In general terms, current conduction is due to the drift of mobile charge carriers through a solid by the application of

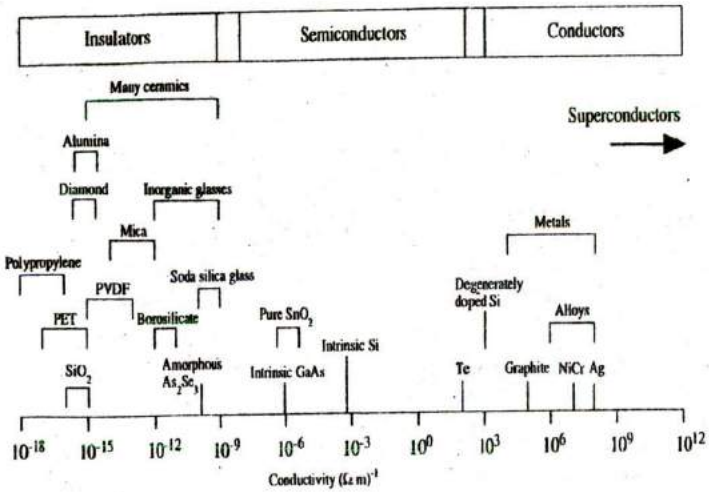


Figure 2.25 Range of conductivities exhibited by various materials.

an electric field. Each of the drifting species of charge carriers contributes to the observed current. In metals, there are only free electrons. In nonmetals there are other types of charge carriers that can drift.

2.7.1 SEMICONDUCTORS

A perfect Si crystal has each Si atom bonded to four neighbors, and each covalent bond has two shared electrons as we had shown in Figure 1.59a. We know from classical physics (the kinetic molecular theory and Boltzmann distribution) that all the atoms in the crystal are executing vibrations with a distribution of energies. As the temperature increases, the distribution spreads to higher energies. Statistically some of the atomic vibrations will be sufficiently energetic to rupture a bond as indicated in Figure 2.26a. This releases an electron from the bond which is *free* to wander inside the crystal. The free electron can drift in the presence of an applied field; it is called a **conduction electron.** As an electron has been removed from a region of the crystal that is otherwise neutral, the broken-bond region has a *net positive charge*. This broken-bond region is called a **hole** (h^+). An electron in a neighboring bond can jump and repair this bond and thereby create a hole in its original site as shown in Figure 2.26b. Effectively, the hole has been displaced in the opposite direction to the electron jump by this *bond switching*. Holes can also wander in the crystal by the repetition of bond switching. When a field is applied, both holes and electrons contribute to electrical conduction as in Figure 2.26c. For all practical purposes, these holes behave as if they were *free* positively charged particles (independent of the original electrons) inside the crystal. In the presence of an applied field, holes drift along the field direction and contribute to conduction just as the free electrons

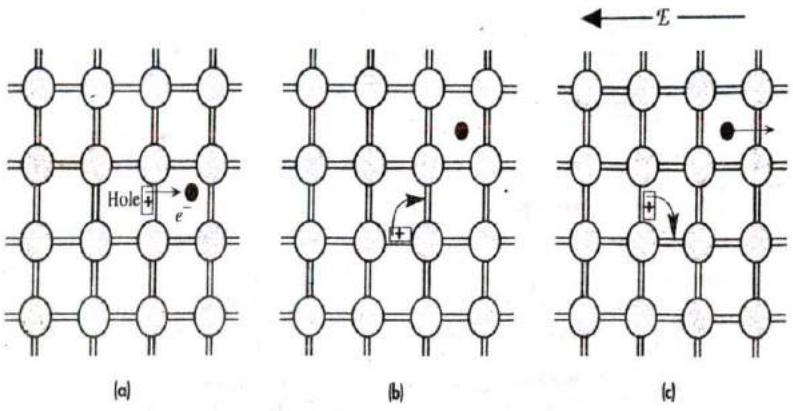


Figure 2.26

(a) Thermal vibrations of the atoms rupture a bond and release a free electron into the crystal. A hole is left in the broken bond, which has an effective positive charge.

(b) An electron in a neighboring bond can jump and repair this bond and thereby create a hole in its original site; the hole has been displaced.

(c) When a field is applied, both holes and electrons contribute to electrical conduction.

released from the broken bonds drift in the opposite direction and contribute to conduction.

It is also possible to create free electrons or holes by intentionally doping a semiconductor crystal, that is substituting impurity atoms for some of the Si atoms. Defects can also generate free carriers. The simplest example is nonstoichiometric ZnO that is shown in Figure 1.55b which has excess Zn. The electrons from the excess Zn are free to wander in the crystal and hence contribute to conduction.

Suppose that n and p are the concentrations of electrons and holes in a semiconductor crystal. If electrons and holes have drift mobilities of μ_e and μ_h, respectively, then the overall conductivity of the crystal is given by

Conductivity of a semiconductor

$$\sigma = ep\mu_h + en\mu_e \qquad [2.41]$$

Unless a semiconductor has been heavily doped, the concentrations n and p are much smaller than the electron concentration in a metal. Even though carrier drift mobilities in most semiconductors are higher than electron drift mobilities in metals, semiconductors have much lower conductivities due to their lower concentration of free charge carriers.

EXAMPLE 2.21 **HALL EFFECT IN SEMICONDUCTORS** The Hall effect in a sample where there are both negative and positive charge carriers, for example, electrons and holes in a semiconductor, involves not only the concentrations of electrons and holes, n and p, respectively, but also the electron and hole drift mobilities, μ_e and μ_h. We first have to reinterpret the relationship between the drift velocity and the electric field $\mathcal{E}$.

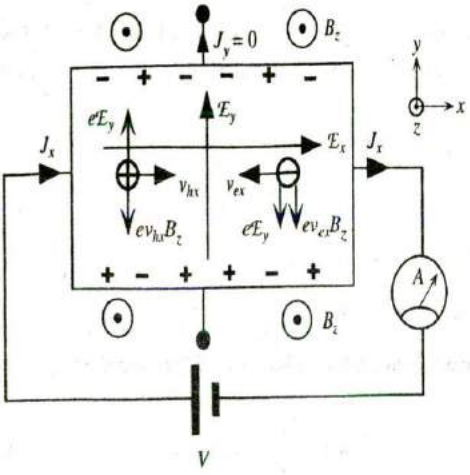


Figure 2.27 Hall effect for ambipolar conduction as in a semiconductor where there are both electrons and holes.

The magnetic field B_z is out from the plane of the paper. Both electrons and holes are deflected toward the bottom surface of the conductor and consequently the Hall voltage depends on the relative mobilities and concentrations of electrons and holes.

If μ_e is the drift mobility and v_e is the drift velocity of the electrons, then we already know that $v_e = \mu_e \mathcal{E}$. This has been derived by considering the *net electrostatic force* $e\mathcal{E}$ acting on a single electron and the imparted acceleration $a = e\mathcal{E}/m_e$. The drift is therefore due to the net force $F_{net} = e\mathcal{E}$ experienced by a conduction electron. If we were to keep $e\mathcal{E}$ as the *net force* F_{net} acting on a single electron, then we would have found

$$v_e = \frac{\mu_e}{e} F_{net} \qquad [2.42]$$

Drift velocity and net force

Equation 2.42 emphasizes the fact that drift is due to a net force F_{net} acting on an electron. A similar expression would also apply to the drift of a hole in a semiconductor.

When both electrons and holes are present in a semiconductor sample, both charge carriers experience a Lorentz force in the same direction since they would be drifting in the opposite directions as illustrated in Figure 2.27. Thus, both holes and electrons tend to pile near the bottom surface. The magnitude of the Lorentz force, however, will be different since the drift mobilities and hence drift velocities will be different in general. Once equilibrium is reached, there should be no current flowing in the y direction as we have an open circuit. Suppose that more holes have accumulated near the bottom surface so there is a built-in electric field $\mathcal{E}_y$ along y as shown in Figure 2.27. Suppose that v_{ey} and v_{hy} are the *usual* electron and hole drift velocities in the $-y$ and $+y$ directions, respectively, as if the electric field $\mathcal{E}_y$ existed alone in the $+y$ direction. The net current along y is zero, which means that

$$J_y = J_h + J_e = ep v_{hy} + en v_{ey} = 0 \qquad [2.43]$$

From Equation 2.43 we obtain

$$p v_{hy} = -n v_{ey} \qquad [2.44]$$

We note that either the electron or the hole drift velocity must be reversed from its usual direction; for example, holes drifting in the opposite direcion to $\mathcal{E}_y$ The net force acting on the charge carriers cannot be zero. This is impossible when two types of carriers are involved and both carriers are drifting along y to give a net current J_y that is zero. This is what Equation 2.43 represents. We therefore conclude that, along y, both the electron and the hole must experience a

driving force to drift them. The net force experienced by the carriers, as shown in Figure 2.27, is

$$F_{hy} = e\mathcal{E}_y - ev_{hx}B_z \quad \text{and} \quad -F_{ey} = e\mathcal{E}_y + ev_{ex}B_z \tag{2.45}$$

where v_{hx} and v_{ex} are the hole and electron drift velocities, respectively, along x. In general, the drift velocity is determined by the net force acting on a charge carrier; that is, from Equation 2.42

$$F_{hy} = \frac{ev_{hy}}{\mu_h} \quad \text{and} \quad -F_{ey} = \frac{ev_{ey}}{\mu_e}$$

so that Equation 2.45 becomes,

$$\frac{ev_{hy}}{\mu_h} = e\mathcal{E}_y - ev_{hx}B_z \quad \text{and} \quad \frac{ev_{ey}}{\mu_e} = e\mathcal{E}_y + ev_{ex}B_z$$

where v_{hy} and v_{ey} are the hole and electron drift velocities along y. Substituting $v_{hx} = \mu_h\mathcal{E}_x$ and $v_{ex} = \mu_e\mathcal{E}_x$, these become

$$\frac{v_{hy}}{\mu_h} = \mathcal{E}_y - \mu_h\mathcal{E}_xB_z \quad \text{and} \quad \frac{v_{ey}}{\mu_e} = \mathcal{E}_y + \mu_e\mathcal{E}_xB_z \tag{2.46}$$

From Equation 2.46 we can substitute for v_{hy} and v_{ey} in Equation 2.44 to obtain

$$p\mu_h\mathcal{E}_y - p\mu_h^2\mathcal{E}_xB_z = -n\mu_e\mathcal{E}_y - n\mu_e^2\mathcal{E}_xB_z$$

or

$$\mathcal{E}_y(p\mu_h + n\mu_e) = B_z\mathcal{E}_x(p\mu_h^2 - n\mu_e^2) \tag{2.47}$$

We now consider what happens along the x direction. The total current density is finite and is given by the usual expression,

Current density along x

$$J_x = epv_{hx} + env_{ex} = (p\mu_h + n\mu_e)e\mathcal{E}_x \tag{2.48}$$

We can use Equation 2.48 to substitute for $\mathcal{E}_x$ in Equation 2.47, to obtain

$$e\mathcal{E}_y(n\mu_e + p\mu_h)^2 = B_zJ_x(p\mu_h^2 - n\mu_e^2)$$

The Hall coefficient, by definition, is $R_H = \mathcal{E}_y/J_xB_z$, so

Hall effect for ambipolar conduction

$$R_H = \frac{p\mu_h^2 - n\mu_e^2}{e(p\mu_h + n\mu_e)^2} \tag{2.49}$$

or

Hall effect for ambipolar conduction

$$R_H = \frac{p - nb^2}{e(p + nb)^2} \tag{2.50}$$

where $b = \mu_e/\mu_h$. It is clear that the Hall coefficient depends on both the drift mobility ratio and the concentrations of holes and electrons. For $p > nb^2$, R_H will be positive and for $p < nb^2$, it will be negative. We should note that when only one type of carrier is involved, for example, electrons only, the $J_y = 0$ requirement means that $J_y = env_{ey} = 0$, or $v_{ey} = 0$. The drift velocity along y can only be zero, if the net driving force F_{ey} along y is zero. This occurs when $e\mathcal{E}_y - ev_{ex}B_z = 0$, that is, when the Lorentz force just balances the force due to the built-in field.

EXAMPLE 2.22 **HALL COEFFICIENT OF INTRINSIC SILICON** At room temperature, a pure silicon crystal (called **intrinsic silicon**) has electron and hole concentrations $n = p = n_i = 1.5 \times 10^{10}$ cm^{-3}, and electron and hole drift mobilities $\mu_e = 1350$ cm^2 V^{-1} s^{-1} and $\mu_h = 450$ cm^2 V^{-1} s^{-1}. Calculate the Hall coefficient and compare it with a typical metal.

SOLUTION

Given $n = p = n_i = 1.5 \times 10^{10}$ cm^{-3}, $\mu_e = 1350$ cm^2 V^{-1} s^{-1}, and $\mu_h = 450$ cm^2 V^{-1} s^{-1}, we have

$$b = \frac{\mu_e}{\mu_h} = \frac{1350}{450} = 3$$

Then from Equation 2.50,

$$R_H = \frac{(1.5 \times 10^{16}\ \text{m}^{-3}) - (1.5 \times 10^{16}\ \text{m}^{-3})(3)^2}{(1.6 \times 10^{-19}\ \text{C})[(1.5 \times 10^{16}\ \text{m}^{-3}) + (1.5 \times 10^{16}\ \text{m}^{-3})(3)]^2}$$

$$= -208\ \text{m}^3\ \text{A}^{-1}\ \text{s}^{-1}$$

which is orders of magnitude larger than that for a typical metal. All Hall-effect devices use a semiconductor rather than a metal sample.

2.7.2 IONIC CRYSTALS AND GLASSES

Figure 2.28a shows how crystal defects in an ionic crystal lead to mobile charges that can contribute to the conduction process. All crystalline solids possess vacancies and interstitial atoms as a requirement of thermal equilibrium. Many solids have interstitial impurities which are often ionized or charged. These interstitial ions can jump, *i.e.*, diffuse, from one interstitial site to another and hence drift by diffusion in the presence of a field. A positive ion at an interstitial site such as that shown in Figure 2.28a always prefers to jump into a neighboring interstitial site along the direction of the field because it experiences an effective force in this direction. When an ion with

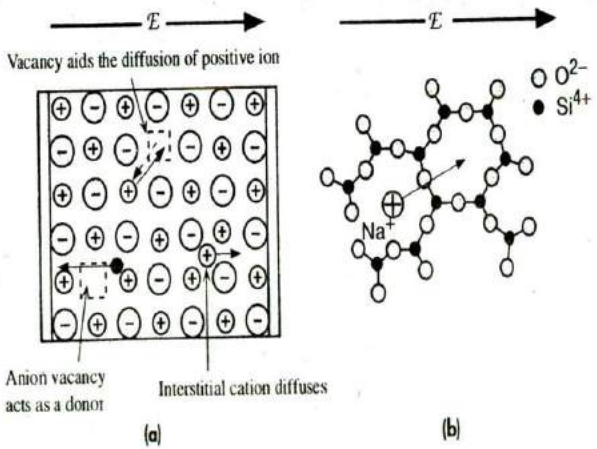
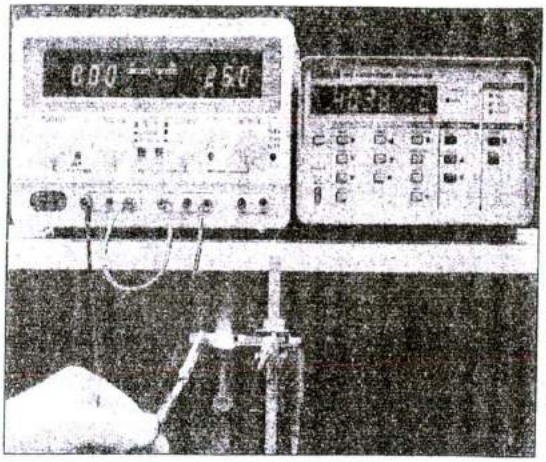


Figure 2.28 Possible contributions to the conductivity of ceramic and glass insulators.
(a) Possible mobile charges in a ceramic.
(b) An Na^+ ion in the glass structure diffuses and therefore drifts in the direction of the field.



This soda glass rod when heated under a torch becomes electrically conducting. It passes 4 mA when the voltage is 50 V (2 × 25 V); a resistance of 12.5 kΩ! Ordinary soda glass at room temperature is an insulator but can be quite conducting at sufficiently high temperatures.

charge q_{ion} jumps a distance d along the field, its potential energy decreases by $q_{ion}\mathcal{E}d$. If it tries to jump in the opposite direction, it has to do work $q_{ion}\mathcal{E}d$ against the force of the field.

Deviations from stoichiometry in compound solids often lead to the generation of mobile electrons (or holes) and point defects such as vacancies. Therefore, there are electrons, holes, and various mobile ions available for conduction under an applied field as depicted in Figure 2.28a. Many glasses and polymers contain a certain concentration of mobile ions in the structure. An example of a Na^+ ion in silica glass is shown in Figure 2.28b. Aided by the field, the Na^+ can jump from one interstice to a neighboring interstice along the field and thereby drift in the glass and contribute to current conduction. The conduction process is then essentially field-directed diffusion. Ordinary window glass, in fact, has a high concentration of Na^+ ions in the structure and becomes reasonably conducting above 300–400 °C. Some polymers may contain ions derived from the polymerization process, from the local degradation (dissociation) of the polymer itself, or from water absorption.

Conductivity σ of the material depends on all the conduction mechanisms with each species of charge carrier making a contribution, so it is given by

General conductivity

$$\sigma = \sum q_i n_i \mu_i \qquad [2.51]$$

where n_i is the concentration, q_i is the charge carried by the charge carrier species of type i (for electrons and holes $q_i = e$), and μ_i is the drift mobility of these carriers. The dominant conduction mechanism in Equation 2.51 is often quite difficult to uniquely identify. Further, it may change with temperature, composition, and ambient conditions such as the air pressure as in some oxide ceramics. For many insulators, whether ceramic, glass, or polymer, it has been found that, in the majority of cases, the conductivity follows an exponential or Arrhenius-type temperature dependence so that σ is

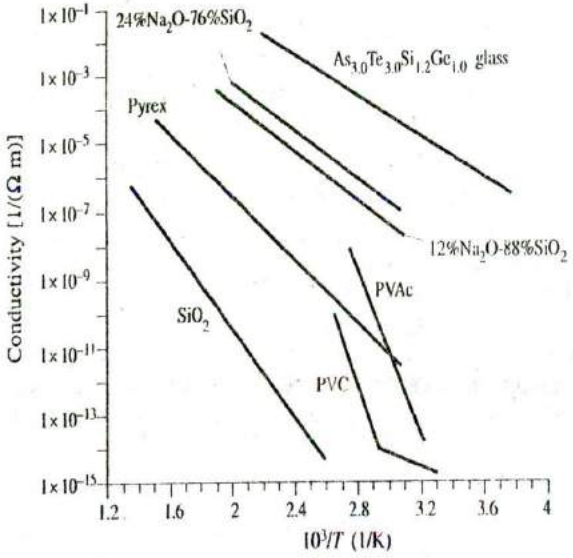


Figure 2.29 Conductivity versus reciprocal temperature for various low-conductivity solids.
[SOURCE: Data selectively combined from numerous sources.

thermally activated,

$$\sigma = \sigma_o \exp\left(-\frac{E_\sigma}{kT}\right) \qquad [2.52]$$

Temperature dependence of conductivity

where E_σ is the **activation energy for conductivity.**

Figure 2.29 shows examples of the temperature dependence of conductivity for various high-resistivity solids: oxide ceramics, glasses, and polymers. When Equation 2.51 is plotted as $\log(\sigma)$ versus $1/T$, the result is a straight line with a negative slope that indicates the activation energy E_σ. Equation 2.52 is useful in predicting the conductivity at different temperatures and evaluating the temperature stability of the insulator.

EXAMPLE 2.23

CONDUCTIVITY OF A SODA–SILICATE GLASS Figure 2.29 shows the temperature dependence of 12% Na_2O–88% SiO_2, soda–silicate glass which has 12 mol% Na_2O and 88 mol% SiO_2. Calculate the activation energy of conductivity and compare this with the activation energy for the diffusion of Na^+ ions in the soda–silicate glass structure which is in the range 0.65–0.75 eV.

SOLUTION

According to Equation 2.52 when $\ln(\sigma)$ is plotted against $1/T$, the slope should be $-E_\sigma/k$. If the conductivity at temperatures T_1 and T_2 are σ_1 and σ_2, respectively, then the slope of the straight

line for 12% Na_2O–88% SiO_2 in Figure 2.29 is

$$\text{Slope} = \frac{\ln(\sigma_2/\sigma_1)}{(1/T_2 - 1/T_1)} = -\frac{E_\sigma}{k}$$

Taking $\sigma_1 = 10^{-4}\ \Omega^{-1}\ m^{-1}$ and $\sigma_2 = 10^{-6}\ \Omega^{-1}\ m^{-1}$ in Figure 2.29, we find $1/T_1 = 0.00205$ and $1/T_2 = 0.00261$. Then, E_σ as eV is

$$E_\sigma = -\frac{\ln(\sigma_2/\sigma_1)}{(1/T_2 - 1/T_1)}\frac{k}{e} = -\frac{\ln(10^{-6}/10^{-4})}{(0.00261 - 0.00205)}\frac{1.38 \times 10^{-23}}{1.602 \times 10^{-19}} = 0.71\ \text{eV}$$

A similar calculation for the 24% Na_2O–76% SiO_2 gives an activation energy of 0.69 eV.

Both of these activation energies are comparable with the activation energy for the diffusion of Na^+ ions in the structure. Thus, Na^+ diffusion is responsible for the conductivity.

EXAMPLE 2.24 **DRIFT MOBILITY DUE TO IONIC CONDUCTION** The soda–silicate glass of composition 20% Na_2O–80% SiO_2 and density of approximately 2.4 g cm^{-3} has a conductivity of $8.25 \times 10^{-6}\ \Omega^{-1}\ m^{-1}$ at 150 °C. If conduction occurs by the diffusion of Na^+ ions, what is their drift mobility?

SOLUTION

We can calculate the drift mobility μ_i of the Na^+ ions from the conductivity expression $\sigma = q_i n_i \mu_i$ where q_i is the charge of the ion Na^+, so that it is $+e$, and n_i is the concentration of Na^+ ions in the structure. For simplicity we can take the glass to be made of $(Na_2O)_{0.2}(SiO_2)_{0.8}$ units. The atomic masses of Na, O, and Si are 23, 16, and 28.1, respectively. The atomic mass of $(Na_2O)_{0.2}(SiO_2)_{0.8}$ is

$$M_{at} = 0.2[2(23) + 1(16)] + 0.8[1(28.1) + 2(16)]$$
$$= 60.48\ \text{g mol}^{-1}\ \text{of}\ (Na_2O)_{0.2}(SiO_2)_{0.8}$$

The number of $(Na_2O)_{0.2}(SiO_2)_{0.8}$ units per unit volume can be found from the density d by

$$n = \frac{dN_A}{M_{at}} = \frac{(2.4 \times 10^3\ \text{kg m}^{-3})(6.02 \times 10^{23}\ \text{mol}^{-1})}{(10^{-3}\ \text{kg/g})(60.48\ \text{g mol}^{-1})}$$
$$= 2.39 \times 10^{28}(Na_2O)_{0.2}(SiO_2)_{0.8}\ \text{units m}^{-3}$$

The concentration n_i of Na^+ ions is the concentration of Na atoms as each would be ionized. Then n_i can be expressed as $n_i = n_{Na} = [\text{atomic fraction of Na in } (Na_2O)_{0.2}(SiO_2)_{0.8}] \times n$.

$$n_i = \left[\frac{0.2(2)}{0.2(2+1) + 0.8(1+2)}\right](2.39 \times 10^{28}\ \text{m}^{-3}) = 3.186 \times 10^{27}\ \text{m}^{-3}$$

and

$$\mu_i = \frac{\sigma}{en_i} = \frac{(8.25 \times 10^{-6}\ \Omega^{-1}\ \text{m}^{-1})}{(1.60 \times 10^{-19}\ \text{C})(3.186 \times 10^{27}\ \text{m}^{-3})} = 1.62 \times 10^{-14}\ \text{m}^2\ \text{V}^{-1}\ \text{s}^{-1}$$

This is an extremely small drift mobility, by orders of magnitude, compared with the typical electron drift mobility in metals and semiconductors. The reason is that the drift involves the Na^+ ion jumping from one site to another by a diffusion process. This diffusion requires overcoming a potential energy barrier, typically 0.5 to 1 eV, which limits drastically the rate of diffusion by virtue of the Boltzmann factor.

ADDITIONAL TOPICS

2.8 SKIN EFFECT: HF RESISTANCE OF A CONDUCTOR

Consider the cylindrical conductor shown in Figure 2.30a, which is carrying a current I into the paper (×). The magnetic field B of I is clockwise. Consider two magnetic field values B_1 and B_2, which are shown in Figure 2.30a. B_1 is inside the core and B_2 is just outside the conductor.

Assume that the conductor is divided into two conductors. The hypothetical cut is taken just outside of B_1. The conductor in Figure 2.30a is now cut into a hollow cylinder and a smaller solid cylinder, as shown in Figure 2.30b and c, respectively. The currents I_1 and I_2 in the solid and hollow cylinders sum to I. We can arrange things and choose B_1 such that our cut gives $I_1 = I_2 = \frac{1}{2}I$. Obviously, I_1 flowing in the inner conductor is threaded (or linked) by both B_1 and B_2. (Remember that B_1 is just inside the conductor in Figure 2.30b, so it threads at least 99% of I_1.) On the other hand, the outer conductor is only threaded by B_2, simply because I_2 flows in the hollow cylinder and there is no current in the hollow, which means that B_1 is not threaded by I_2. Clearly, I_1 threads more magnetic field than I_2 and thus conductor (c) has a higher inductance than (b). Recall that **inductance** is defined as the *total magnetic flux threaded per unit current*. Consequently, an ac current will prefer paths near the surface where the inductive impedance is smaller. As the frequency increases, the current is confined more and more to the surface region.

For a given conductor, we can assume that most of the current flows in a surface region of depth δ, called the **skin depth,** as indicated in Figure 2.31. In the central region,

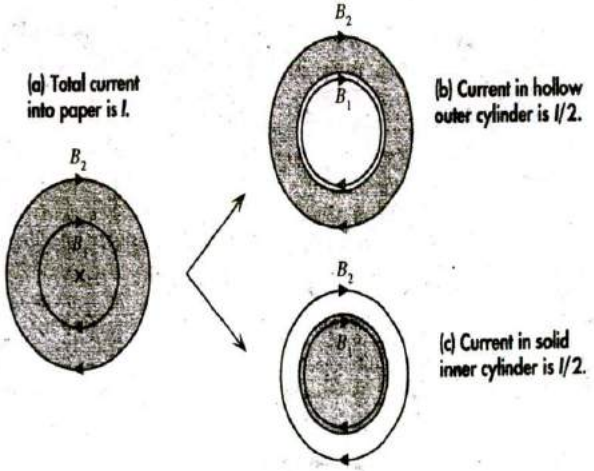


Figure 2.30 Illustration of the skin effect.
A hypothetical cut produces a hollow outer cylinder and a solid inner cylinder. Cut is placed where it would give equal current in each section. The two sections are in parallel so that the currents in (b) and (c) sum to that in (a).

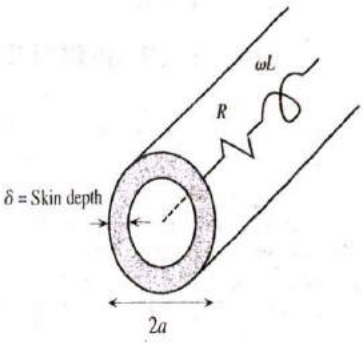


Figure 2.31 At high frequencies, the core region exhibits more inductive impedance than the surface region, and the current flows in the surface region of a conductor defined approximately by the skin depth, δ.

the current will be negligibly small. The skin depth will obviously depend on the frequency ω. To find δ, we must solve Maxwell's equations in a conductive medium, a tedious task that, fortunately, has been done by others. We can therefore simply take the result that the skin depth δ is given by

Skin depth for conduction

$$\delta = \frac{1}{\sqrt{\frac{1}{2}\omega\sigma\mu}} \qquad [2.53]$$

where ω is the angular frequency of the current, σ is the conductivity (σ is constant from dc up to $\sim 10^{14}$ Hz in metals), and μ is the magnetic permeability of the medium, which is the product of the absolute (free space) permeability μ_o and the relative permeability μ_r.

We can imagine the central conductor as a resistance R in series with an inductance L. Intuitively, those factors that enhance the inductive impedance ωL over the resistance R will also tend to emphasize the skin effect and will hence tend to decrease the skin depth. For example, the greater the permeability of the conducting medium, the stronger the magnetic field inside the conductor, and hence the larger the inductance of the central region. The higher the frequency of the current, the greater the inductive impedance ωL compared with R and the more significant is the skin effect. The greater is the conductivity σ, the smaller is R compared with ωL and hence the more important is the skin effect. All these dependences are accounted for in Equation 2.53.

With the skin depth known, the effective cross-sectional area is given approximately by

$$A = \pi a^2 - \pi(a - \delta)^2 \approx 2\pi a\delta$$

where δ^2 is neglected ($\delta \ll a$). The ac resistance r_{ac} of the conductor per unit length is therefore

HF resistance per unit length due to skin effect

$$r_{ac} = \frac{\rho}{A} \approx \frac{\rho}{2\pi a\delta} \qquad [2.54]$$

where ρ is the ac resistivity at the frequency of interest, which for all practical purposes is equal to the dc resistivity of the metal. Equation 2.54 clearly shows that as ω increases, δ decreases, by virtue of $\delta \propto \omega^{-1/2}$ and, as a result, r_{ac} increases.

From this discussion, it is obvious that the skin effect arises because the magnetic field of the ac current in the conductor restricts the current flow to the surface region within a depth of $\delta < a$. Since the current can only flow in the surface region, there is an effective increase in the resistance due to a decrease in the cross-sectional area for current flow. Taking this effective area for current flow as $2\pi a\delta$ leads to Equation 2.54.

The skin effect plays an important role in electronic engineering because it limits the use of solid-core conductors in high-frequency applications. As the signal frequencies reach and surpass the gigahertz (10^9 Hz) range, the transmission of the signal over a long distance becomes almost impossible through an ordinary, solid-metal conductor. We must then resort to pipes (or waveguides).

SKIN EFFECT FROM DIMENSIONAL ANALYSIS Using dimensional analysis, obtain the general form of the equation for the skin depth δ in terms of the angular frequency of the current ω, conductivity σ, and permeability μ. **EXAMPLE 2.25**

SOLUTION

The skin effect depends on the angular frequency ω of the current, the conductivity σ, and the magnetic permeability μ of the conducting medium. In the most general way, we can group these effects as

$$[\delta] = [\omega]^x[\sigma]^y[\mu]^z$$

where the indices x, y, and z are to be determined. We then substitute the dimensions of each quantity in this expression. The dimensions of each, in terms of the fundamental units, are as follows:

Quantity	Units	Fundamental Units	Comment
δ	m	m	
ω	s^{-1}	s^{-1}	
σ	$\Omega^{-1}\,m^{-1}$	$C^2\,s\,kg^{-1}\,m^{-3}$	$\Omega = V\,A^{-1} = (J\,C^{-1})(C\,s^{-1})^{-1}$ $= N\,m\,s\,C^{-2} = (kg\,m\,s^{-2})(m\,s\,C^{-2})$
μ	$Wb\,A^{-1}\,m^{-1}$	$kg\,m\,C^{-2}$	$Wb = T\,m^2 = (N\,A^{-1}\,m^{-1})(m^2)$ $= (kg\,m\,s^{-2})(C^{-1}\,s)(m)$

Therefore,

$$[m] = [s^{-1}]^x[C^2\,s\,kg^{-1}\,m^{-3}]^y[kg\,m\,C^{-2}]^z$$

Matching the dimensions of both sides, we see that $y = z$; otherwise C and kg do not cancel.

For m	$1 = -3y + z$		
For s	$0 = -x + y$		
For C or kg	$0 = 2y - 2z$	or	$0 = -y + z$

Clearly, $x = y = z = -\frac{1}{2}$ is the only possibility. Then, $\delta \propto [\omega\sigma\mu]^{-1/2}$. It should be reemphasized that the dimensional analysis is not a proof of the skin depth expression, but a consistency check that assures confidence in the equation.

EXAMPLE 2.26 **SKIN EFFECT IN AN INDUCTOR** What is the change in the dc resistance of a copper wire of radius 1 mm for an ac signal at 10 MHz? What is the change in the dc resistance at 1 GHz? Copper has $\rho_{dc} = 1.70 \times 10^{-8}\ \Omega$ m or $\sigma_{dc} = 5.9 \times 10^{7}\ \Omega^{-1}\ m^{-1}$ and a relative permeability near unity.

SOLUTION

Per unit length, $r_{dc} = \rho_{dc}/\pi a^2$ and at high frequencies, from Equation 2.54, $r_{ac} = \rho_{dc}/2\pi a\delta$. Therefore, $r_{ac}/r_{dc} = a/2\delta$.

We need to find δ. From Equation 2.53, at 10 MHz we have

$$\delta = \left[\tfrac{1}{2}\omega\sigma_{dc}\mu\right]^{-1/2} = \left[\tfrac{1}{2} \times 2\pi \times 10 \times 10^{6} \times 5.9 \times 10^{7} \times 1.257 \times 10^{-6}\right]^{-1/2}$$
$$= 2.07 \times 10^{-5}\ \text{m} = 20.7\ \mu\text{m}$$

Thus

$$\frac{r_{ac}}{r_{dc}} = \frac{a}{2\delta} = \frac{(10^{-3}\ \text{m})}{(2 \times 2.07 \times 10^{-5}\ \text{m})} = 24.13$$

The resistance has increased by 24 times. At 1 GHz, the increase is 240 times. Furthermore, the current is confined to a surface region of about $\sim 2 \times 10^{-5}$ (20 μm) at 10 MHz and $\sim 2 \times 10^{-6}$ m (2 μm) at 1 GHz, so most of the material is wasted. This is exactly the reason why solid conductors would not be used for high-frequency work. As very high frequencies, in the gigahertz range and above, are reached, the best bet would be to use pipes (waveguides).

One final comment is appropriate. An inductor wound from a copper wire would have a certain Q (quality factor) value[9] that depends inversely on its resistance. At high frequencies, Q would drop, because the current would be limited to the surface of the wire. One way to overcome this problem is to use a thick conductor that has a surface coating of higher-conductivity metal, such as silver. This is what the early radio engineers practiced. In fact, tank circuits of high-power radio transmitters often have coils made from copper tubes with a coolant flowing inside.

2.9 THIN METAL FILMS

2.9.1 CONDUCTION IN THIN METAL FILMS

The resistivity of a material, as listed in materials tables and in our analysis of conduction, refers to the resistivity of the material in bulk form; that is, any dimension of the specimen is much larger than the mean free path for electron scattering. In such cases resistivity is determined by scattering from lattice vibrations and, if significant, scattering from various impurities and defects in the crystal. In certain applications,

[9] The Q value refers to the quality factor of an inductor, which is defined by $Q = \omega_o L/R$, where ω_o is the resonant frequency, L is the inductance, and R is the resistance due to the losses in the inductor.

notably microelectronics, metal films are widely used to provide electrical conduction paths to and from the semiconductor devices. Various methods are used to deposit thin films. In many applications, the metal film is simply deposited onto a substrate, such as a semiconductor or an insulator (*e.g.*, SiO_2), by **physical vapor deposition** (PVD), that is, by **vacuum deposition,** which typically involves either evaporation or sputtering. In **thermal evaporation,** the metal is evaporated from a heated source in a vacuum chamber as depicted in Figure 1.74. As the metal atoms, evaporated from the source, impinge and adhere to the semiconductor surface, they form a metal film which is often highly polycrystalline. Stated differently, the metal atoms in the vapor condense to form a metal film on a suitably placed substrate. In **electron beam deposition,** an energetic electron beam is used to melt and evaporate the metal. **Sputtering** is a vacuum deposition process that involves bombarding a metal target material with energetic Ar ions, which dislodges the metal atoms and then condenses them onto a substrate. The use of sputtering is quite common in microelectronic fabrication. Copper metal interconnect films used in microelectronics are usually grown by **electrodeposition,** that is, using electroplating, an electrochemical process, to deposit the metal film onto the required chip areas. In many applications, especially in microelectronics, we are interested in the resistivity of a metal film in which the thickness of the film or the average size of the grains is comparable to the mean distance between scattering events ℓ_{bulk} (the mean free path) in the bulk material. In such cases, the resistivity of the metal film is greater than the corresponding resistivity of the bulk crystal. A good example is the resistivity of interconnects and various metal films used in the "shrinking" world of microelectronics, in which more and more transistors are packed into a single Si crystal, and various device dimensions are scaled down.

2.9.2 RESISTIVITY OF THIN FILMS

Polycrystalline Films and Grain Boundary Scattering In a highly polycrystalline sample the conduction electrons are more likely to be scattered by grain boundaries than by other processes as depicted in Figure 2.32a. Consider the resistivity due to scattering from grain boundaries alone as shown in Figure 2.32b. The conduction electron is free within a grain, but becomes scattered at the grain boundary. Its mean free path ℓ_{grains} is therefore roughly equal to the average grain size d. If $\lambda = \ell_{crystal}$ is

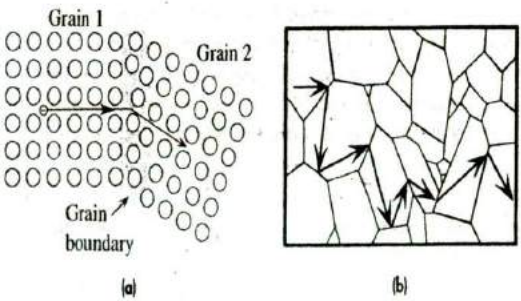


Figure 2.32
(a) Grain boundaries cause scattering of the electron and therefore add to the resistivity by Matthiessen's rule.
(b) For a very grainy solid, the electron is scattered from grain boundary to grain boundary and the mean free path is approximately equal to the mean grain diameter.

the mean free path of the conduction electrons in the *single crystal* (no grain boundaries), then

Mean free path in polycrystalline sample

$$\frac{1}{\ell} = \frac{1}{\ell_{\text{crystal}}} + \frac{1}{\ell_{\text{grains}}} = \frac{1}{\lambda} + \frac{1}{d} \qquad [2.55]$$

The resistivity is inversely proportional to the mean free path which means that the resistivity of the bulk single crystal $\rho_{\text{crystal}} \propto 1/\lambda$ and the resistivity of the polycrystalline sample $\rho \propto 1/\ell$. Thus,

Resistivity of a polycrystalline sample

$$\frac{\rho}{\rho_{\text{crystal}}} = 1 + \left(\frac{\lambda}{d}\right) \qquad [2.56]$$

Polycrystalline metal films with a smaller grain diameter d (*i.e.*, more grainy films) will have a higher resistivity.

In a more rigorous theory we have to consider a number of effects. It may take more than one scattering at a grain boundary to totally randomize the velocity, so we need to calculate the effective mean free path that accounts for how many collisions are needed to randomize the velocity. There is a possibility that the electron may be totally reflected back at a grain boundary (bounce back). Suppose that the probability of reflection at a grain boundary is R. If d is the average grain size (diameter), then the popular **Mayadas–Shatkez formula** is approximately given by[10]

Resistivity due to grain boundary scattering

$$\frac{\rho}{\rho_{\text{crystal}}} \approx 1 + 1.33\beta \qquad [2.57a]$$

where

$$\beta = \frac{\lambda}{d}\left(\frac{R}{1-R}\right) \qquad [2.57b]$$

Equation 2.57a is in the form of Matthiessen's rule and indicates that the grain boundary scattering contribution ρ_{grains} to the overall resistivity is $(1.33\beta)\rho_{\text{crystal}}$. The approximate sign in Equation 2.57 implies that Matthiessen's rule is "approximately," though reasonably well, obeyed. For copper, typical R values are 0.24 to 0.40, and R is somewhat smaller for Al. Equation 2.57 for a Cu film with $R \approx 0.3$ predicts $\rho/\rho_{\text{crystal}} \approx 1.20$ for $d \approx 3\lambda$ or a grain size $d \approx 120$ nm since the bulk crystal $\lambda \approx 40$ nm.

Surface Scattering Consider the scattering of electrons from the surfaces of a conducting film as in Figure 2.33. Take the film thickness as D. Assume that the scattering from the surface is *inelastic;* that is, the electron loses the gained velocity from the field. Put differently, the direction of the electron after the scattering process is *independent* of the direction before the scattering process. This type of scattering is called *nonspecular*. (If the electron is elastically reflected from the surface just like a rubber ball bouncing off a wall, then there is no increase in the resistivity.) It is unlikely that one surface scattering will completely randomize the electron's velocity. The mean free path ℓ_{surf} of the electron will depend on its direction right after the scattering

[10] This is obtained by expanding the original long expression about $\beta = 1$ to the first term. To two decimal places, the expansion is $1 + 1.33\beta$.

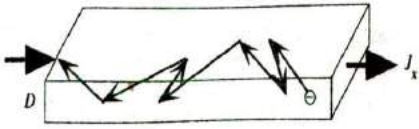


Figure 2.33 Conduction in thin films may be controlled by scattering from the surfaces.

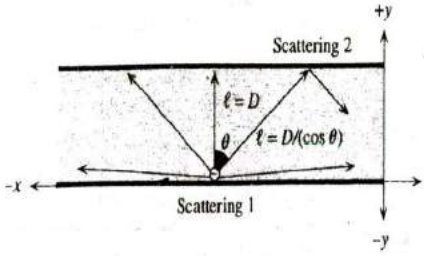


Figure 2.34 The mean free path of the electron depends on the angle θ after scattering.

process as depicted in Figure 2.34. For example, if the angle θ after surface scattering is zero, (the electron moves transversely to the film length), then $\ell_{surf} = D$. In general, the mean free path ℓ_{surf} will be $D/(\cos\theta)$ as illustrated in Figure 2.34.

Consider the surface scattering example in Figure 2.34 where the electron is scattered from the bottom surface. If the scattering of the electron were truly random, then the probability of being scattered in a direction back into the film, that is, in the $+y$ direction, would be 0.5 on average. However, the electron's direction right after the surface scattering is not totally random because we know that the electron cannot leave the film; thus θ is between $-\pi/2$ and $+\pi/2$ and cannot be between $-\pi$ and $+\pi$. The electron's velocity after the first surface scattering must have a y component along $+y$ and not along $-y$. The electron can only acquire a velocity component along $-y$ again after the second surface scattering as shown in Figure 2.34. It therefore takes two collisions to randomize the velocity, which means that the *effective mean free path* must be twice as long, that is $2D/\cos\theta$. To find the overall mean free path ℓ for calculating the resistivity we must use Matthiessen's rule. If λ is the mean free path of the conduction electrons in the *bulk crystal* (no surface scattering), then

$$\frac{1}{\ell} = \frac{1}{\lambda} + \frac{1}{\ell_{surf}} = \frac{1}{\lambda} + \frac{\cos\theta}{2D} \qquad [2.58]$$

Mean free path in a film

We have to average for all possible θ values per scattering, that is, θ from $-\pi/2$ to $+\pi/2$. Once this is done we can relate ℓ to λ as follows:

$$\frac{\lambda}{\ell} = 1 + \frac{\lambda}{\pi D}$$

Averaged mean free path in a film

The resistivity of the bulk crystal is $\rho_{bulk} \propto 1/\lambda$, and the resistivity of the film is $\rho \propto 1/\ell$. Thus,

$$\frac{\rho}{\rho_{bulk}} = 1 + \frac{1}{\pi}\left(\frac{\lambda}{D}\right) \qquad [2.59]$$

Resistivity of a conducting thin film

Table 2.6 Resistivities of some thin Cu and Au films at room temperature

Film	D (nm)	d (nm)	ρ (nΩ m)	Comment
		Cu films (Polycrystalline)		
Cu on TiN, W, and TiW [1]	>250	186	21	Chemical vapor deposition (CVD).
		45	32	Substrate temperature 200 °C. ρ depends on d not D = 250–900 nm.
Cu on 500 nm SiO_2 [2]	20.5		35	Thermal evaporation. Substrate at RT.
	37		27	
Cu on Si (100) [3]	52		38	Sputtered Cu films. Annealing at 150 °C has no effect. $R \approx 0.40$ and $p \approx 0$.
	100		22	
Cu on glass [4]	40		50	As deposited
	40		29	Annealed at 200 °C
	40		25	Annealed at 250 °C
				All thermal evaporated and PC.
		Au films		
Au epitaxial film on mica	30		25	Single crystal on mica. $p \approx 0.8$. Specular scattering.
Au PC film on mica	30		54	PC. Sputtered on mica. p is small.
Au film on glass	30		70	PC. Evaporated onto glass. p is small. Nonspecular scattering.
Au on glass [5]	40	8.5	92	PC. Sputtered films. R = 0.27–0.33.
	40	3.8	189	

NOTE: PC-polycrystalline film, RT-room temperature, D = film thickness, d = average grain size. At RT for Cu, λ = 38–40 nm, and for Au, λ = 36–38 nm.

SOURCES: Data selectively combined from various sources, including [1] S. Riedel *et al.*, *Microelec. Engin.* **33**, 165, 1997; [2] H. D. Liu *et al.*, *Thin Solid Films*, **34**, 151, 2001; [3] J. W. Lim *et al.*, *Appl. Surf. Sci.* **217**, 95, 2003. [4] R. Suri *et al.*, *J. Appl. Phys.*, **46**, 2574, 1975; [5] R. H. Cornely and T. A. Ali, *J. Appl. Phys.*, **49**, 4094, 1978.

Surface scattering resistivity

A more rigorous calculation modifies the numerical factor $1/\pi$ and also considers what fraction p of surface collisions is specular and results in[11]

$$\frac{\rho}{\rho_{\text{bulk}}} \approx 1 + \frac{3\lambda}{8D}(1-p) \qquad \frac{D}{\lambda} > 0.3 \tag{2.60}$$

which is valid down to about $D \approx 0.3\lambda$. Equation 2.60 is in Matthiessen's rule format, which means that the second term is the fractional contribution of the surfaces to the resistivity. It can be seen that for elastic or specular scattering $p = 1$ and there is no change in the resistivity. For $p = 0$, Equation 2.60 predicts $\rho/\rho_{\text{bulk}} \approx 1.20$ for roughly $D \approx 1.9\lambda$ or a thickness $D \approx 75$ nm for Cu for which $\lambda \approx 40$ nm. The value of p depends on the film preparation method (*e.g.*, sputtering, epitaxial growth) and the substrate on which the film has been deposited.

Equation 2.60 involves scattering from two surfaces, that is, from the two interfaces of the film. In general the two interfaces will not be identical and hence will have different p coefficients; p in Equation 2.60 is some mean p value. Table 2.6

[11] This is known as the **Fuchs-Sondheimer equation** in a simplified form.

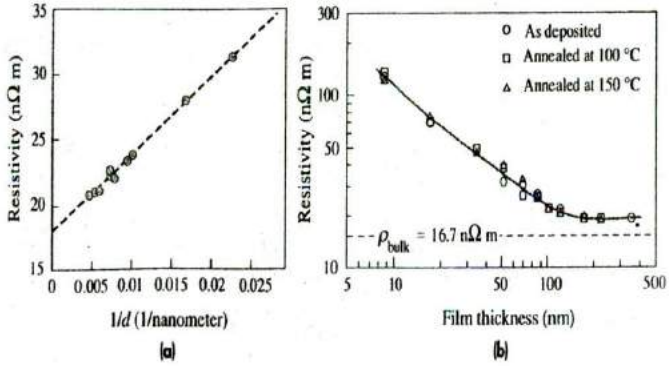


Figure 2.35

(a) ρ_{film} of Cu polycrystalline films versus reciprocal mean grain size (diameter) $1/d$. Film thickness $D = 250$–900 nm does not affect the resistivity. The straight line is $\rho_{film} = 17.8\ n\Omega\ m + (595\ n\Omega\ m\ nm)(1/d)$.

(b) ρ_{film} of thin Cu polycrystalline films versus film thickness D. In this case, annealing (heat treating) the films to reduce the polycrystallinity does not significantly affect the resistivity because ρ_{film} is controlled mainly by surface scattering.

SOURCES: Data extracted from (a) S. Riedel *et al., Microelec. Engin.* **33,** 165, 1997, and (b) W. Lim *et al., Appl. Surf. Sci.*, **217,** 95, 2003.

summarizes the resistivity of thin Cu and Au gold films deposited by various preparation techniques. Notice the large difference between the Au films deposited on a noncrystalline glass substrate and on a crystalline mica substrate. Such differences between films are typically attributed to different values of p. The p value can also change (increase) when the film is annealed. Obviously, the polycrystallinity of the film will also affect the resistivity as discussed previously. Typically, most epitaxial thin films, unless very thin ($D \ll \lambda$), deposited onto heated crystalline substrates exhibit highly specular scattering with $p = 0.9$–1.

It is generally very difficult to separate the effects of surface and grain boundary scattering in thin polycrystalline films; the contribution from grain boundary scattering is likely to exceed that from the surfaces. In any event, both contributions, by Matthiessen's general rule, increase the overall resistivity. Figure 2.35a shows an example in which the resistivity ρ_{film} of thin Cu polycrystalline films is due to grain boundary scattering, and thickness has no effect (D was 250–900 nm and much greater than λ). The resistivity ρ_{film} is plotted against the reciprocal mean grain size $1/d$, which then follows the expected linear behavior in Equation 2.57a. On the other hand, Figure 2.35b shows the resistivity of Cu films as a function of film thickness D. In this case, annealing (heat treating) the films to reduce the polycrystallinity does not significantly affect the resistivity because ρ_{film} is controlled primarily by surface scattering and is given by Equation 2.60.

EXAMPLE 2.27

THIN-FILM RESISTIVITY Consider the data presented in Figure 2.35a. What can you conclude from the plot given that the mean free path $\lambda \approx 40$ nm in Cu?

SOLUTION

Consider the results in Figure 2.35a. It is stated that the film thickness $D = 250$–900 nm does not affect the resistivity, which implies that ρ_{film} is controlled only by the grain size d. From Equation 2.57a and b we expect

$$\rho_{film} \approx \rho_{crystal}(1 + 1.33\beta) \approx \rho_{crystal} + 1.33\rho_{crystal}\left(\frac{R}{1-R}\right)\frac{\lambda}{d}$$

This equation represents the observed line when ρ_{film} is plotted against $1/d$ as in Figure 2.35a. The $\rho_{film} - 1/d$ line has an intercept given by 17.8 nΩ m and a slope given by 595 (nΩ m) (nm). The intercept approximately matches the bulk resistivity $\rho_{crystal}$ of Cu. The slope is

$$\text{Slope} \approx 1.33\rho_{crystal}\left(\frac{R}{1-R}\right)\lambda$$

or

$$595(\text{n}\Omega\ \text{m})(\text{nm}) \approx 1.33(17.8\ \text{n}\Omega\ \text{m})\left(\frac{1}{R^{-1}-1}\right)(40\ \text{nm})$$

Solving this equation yields $R \approx 0.39$ for these copper films.

2.10 INTERCONNECTS IN MICROELECTRONICS

An integrated circuit (IC) is a single crystal of Si that contains millions of transistors that have been fabricated within this one crystal. **Interconnects** are simply metal conductors that are used to wire the devices together to implement the desired overall operation of the IC; see the photographs in Figure 2.36. Aluminum and Al alloys, or Al silicides, have been the workhouse of the interconnects, but today's fast chips rely on copper interconnects, which have three distinct advantages. First, copper has a resistivity that is about 40 percent lower than that of Al. In high-transistor-density chips in which various voltages are switched on and off, what limits the speed of operation is the RC time constant, that is, the time constant that is involved in charging and discharging the capacitance between the interconnects, and the input capacitance of the transistor; usually the former dominates. The RC is substantially reduced with Cu replacing Al so that the chip speed is faster. The second advantage is that a lower overall interconnect resistance leads to a lower power consumption, lower I^2R.

The third advantage is that copper has superior resistance to **electromigration,** a process in which metal atoms are forced to migrate by a large current density. Such electromigration can eventually lead to a failure of the interconnect. The current density in interconnects with a small cross-sectional area can be very high, and hence the electron drift velocities can also be very high. As these fast electrons collide with the metal ions there is a momentum transfer that slowly drifts the metal ions. Thus, the metal ions are forced to slowly migrate as a result of being bombarded by drifting electrons; the migration is in the direction of electron flow (not current flow). This atomic migration can deplete or accumulate material in certain local regions of the interconnect structure. The result is that electromigration can lead to voids (material depletion) or hillocks (material accumulation), and eventually there may be a break or a short between interconnects (an interconnect failure). The electromigration effects are reduced

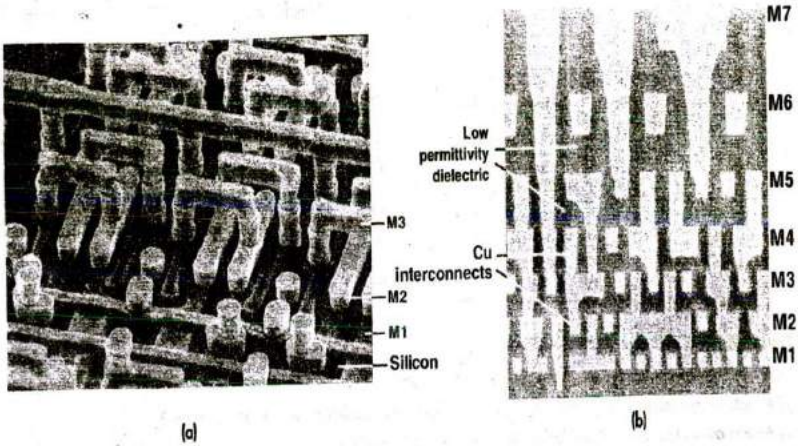


Figure 2.36

(a) Metal interconnects wiring devices on a silicon crystal. Three different metallization levels M1, M2, and M3 are used. The dielectric between the interconnects has been etched away to expose the interconnect structure.

(b) Cross section of a chip with seven levels of metallization, M1 to M7. The image is obtained with a scanning electron microscope (SEM).

| SOURCES: (a) Courtesy of IBM. (b) Courtesy of Mark Bohr, Intel..

in Cu interconnects because the Cu atoms are heavier and cannot be as easily migrated by an electric current as are Al atoms.

There is a relatively simple expression for estimating the ***RC* time constant** of multilevel interconnects that is useful in comparing various interconnect technologies and the effects of interconnect metal resistance ρ, the relative permittivity ε_r of the interlevel dielectric (insulation) between the interconnects, and the geometry of the whole interconnect wiring. First consider a simple interconnect line, as in Figure 2.37a, whose thickness is T, width is W, and length is L. Its resistance R is simply $\rho L/(TW)$. In the chip, this interconnect will have other interconnects around it as shown in a simplified way in Figure 2.37b. It will couple with all these different conductors around it and will have an overall (effective) capacitance C_{eff}. RC_{eff} is what we know as the RC time constant associated with the interconnect line in Figure 2.37b.

Suppose that the interconnect is an Mth-level metallization. It will have a series of many "horizontal" neighbors along this Mth level. Let X be the nearest edge-to-edge *separation* and P be the *pitch* of these horizontal neighbors at the Mth level. The pitch P refers to the separation from center to center, or the periodicity of interconnects; $P = W + X$. At a height H above the interconnect there will be a line running at the $(M + 1)$ level. Similarly there will be an interconnect line at a distance H below at the $(M - 1)$ level. We can identify two sets of capacitances. C_V represents the capacitance in the vertical direction, between the interconnect and its upper or lower neighbor. C_H is the lateral capacitance in the horizontal direction, between a neighbor on the right or left. Both are shown in Figure 2.37c. The interconnect therefore has two C_V and two

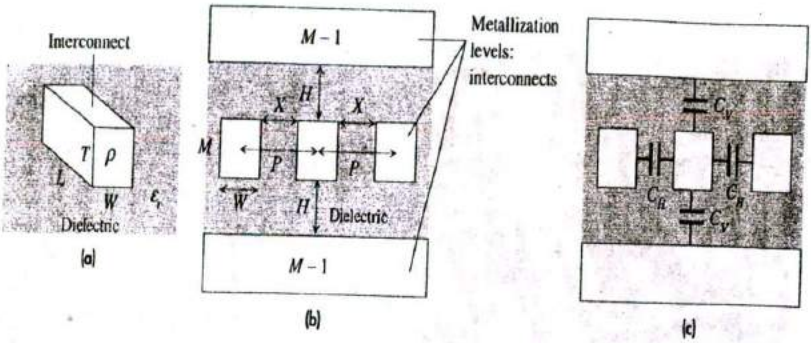


Figure 2.37

(a) A single line interconnect surrounded by dielectric insulation.

(b) Interconnects crisscross each other. There are three levels of interconnect: $M - 1$, M, and $M + 1$.

(c) An interconnect has vertical and horizontal capacitances C_V and C_H.

C_H, four capacitances in total, and all are in parallel as shown in Figure 2.37c. From the simple parallel plate capacitance formula we can write

$$C_H = \frac{\varepsilon_o \varepsilon_r TL}{X} \quad \text{and} \quad C_V = \frac{\varepsilon_o \varepsilon_r WL}{H}$$

Effective capacitance in multilevel interconnect structures

Usually C_H is greater than C_V. From Figure 2.37c, the effective capacitance $C_{\text{eff}} = 2(C_H + C_V)$,

$$C_{\text{eff}} = 2\varepsilon_o \varepsilon_r L\left(\frac{T}{X} + \frac{W}{H}\right) \qquad [2.61]$$

RC time constant in multilevel interconnect structures

which is the **effective multilevel interconnect capacitance.** We now multiply this with $R = \rho L/(TW)$ to obtain the RC time constant,

$$RC = 2\varepsilon_o \varepsilon_r \rho\left(\frac{L^2}{TW}\right)\left(\frac{T}{X} + \frac{W}{H}\right) \qquad [2.62]$$

Equation 2.62 is only an approximate first-order calculation, but, nonetheless, it turns out to be quite a useful equation for roughly predicting the RC time constant and hence the speed of multilevel interconnect based high-transistor-density chips.[12] Most significantly, it highlights the importance of *three* influencing effects: the resistivity of the interconnect metal; relative permittivity ε_r of the dielectric insulation between the conductors; and the geometry or "architecture" of the interconnects L, T, W, X, and H. Notice that L appears as L^2 in Equation 2.62 and has

[12] A more rigorous theory would consider the interconnect system as having a distributed resistance and a distributed capacitance, similar to a transmission line; a topical research area. The treatment here is more than sufficient to obtain approximate results and understand the factors that control the interconnect delay time.

significant control on the overall RC. Equation 2.62 does not obviously include the time it takes to turn on and off the individual transistors connected to the interconnects. In a high-transistor-density chip, the latter is smaller than the interconnect RC time constant.

The reduction in the interconnect resistivity ρ by the use of Cu instead of Al has been a commendable achievement, and cuts down RC significantly. Further reduction in ρ is limited because Cu already has a very small resistivity; the smallest ρ is for Ag which is only about 5 percent lower. Current research efforts for reducing RC further are concentrated on mainly two factors. First is the reduction of ε_r as much as possible by using dielectrics such as *fluorinated* SiO_2 (known as FSG) for which $\varepsilon_r = 3.6$, or, more importantly, using what are called **low-k dielectric materials** (k stands for ε_r) such as various polymers or porous dielectrics[13] that have a lower ε_r, typically 2–3, which is a substantial reduction from 3.6. The second is the development of optimized interconnect geometries that reduce L^2 in Equation 2.62. (T, W, X, and H are all of comparable size, so L^2 is the most dominant geometric factor.)

The ratio of the thickness T to width W of an interconnect is called the **aspect ratio,** $A_r = T/W$. This ratio is typically between 1 to 2. Very roughly, in many cases, X and W are the same, $X \approx W$ and $X \approx P/2$ (see Figure 2.37b). Then Equation 2.62 simplifies further,

RC time constant in multilevel interconnect structures

$$RC \approx 2\varepsilon_o\varepsilon_r\rho L^2\left(\frac{4}{P^2} + \frac{1}{T^2}\right) \qquad [2.63]$$

The signal delays between the transistors on a chip arise from the interconnect RC time constant. Equations 2.62 and 2.63 are often also used to calculate the **multilevel interconnect delay time.** Suppose that we take some typical values, $L \approx 10$ mm, $T \approx 1$ μm, $P \approx 1$ μm, $\rho = 17$ nΩ m for a Cu interconnect, and $\varepsilon_r \approx 3.6$ for FSG; then $RC \approx 0.43$ ns, not a negligible value in today's speed hungry computing.

EXAMPLE 2.28

MULTILEVEL INTERCONNECT RC TIME CONSTANT In a particular high-transistor-density IC where copper is used as the interconnect, one level of the multilevel interconnects has the following characteristics: pitch $P = 0.45$ μm, $T = 0.36$ μm, $A_R = 1.6$, $H = X$, and $\varepsilon_r \approx 3.6$. Find the effective capacitance per millimeter of interconnect length, and the RC delay time per L^2 as ps/mm^2 (as normally used in industry).

SOLUTION

Since $A_R = T/W$, $W = T/A_R = 0.36/1.6 = 0.225$ μm. Further, from Figure 2.37b, $P = W + X$, so that $X = P - W = 0.45 - 0.225 = 0.225$ μm. $H = X = 0.225$ μm. Thus, Equation 2.61 for $L = 1$ mm $= 10^{-3}$ m gives

$$C_{\text{eff}} = 2\varepsilon_o\varepsilon_r L\left(\frac{T}{X} + \frac{W}{H}\right) = 2(8.85 \times 10^{-12})(3.6)(10^{-3})\left[\frac{0.36}{0.225} + \frac{0.225}{0.225}\right] = 0.17 \text{ pF}$$

---

[13] The mixture rules mentioned in this chapter turn up again in a different but recognizable form for predicting the overall relative permittivity of porous dielectrics.

which is about 0.2 pF per millimeter of interconnect. The RC time constant per L^2 is

$$\frac{RC}{L^2} = 2\varepsilon_o\varepsilon_r\rho\left(\frac{1}{TW}\right)\left(\frac{T}{X}+\frac{W}{H}\right) = 2\varepsilon_o\varepsilon_r\rho\left(\frac{1}{WX}+\frac{1}{TH}\right)$$

$$= 2(8.85 \times 10^{-12})(3.6)(17 \times 10^{-9})$$

$$\left[\frac{1}{(0.225 \times 10^{-6})(0.225 \times 10^{-6})} + \frac{1}{(0.36 \times 10^{-6})(0.225 \times 10^{-6})}\right]$$

$$= 3.4 \times 10^{-5}\ \text{s m}^{-2} \quad \text{or} \quad 34\ \text{ps mm}^{-2}$$

2.11 ELECTROMIGRATION AND BLACK'S EQUATION

Interconnects have small cross-sectional dimensions, and consequently the current densities can be quite large. Figure 2.38a depicts how the continual bombardment of lattice atoms (metal ions) by many "fast" conduction electrons in high-current-density regions can transfer enough momentum to a host metal atom to migrate it, that is, diffuse it along a suitable path in the crystal. The bombarded metal atom has to jump to a suitable lattice location to migrate, which is usually easiest along grain boundaries or surfaces where there is sufficient space as depicted in Figure 2.38a and b. Grain boundaries that are parallel to the electron flow therefore can migrate atoms more efficiently than grain boundaries in other directions. Atomic diffusion can also occur along a surface of the interconnect, that is, along an interface between the interconnect metal and the neighboring material. The final result of atomic migration is usually either material depletion or accumulation as depicted in Figure 2.38c. The depletion of material

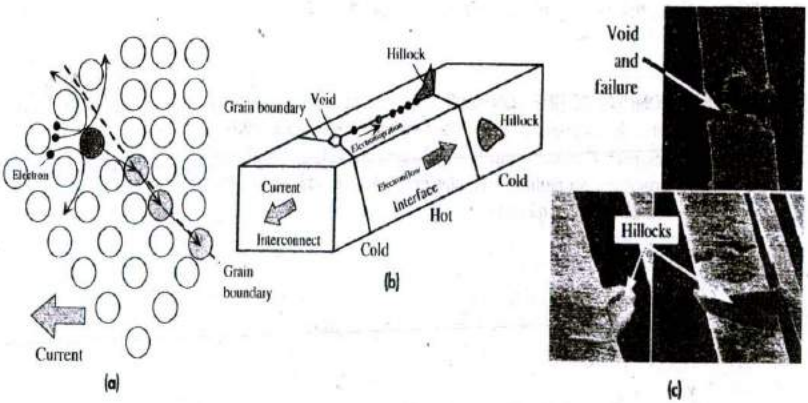


Figure 2.38

(a) Electrons bombard the metal ions and force them to slowly migrate.

(b) Formation of voids and hillocks in a polycrystalline metal interconnect by the electromigration of metal ions along grain boundaries and interfaces.

(c) Accelerated tests on a 3 μm chemical vapor deposited Cu line: $T = 200\ ^\circ\text{C}$ and $J = 6\ \text{MA cm}^{-2}$. The photos show void formation and fatal failure (break), and hillock formation.

SOURCE: Courtesy of L. Arnaud *et al.*, *Microelectronics Reliability*, **40**, 86, 2000.

leads to a **void** and a possible eventual break in the interconnect. The accumulation of material leads to a **hillock** and a short between lines. Interconnect failure by electromigration is measured by the **mean time to 50 percent failure** t_{MTF}. There are two factors that control the rate of electromigration R_{EM}. First is the activation energy E_A involved in migrating (diffusing) the metal atom, and the second is the rate at which the atoms are bombarded with electrons, which depends on the current density J. Thus,

Electromigration rate

$$R_{EM} \propto J^n \exp\left(-\frac{E_A}{kT}\right)$$

in which the rate is proportional to J^n, instead of just J because it is found experimentally that $n \geq 1$. From the electromigration rate we can find the average time t_{MTF} it takes for 50 percent failure of interconnects because this time is inversely proportional to the electromigration rate just given:

Black's electromigration failure equation

$$t_{MTF} = A_B J^{-n} \exp\left(\frac{E_A}{kT}\right) \qquad [2.64]$$

where A_B is a constant. Equation 2.64 is known as **Black's equation,** and it is extremely useful in extrapolating high-temperature failure tests to normal operating temperatures. Electromigration-induced interconnect failures are typically examined at elevated temperatures where the failure times are over a measurable time scale in the laboratory (perhaps several hours or a few days). These experiments are called **accelerated failure** tests because they make use of the fact that at high temperatures the electromigration failure occurs more quickly. The results are then extrapolated to room temperature using Black's equation.

Typically electromigration occurs along grain boundaries or along various interfaces that the interconnect has with its surroundings, the semiconductor, dielectric material, etc. The diffusion coefficient has a lower activation energy E_A for these migration paths than for diffusion within the volume of the crystal. The electromigration process therefore depends on the microstructure of the interconnect metal, and its interfaces. Usually another metal, called a **barrier,** is deposited to occupy the interface space between the interconnect and the semiconductor or the oxide. The barrier *passivates* the interface, rendering it relatively inactive in terms of providing an electromigration path. An interconnect can also have a temperature gradient along it. (The heat generated by I^2R may be conducted away faster at the ends of the interconnect, leaving the central region hotter.) Electromigration would be faster in the hot region and very slow (almost stationary) in the cold region since it is a thermally activated process. Consequently a pileup of electromigrated atoms can occur as atoms are migrated from hot to cold regions along the interconnect, leading to a hillock.[14]

Pure Al suffers badly from electromigration problems and is usually alloyed with small amounts of Cu, called Al(Cu), to reduce electromigration to a tolerable level. But the resistivity increases. (Why?) In recent Cu interconnects, the most important diffusion path seems to be the interface between the Cu surface and the dielectric. Surface coating of these Cu interconnects provides control over electromigration failures.

[14] Somewhat like a traffic accident pileup in which speeding cars run into stationary cars ahead of them.

CD Selected Topics and Solved Problems

Selected Topics

Conduction in Metals: Electrical and Thermal Conduction
Joule's Law
Hall Effect
Heat Transfer in Electrical Engineering: Fundamentals and Applications
Thermal Conductivity

Solved Problems

Radiation Theory of the Fuse
The Strain Gauge

DEFINING TERMS

Alloy is a metal that contains more than one element.

Brass is a copper-rich Cu–Zn alloy.

Bronze is a copper-rich Cu–Sn alloy.

Drift mobility is the drift velocity per unit applied field. If μ_d is the drift mobility, then the defining equation is $v_d = \mu_d \mathcal{E}$, where v_d is the drift velocity and $\mathcal{E}$ is the field.

Drift velocity is the average electron velocity, over all the conduction electrons in the conductor, in the direction of an applied electrical force ($F = -e\mathcal{E}$ for electrons). In the absence of an applied field, all the electrons move around randomly, and the average velocity over all the electrons in any direction is zero. With an applied field $\mathcal{E}_x$, there is a net velocity per electron v_{dx}, in the direction opposite to the field, where v_{dx} depends on $\mathcal{E}_x$ by virtue of $v_{dx} = \mu_d \mathcal{E}_x$, where μ_d is the drift mobility.

Electrical conductivity (σ) is a property of a material that quantifies the ease with which charges flow inside the material along an applied electric field or a voltage gradient. The conductivity is the inverse of electrical resistivity ρ. Since charge flow is caused by a voltage gradient, σ is the rate of charge flow across a unit area per unit voltage gradient, $J = \sigma \mathcal{E}$.

Electromigration is current density–induced diffusion of host metal atoms due to their repeated bombardment by conduction electrons at high current densities; the metal atoms migrate in the direction of electron flow. **Black's equation** describes the mean time to failure of metal film interconnects due to electromigration failure.

Fourier's law states that the rate of heat flow Q' through a sample, due to thermal conduction, is proportional to the temperature gradient dT/dx and the cross-sectional area A, that is, $Q' = -\kappa A(dT/dx)$, where κ is the thermal conductivity.

Hall coefficient (R_H) is a parameter that gauges the magnitude of the Hall effect. If $\mathcal{E}_y$ is the electric field in the y direction, due to a current density J_x along x and a magnetic field B_z along z, then $R_H = \mathcal{E}_y / J_x B_z$.

Hall effect is a phenomenon that occurs in a conductor carrying a current when the conductor is placed in a magnetic field perpendicular to the current. The charge carriers in the conductor are deflected by the magnetic field, giving rise to an electric field (Hall field) that is perpendicular to both the current and the magnetic field. If the current density J_x is along x and the magnetic field B_z is along z, then the Hall field is along either $+y$ or $-y$, depending on the polarity of the charge carriers in the material.

Heterogeneous mixture is a mixture in which the individual components remain physically separate and possess different chemical and physical properties; that is, a mixture of different phases.

Homogeneous mixture is a mixture of two or more chemical species in which the chemical properties (*e.g.*, composition) and physical properties (*e.g.*, density,

heat capacity) are uniform throughout. A homogeneous mixture is a solution.

Interconnects are various thin metal conductors in a Si integrated circuit that connect various devices to implement the required wiring of the devices. In modern ICs, these interconnects are primarily electrodeposited Cu films.

Ionic conduction is the migration of ions in the material as a result of field-directed diffusion. When a positive ion in an interstitial site jumps to a neighboring interstitial site in the direction of the field, it lowers its potential energy which is a favorable process. If it jumps in the opposite direction, then it has to do work against the force of the field which is undesirable. Thus the diffusion of the positive ion is directed along the field.

Isomorphous phase diagram is a phase diagram for an alloy that has unlimited solid solubility.

Joule's law relates the power dissipated per unit volume P_{vol} by a current-carrying conductor to the applied field $\mathcal{E}$ and the current density J, such that $P_{vol} = J\mathcal{E} = \sigma\mathcal{E}^2$.

Lorentz force is the force experienced by a moving charge in a magnetic field. When a charge q is moving with a velocity $\mathbf{v}$ in a magnetic field $\mathbf{B}$, the charge experiences a force $\mathbf{F}$ that is proportional to the magnitude of its charge q, its velocity $\mathbf{v}$, and the field $\mathbf{B}$, such that $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$.

Magnetic field, magnetic flux density, or magnetic induction (B) is a vector field quantity that describes the magnitude and direction of the *magnetic force* exerted on a moving charge or a current-carrying conductor. The magnetic force is essentially the Lorentz force and excludes the electrostatic force $q\mathcal{E}$.

Magnetic permeability (μ) or simply permeability is a property of the medium that characterizes the effectiveness of a medium in generating as much magnetic field as possible for given external currents. It is the product of the permeability of free space (vacuum) or absolute permeability (μ_o) and relative permeability of the medium (μ_r), *i.e.*, $\mu = \mu_o\mu_r$.

Magnetometer is an instrument for measuring the magnitude of a magnetic field.

Matthiessen's rule gives the overall resistivity of a metal as the sum of individual resistivities due to scattering from thermal vibrations, impurities, and crystal defects. If the resistivity due to scattering from thermal vibrations is denoted ρ_T and the resistivities due to scattering from crystal defects and impurities can be lumped into a single resistivity term called the residual resistivity ρ_R, then $\rho = \rho_T + \rho_R$.

Mean free path is the mean distance traversed by an electron between scattering events. If τ is the mean free time between scattering events and u is the mean speed of the electron, then the mean free path is $\ell = u\tau$.

Mean free time is the average time it takes to scatter a conduction electron. If t_i is the free time between collisions (between scattering events) for an electron labeled i, then $\tau = \bar{t}_i$ averaged over all the electrons. The drift mobility is related to the mean free time by $\mu_d = e\tau/m_e$. The reciprocal of the mean free time is the mean probability per unit time that a conduction electron will be scattered; in other words, the mean frequency of scattering events.

Nordheim's rule states that the resistivity of a solid solution (an isomorphous alloy) due to impurities ρ_I is proportional to the concentrations of the solute X and the solvent $(1 - X)$.

Phase (in materials science) is a physically homogeneous portion of a materials system that has uniform physical and chemical characteristics.

Relaxation time is an equivalent term for the mean free time between scattering events.

Residual resistivity (ρ_R) is the contribution to the resistivity arising from scattering processes other than thermal vibrations of the lattice, for example, impurities, grain boundaries, dislocations, point defects.

Skin effect is an electromagnetic phenomenon that, at high frequencies, restricts ac current flow to near the surface of a conductor to reduce the energy stored in the magnetic field.

Solid solution is a crystalline material that is a homogeneous mixture of two or more chemical species. The mixing occurs at the atomic scale, as in mixing alcohol and water. Solid solutions can be substitutional (as in Cu–Ni) or interstitial (for example, C in Fe).

Stefan's law is a phenomenological description of the energy radiated (as electromagnetic waves) from a surface per second. When a surface is heated to a

temperature T, it radiates net energy at a rate given by $P_{\text{radiated}} = \epsilon\sigma_S A(T^4 - T_0^4)$, where σ_S is Stefan's constant (5.67×10^{-8} W m^{-2} K^{-4}), ϵ is the emissivity of the surface, A is the surface area, and T_0 is the ambient temperature.

Temperature coefficient of resistivity (TCR) (α_0) is defined as the fractional change in the electrical resistivity of a material per unit increase in the temperature with respect to some reference temperature T_0.

Thermal conductivity (κ) is a property of a material that quantifies the ease with which heat flows along the material from higher to lower temperature regions. Since heat flow is due to a temperature gradient, κ is the rate of heat flow across a unit area per unit temperature gradient.

Thermal resistance (θ) is a measure of the difficulty with which heat conduction takes place along a material sample. The thermal resistance is defined as the temperature drop per unit heat flow, $\theta = \Delta T/Q'$. It depends on both the material and its geometry. If the heat losses from the surfaces are negligible, then $\theta = L/\kappa A$, where L is the length of the sample (along heat flow) and A is the cross-sectional area.

Thermally activated conductivity means that the conductivity increases in an exponential fashion with temperature as in $\sigma = \sigma_o \exp(-E_a/kT)$ where E_a is the activation energy.

Thin film is a conductor whose thickness is typically less than ~1 micron; the thickness is also much less than the width and length of the conductor. Typically thin films have a higher resistivity than the corresponding bulk material due to the grain boundary and surface scattering.

QUESTIONS AND PROBLEMS

2.1 **Electrical conduction** Na is a monovalent metal (BCC) with a density of 0.9712 g cm^{-3}. Its atomic mass is 22.99 g mol^{-1}. The drift mobility of electrons in Na is 53 cm^2 V^{-1} s^{-1}.

a. Consider the collection of conduction electrons in the solid. If each Na atom donates one electron to the electron sea, *estimate* the mean separation between the electrons. (Note: If n is the concentration of particles, then the particles' mean separation $d = 1/n^{1/3}$.)

b. Estimate the mean separation between an electron (e^-) and a metal ion (Na^+), assuming that most of the time the electron prefers to be between two neighboring Na^+ ions. What is the approximate Coulombic interaction energy (in eV) between an electron and an Na^+ ion?

c. How does this electron/metal-ion interaction energy compare with the average thermal energy per particle, according to the kinetic molecular theory of matter? Do you expect the kinetic molecular theory to be applicable to the conduction electrons in Na? If the mean electron/metal-ion interaction energy is of the same order of magnitude as the mean *KE* of the electrons, what is the mean speed of electrons in Na? Why should the mean kinetic energy be comparable to the mean electron/metal-ion interaction energy?

d. Calculate the electrical conductivity of Na and compare this with the experimental value of $2.1 \times 10^7\ \Omega^{-1}\ \text{m}^{-1}$ and comment on the difference.

2.2 **Electrical conduction** The resistivity of aluminum at 25 °C has been measured to be $2.72 \times 10^{-8}\ \Omega$ m. The thermal coefficient of resistivity of aluminum at 0 °C is 4.29×10^{-3} K^{-1}. Aluminum has a valency of 3, a density of 2.70 g cm^{-3}, and an atomic mass of 27.

a. Calculate the resistivity of aluminum at −40 °C.

b. What is the thermal coefficient of resistivity at −40 °C?

c. Estimate the mean free time between collisions for the conduction electrons in aluminum at 25 °C and hence estimate their drift mobility.

d. If the mean speed of the conduction electrons is about 2.0×10^6 m s^{-1}, calculate the mean free path and compare this with the interatomic separation in Al (Al is FCC). What should be the

thickness of an Al film that is deposited on an IC chip such that its resistivity is the same as that of bulk Al?

e. What is the percentage change in the power loss due to Joule heating of the aluminum wire when the temperature drops from 25 °C to −40 °C?

2.3 **Conduction in gold** Gold is in the same group as Cu and Ag. Assuming that each Au atom donates one conduction electron, calculate the drift mobility of the electrons in gold at 22 °C. What is the mean free path of the conduction electrons if their mean speed is 1.4×10^6 m s^{-1}? (Use ρ_0 and α_ρ in Table 2.1.)

2.4 **Effective number of conduction electrons per atom**

a. Electron drift mobility in tin (Sn) is 3.9 cm^2 V^{-1} s^{-1}. The room temperature (20 °C) resistivity of Sn is about 110 nΩ m. Atomic mass M_{at} and density of Sn are 118.69 g mol^{-1} and 7.30 g cm^{-3}, respectively. How many "free" electrons are donated by each Sn atom in the crystal? How does this compare with the position of Sn in Group IVB of the Periodic Table?

b. Consider the resistivity of few selected metals from Groups I to IV in the Periodic Table in Table 2.7. Calculate the number of conduction electrons contributed per atom and compare this with the location of the element in the Periodic Table. What is your conclusion?

Table 2.7 Selection of metals from Groups I to IV in the Periodic Table

Metal	Periodic Group	Valency	Density (g cm^{-3})	Resistivity (nΩ m)	Mobility (cm^2 V^{-1}s^{-1})
Na	IA	1	0.97	42.0	53
Mg	IIA	2	1.74	44.5	17
Ag	IB	1	10.5	15.9	56
Zn	IIB	2	7.14	59.2	8
Al	IIIB	3	2.7	26.5	12
Sn	IVB	4	7.30	110	3.9
Pb	IVB	4	11.4	206	2.3

NOTE: Mobility from Hall-effect measurements.

2.5 **TCR and Matthiessen's rule** Determine the temperature coefficient of resistivity of pure iron and of electrotechnical steel (Fe with 4% C), which are used in various electrical machinery, at two temperatures: 0 °C and 500 °C. Comment on the similarities and differences in the resistivity versus temperature behavior shown in Figure 2.39 for the two materials.

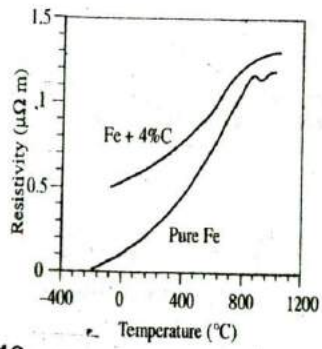


Figure 2.39 Resistivity versus temperature for pure iron and 4% C steel.

***2.6 TCR of isomorphous alloys**

a. Show that for an isomorphous alloy $A\%$–$B\%$ ($B\%$ solute in $A\%$ solvent), the temperature coefficient of resistivity α_{AB} is given by

$$\alpha_{AB} \approx \frac{\alpha_A \rho_A}{\rho_{AB}}$$

where ρ_{AB} is the resistivity of the alloy (AB) and ρ_A and α_A are the resistivity and TCR of pure A. What are the assumptions behind this equation?

b. Determine the composition of the Cu–Ni alloy that will have a TCR of $4 \times 10^{-4}\ \mathrm{K}^{-1}$, that is, a TCR that is an order of magnitude less than that of Cu. Over the composition range of interest, the resistivity of the Cu–Ni alloy can be calculated from $\rho_{CuNi} \approx \rho_{Cu} + C_{eff} X(1 - X)$, where C_{eff}, the effective Nordheim coefficient, is about 1310 nΩ m.

2.7 Resistivity of isomorphous alloys and Nordheim's rule What are the maximum atomic and weight percentages of Cu that can be added to Au without exceeding a resistivity that is twice that of pure gold? What are the maximum atomic and weight percentages of Au that can be added to pure Cu without exceeding twice the resistivity of pure copper? (Alloys are normally prepared by mixing the elements in weight.)

2.8 Nordheim's rule and brass Brass is a Cu–Zn alloy. Table 2.8 shows some typical resistivity values for various Cu–Zn compositions in which the alloy is a solid solution (up to 30% Zn).

a. Plot ρ versus $X(1 - X)$. From the slope of the best-fit line find the mean (effective) Nordheim coefficient $\overline{C}$ for Zn dissolved in Cu over this compositional range.

b. Since X is the atomic fraction of Zn in brass, for each atom in the alloy, there are X Zn atoms and $(1 - X)$ Cu atoms. The conduction electrons consist of each Zn donating two electrons and each copper donating one electron. Thus, there are $2(X) + 1(1 - X) = 1 + X$ conduction electrons per atom. Since the conductivity is proportional to the electron concentration, the combined Nordheim–Matthiessens rule must be scaled up by $(1 + X)$,

$$\rho_{brass} = \frac{\rho_0 + CX(1 - X)}{(1 + X)}$$

Plot the data in Table 2.8 as $\rho(1 + X)$ versus $X(1 - X)$. From the best-fit line find C and ρ_0. What is your conclusion? (Compare the correlation coefficients of the best-fit lines in your two plots.[15])

Table 2.8 Cu–Zn brass alloys

Zn at.% in Cu–Zn	0	0.34	0.5	0.93	3.06	4.65	9.66	15.6	19.59	29.39
Resistivity nΩ m	17	18.1	18.84	20.7	26.8	29.9	39.1	49.0	54.8	63.5

SOURCE: H. A. Fairbank, *Phys. Rev.*, **66**, 274, 1944.

2.9 Resistivity of solid solution metal alloys: testing Nordheim's rule Nordheim's rule accounts for the increase in the resistivity resulting from the scattering of electrons from the random distribution of impurity (solute) atoms in the host (solvent) crystal. It can nonetheless be quite useful in approximately

[15] More rigorously, $\rho_{brass} = \rho_{matrix} + C_{eff} X(1 - X)$, in which ρ_{matrix} is the resistivity of the perfect matrix. Accounting for the extra electrons, $\rho_{matrix} \approx \rho_0/(1 + X)$, where ρ_0 is the pure metal matrix resistivity and C_{eff} is the Nordheim coefficient at the composition of interest, given by $C_{eff} \approx C/(1 + X)^{2/3}$. (It is assumed that the atomic concentration does not change significantly.) As always, there are also other theories; part *b* is more than sufficient for most practical purposes.

predicting the resistivity at one composition of a solid solution metal alloy, given the value at another composition. Table 2.9 lists some solid solution metal alloys and gives the resistivity ρ at one composition X and asks for a prediction ρ' based on Nordheim's rule at another composition X'. Fill in the table for ρ' and compare the predicted values with the experimental values, and comment.

Table 2.9 Resistivities of some solid solution metal alloys

	Alloy							
	Ag–Au	Au–Ag	Cu–Pd	Ag–Pd	Au–Pd	Pd–Pt	Pt–Pd	Cu–Ni
X (at.%)	8.8% Au	8.77% Ag	6.2% Pd	10.1% Pd	8.88% Pd	7.66% Pt	7.1% Pd	2.16% Ni
ρ_0 (nΩ m)	16.2	22.7	17	16.2	22.7	108	105.8	17
ρ at X (nΩ m)	44.2	54.1	70.8	59.8	54.1	188.2	146.8	50
C_{eff}								
X'	15.4% Au	24.4% Ag	13% Pd	15.2% Pd	17.1% Pd	15.5% Pt	13.8% Pd	23.4% Ni
ρ' at X' (nΩ m)								
ρ' at X' (nΩ m) Experimental	66.3	107.2	121.6	83.8	82.2	244	181	300

NOTE: First symbol (e.g., Ag in AgAu) is the matrix (solvent) and the second (Au) is the added solute. X is in at.%, converted from traditional weight percentages reported with alloys. C_{eff} is the effective Nordheim coefficient in $\rho = \rho_0 + C_{\text{eff}} X[1 - X]$.

***2.10 TCR and alloy resistivity** Table 2.10 shows the resistivity and TCR (α) of Cu–Ni alloys. Plot TCR versus $1/\rho$, and obtain the best-fit line. What is your conclusion? Consider the Matthiessen rule, and explain why the plot should be a straight line. What is the relationship between ρ_{Cu}, α_{Cu}, ρ_{CuNi}, and α_{CuNi}? Can this be generalized?

Table 2.10 Cu–Ni alloys, resistivity, and TCR

	Ni wt.% in Cu–Ni				
	0	2	6	11	20
Resistivity (nΩ m)	17	50	100	150	300
TCR (ppm °C^{-1})	4270	1350	550	430	160

NOTE: ppm-parts per million, *i.e.*, 10^{-6}.

2.11 Electrical and thermal conductivity of In Electron drift mobility in indium has been measured to be 6 cm^2 V^{-1} s^{-1}. The room temperature (27 °C) resistivity of In is 8.37×10^{-8} Ω m; and its atomic mass and density are 114.82 amu or g mol^{-1} and 7.31 g cm^{-3}, respectively.

a. Based on the resistivity value, determine how many free electrons are donated by each In atom in the crystal. How does this compare with the position of In in the Periodic Table (Group IIIB)?

b. If the mean speed of conduction electrons in In is 1.74×10^8 cm s^{-1}, what is the mean free path?

c. Calculate the thermal conductivity of In. How does this compare with the experimental value of 81.6 W m^{-1} K^{-1}?

2.12 Electrical and thermal conductivity of Ag The electron drift mobility in silver has been measured to be 56 cm^2 V^{-1} s^{-1} at 27 °C. The atomic mass and density of Ag are given as 107.87 amu or g mol^{-1} and 10.50 g cm^{-3}, respectively.

a. Assuming that each Ag atom contributes one conduction electron, calculate the resistivity of Ag at 27 °C. Compare this value with the measured value of 1.6×10^{-8} Ω m at the same temperature and suggest reasons for the difference.

b. Calculate the thermal conductivity of silver at 27 °C and at 0 °C.

2.13 **Mixture rules** A 70% Cu–30% Zn brass electrical component has been made of powdered metal and contains 15 vol.% porosity. Assume that the pores are dispersed randomly. Given that the resistivity of 70% Cu–30% Zn brass is 62 nΩ m, calculate the effective resistivity of the brass component using the simple conductivity mixture rule, Equation 2.26, and the Reynolds and Hough rule.

2.14 **Mixture rules**

a. A certain carbon electrode used in electrical arcing applications is 47 percent porous. Given that the resistivity of graphite (in polycrystalline form) at room temperature is about 9.1 $\mu\Omega$ m, estimate the effective resistivity of the carbon electrode using the appropriate Reynolds and Hough rule and the simple conductivity mixture rule. Compare your estimates with the measured value of 18 $\mu\Omega$ m and comment on the differences.

b. Silver particles are dispersed in a graphite paste to increase the effective conductivity of the paste. If the volume fraction of dispersed silver is 30 percent, what is the effective conductivity of this paste?

2.15 **Ag–Ni alloys (contact materials) and the mixture rules** Silver alloys, particularly Ag alloys with the precious metals Pt, Pd, Ni, and Au, are extensively used as contact materials in various switches. Alloying Ag with other metals generally increases the hardness, wear resistance, and corrosion resistance at the expense of electrical and thermal conductivity. For example, Ag–Ni alloys are widely used as contact materials in switches in domestic appliances, control and selector switches, circuit breakers, and automotive switches up to several hundred amperes of current. Table 2.11 shows the resistivities of four Ag–Ni alloys used in make-and-break as well as disconnect contacts with current ratings up to ~100 A.

a. Ag–Ni is a two-phase alloy, a mixture of Ag-rich and Ni-rich phases. Using an appropriate mixture rule, predict the resistivity of the alloy and compare with the measured values in Table 2.11. Explain the difference between the predicted and experimental values.

b. Compare the resistivity of Ag–10% Ni with that of Ag–10% Pd in Table 2.9. The resistivity of the Ag–Pd alloy is almost a factor of 5 greater. Ag–Pd is an isomorphous solid solution, whereas Ag–Ni is a two-phase mixture. Explain the difference in the resistivities of Ag–Ni and Ag–Pd.

Table 2.11 Resistivity of Ag–Ni contact alloys for switches

	Ni % in Ag–Ni					
	0	10	15	20	30	100
ρ(nΩ m)	16.9	20.9	23.6	25	31.1	71.4
d(g cm^{-3})	10.5	10.3	9.76	9.4	9.47	8.9
Hardness VHN	30	50	55	60	65	80

NOTE: Compositions are in wt.%. Ag–10% Ni means 90% Ag–10% Ni. Vickers hardness number (VHN) is a measure of the hardness or strength of the alloy, and d is density.

2.16 **Ag–W alloys (contact materials) and the mixture rule** Silver–tungsten alloys are frequently used in heavy-duty switching applications (*e.g.*, current-carrying contacts and oil circuit breakers) and in arcing tips. Ag–W is a two-phase alloy, a mixture of Ag-rich and W-rich phases. The measured resistivity and density for various Ag–W compositions are summarized in Table 2.12.

a. Plot the resistivity and density of the Ag–W alloy against the W content (wt.%)

b. Show that the density of the mixture, d, is given by

$$d^{-1} = w_\alpha d_\alpha^{-1} + w_\beta d_\beta^{-1}$$

where w_α is the weight fraction of phase α, w_β is the weight fraction of phase β, d_α is the density of phase α, and d_β is the density of phase β.

c. Show that the resistivity mixture rule is

$$\rho = \rho_\alpha \frac{dw_\alpha}{d_\alpha} + \rho_\beta \frac{dw_\beta}{d_\beta}$$

Mixture rule and weight fractions

where ρ is the resistivity of the alloy (mixture), d is the density of the alloy (mixture), and subscripts α and β refer to phases α and β, respectively.

d. Calculate the density d and the resistivity ρ of the mixture for various values of W content (in wt.%) and plot the calculated values in the same graph as the experimental values. What is your conclusion?

Table 2.12 Dependence of resistivity in Ag–W alloy on composition as a function of wt.% W

	W(wt.%)												
	0	10	15	20	30	40	65	70	75	80	85	90	100
ρ (nΩ m)	16.2	18.6	19.7	20.9	22.7	27.6	35.5	38.3	40	46	47.9	53.9	55.6
d (g cm^{-3})	10.5	10.75	10.95	11.3	12	12.35	14.485	15.02	15.325	16.18	16.6	17.25	19.1

NOTE: ρ = resistivity and d = density.

2.17 **Thermal conduction** Consider brass alloys with an X atomic fraction of Zn. Since Zn addition increases the number of conduction electrons, we have to scale the final alloy resistivity calculated from the simple Matthiessen–Nordheim rule in Equation 2.22 down by a factor $(1 + X)$ (see Question 2.8) so that the resistivity of the alloy is $\rho \approx [\rho_0 + CX(1 - X)]/(1 + X)$ in which $C = 300$ nΩ m and $\rho_0 = \rho_{Cu} = 17$ nΩ m.

a. An 80 at.% Cu–20 at.% Zn brass disk of 40 mm diameter and 5 mm thickness is used to conduct heat from a heat source to a heat sink.
   (1) Calculate the thermal resistance of the brass disk.
   (2) If the disk is conducting heat at a rate of 100 W, calculate the temperature drop along the disk.

b. What should be the composition of brass if the temperature drop across the disk is to be halved?

2.18 **Thermal resistance** Consider a thin insulating disk made of mica to electrically insulate a semiconductor device from a conducting heat sink. Mica has $\kappa = 0.75$ W m^{-1} K^{-1}. The disk thickness is 0.1 mm, and the diameter is 10 mm. What is the thermal resistance of the disk? What is the temperature drop across the disk if the heat current through it is 25 W?

***2.19** **Thermal resistance** Consider a coaxial cable operating under steady-state conditions when the current flow through the inner conductor generates Joule heat at a rate $P = I^2R$. The heat generated per second by the core conductor flows through the dielectric: $Q' = I^2R$. The inner conductor reaches a temperature T_i, whereas the outer conductor is at T_o. Show that the thermal resistance θ of the hollow cylindrical insulation for heat flow in the radial direction is

$$\theta = \frac{(T_i - T_o)}{Q'} = \frac{\ln(b/a)}{2\pi\kappa L} \qquad [2.65]$$

Thermal resistance of hollow cylinder

where a is the inside (core conductor) radius, b is the outside radius (outer conductor), κ is the thermal conductivity of the insulation, and L is the cable length. Consider a coaxial cable that has a copper core conductor and polyethylene (PE) dielectric with the following properties: Core conductor resistivity $\rho = 19$ nΩ m, core radius $a = 4$ mm, dielectric thickness $b - a = 3.5$ mm, dielectric thermal conductivity $\kappa = 0.3$ W m^{-1} K^{-1}. The outside temperature T_o is 25 °C. The cable is carrying a current of 500 A. What is the temperature of the inner conductor?

2.20 **The Hall effect** Consider a rectangular sample, a metal or an n-type semiconductor, with a length L, width W, and thickness D. A current I is passed along L, perpendicular to the cross-sectional

area WD. The face $W \times L$ is exposed to a magnetic field density B. A voltmeter is connected across the width, as shown in Figure 2.40, to read the Hall voltage V_H.

a. Show that the Hall voltage recorded by the voltmeter is

Hall voltage

$$V_H = \frac{IB}{Den}$$

b. Consider a 1-micron-thick strip of gold layer on an insulating substrate that is a candidate for a Hall probe sensor. If the current through the film is maintained at constant 100 mA, what is the magnetic field that can be recorded per μV of Hall voltage?

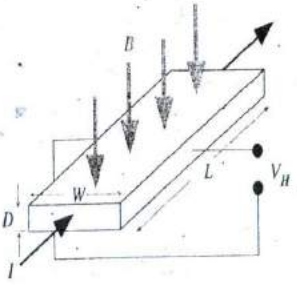


Figure 2.40 Hall effect in a rectangular material with length L, width W, and thickness D. The voltmeter is across the width W.

2.21 **The strain gauge.** A **strain gauge** is a transducer attached to a body to measure its fractional elongation $\Delta L/L$ under an applied load (force) F. The gauge is a grid of many folded runs of a thin, resistive wire glued to a flexible backing, as depicted in Figure 2.41. The gauge is attached to the body under test such that the resistive wire length is parallel to the strain.

a. Assume that the elongation does not change the resistivity and show that the change in the resistance ΔR is related to the strain $\varepsilon = \Delta L/L$ by

Strain gauge equation

$$\Delta R \approx R(1 + 2\nu)\varepsilon \qquad \textbf{[2.66]}$$

where ν is the **Poisson ratio,** which is defined by

Poisson ratio

$$\nu = -\frac{\text{Transverse strain}}{\text{Longitudinal strain}} = -\frac{\varepsilon_t}{\varepsilon_l} \qquad \textbf{[2.67]}$$

where ε_l is the strain along the applied load, that is, $\varepsilon_l = \Delta L/L = \varepsilon$, and ε_t is the strain in the transverse direction, that is, $\varepsilon_t = \Delta D/D$, where D is the diameter (thickness) of the wire.

b. Explain why a nichrome wire would be a better choice than copper for the strain gauge (consider the TCR).

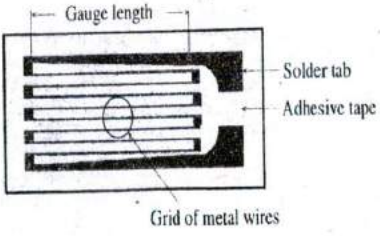


Figure 2.41 The strain gauge consists of a long, thin wire folded several times along its length to form a grid as shown and embedded in a self-adhesive tape. The ends of the wire are attached to terminals (solder pads) for external connections. The tape is stuck on the component for which the strain is to be measured.

c. How do temperature changes affect the response of the gauge? Consider the effect of temperature on ρ. Also consider the differential expansion of the specimen with respect to the gauge wire such that even if there is no applied load, there is still strain, which is determined by the differential expansion coefficient, $\lambda_{specimen} - \lambda_{gauge}$, where λ is the thermal coefficient of linear expansion: $L = L_0[1 + \lambda(T - T_0)]$, where T_0 is the reference temperature.

d. The gauge factor for a transducer is defined as the fractional change in the measured property $\Delta R/R$ per unit input signal (ε). What is the gauge factor for a metal-wire strain gauge, given that for most metals, $\nu \approx \frac{1}{3}$?

e. Consider a strain gauge that consists of a nichrome wire of resistivity 1 μΩ m, a total length of 1 m, and a diameter of 25 μm. What is ΔR for a strain of 10^{-3}? Assume that $\nu \approx \frac{1}{3}$.

f. What will ΔR be if constantan wire with a resistivity of 500 nΩ m is used?

2.22 Thermal coefficients of expansion and resistivity

a. Consider a thin metal wire of length L and diameter D. Its resistance is $R = \rho L/A$, where $A = \pi D^2/4$. By considering the temperature dependence of L, A, and ρ individually, show that

Change in R with temperature

$$\frac{1}{R}\frac{dR}{dT} = \alpha_0 - \lambda_0$$

where α_0 is the temperature coefficient of resistivity (TCR), and λ_0 is the temperature coefficient of linear expansion (thermal expansion coefficient or expansivity), that is,

$$\lambda_0 = L_0^{-1}\left(\frac{dL}{dT}\right)_{T=T_0} \quad \text{or} \quad \lambda_0 = D_0^{-1}\left(\frac{dD}{dT}\right)_{T=T_0}$$

Note: Consider differentiating $R = \rho L/[(\pi D^2)/4]$ with respect to T with each parameter, ρ, L, and D, having a temperature dependence.

Given that typically, for most pure metals, $\alpha_0 \approx 1/273\ \text{K}^{-1}$ and $\lambda_0 \approx 2 \times 10^{-5}\ \text{K}^{-1}$, confirm that the temperature dependence of ρ controls R, rather than the temperature dependence of the geometry. Is it necessary to modify the given equation for a wire with a noncircular cross section?

b. Is it possible to design a resistor from a suitable alloy such that its temperature dependence is almost nil? Consider the TCR of an alloy of two metals A and B, for which $\alpha_{AB} \approx \alpha_A \rho_A / \rho_{AB}$.

2.23 Temperature of a light bulb filament

a. Consider a 100 W, 120 V incandescent bulb (lamp). The tungsten filament has a length of 0.579 m and a diameter of 63.5 μm. Its resistivity at room temperature is 56 nΩ m. Given that the resistivity of the filament can be represented as

$$\rho = \rho_0\left[\frac{T}{T_0}\right]^n \qquad [2.68]$$

Resistivity of W

where T is the temperature in K, ρ_0 is the resistance of the filament at T_0 K, and $n = 1.2$, estimate the temperature of the bulb when it is operated at the rated voltage, that is, directly from the main outlet. Note that the bulb dissipates 100 W at 120 V.

b. Suppose that the electrical power dissipated in the tungsten wire is totally radiated from the surface of the filament. The radiated power at the absolute temperature T can be described by Stefan's law

$$P_{\text{radiated}} = \epsilon\sigma_S A\left(T^4 - T_0^4\right) \qquad [2.69]$$

Radiated power

where σ_S is Stefan's constant ($5.67 \times 10^{-8}\ \text{W m}^{-2}\ \text{K}^{-4}$), ϵ is the emissivity of the surface (0.35 for tungsten), A is the surface area of the tungsten filament, and T_0 is room temperature (293 K). Obviously, for $T > T_0$, $P_{\text{radiated}} = \epsilon\sigma_S A T^4$.

Assuming that all the electrical power is radiated from the surface, estimate the temperature of the filament and compare it with your answer in part (a).

c. If the melting temperature of W is 3407 °C, what is the voltage that guarantees that the light bulb will blow?

2.24 **Einstein relation and ionic conductivity** In the case of ionic conduction, ions have to jump from one interstice to the neighboring one. This process involves overcoming a potential energy barrier just like atomic diffusion, and drift and diffusion are related. The drift mobility μ of ions is proportional to the diffusion coefficient D because drift is limited by the atomic diffusion process. The **Einstein relation** relates the two by

Einstein relation

$$\frac{D}{\mu} = \frac{kT}{e} \qquad [2.70]$$

Diffusion coefficient of the Na^+ ion in sodium silicate (Na_2O–SiO_2) glasses at 400 °C is 3.4×10^{-9} $cm^2\ s^{-1}$. The density of such glasses is approximately 2.4 g cm^{-3}. Calculate the ionic conductivity and resistivity of (17.5 mol% Na_2O)(82.5 mol% SiO_2) sodium silicate glass at 400 °C and compare your result with the experimental values of the order of 10^4 Ω cm for the resistivity.

2.25 **Skin effect**

a. What is the skin depth for a copper wire carrying a current at 60 Hz? The resistivity of copper at 27 °C is 17 nΩ m. Its relative permeability $\mu_r \approx 1$. Is there any sense in using a conductor for power transmission which has a diameter more than 2 cm?

b. What is the skin depth for an iron wire carrying a current at 60 Hz? The resistivity of iron at 27 °C is 97 nΩ m. Assume that its relative permeability $\mu_r \approx 700$. How does this compare with the copper wire? Discuss why copper is preferred over iron for power transmission even though iron is nearly 100 times cheaper than copper.

2.26 **Thin films**

a. Consider a polycrystalline copper film that has $R = 0.40$. What is the approximate mean grain size d in terms of the mean free path λ in the bulk that would lead to the polycrystalline Cu film having a resistivity that is $1.5\rho_{bulk}$. If the mean free path in the crystal is about 40 nm at room temperature, what is d?

b. What is the thickness D of a copper film in terms of λ in which surface scattering increases the film resistivity to $1.2\rho_{bulk}$ if the specular scattering fraction p is 0.5?

c. Consider the data of Lim *et al.* (2003) presented in Table 2.13. Show that the excess resistivity, *i.e.* resistivity above that of bulk Cu, is roughly proportional to the reciprocal film thickness.

Table 2.13 Resistivity ρ_{film} of a copper film as a function of thickness D.

D (nm)	8.61	17.2	34.4	51.9	69	85.8	102.6	120.3	173.2	224.3
ρ_{film} (nΩ m)	121.8	75.3	46.1	38.5	32.1	25.2	22.0	20.5	19.9	18.8

NOTE: Film annealed at 150 °C.
SOURCE: Data extracted from J. W. Lim *et al.*, *Appl. Surf. Sci.* **217**, 95, 2003.

2.27 **Interconnects** Consider a high-transistor-density CMOS chip in which the interconnects are copper with a pitch P of 500 nm, interconnect thickness T of 400 nm, aspect ratio 1.4, and $H = X$. The dielectric is FSG with $\varepsilon_r = 3.6$. Consider two cases, $L = 1$ mm and $L = 10$ mm, and calculate the overall effective interconnect capacitance C_{eff} and the RC delay time. Suppose that Al, which is normally Al with about 4 wt.% Cu in the microelectronics industry with a resistivity 31 nΩ m, is used as the interconnect. What is the corresponding RC delay time?

***2.28** **Thin 50 nm interconnects** Equation 2.60 is for conduction in a thin film of thickness D and assumes scattering from two surfaces, which yields an additional resistivity $\rho_2 = \rho_{bulk}\frac{3}{8}(\lambda/D)(1-p)$. An interconnect line in an IC is not quite a thin film and has four surfaces (interfaces), because the thickness T of the conductor is comparable to the width W. If we assume $T = W$, we can very roughly take $\rho_4 \approx \rho_2 + \rho_2 \approx \rho_{bulk}\frac{3}{4}(\lambda/D)(1-p)$ in which $D = T$. (The exact expression is more complicated, but the latter will suffice for this problem.) In addition there will be a contribution from grain boundary

scattering, (Equation 2.57a). For simplicity assume $T \approx W \approx X \approx H \approx 60$ nm, $\lambda = 40$ nm, $p = 0.5$ and $\varepsilon_r = 3.6$. If the mean grain size d is roughly 40 nm and $R = 0.4$, estimate the resistivity of the interconnect and hence the RC delay for a 1 mm interconnect.

2.29 **TCR of thin films** Consider Matthiessen's rule applied to a thin film. Show that, very approximately, the product of the thermal coefficient of resistivity (TCR) α_{film} and the resistivity ρ_{film} is equivalent to the product of the bulk TCR and resistivity:

$$\alpha_{film}\rho_{film} \approx \alpha_{bulk}\rho_{bulk}$$

2.30 **Electromigration** Although electromigration-induced failure in Cu metallization is less severe than in Al metallization, it can still lead to interconnect failure depending on current densities and the operating temperature. In a set of experiments carried out on electroplated Cu metallization lines, failure of the Cu interconnects have been examined under accelerated tests (at elevated temperatures). The mean lifetime t_{50} (time for 50 percent of the lines to break) have been measured as a function of current density J and temperature T at a given current density. The results are summarized in Table 2.14.

a. Plot semilogarithmically t_{50} versus $1/T$ (T in Kelvins) for the first three interconnects. Al(Cu) and Cu ($1.3 \times 0.7\mu m^2$) have single activation energies E_A. Calculate E_A for these interconnects. Cu ($1.3 \times 0.7\mu m^2$) exhibits different activation energies for the high-and low-temperature regions. Estimate these E_A.

b. Plot on a log-log plot t_{50} versus J at 370 °C. Show that at low J, $n \approx 1.1$ and at high J, $n \approx 1.8$.

Table 2.14 Results of electromigration failure experiments on various Al and Cu interconnects

Al(Cu) $[J = 25\ mA/\mu m^2,$ $A = 0.35 \times 0.2\ (\mu m)^2]$		Cu $[J = 25\ mA/\mu m^2,$ $A = 0.24 \times 0.28\ (\mu m)^2]$		Cu $[J = 25\ mA/\mu m^2,$ $A = 1.3 \times 0.7\ (\mu m)^2]$		Cu $(T = 370\ °C)$	
T (°C)	t_{50} (hr)	T (°C)	t_{50} (hr)	T (°C)	t_{50} (hr)	J mA μm^{-2}	t_{50} (hr)
365	0.11	397	2.87	395	40.3	3.54	131.5
300	0.98	354	12.8	360	196	11.7	25.2
259	5.73	315	70.53	314	825	24.8	14.9
233	15.7	269	180	285	2098	49.2	4.28
		232	899			74.1	2.29
						140	0.69

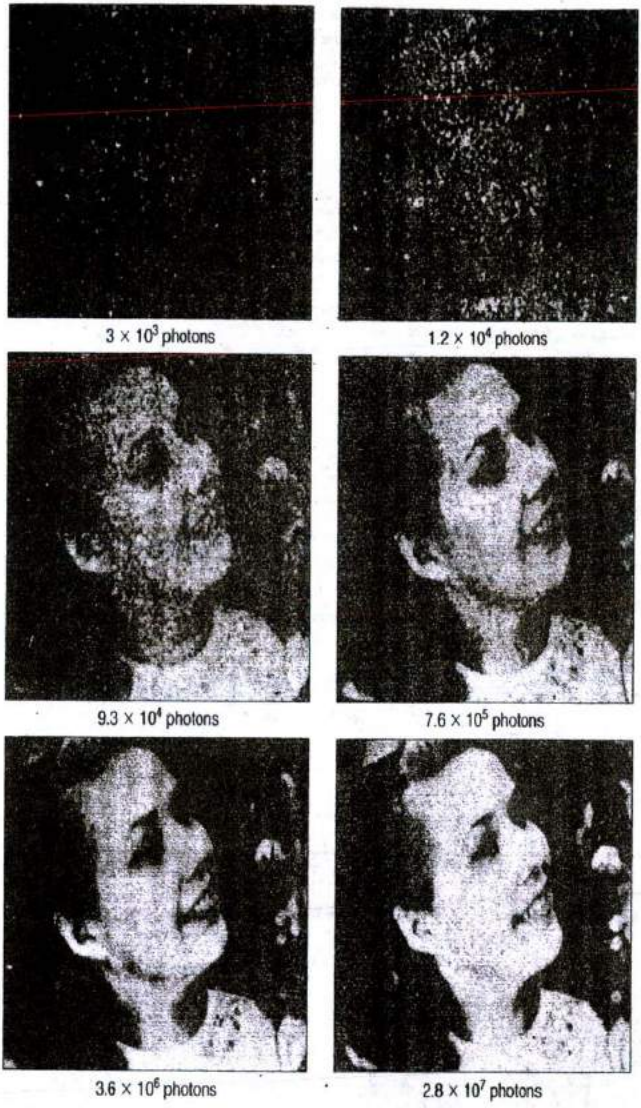
NOTE: A = width × height in micron^2.

SOURCE: Data extracted from R. Rosenberg *et al.*, (IBM, T. J. Watson Research Center, *Annu. Rev. Mater. Sci.*, **30**, 229, 2000, figures 29 and 31, and subject to small extraction errors.)



Gordon Teal (Left) and Morgan Sparks fabricated the first grown-junction Ge transistor in 1950–1951 at Bell Labs. Gordon Teal started at Bell Labs but later moved to Texas Instruments where he lead the development of the first commercial Si transistor; the first Si transistor was made at Bell Labs by Morris Tanenbaum.

SOURCE: Courtesy of Bell Laboratories, lucent Technologies



These electronic images were made with the number of photons indicated. The discrete nature of photons means that a large number of photons are needed to constitute an image with satisfactorily discernable details.

SOURCE: A. Rose, "Quantum and noise limitations of the visual process" *J. Opt. Soc. of America*, vol. 43, 715, 1953. (Courtesy of OSA.)

CHAPTER

3

Elementary Quantum Physics

The triumph of modern physics is the triumph of quantum mechanics. Even the simplest experimental observation that the resistivity of a metal depends linearly on the temperature can only be explained by quantum physics, simply because we must take the mean speed of the conduction electrons to be nearly independent of temperature. The modern definitions of voltage and ohm, adopted in January 1990 and now part of the IEEE standards, are based on Josephson and quantum Hall effects, both of which are quantum mechanical phenomena.

One of the most important discoveries in physics has been the wave–particle duality of nature. The electron, which we have so far considered to be a particle and hence to be obeying Newton's second law ($F = ma$), can also exhibit wave-like properties quite contrary to our intuition. An electron beam can give rise to diffraction patterns and interference fringes, just like a light wave. Interference and diffraction phenomena displayed by light can only be explained by treating light as an electromagnetic wave. But light can also exhibit particle-like properties in which it behaves as if it were a stream of discrete entities ("photons"), each carrying a linear momentum and each interacting discretely with electrons in matter (just like a particle colliding with another particle).

3.1. PHOTONS

3.1.1 LIGHT AS A WAVE

In introductory physics courses, light is considered to be a wave. Indeed, such phenomena as interference, diffraction, refraction, and reflection can all be explained by the theory of waves. In all these phenomena, a ray of light is considered to be an **electromagnetic (EM) wave** with a given frequency, as depicted in Figure 3.1. The electric and magnetic fields, $\mathcal{E}_y$ and B_z, of this wave are perpendicular to each other and to the direction of propagation x. The electric field $\mathcal{E}_y$ at position x at time t may be

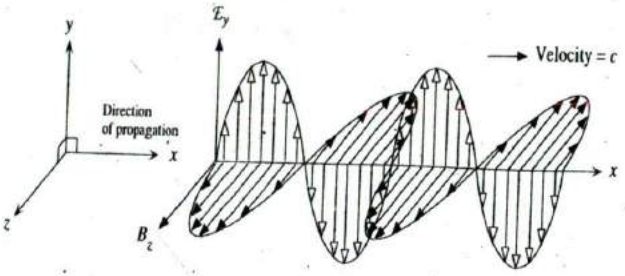


Figure 3.1 The classical view of light as an electromagnetic wave.
An electromagnetic wave is a traveling wave with time-varying electric and magnetic fields that are perpendicular to each other and to the direction of propagation.

described by

Traveling wave

$$\mathcal{E}_y(x, t) = \mathcal{E}_o \sin(kx - \omega t) \qquad [3.1]$$

where k is the wavenumber (propagation constant) related to the wavelength λ by $k = 2\pi/\lambda$, and ω is the angular frequency of the wave (or $2\pi\nu$, where ν is the frequency). A similar equation describes the variation of the magnetic field B_z (directed along z) with x at any time t. Equation 3.1 represents a traveling wave in the x direction, which, in the present example, is a sinusoidally varying function (Figure 3.1). The velocity of the wave (strictly the phase velocity) is

$$c = \frac{\omega}{k} = \nu\lambda$$

where ν is the frequency. The intensity I, that is, the energy flowing per unit area per second, of the wave represented by Equation 3.1 is given by

Intensity of light wave

$$I = \frac{1}{2} c\varepsilon_o \mathcal{E}_o^2 \qquad [3.2]$$

where ε_o is the absolute permittivity.

Understanding the wave nature of light is fundamental to understanding interference and diffraction, two phenomena that we experience with sound waves almost on a daily basis. Figure 3.2 illustrates how the interference of secondary waves from the two slits S_1 and S_2 gives rise to the dark and bright fringes (called **Young's fringes**) on a screen placed at some distance from the slits. At point P on the screen, the waves emanating from S_1 and S_2 interfere constructively, if they are in phase. This is the case if the path difference between the two rays is an integer multiple of the wavelength λ, or

$$S_1P - S_2P = n\lambda$$

where n is an integer. If the two waves are out of phase by a path difference of $\lambda/2$, or

$$S_1P - S_2P = \left(n + \frac{1}{2}\right)\lambda$$

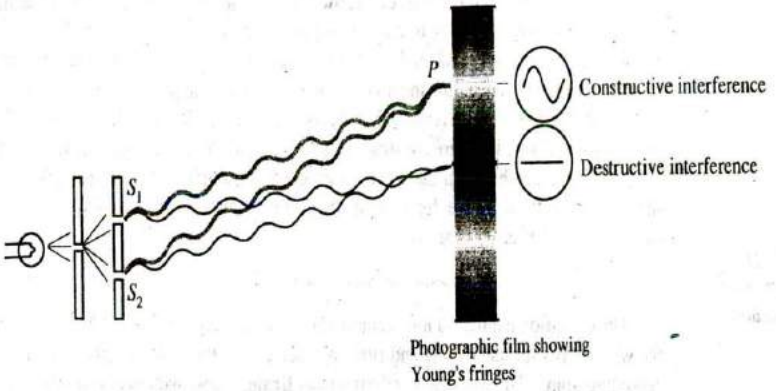


Figure 3.2 Schematic illustration of Young's double-slit experiment.

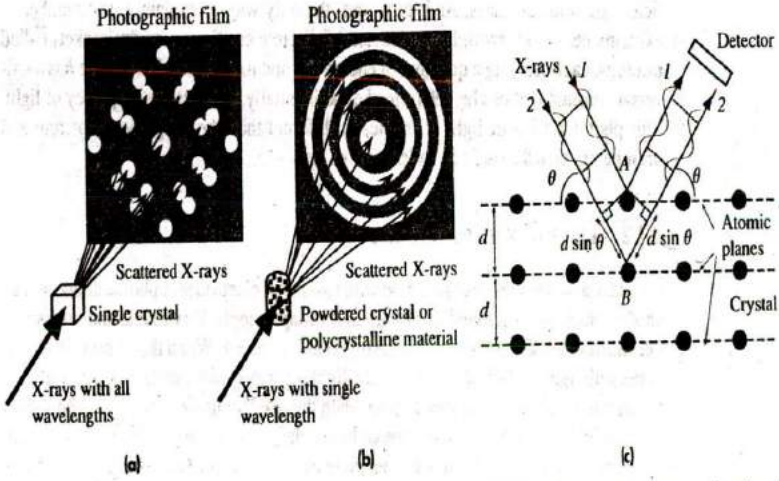


Figure 3.3 Diffraction patterns obtained by passing X-rays through crystals can only be explained by using ideas based on the interference of waves.

(a) Diffraction of X-rays from a single crystal gives a diffraction pattern of bright spots on a photographic film.

(b) Diffraction of X-rays from a powdered crystalline material or a polycrystalline material gives a diffraction pattern of bright rings on a photographic film.

(c) X-ray diffraction involves the constructive interference of waves being "reflected" by various atomic planes in the crystal.

then the waves interfere destructively and the intensity at point P vanishes. Thus, in the y direction, the observer sees a pattern of bright and dark fringes.

When X-rays are incident on a crystalline material, they give rise to typical diffraction patterns on a photographic plate, as shown in Figure 3.3a and b, which can only be explained by using wave concepts. For simplicity, consider two waves, *1* and *2*, in an X-ray beam. The waves are initially in phase, as shown in Figure 3.3c. Suppose that wave *1* is "reflected" from the first plane of atoms in the crystal, whereas

wave 2 is "reflected" from the second plane.[1] After reflection, wave 2 has traveled an additional distance equivalent to $2d \sin\theta$ before reaching wave 1. The path difference between the two waves is $2d \sin\theta$, where d is the separation of the atomic planes. For constructive interference, this must be $n\lambda$, where n is an integer. Otherwise, waves 1 and 2 will interfere destructively and will cancel each other. Waves reflected from adjacent atomic planes interfere constructively to constitute a diffracted beam *only* when the path difference between the waves is an integer multiple of the wavelength, and this will only be the case for certain directions. Therefore the *condition* for the existence of a diffracted beam is

Bragg diffraction condition

$$2d \sin\theta = n\lambda \qquad n = 1, 2, 3, \ldots \qquad [3.3]$$

The condition expressed in Equation 3.3, for observing a diffracted beam, forms the whole basis for identifying and studying various crystal structures (the science of crystallography). The equation is referred to as **Bragg's law,** and arises from the constructive interference of waves.

Aside from exhibiting wave-like properties, light can behave like a stream of "particles" of zero rest-mass. As it turns out, the only way to explain a vast number of experiments is to view light as a stream of discrete entities or energy packets called **photons,** each carrying a quantum of energy $h\nu$, and momentum h/λ, where h is a universal constant that can be determined experimentally, and ν is the frequency of light. This photonic view of light is drastically different than the simple wave picture and must be examined closely to understand its origin.

3.1.2 THE PHOTOELECTRIC EFFECT

Consider a quartz glass vacuum tube with two metal electrodes, a photocathode and an anode, which are connected externally to a voltage supply V (variable and reversible) via an ammeter, as schematically illustrated in Figure 3.4. When the cathode is illuminated with light, if the frequency ν of the light is greater than a certain critical value ν_0, the ammeter registers a current I, even when the anode voltage is zero (*i.e.*, the supply is bypassed). When light strikes the cathode, electrons are emitted with sufficient kinetic energy to reach the opposite electrode. Applying a positive voltage to the anode helps to collect more of the electrons and thus increases the current, until it saturates because all the photoemitted electrons have been collected. The current, then, is limited by the rate of supply of photoemitted electrons. If, on the other hand, we apply a negative voltage to the anode, we can "push" back the photoemitted electrons and hence reduce the current I. Figure 3.5a shows the dependence of the photocurrent on the anode voltage, for one particular frequency of light.

Recall that when an electron traverses a voltage difference V, its potential energy changes by eV (potential difference is defined as work done per unit charge). When a negative voltage is applied to the anode, the electron has to do work to get to this electrode, and this work comes from its kinetic energy just after photoemission. When the negative anode voltage V is equal to V_0, which just "extinguishes" the current I, we

[1] Strictly, one must consider the scattering of waves from the electrons in individual atoms [e.g., atoms A and B in Figure 3.3c] and examine the constructive interference of these scattered waves, which leads to the same condition as that derived in Equation 3.3.

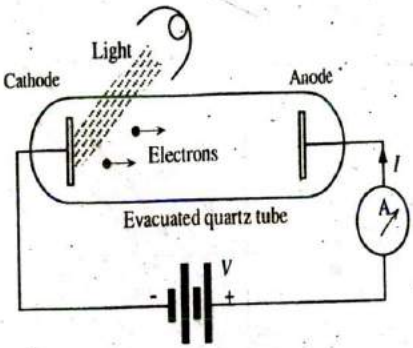
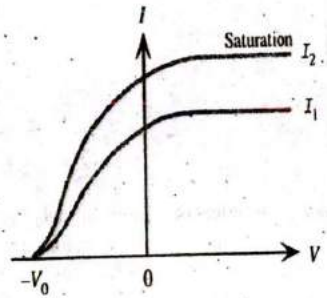
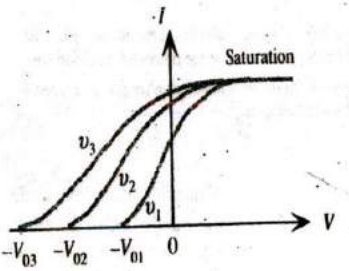


Figure 3.4 The photoelectric effect.



(a) Photoelectric current versus voltage when the cathode is illuminated with light of identical wavelength but different intensities (I). The saturation current is proportional to the light intensity.



(b) The stopping voltage and therefore the maximum kinetic energy of the emitted electron increases with the frequency of light, υ. (The light intensity is not the same; it is adjusted to keep the saturation current the same.)

Figure 3.5 Results from the photoelectric experiment.

know that the potential energy "gained" by the electron is just the kinetic energy lost by the electron, or

$$eV_0 = \frac{1}{2}m_e v^2 = KE_m$$

where v is the velocity and KE_m is the kinetic energy of the electron just after photoemission. Therefore, we can conveniently measure the maximum kinetic energy KE_m of the emitted electrons.

For a given frequency of light, increasing the intensity of light I requires the *same* voltage V_0 to extinguish the current; that is, the KE_m of emitted electrons is independent of the light intensity I. This is quite surprising. However, increasing the intensity does increase the saturation current. Both of these effects are noted in the I–V results shown in Figure 3.5a.

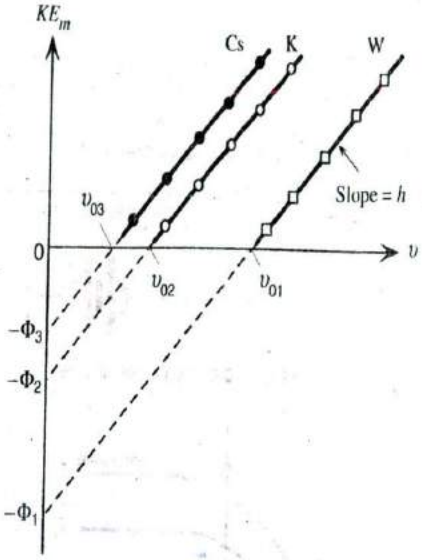


Figure 3.6 The effect of varying the frequency of light and the cathode material in the photoelectric experiment. The lines for the different materials have the same slope h but different intercepts.

Since the magnitude of the saturation photocurrent depends on the light intensity I, whereas the KE of the emitted electron is independent of I, we are forced to conclude that only the *number* of electrons ejected depends on the light intensity. Furthermore, if we plot KE_m (from the V_0 value) against the light frequency ν for different electrode metals for the cathode, we find the typical behavior shown in Figure 3.6. This shows that the KE of the emitted electron depends on the frequency of light. The experimental results shown in Figure 3.6 can be summarized by a statement that relates the KE_m of the electron to the frequency of light and the electrode metal, as follows:

Photoemitted electron maximum KE

$$KE_m = h\nu - h\nu_0 \qquad [3.4]$$

where h is the slope of the straight line and is independent of the type of metal, whereas ν_0 depends on the electrode material for the photocathode (*e.g.*, ν_{01}, ν_{02}, etc.). Equation 3.4 is essentially a succinct statement of the experimental observations of the photoelectric effect as exhibited in Figure 3.6. The constant h is called **Planck's constant,** which, from the slope of the straight lines in Figure 3.6, can be shown to be about 6.6×10^{-34} J s. This was beautifully demonstrated by Millikan in 1915, in an excellent series of photoelectric experiments using different photocathode materials.

The successful interpretation of the photoelectric effect was first given in 1905 by Einstein, who proposed that light consists of "energy packets," each of which has the magnitude $h\nu$. We can call these energy quanta **photons.** When one photon strikes an electron, its energy is transferred to the electron. The whole photon becomes absorbed by the electron. Yet, an electron in a metal is in a lower state of potential energy (PE) than in vacuum, by an amount Φ, which we call the **work function** of the metal, as illustrated in Figure 3.7. The lower PE is what keeps the electron in the metal; otherwise, it would "drop out."

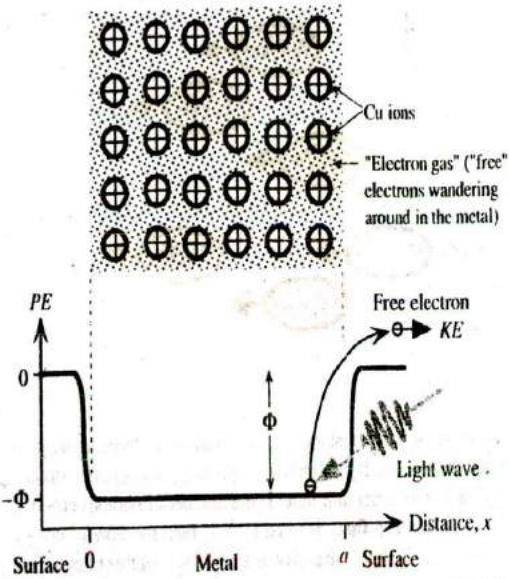


Figure 3.7 The *PE* of an electron inside the metal is lower than outside by an energy called the workfunction of the metal.
Work must be done to remove the electron from the metal.

This lower *PE* is a result of the Coulombic attraction interaction between the electron and the positive metal ions. Some of the photon energy $h\nu$ therefore goes toward overcoming this *PE* barrier. The energy that is left ($h\nu - \Phi$) gives the electron its *KE*. The work function Φ changes from one metal to another. Photoemission only occurs when $h\nu$ is greater than Φ. This is clearly borne out by experiment, since a critical frequency ν_0 is needed to register a photocurrent. When ν is less than ν_0, even if we use an extremely intense light, no current exists because no photoemission occurs, as demonstrated by the experimental results in Figure 3.6. Inasmuch as Φ depends on the metal, so does ν_0. Therefore, in Einstein's interpretation $h\nu_0 = \Phi$. In fact, the measurement of ν_0 constitutes one method of determining the work function of the metal.

This explanation for the photoelectric effect is further supported by the fact that the work function Φ from $h\nu_0$ is in good agreement with that from thermionic emission experiments. There is an apparent similarity between the *I–V* characteristics of the phototube and that of the vacuum tube used in early radios. The only difference is that in the vacuum tube, the emission of electrons from the cathode is achieved by heating the cathode. Thermal energy ejects some electrons over the *PE* barrier Φ. The measurement of Φ by this thermionic emission process agrees with that from photoemission experiments.

In the photonic interpretation of light, we still have to resolve the meaning of the intensity of light, because the classical intensity in Equation 3.2

$$I = \frac{1}{2} c\varepsilon_o E_o^2$$

Classical light intensity

is obviously not acceptable. Increasing the intensity of illumination in the photoelectric experiment increases the saturation current, which means that more electrons are

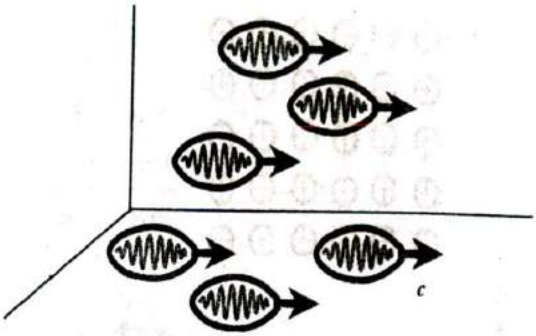


Figure 3.8 Intuitive visualization of light consisting of a stream of photons (not to be taken too literally).
SOURCE: R. Serway, C. J. Moses, and C. A. Moyer, *Modern Physics*, Saunders College Publishing, 1989, p. 56, figure 2.16(b).

emitted per unit time. We therefore infer that the cathode must be receiving more photons per unit time at higher intensities. By definition, "intensity" refers to the amount of energy flowing through a unit area per unit time. If the number of photons crossing a unit area per unit time is the **photon flux,** denoted by Γ_{ph}, then the flow of energy through a unit area per unit time, the **light intensity,** is the product of this photon flux and the energy per photon, that is,

Light intensity

$$I = \Gamma_{ph} h\nu \qquad [3.5]$$

where

Photon flux

$$\Gamma_{ph} = \frac{\Delta N_{ph}}{A\,\Delta t} \qquad [3.6]$$

in which ΔN_{ph} is the net number of photons crossing an area A in time Δt. With the energy of a photon given as $h\nu$ and the intensity of light defined as $\Gamma_{ph} h\nu$, the explanation for the photoelectric effect becomes self-consistent. The interpretation of light as a stream of photons can perhaps be intuitively imagined as depicted in Figure 3.8.

EXAMPLE 3.1 **ENERGY OF A BLUE PHOTON** What is the energy of a blue photon that has a wavelength of 450 nm?

SOLUTION

The energy of the photon is given by

$$E_{ph} = h\nu = \frac{hc}{\lambda} = \frac{(6.6 \times 10^{-34}\ \text{J s})(3 \times 10^{8}\ \text{m s}^{-1})}{450 \times 10^{-9}\ \text{m}} = 4.4 \times 10^{-19}\ \text{J}$$

Generally, with such small energy values, we prefer electron–volts (eV), so the energy of the photon is

$$\frac{4.4 \times 10^{-19}\ \text{J}}{1.6 \times 10^{-19}\ \text{J/eV}} = 2.75\ \text{eV}$$

EXAMPLE 3.2

THE PHOTOELECTRIC EXPERIMENT In the photoelectric experiment, green light, with a wavelength of 522 nm, is the longest-wavelength radiation that can cause the photoemission of electrons from a clean sodium surface.

a. What is the work function of sodium, in electron–volts?

b. If UV (ultraviolet) radiation of wavelength 250 nm is incident to the sodium surface, what will be the kinetic energy of the photoemitted electrons, in electron–volts?

c. Suppose that the UV light of wavelength 250 nm has an intensity of 20 mW cm^{-2}. If the emitted electrons are collected by applying a positive bias to the opposite electrode, what will be the photoelectric current density?

SOLUTION

a. At threshold, the photon energy just causes photoemissions; that is, the electron just overcomes the potential barrier Φ. Thus, $hc/\lambda_0 = e\Phi$, where Φ is the work function in eV, and λ_0 is the longest wavelength.

$$\Phi = \frac{hc}{e\lambda_0} = \frac{(6.626 \times 10^{-34}\,\text{J s})(3 \times 10^{8}\,\text{m s}^{-1})}{(1.6 \times 10^{-19}\,\text{J/eV})(522 \times 10^{-9}\,\text{m})} = 2.38\ \text{eV}$$

b. The energy of the incoming photon E_{ph} is (hc/λ), so the excess energy over $e\Phi$ goes to the kinetic energy of the electron. Thus,

$$KE = \frac{hc}{e\lambda} - \Phi = \frac{(6.626 \times 10^{-34}\,\text{J s})(3 \times 10^{8}\,\text{m s}^{-1})}{(1.6 \times 10^{-19}\,\text{J/eV})(250 \times 10^{-9}\,\text{m})} - 2.38\ \text{eV} = 2.58\ \text{eV}$$

c. The light intensity (defined as energy flux) is given by $I = \Gamma_{ph}(hc/\lambda)$, where Γ_{ph} is the number of photons arriving per unit area per unit time; that is, photon flux and (hc/λ) is the energy per photon. Thus, if each photon releases one electron, the electron flux will be equal to the photon flux, and the current density, which is the charge flux, will be

$$J = e\Gamma_{ph} = \frac{eI\lambda}{hc} = \frac{(1.6 \times 10^{-19}\,\text{C})(20 \times 10^{-3} \times 10^{4}\,\text{J s}^{-1}\,\text{m}^{-2})(250 \times 10^{-9}\,\text{m})}{(6.626 \times 10^{-34}\,\text{J s})(3 \times 10^{8}\,\text{m s}^{-1})}$$

$$= 40.3\ \text{A m}^{-2} \quad \text{or} \quad 4.0\ \text{mA cm}^{-2}$$

3.1.3 COMPTON SCATTERING

When an X-ray strikes an electron, it is deflected, or "scattered." In addition, the electron moves away after the interaction, as depicted in Figure 3.9. The wavelength of the incoming and scattered X-rays can readily be measured. The frequency ν' of the scattered X-ray is less than the frequency ν of the incoming X-ray. When the KE of the electron is determined, we find that

$$KE = h\nu - h\nu'$$

Since the electron now also has a momentum p_e, then from the conservation of linear momentum law, we are forced to accept that the X-ray also has a momentum. The Compton effect experiments showed that the momentum of the photon is related to its wavelength by

$$p = \frac{h}{\lambda} \qquad [3.7]$$

Momentum of a photon

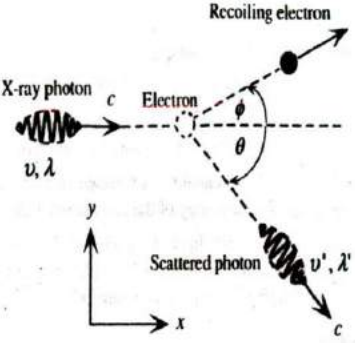


Figure 3.9 Scattering of an X-ray photon by a "free" electron in a conductor.

We see that a photon not only has an energy $h\nu$, but also a momentum p, and it interacts as if it were a discrete entity like a particle. Therefore, when discussing the properties of a photon, we must consider its energy and momentum as if it were a particle.

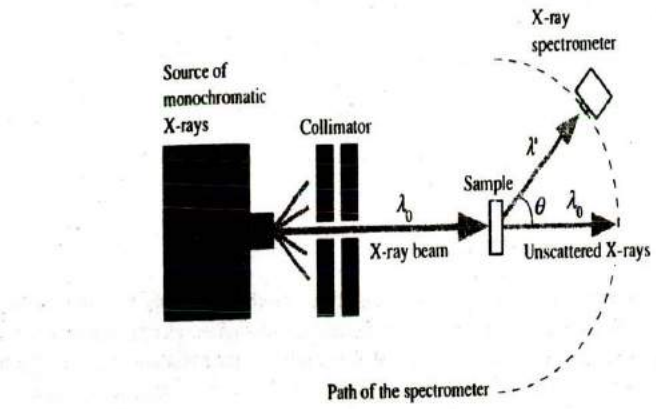
We should mention that the description of the Compton effect shown in Figure 3.9 is, in fact, the inference from a more practical experiment involving the scattering of X-rays from a metal target. A collimated monochromatic beam of X-rays of wavelength λ_0 strikes a conducting target, such as graphite, as illustrated in Figure 3.10a. A conducting target contains a large number of nearly "free" electrons (conduction electrons), which can scatter the X-rays. The scattered X-rays are detected at various angles θ with respect to the original direction, and their wavelength λ' is measured. The result of the experiment is therefore the scattered wavelength λ' measured at various scattering angles θ, as shown in Figure 3.10b. It turns out that the λ' versus θ results agree with the conservation of linear momentum law applied to an X-ray photon colliding with an electron with the momentum of the photon given precisely by Equation 3.7.

The photoelectric experiment and the Compton effect are just two convincing experiments in modern physics that force us to accept that light can have particle-like properties. We already know that it can also exhibit wave-like properties, in such experiments as Young's interference fringes. We are then faced with what is known as the wave–particle dilemma. How do we know whether light is going to behave like a wave or a particle? The properties exhibited by light depend very much on the nature of the experiment. Some experiments will require the wave model, whereas others may use the particulate interpretation of light. We should perhaps view the two interpretations as two complementary ways of modeling the behavior of light when it interacts with matter, accepting the fact that light has a dual nature. Both models are needed for a full description of the behavior of light.

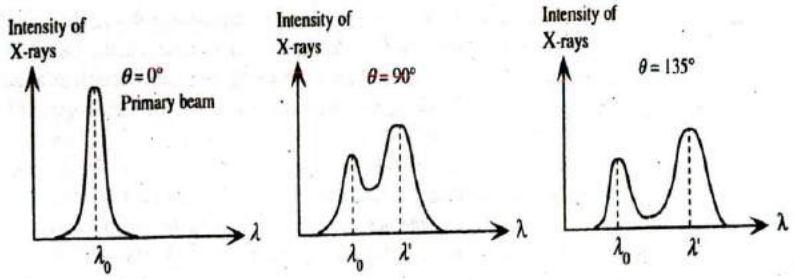
The expressions for the energy and momentum of the photon, $E = h\nu$ and $p = h/\lambda$, can also be written in terms of the angular frequency $\omega (= 2\pi\nu)$ and the wave number k, defined as $k = 2\pi/\lambda$. If we define $\hbar = h/2\pi$, then

Photon energy and momentum

$$E = h\nu = \hbar\omega \quad \text{and} \quad p = \frac{h}{\lambda} = \hbar k \tag{3.8}$$



(a) A schematic diagram of the Compton experiment



(b) Results from the Compton experiment

Figure 3.10 The Compton experiment and its results.

EXAMPLE 3.3

X-RAY PHOTON ENERGY AND MOMENTUM X-rays are photons with very short wavelengths that can penetrate or pass through objects, hence their use in medical imaging, security scans at airports, and many other applications including X-ray diffraction studies of crystal structures. Typical X-rays used in mammography (medical imaging of breasts) have a wavelength of about 0.6 angstrom ($1\ \text{Å} = 10^{-10}\ \text{m}$). Calculate the energy and momentum of an X-ray photon with this wavelength, and the velocity of a *corresponding* electron that has the same momentum.

SOLUTION

The photon energy E_{ph} is given by

$$E_{ph} = h\nu = \frac{hc}{\lambda} = \frac{(6.6 \times 10^{-34}\ \text{J s})(3 \times 10^{8}\ \text{m s}^{-1})}{0.6 \times 10^{-10}\ \text{m}} \times \frac{\text{eV J}^{-1}}{1.6 \times 10^{-19}}$$

$$= 2.06 \times 10^{4}\ \text{eV} \quad \text{or} \quad 20.6\ \text{keV}$$

The momentum p of this X-ray photon is

$$p = \frac{h}{\lambda} = \frac{6.6 \times 10^{-34}\ \text{J s}}{0.6 \times 10^{-10}\ \text{m}} = 1.1 \times 10^{-23}\ \text{kg m s}^{-1}$$

A corresponding electron with the same momentum, $m_e v_{electron} = p$, would have a velocity

$$v_{electron} = \frac{p}{m_e} = \frac{1.1 \times 10^{-23}\ \text{kg m s}^{-1}}{9.1 \times 10^{-31}\ \text{kg}} = 1.2 \times 10^{7}\ \text{m s}^{-1}$$

This is much greater than the average speed of conduction (free) electrons whizzing around inside a metal, which is $\sim 10^{6}\ \text{m s}^{-1}$.

3.1.4 BLACK BODY RADIATION

Experiments indicate that all objects emit and absorb energy in the form of radiation, and the intensity of this radiation depends on the radiation wavelength and temperature of the object. This radiation is frequently termed **thermal radiation.** When the object is in thermal equilibrium with its surroundings, that is, at the same temperature, the object absorbs as much radiation energy as it emits. On the other hand, when the temperature of the object is above the temperature of its surroundings, there is a net emission of radiation energy. The maximum amount of radiation energy that can be emitted by an object is called the **black body radiation.** Although, in general, the intensity of the radiated energy depends on the material's surface, the radiation emitted from a cavity with a small aperture is independent of the material of the cavity and corresponds very closely to black body radiation.

The intensity of the emitted radiation has the spectrum (*i.e.*, intensity vs. wavelength characteristic), and the temperature dependence illustrated in Figure 3.11. It is useful to define a **spectral irradiance** I_λ as the emitted radiation intensity (power per unit area) per unit wavelength, so that $I_\lambda\,\delta\lambda$ is the intensity in a small range of wavelengths $\delta\lambda$. Figure 3.11 shows the typical I_λ versus λ behavior of black body radiation at two temperatures. We assume that the characteristics of the radiation emerging from the aperture represent those of the radiation within the cavity.

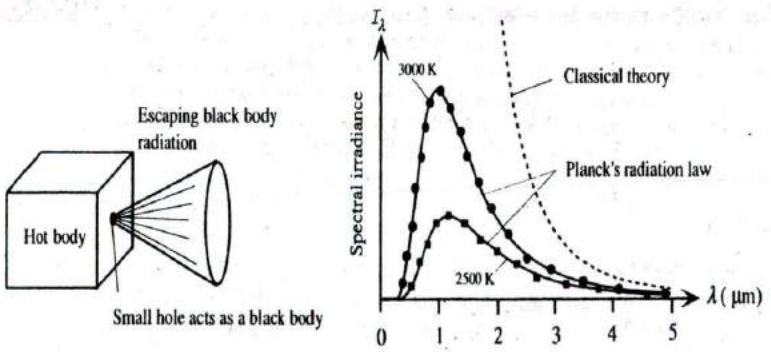


Figure 3.11 Schematic illustration of black body radiation and its characteristics.
Spectral irradiance versus wavelength at two temperatures (3000 K is about the temperature of the incandescent tungsten filament in a light bulb).

Classical physics predicts that the acceleration and deceleration of the charges due to various thermal vibrations, oscillations, or motions of the atoms in the surface region of the cavity material result in electromagnetic waves of the emissions. These waves then interfere with each other, giving rise to many types of standing electromagnetic waves with different wavelengths in the cavity. Each wave contributes an energy kT to the emitted intensity. If we calculate the number of standing waves within a small range of wavelength, the classical prediction leads to the **Rayleigh–Jeans law** in which $I_\lambda \propto 1/\lambda^4$ and $I_\lambda \propto T$, which are not in agreement with the experiment, especially in the short-wavelength range (see Figure 3.11).

Max Planck (1900) was able to show that the experimental results can be explained if we assume that the radiation within the cavity involves the emission and absorption of discrete amounts of light energy by the oscillation of the molecules of the cavity material. He assumed that oscillating molecules emit and absorb a quantity of energy that is an integer multiple of a discrete energy quantum that is determined by the frequency ν of the radiation and given by $h\nu$. This is what we now call a photon. He then considered the energy distribution (the statistics) in the molecular oscillations and took the probability of an oscillator possessing an energy $nh\nu$ (where n is an integer) to be proportional to the Boltzmann factor, $\exp(-nh\nu/kT)$. He eventually derived the mathematical form of the black body radiation characteristics in Figure 3.11. Planck's black body radiation formula for I_λ is generally expressed as

$$I_\lambda = \frac{2\pi hc^2}{\lambda^5\left[\exp\left(\frac{hc}{\lambda kT}\right) - 1\right]} \qquad [3.9]$$

Planck's radiation law

where k is the Boltzmann constant. Planck's radiation law based on the emission and absorption of photons is in excellent agreement with all observed black body radiation characteristics as depicted in Figure 3.11.

Planck's radiation law is undoubtedly one of the major successes of modern physics. We can take Equation 3.9 one step further and derive *Stefan's black body radiation law* that was used in Chapter 2 to calculate the rate of radiation energy emitted from the hot filament of a light bulb. If we integrate I_λ over all wavelengths,[2] we will obtain the total radiative power P_S emitted by a black body per unit surface area at a temperature T,

$$P_S = \int_0^\infty I_\lambda \, d\lambda = \left(\frac{2\pi^5 k^4}{15c^2h^3}\right)T^4 = \sigma_S T^4 \qquad [3.10]$$

Stefan's black body radiation law

where

$$\sigma_S = \frac{2\pi^5 k^4}{15c^2h^3} = 5.670 \times 10^{-8}\ \mathrm{W\ m^{-2}\ K^{-4}} \qquad [3.11]$$

Stefan's constant

[2] The integration of Equation 3.9 can be done by looking up definite integral tables in math handbooks—we only need the result of the mathematics, which is Equation 3.10. The P_S in Equation 3.10 is sometimes called the *radiant emittance*. Stefan's law is also known as the Stefan-Boltzmann law.

Equation 3.10 in which $P_S = \sigma_S T^4$ is **Stefan's law** for black body radiation, and the σ_S in Equation 3.11 is the **Stefan constant** with a value of approximately $5.67 \times 10^{-8}\ \text{W m}^{-2}\ \text{K}^{-4}$. Stefan's law was known before Planck used quantum physics to derive his black body radiation law embedded in I_λ. A complete explanation of Stefan's law and the value for σ_S however had to wait for Planck's law. The h in Equation 3.10 or 3.11 is a clear pointer that the origin of Stefan's law lies in quantum physics.

EXAMPLE 3.4 **STEFAN'S LAW AND THE LIGHT BULB** Stefan's law as stated in Equation 3.10 applies to a perfect black body that is emitting radiation into its environment which is at absolute zero. If the environment or the surroundings of the black body is at a finite temperature T_o, than the surroundings would also be emitting radiation. The same black body will then also absorb radiation from its environment. By definition, a black body is not only a perfect emitter of radiation but also a perfect absorber of radiation. The rate of radiation absorbed from the environment per unit surface is again given by Equation 3.10 but with T_o instead of T since it is the surroundings that are emitting the radiation. Thus, $\sigma_S T_o^4$ is the absorbed radiation rate from the surroundings, so

$$\text{Net rate of radiative power emission per unit surface} = \sigma_S T^4 - \sigma_S T_o^4$$

Further, not all surfaces are perfect black bodies. Black body emission is the maximum possible emission from a surface at a given temperature. A real surface emits less than a black body. **Emissivity** ε of a surface measures the efficiency of a surface in terms of a black body emitter; it is the ratio of the emitted radiation from a real surface to that emitted from a black body at a given temperature and over the same wavelength range. The *total* net rate of radiative power emission becomes

Stefan's law for a real surface

$$P_{\text{radiation}} = S\varepsilon\sigma_S\left(T^4 - T_o^4\right) \qquad \textbf{[3.12]}$$

where S is the surface area that is emitting the radiation. Consider the tungsten filament of a 100 W light bulb in a lamp. When we switch the lamp on, the current through the filament generates heat which quickly heats up the filament to an operating temperature T_f. At this temperature, the electric energy that is input into the bulb is radiated away from the filament as radiation energy. A typical 100 W bulb filament has a length of 57.9 cm and a diameter of 63.5 μm. Its surface area is then

$$S = \pi(63.5 \times 10^{-6}\ \text{m})(0.579\ \text{m}) = 1.155 \times 10^{-4}\ \text{m}^2$$

The emissivity ε of tungsten is about 0.35. Assuming that under steady-state operation all the electric power that is input into the bulb's filament is radiated away,

$$\begin{aligned} 100\ \text{W} = P_{\text{radiation}} &= S\varepsilon\sigma_S\left(T_f^4 - T_o^4\right) \\ &= (1.155 \times 10^{-4}\ \text{m}^2)(0.35)(5.67 \times 10^{-8}\ \text{W m}^{-2}\ \text{K}^{-4})\left(T_f^4 - 300^4\right) \end{aligned}$$

Solving we find,

$$T_f = 2570\ \text{K} \quad \text{or} \quad 2297\ ^\circ\text{C}$$

which is well below the melting temperature of tungsten which is 3422 °C. The second term that has T_o^4 has very little effect on the calculation as radiation absorption from the environment is practically nil compared with the emitted radiation at T_f.

The shift in the spectral intensity emitted from a black body with temperature is of particular interest to many photoinstrumention engineers. The peak spectral intensity in Figure 3.11 occurs at a wavelength λ_{max}, which, by virtue of Equation 3.9, depends on the temperature of

the black body. By substituting a new variable $x = hc/(kT\lambda)$ into Equation 3.9 and differentiating it, or plotting it against x, we can show that the peak occurs when

$$\lambda_{max} T \approx 2.89 \times 10^{-3} \text{ m K}$$

Wien's displacement law

which is known as **Wien's displacement law.** The peak emission shifts to lower wavelengths as the temperature increases. We can calculate the wavelength λ_{max} corresponding to the peak in the spectral distribution of emitted radiation from our 100 W lamp: $\lambda_{max} = (2.89 \times 10^{-3} \text{ m K})/(2570 \text{ K}) = 1.13 \ \mu\text{m}$ (in the infrared).

3.2 THE ELECTRON AS A WAVE

3.2.1 DE BROGLIE RELATIONSHIP

It is apparent from the photoelectric and Compton effects that light, which we thought was a wave, can behave as if it were a stream of particulate-like entities called photons. Can electrons exhibit wave-like properties? Again, this depends on the experiment and on the energy of the electrons.

When the interference and diffraction experiments in Figures 3.2 and 3.3 are repeated with an electron beam, very similar results are found to those obtainable with light and X-rays. When we use an electron beam in Young's double-slit experiment, we observe high- and low-intensity regions (*i.e.*, Young's fringes), as illustrated in Figure 3.12. The interference pattern is viewed on a fluorescent TV screen. When an energetic electron beam hits an Al polycrystalline sample, it produces diffraction rings on a fluorescent screen (Figure 3.13), just like X-rays do on a photographic

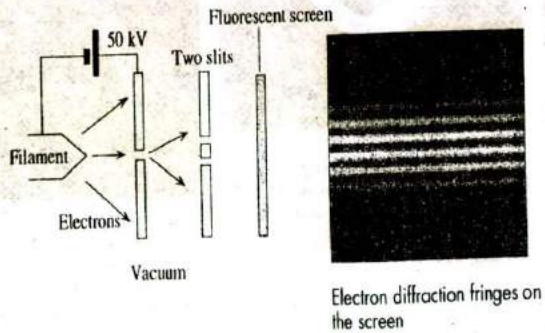
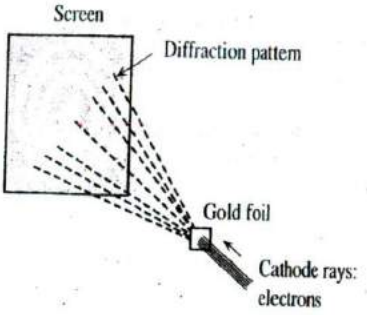


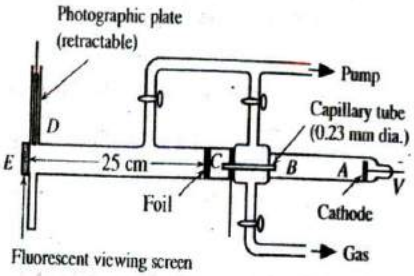
Figure 3.12 Young's double-slit experiment with electrons involves an electron gun and two slits in a cathode ray tube (CRT) (hence, in vacuum).

Electrons from the filament are accelerated by a 50 kV anode voltage to produce a beam that is made to pass through the slits. The electrons then produce a visible pattern when they strike a fluorescent screen (e.g., a TV screen), and the resulting visual pattern is photographed.

SOURCE: Pattern from C. Jönsson, D. Brandt, and S. Hirschi, *Am. J. Physics*, **42**, 1974, p. 9, figure 8. Used with permission.



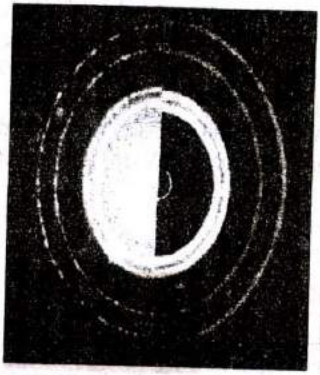
(a) Thomson diffracted electrons by using a thin gold foil and produced a diffraction pattern on the screen of his apparatus in (b). The foil was polycrystalline, so the diffraction pattern was circular rings.



(b) In Thomson's electron diffraction apparatus a beam of electrons is generated in tube *A*, passed through collimating tube *B*, and made to impinge on a thin gold foil *C*. The transmitted electrons impinge on the fluorescent screen *E*, or a photographic plate *D*, which could be lowered into the path. The entire apparatus was evacuated during the experiment.



(c) Electron diffraction pattern obtained by G. P. Thomson using a gold foil target.



(d) Composite photograph showing diffraction patterns produced with an aluminum foil by X-rays and electrons of similar wavelength. Left: X-rays of $\lambda = 0.071$ nm. Right: Electrons of energy 600 eV.



(e) Diffraction pattern produced by 40 keV electrons passing through zinc oxide powder. The distortion of the pattern was produced by a small magnet placed between the sample and the photographic plate. An X-ray diffraction pattern would not be affected by a magnetic field.

Figure 3.13 The diffraction of electrons by crystals gives typical diffraction patterns that would be expected if waves were being diffracted, as in X-ray diffraction with crystals.

SOURCE: (b) from G. P. Thomson, *Proceedings of the Royal Society*, A117, no. 600, 1928; (c) and (d) from A. P. French and F. Taylor, *An Introduction to Quantum Mechanics*, Norton, New York, 1978, p. 75; (e) from R. B. Leighton, *Principles of Modern Physics*, New York: McGraw-Hill, 1959, p. 84.

plate. The diffraction pattern obtained with an electron beam (Figure 3.13) means that the electrons are obeying the Bragg diffraction condition $2d \sin\theta = n\lambda$ just as much as the X-ray waves.

Since we know the interatomic spacing d and we can measure the angle of diffraction 2θ, we can readily evaluate the wavelength λ associated with the wave-like behavior of the electrons. Furthermore, from the accelerating voltage V in the electron tube, we can also determine the momentum of the electrons, because the kinetic energy gained by the electrons, $(p^2/2m_e)$, is equal to eV. Simply by adjusting the accelerating voltage V, we can therefore study how the wavelength of the electron depends on the momentum.

As a result of such studies and other similar experiments, it has been found that an electron traveling with a momentum p behaves like a wave of wavelength λ given by

$$\lambda = \frac{h}{p} \qquad [3.13]$$

Wavelength of the electron

This is just the reverse of the equation for the momentum of a photon given its wavelength. The same equation therefore relates wave-like and particle-like properties to and from each other. Thus,

$$\lambda = \frac{h}{p} \quad \text{or} \quad p = \frac{h}{\lambda}$$

De Broglie relations

is an equation that exposes the wave–particle duality of nature. It was first hypothesized by De Broglie in 1924. As an example, we can calculate the wavelengths of a number of particle-like objects:

a. A 50 gram golf ball traveling at a velocity of 20 m s^{-1}.

The wavelength is

$$\lambda = \frac{h}{mv} = \frac{6.63 \times 10^{-34}\ \text{J s}}{(50 \times 10^{-3}\ \text{kg})(20\ \text{m s}^{-1})} = 6.63 \times 10^{-34}\ \text{m}$$

The wavelength is so small that this golf ball will not exhibit any wave effects. Firing a stream of golf balls at a wall will not result in "diffraction rings" of golf balls.

b. A proton traveling at 2200 m s^{-1}.

Using $m_p = 1.67 \times 10^{-27}$ kg, we have $\lambda = (h/mv) \approx 0.18$ nm. This is only slightly smaller than the interatomic distance in crystals, so firing protons at a crystal can result in diffraction. (Recall that to get a diffraction peak, we must satisfy the Bragg condition, $2d \sin\theta = n\lambda$.) Protons, however, are charged, so they can penetrate only a small distance into the crystal. Hence, they are not used in crystal diffraction studies.

c. Electron accelerated by 100 V.

This voltage accelerates the electron to a KE equal to eV. From $KE = p^2/2m_e = eV$, we can calculate p and hence $\lambda = h/p$. The result is $\lambda = 0.123$ nm. Since this is comparable to typical interatomic distances in solids, we would see a diffraction pattern when an electron beam strikes a crystal. The actual pattern is determined by the Bragg diffraction condition.

3.2 TIME-INDEPENDENT SCHRÖDINGER EQUATION

The experiments in which electrons exhibit interference and diffraction phenomena show quite clearly that, under certain conditions, the electron can behave as a wave; in other words, it can exhibit wave-like properties. There is a general equation that describes this wave-like behavior and, with the appropriate potential energy and boundary conditions, will predict the results of the experiments. The equation is called **the Schrödinger equation** and it forms the foundations of quantum theory. Its fundamental nature is analogous to the classical physics assertion of Newton's second law, $F = ma$, which of course cannot be proved. As a fundamental equation, Schrödinger's has been found to successfully predict every observable physical phenomenon at the atomic scale. Without this equation, we will not be able to understand the properties of electronic materials and the principles of operation of many semiconductor devices. We introduce the equation through an analogy.

A traveling electromagnetic wave resulting from sinusoidal current oscillations, or the traveling voltage wave on a long transmission line, can generally be described by a traveling-wave equation of the form

$$\mathcal{E}(x, t) = \mathcal{E}_o \exp\, j(kx - \omega t) = \mathcal{E}(x) \exp(-j\omega t) \qquad [3.14]$$

where $\mathcal{E}(x) = \mathcal{E}_0 \exp(jkx)$ represents the spatial dependence, which is separate from the time variation. We assume that no transients exist to upset this perfect sinusoidal propagation. We note that the time dependence is harmonic and therefore predictable. For this reason, in ac circuits we put aside the $\exp(-j\omega t)$ term until we need the instantaneous magnitude of the voltage.

The average intensity $I_{av} = \frac{1}{2}c\epsilon_o\mathcal{E}_o^2$ depends on the square of the amplitude. In Young's double-slit experiment, the intensity varies along the y direction, which means that $\mathcal{E}_o^2$ for the resultant wave depends on y. In the electron version of this experiment in Figure 3.12, what changes in the y direction is the probability of observing electrons; that is, there are peaks and troughs in the probability of finding electrons along y, just like the $\mathcal{E}_o^2$ variation along y. We should therefore attach some probability interpretation to the wave description of the electron.

In 1926, Max Born suggested a probability wave interpretation for the wave-like behavior of the electron.

$$\mathcal{E}(x, t) = \mathcal{E}_o \sin(kx - \omega t)$$

is a plane traveling **wavefunction** for an electric field; experimentally, we measure and interpret the *intensity* of a wave, namely $|\mathcal{E}(x, t)|^2$. There may be a similar wave function for the electron, which we can represent by a function $\Psi(x, t)$. According to Born, the significance of $\Psi(x, t)$ is that its amplitude squared represents the probability of finding the electron per unit distance. Thus, in three dimensions, if $\Psi(x, y, z, t)$ represents the wave property of the electron, it must have one of the following interpretations:

$|\Psi(x, y, z, t)|^2$ is the probability of finding the electron per unit volume at x, y, z at time t.

$|\Psi(x, y, z, t)|^2\, dx\, dy\, dz$ is the probability of finding the electron in a small elemental volume $dx\, dy\, dz$ at x, y, z at time t.

If we are just considering one dimension, then the wavefunction is $\Psi(x,t)$, and $|\Psi(x,t)|^2\,dx$ is the probability of finding the electron between x and $(x+dx)$ at time t.

We should note that since only $|\Psi|^2$ has meaning, not Ψ, the latter function need not be real; it can be a complex function with real and imaginary parts. For this reason, we tend to use $\Psi^*\Psi$, where Ψ^* is the complex conjugate of Ψ, instead of $|\Psi|^2$, to represent the probability per unit volume.

To obtain the wavefunction $\Psi(x,t)$ for the electron, we need to know how the electron interacts with its environment. This is embodied in its potential energy function $V = V(x,t)$, because the net force the electron experiences is given by

$$F = -dV/dx.$$

For example, if the electron is attracted by a positive charge (*e.g.*, the proton in a hydrogen atom), then it clearly has an electrostatic potential energy given by

$$V(r) = -\frac{e^2}{4\pi\varepsilon_o r}$$

where $r = \sqrt{x^2+y^2+z^2}$ is the distance between the electron and the proton.

If the *PE* of the electron is time independent, which means that $V = V(x)$ in one dimension, then the spatial and time dependences of $\Psi(x,t)$ can be separated, just as in Equation 3.14, and the **total wavefunction** $\Psi(x,t)$ of the electron can be written as

Steady-state total wave function

$$\Psi(x,t) = \psi(x)\exp\left(-\frac{jEt}{\hbar}\right) \qquad [3.15]$$

where $\psi(x)$ is the electron wavefunction that describes only the spatial behavior, and E is the energy of the electron. The temporal behavior is simply harmonic, by virtue of $\exp(-jEt/\hbar)$, which corresponds to $\exp(-j\omega t)$ with an angular frequency $\omega = E/\hbar$. The fundamental equation that describes the electron's behavior by determining $\psi(x)$ is called the **time-independent Schrödinger equation.** It is given by the famous equation

Schrödinger's equation for one dimension

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}(E-V)\psi = 0 \qquad [3.16a]$$

where m is the mass of the electron.

This is a second-order differential equation. It should be reemphasized that the potential energy V in Equation 3.16a depends only on x. If the potential energy of the electron depends on time as well, that is, if $V = V(x,t)$, then in general $\Psi(x,t)$ cannot be written as $\psi(x)\exp(-jEt/\hbar)$. Instead, we must use the full version of the Schrödinger equation, which is discussed in more advanced textbooks.

In three dimensions, there will be derivatives of ψ with respect to x, y, and z. We use the calculus notation $(\partial\psi/\partial x)$, differentiating $\psi(x,y,z)$ with respect to x but keeping y and z constant. Similar notations $\partial\psi/\partial y$ and $\partial\psi/\partial z$ are used for derivatives with respect to y alone and with respect to z alone, respectively. In three dimensions, Equation 3.16a becomes

Schrödinger's equation for three dimensions

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} + \frac{2m}{\hbar^2}(E-V)\psi = 0 \qquad [3.16b]$$

where $V = V(x,y,z)$ and $\psi = \psi(x,y,z)$.

Equation 3.16b is a fundamental equation, called the time-independent Schrödinger equation, the solution of which gives the steady-state behavior of the electron in a time-independent potential energy environment described by $V = V(x, y, z)$. By solving Equation 3.16b, we will know the probability distribution and the energy of the electron. Once $\psi(x, y, z)$ has been determined, the total wavefunction for the electron is given by Equation 3.15 so that

$$|\Psi(x, y, z, t)|^2 = |\psi(x, y, z)|^2$$

which means that the steady-state probability distribution of the electron is simply $|\psi(x, y, z)|^2$.

The time-independent Schrödinger equation can be viewed as a "mathematical crank." We input the potential energy of the electron and the boundary conditions, turn the crank, and get the probability distribution and the energy of the electron under steady-state conditions.

Two important boundary conditions are often used to solve the Schrödinger equation. First, as an analogy, when we stretch a string between two fixed points and put it into a steady-state vibration, there are no discontinuities or kinks along the string. We can therefore intelligently guess that because $\psi(x)$ represents wave-like behavior, it must be a smooth function without any discontinuities.

The first boundary condition is that Ψ must be continuous, and the second is that $d\Psi/dx$ must be continuous. In the steady state, these two conditions translate directly to ψ and $d\psi/dx$ being continuous. Since the probability of finding the electron is represented by $|\psi|^2$, this function must be single-valued and smooth, without any discontinuities, as illustrated in Figure 3.14. The enforcement of these boundary conditions results in strict requirements on the wavefunction $\psi(x)$, as a result of which only certain wavefunctions are acceptable. These wavefunctions are called the **eigenfunctions** (characteristic functions) of the system, and they determine the behavior and energy of the electron under steady-state conditions. The eigenfunctions $\psi(x)$ are also called **stationary states,** inasmuch as we are only considering steady-state behavior.

It is important to note that the Schrödinger equation is generally applicable to all matter, not just the electron. For example, the equation can also be used to describe the behavior of a proton, if the appropriate potential energy $V(x, y, z)$ and mass (m_{proton}) are used. Wavefunctions associated with particles are frequently called **matter waves.**

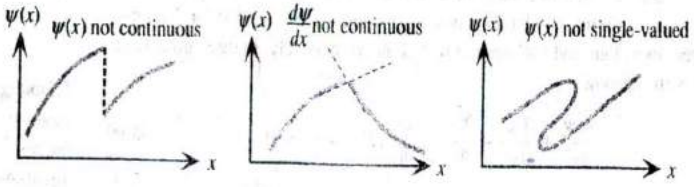


Figure 3.14 Unacceptable forms of $\psi(x)$.

EXAMPLE 3.5

THE FREE ELECTRON Solve the Schrödinger equation for a free electron whose energy is E. What is the uncertainty in the position of the electron and the uncertainty in the momentum of the electron?

SOLUTION

Since the electron is free, its potential energy is zero, $V = 0$. In the Schrödinger equation, this leads to

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}E\psi = 0$$

We can write this as

$$\frac{d^2\psi}{dx^2} + k^2\psi = 0$$

where we defined $k^2 = (2m/\hbar^2)E$. Solving the differential equation, we get

$$\psi(x) = A\exp(jkx) \quad \text{or} \quad B\exp(-jkx)$$

The total wavefunction is obtained by multiplying $\psi(x)$ by $\exp(-jEt/\hbar)$. We can define a fictitious frequency for the electron by $\omega = E/\hbar$ and multiply $\psi(x)$ by $\exp(-j\omega t)$:

$$\Psi(x,t) = A\exp j(kx - \omega t) \quad \text{or} \quad B\exp j(-kx - \omega t)$$

Each of these is a traveling wave. The first solution is a traveling wave in the $+x$ direction, and the second one is in the $-x$ direction. Thus, the free electron has a traveling wave solution with a wavenumber $k = 2\pi/\lambda$, that can have any value. The energy E of the electron is simply KE, so

$$KE = E = \frac{(\hbar k)^2}{2m}$$

When we compare this with the classical physics expression $KE = (p^2/2m)$, we see that the momentum is given by

$$p = \hbar k \quad \text{or} \quad p = \frac{h}{\lambda}$$

This is the de Broglie relationship. The latter therefore results naturally from the Schrödinger equation for a free electron.

The probability distribution for the electron is

$$|\psi(x)|^2 = |A\exp j(kx)|^2 = A^2$$

which is constant over the entire space. Thus, the electron can be anywhere between $x = -\infty$ and $x = +\infty$. The uncertainty Δx in its position is infinite. Since the electron has a well-defined wavenumber k, its momentum p is also well-defined by virtue of $p = \hbar k$. The uncertainty Δp in its momentum is thus zero.

EXAMPLE 3.6

WAVELENGTH OF AN ELECTRON BEAM Electrons are accelerated through a 100 V potential difference to strike a polycrystalline aluminum sample. The diffraction pattern obtained indicates that the highest intensity and smallest angle diffraction, corresponding to diffraction from the (111) planes, has a diffraction angle of 30.4°. From X-ray studies, the separation of the (111)

planes is 0.234 nm. What is the wavelength of the electron and how does it compare with that from the de Broglie relationship?

SOLUTION

Since we know the angle of diffraction $2\theta (= 30.4^\circ)$ and the interplanar separation $d (= 0.234$ nm), we can readily calculate the wavelength of the electron from the Bragg condition for diffraction, $2d \sin\theta = n\lambda$. With $n = 1$,

$$\lambda = 2d \sin\theta = 2(0.234 \text{ nm}) \sin(15.2^\circ) = 0.1227 \text{ nm}$$

This is the wavelength of the electron.

When an electron is accelerated through a voltage V, it gains KE equal to eV, so $p^2/2m = eV$ and $p = (2meV)^{1/2}$. This is the momentum imparted by the potential difference V. From the de Broglie relationship, the wavelength should be

$$\lambda = \frac{h}{p} = \frac{h}{(2meV)^{1/2}}$$

or

$$\lambda = \left(\frac{h^2}{2meV}\right)^{1/2}$$

Substituting for e, h, and m, we obtain

$$\lambda = \frac{1.226 \text{ nm}}{V^{1/2}}$$

The experiment uses 100 V, so the de Broglie wavelength is

$$\lambda = \frac{1.226 \text{ nm}}{V^{1/2}} = \frac{1.226 \text{ nm}}{100^{1/2}} = 0.1226 \text{ nm}$$

which is in excellent agreement with that determined from the Bragg condition.

3.3 INFINITE POTENTIAL WELL: A CONFINED ELECTRON

Consider the behavior of the electron when it is confined to a certain region, $0 < x < a$. Its PE is zero inside that region and infinite outside, as shown in Figure 3.15. The electron cannot escape, because it would need an infinite PE. Clearly the probability $|\psi|^2$ of finding the electron per unit volume is zero outside $0 < x < a$. Thus, $\psi = 0$ when $x \leq 0$ and $x \geq a$, and ψ is determined by the Schrödinger equation in $0 < x < a$ with $V = 0$. Therefore, in the region $0 < x < a$

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2} E\psi = 0 \qquad [3.17]$$

This is a second-order linear differential equation. As a general solution, we can take

$$\psi(x) = A \exp(jkx) + B \exp(-jkx)$$

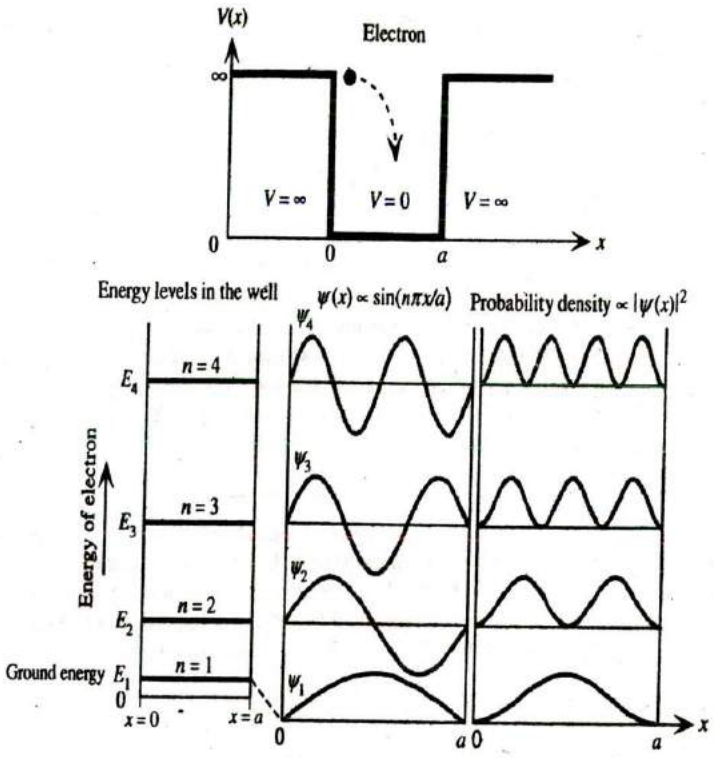


Figure 3.15 Electron in a one-dimensional infinite PE well.
The energy of the electron is quantized. Possible wavefunctions and the probability distributions for the electron are shown.

where k is some constant (to be determined) and substitute this in Equation 3.17 to find k. We first note that $\psi(0) = 0$; therefore, $B = -A$, so that

$$\psi(x) = A[\exp(jkx) - \exp(-jkx)] = 2Aj \sin kx \qquad [3.18]$$

We now substitute this into the Schrödinger Equation 3.17 to relate the energy E to k. Thus, Equation 3.17 becomes

$$-2Ajk^2(\sin kx) + \left(\frac{2m}{\hbar^2}\right)E(2Aj \sin kx) = 0$$

which can be rearranged to obtain the energy of the electron:

$$E = \frac{\hbar^2 k^2}{2m} \qquad [3.19]$$

Since the electron has no *PE* within the well, its total energy E is kinetic energy *KE*, and we can write

$$E = KE = \frac{p_x^2}{2m}$$

where p_x is its momentum. Comparing this with Equation 3.19, we see that the momentum of the electron must be

$$p_x = \pm\hbar k \qquad [3.20]$$

The momentum p_x may be in the $+x$ direction or the $-x$ direction (which is the reason for $\pm$), so the **average momentum** is actually zero, $p_{av} = 0$.

We have already seen this relationship, when we defined k as $2\pi/\lambda$ (wavenumber) for a free traveling wave. So the constant k here is a wavenumber-type quantity even though there is no distinct traveling wave. Its value is determined by the boundary condition at $x = a$ where $\psi = 0$, or

$$\psi(a) = 2Aj \sin ka = 0$$

The solution to $\sin ka = 0$ is simply $ka = n\pi$, where $n = 1, 2, 3, \ldots$ is an integer. We exclude $n = 0$ because it will result in $\psi = 0$ everywhere (no electron at all).

We notice immediately that k, and therefore the energy of the electron, can only have certain values; they are **quantized** by virtue of n being an integer. Here, n is called a **quantum number.** For each n, there is a special wavefunction

Wavefunction in infinite PE well

$$\psi_n(x) = 2Aj \sin\left(\frac{n\pi x}{a}\right) \qquad [3.21]$$

which is called an eigenfunction.[3] All ψ_n for $n = 1, 2, 3 \ldots$ constitute the eigenfunctions of the system. Each eigenfunction identifies a possible state for the electron. For each n, there is one special k value, $k_n = n\pi/a$, and hence a special energy value E_n, since

$$E_n = \frac{\hbar^2 k_n^2}{2m}$$

that is,

Electron energy in infinite PE well

$$E_n = \frac{\hbar^2(\pi n)^2}{2ma^2} = \frac{h^2 n^2}{8ma^2} \qquad [3.22]$$

The energies E_n defined by Equation 3.22 with $n = 1, 2, 3 \ldots$ are called eigenenergies of the system.

We still have not completely solved the problem, because A has yet to be determined. To find A, we use what is called **the normalization condition.** The total probability of finding the electron in the whole region $0 < x < a$ is unity, because we know the electron is somewhere in this region. Thus, $|\psi|^2\,dx$ summed between $x = 0$ and

[3] From the German meaning "characteristic function."

$x = a$ must be unity, or

$$\int_{x=0}^{x=a} |\psi(x)|^2 \, dx = \int_{x=0}^{x=a} \left| 2Aj \sin\left(\frac{n\pi x}{a}\right) \right|^2 dx = 1$$

Normalization condition

Carrying out the simple integration, we find

$$A = \left(\frac{1}{2a}\right)^{1/2}$$

The resulting wavefunction for the electron is thus

$$\psi_n(x) = j\left(\frac{2}{a}\right)^{1/2} \sin\left(\frac{n\pi x}{a}\right) \qquad [3.23]$$

We can now summarize the behavior of an electron in a one-dimensional *PE* well. Its wavefunction and energy, shown in Figure 3.15, are given by Equations 3.23 and 3.22, respectively. Both depend on the quantum number n. The energy of the electron increases with n^2, so the minimum energy of the electron corresponds to $n = 1$. This is called the **ground state,** and the energy of the ground state is the lowest energy the electron can possess. Note also that the energy of the electron in this potential well cannot be zero, even though the *PE* is zero. Thus, the electron always has *KE*, even when it is in the ground state.

The **node** of a wavefunction is defined as the point where $\psi = 0$ inside the well. It is apparent from Figure 3.15 that the ground wavefunction ψ_1 with the lowest energy has no nodes, ψ_2 has one node, ψ_3 has two nodes, and so on. Thus, the energy increases as the number of nodes increases in a wavefunction.

It may seem surprising that the energy of the electron is quantized; that is, that it can only have finite values, given by Equation 3.22. The electron cannot be made to take on any value of energy, as in the classical case. If the electron behaved like a particle, then an applied force F could impart any value of energy to it, because $F = dp/dt$ (Newton's second law), or $p = \int F\,dt$. By applying a force F for a time t, we can give the electron a *KE* of

$$E = \frac{p^2}{2m} = \left(\frac{1}{2}m\right)\left[\int F\,dt\right]^2$$

However, Equation 3.22 tells us that, in the microscopic world, the energy can only have quantized values. The two conflicting views can be reconciled if we consider the energy difference between two consecutive energy levels, as follows:

$$\Delta E = E_{n+1} - E_n = \frac{h^2(2n+1)}{8ma^2}$$

Energy separation in infinite PE well

As a increases to macroscopic dimensions, $a \to \infty$, the electron is completely free and $\Delta E \to 0$. Since $\Delta E = 0$, the energy of a completely free electron ($a = \infty$) is continuous. The energy of a confined electron, however, is quantized, and ΔE depends on the dimension (or size) of the potential well confining the electron.

In general, an electron will be "contained" in a spatial region of three dimensions, within which the *PE* will be lower (hence the confinement). We must then solve the

Schrödinger equation in three dimensions. The result is three quantum numbers that characterize the behavior of the electron.

Examination of the wavefunctions ψ_n in Figure 3.15 shows that these are either *symmetric* or *antisymmetric* with respect to the center of the well at $x = \frac{1}{2}a$. The symmetry of a wavefunction is called its **parity.** Whenever the potential energy function $V(x)$ exhibits symmetry about a certain point C, for example, about $x = \frac{1}{2}a$ in Figure 3.15, then the wavefunctions have either **even parity** (such as $\psi_1, \psi_3, \ldots$ that are symmetric) or have odd parity (such as $\psi_2, \psi_4, \ldots$ that are antisymmetric).

EXAMPLE 3.7 **ELECTRON CONFINED WITHIN ATOMIC DIMENSIONS** Consider an electron in an infinite potential well of size 0.1 nm (typical size of an atom). What is the ground energy of the electron? What is the energy required to put the electron at the third energy level? How can this energy be provided?

SOLUTION

The electron is confined in an infinite potential well, so its energy is given by

$$E_n = \frac{h^2 n^2}{8ma^2}$$

We use $n = 1$ for the ground level and $a = 0.1$ nm. Therefore,

$$E_1 = \frac{(6.6 \times 10^{-34}\ \text{J s})^2(1)^2}{8(9.1 \times 10^{-31}\ \text{kg})(0.1 \times 10^{-9}\ \text{m})^2} = 6.025 \times 10^{-18}\ \text{J} \quad \text{or} \quad 37.6\ \text{eV}$$

The frequency of the electron associated with this energy is

$$\omega = \frac{E}{\hbar} = \frac{6.025 \times 10^{-18}\ \text{J}}{1.055 \times 10^{-34}\ \text{J s}} = 5.71 \times 10^{16}\ \text{rad s}^{-1} \quad \text{or} \quad \nu = 9.092 \times 10^{15}\ \text{s}^{-1}$$

The third energy level E_3 is

$$E_3 = E_1 n^2 = (37.6\ \text{eV})(3)^2 = 338.4\ \text{eV}$$

The energy required to take the electron from 37.6 eV to 338.4 eV is 300.8 eV. This can be provided by a photon of exactly that energy; no less, and no more. Since the photon energy is $E = h\nu = hc/\lambda$, or

$$\lambda = \frac{hc}{E} = \frac{(6.6 \times 10^{-34}\ \text{J s})(3 \times 10^{8}\ \text{m s}^{-1})}{300.8\ \text{eV} \times 1.6 \times 10^{-19}\ \text{C}}$$

$$= 4.12\ \text{nm}$$

which is an X-ray photon.

EXAMPLE 3.8 **ENERGY OF AN APPLE IN A CRATE** Consider a macroscopic object of mass 100 grams (say, an apple) confined to move between two rigid walls separated by 1 m (say, a typical size of a large apple crate). What is the minimum speed of the object? What should the quantum number n be if the object is moving with a speed 1 m s^{-1}? What is the separation of the energy levels of the object moving with that speed?

SOLUTION

Since the object is within rigid walls, we take the *PE* outside the walls as infinite and use

$$E_n = \frac{h^2 n^2}{8ma^2}$$

to find the ground-level energy. With $n = 1, a = 1$ m, $m = 0.1$ kg, we have

$$E_1 = \frac{(6.6 \times 10^{-34}\ \text{J s})^2(1)^2}{8(0.1\ \text{kg})(1\ \text{m})^2} = 5.45 \times 10^{-67}\ \text{J} = 3.4 \times 10^{-48}\ \text{eV}$$

Since this is kinetic energy, $\frac{1}{2}mv_1^2 = E_1$, so the minimum speed is

$$v_1 = \sqrt{\frac{2E_1}{m}} = \sqrt{\frac{2(5.45 \times 10^{-67}\ \text{J})}{0.1\ \text{kg}}} = 3.3 \times 10^{-33}\ \text{m s}^{-1}$$

This speed cannot be measured by any instrument; therefore, for all practical purposes, the apple is at rest in the crate (a relief for the fruit grocer). The time required for the object to move a distance of 1 mm is 3×10^{29} s or 10^{21} years, which is more than the present age of the universe!

When the object is moving with a speed 1 m s^{-1},

$$KE = \frac{1}{2}mv^2 = \frac{1}{2}(0.1\ \text{kg})(1\ \text{m s}^{-1})^2 = 0.05\ \text{J}$$

This must be equal to $E_n = h^2n^2/8ma^2$ for some value of n

$$n = \left(\frac{8ma^2E_n}{h^2}\right)^{1/2} = \left[\frac{8(0.1\ \text{kg})(1\ \text{m})^2(0.05\ \text{J})}{(6.6 \times 10^{-34}\ \text{J s})^2}\right]^{1/2} = 3.03 \times 10^{32}$$

which is an enormous number. The separation between two energy levels corresponds to a change in n from 3.03×10^{32} to $3.03 \times 10^{32} + 1$. This is such a negligibly small change in n that for all practical purposes, the energy levels form a continuum. Thus,

$$\Delta E = E_{n+1} - E_n = \frac{h^2(2n+1)}{8ma^2}$$

$$= \frac{[(6.6 \times 10^{-34}\ \text{J s})^2(2 \times 3.03 \times 10^{32} + 1)]}{[8(0.1\ \text{kg})(1\ \text{m})^2]}$$

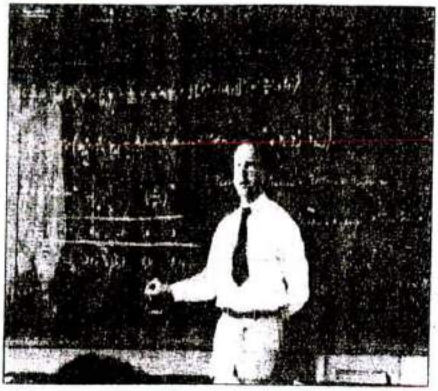
$$= 3.30 \times 10^{-34}\ \text{J} \quad \text{or} \quad 2.06 \times 10^{-15}\ \text{eV}$$

This energy separation is not detectable by any instrument. So for all practical purposes, the energy of the object changes continuously.

We see from this example that in the limit of large quantum numbers, quantum predictions agree with the classical results. This is the essence of **Bohr's correspondence principle.**

3.4 HEISENBERG'S UNCERTAINTY PRINCIPLE

The wavefunction of a free electron corresponds to a traveling wave with a single wavelength λ, as shown in Example 3.5. The traveling wave extends over all space, along all x, with the same amplitude, so the probability distribution function is uniform



Werner Heisenberg (1901–1976) received the Nobel prize in physics in 1932 for the uncertainty principle. This photo was apparently taken in 1936, while he was lecturing on quantum mechanics. "An expert is someone who knows some of the worst mistakes that can be made in his subject, and how to avoid them." W. Heisenberg.

SOURCE: AIP Emilio Segrè Visual Archives.

throughout the whole of space. The uncertainty Δx in the position of the electron is therefore infinite. Yet, the uncertainty Δp_x in the momentum of the electron is zero, because λ is well-defined, which means that we know p_x exactly from the de Broglie relationship, $p_x = h/\lambda$.

For an electron trapped in a one-dimensional infinite *PE* well, the wavefunction extends from $x = 0$ to $x = a$, so the uncertainty in the position of the electron is a. We know that the electron is within the well, but we cannot pinpoint with certainty exactly where it is. The momentum of the electron is either $p_x = \hbar k$ in the $+x$ direction or $-\hbar k$ in the $-x$ direction. The uncertainty Δp_x in the momentum is therefore $2\hbar k$; that is, $\Delta p_x = 2\hbar k$. For the ground-state wavefunction, which corresponds to $n = 1$, we have $ka = \pi$. Thus, $\Delta p_x = 2\hbar\pi/a$. Taking the product of the uncertainties in x and p, we get

$$(\Delta x)(\Delta p_x) = (a)\left(\frac{2\hbar\pi}{a}\right) = h$$

In other words, the product of the position and momentum uncertainties is simply h. This relationship is fundamental; and it constitutes a limit to our knowledge of the behavior of a system. *We cannot exactly and simultaneously know both the position and momentum of a particle along a given coordinate.* In general, if Δx and Δp_x are the respective uncertainties in the simultaneous measurement of the position and momentum of a particle along a particular coordinate (such as x), the **Heisenberg uncertainty principle** states that[4]

Heisenberg uncertainty principle for position and momentum

$$\Delta x\, \Delta p_x \gtrsim \hbar \qquad [3.24]$$

We are therefore forced to conclude that as previously stated, because of the wave nature of quantum mechanics, we are unable to determine exactly and simultaneously the position and momentum of a particle along a given coordinate. There will be an uncertainty Δx in the position and an uncertainty Δp_x in the momentum of the particle

[4] The Heisenberg uncertainty principle is normally written in terms of $\hbar$ rather than h. Further, in some physics texts, $\hbar$ in Equation 3.24 has a factor $\frac{1}{2}$ multiplying it.

and these uncertainties will be related by Heisenberg's uncertainty relationship in Equation 3.24.

These uncertainties are not in any way a consequence of the accuracy of a measurement or the precision of an instrument. Rather, they are the theoretical limits to what we can determine about a system. They are part of the quantum nature of the universe. In other words, even if we build the most perfectly engineered instrument to measure the position and momentum of a particle at one instant, we will still be faced with position and momentum uncertainties Δx and Δp_x such that $\Delta x\, \Delta p_x > \hbar$.

There is a similar uncertainty relationship between the uncertainty ΔE in the energy E (or angular frequency ω) of the particle and the time duration Δt during which it possesses the energy (or during which its energy is measured). We know that the kx part of the wave leads to the uncertainty relation $\Delta x\, \Delta p_x > \hbar$ or $\Delta x\, \Delta k \geq 1$. By analogy we should expect a similar relationship for the ωt part, or $\Delta\omega\, \Delta t \geq 1$. This hypothesis is true, and since $E = \hbar\omega$, we have the uncertainty relation for the particle energy and time:

Heisenberg uncertainty principle for energy and time

$$\Delta E\, \Delta t \gtrsim \hbar \qquad [3.25]$$

Note that the uncertainty relationships in Equations 3.24 and 3.25 have been written in terms of $\hbar$, rather than h, as implied by the electron in an infinite potential energy well ($\Delta x\, \Delta p_x \geq h$). In general there is also a numerical factor of $\frac{1}{2}$ multiplying $\hbar$ in Equations 3.24 and 3.25 which comes about when we consider a Gaussian spread for all possible position and momentum values. The proof is not presented here, but can be found in advanced quantum mechanics books.

It is important to note that the uncertainty relationship applies only when the position and momentum are measured in the same direction (such as the x direction). On the other hand, the exact momentum, along, say, the y direction and the exact position, along, say, the x direction can be determined exactly, since $\Delta x\, \Delta p_y$ need not satisfy the Heisenberg uncertainty relationship (in other words, $\Delta x\, \Delta p_y$ can be zero).

EXAMPLE 3.9

THE MEASUREMENT TIME AND THE FREQUENCY OF WAVES: AN ANALOGY WITH $\Delta E\, \Delta t \geq \hbar$

Consider the measurement of the frequency of a sinusoidal wave of frequency 1000 Hz (or cycles/s). Suppose we can only measure the number of cycles to an accuracy of 1 cycle, because we need to receive a whole cycle to record it as one complete cycle. Then, in a time interval of $\Delta t = 1$ s, we will register 1000 ± 1 cycles. The uncertainty Δf in the frequency is 1 cycle/1 s or 1 Hz. If Δt is 2 s, we will measure 2000 ± 1 cycles, and the uncertainty Δf will be 1 cycle/2 s or $\frac{1}{2}$ cycle/s or $\frac{1}{2}$ Hz. Thus, Δf decreases with Δt.

Suppose that in a time interval Δt, we measure $N \pm 1$ cycles. Since the uncertainty is 1 cycle in a time interval Δt, the uncertainty in f will be

$$\Delta f = \frac{(1 \text{ cycle})}{\Delta t} = \frac{1}{\Delta t} \text{ Hz}$$

Since $\omega = 2\pi f$, we have

$$\Delta\omega\, \Delta t = 2\pi$$

In quantum mechanics, under steady-state conditions, an object has a time-oscillating wavefunction with a frequency ω which is related to its energy E by $\omega = E/\hbar$ (see Equation 3.15).

Substituting this into the previous relationship gives

$$\Delta E\, \Delta t = h$$

The uncertainty in the energy of a quantum object is therefore related, in a fundamental way, to the time duration during which the energy is observed. Notice that we again have h, as for $\Delta x\, \Delta p_x = h$, though the quantum mechanical uncertainty relationship in Equation 3.25 has $\hbar$.

EXAMPLE 3.10 **THE UNCERTAINTY PRINCIPLE ON THE ATOMIC SCALE** Consider an electron confined to a region of size 0.1 nm, which is the typical dimension of an atom. What will be the uncertainty in its momentum and hence its kinetic energy?

SOLUTION

We apply the Heisenberg uncertainty relationship, $\Delta x\, \Delta p_x \approx \hbar$, or

$$\Delta p_x \approx \frac{\hbar}{\Delta x} = \frac{1.055 \times 10^{-34}\,\text{J s}}{0.1 \times 10^{-9}\,\text{m}} = 1.055 \times 10^{-24}\,\text{kg m s}^{-1}$$

The uncertainty in the velocity is therefore

$$\Delta v = \frac{\Delta p_x}{m_e} = \frac{1.055 \times 10^{-24}\,\text{kg m s}^{-1}}{9.1 \times 10^{-31}\,\text{kg}} = 1.16 \times 10^{6}\,\text{m s}^{-1}$$

We can take this uncertainty to represent the order of magnitude of the actual speed. The kinetic energy associated with this momentum is

$$KE = \frac{\Delta p_x^2}{2m_e} = \frac{(1.055 \times 10^{-24}\,\text{kg m s}^{-1})^2}{2(9.1 \times 10^{-31}\,\text{kg})}$$

$$= 6.11 \times 10^{-19}\,\text{J} \quad \text{or} \quad 3.82\,\text{eV}$$

EXAMPLE 3.11 **THE UNCERTAINTY PRINCIPLE WITH MACROSCOPIC OBJECTS** Estimate the minimum velocity of an apple of mass 100 g confined to a crate of size 1 m.

SOLUTION

Taking the uncertainty in the position of the apple as 1 m, the apple is somewhere in the crate,

$$\Delta p_x \approx \frac{\hbar}{\Delta x} = \frac{1.05 \times 10^{-34}\,\text{J s}}{1\,\text{m}} = 1.05 \times 10^{-34}\,\text{kg m s}^{-1}$$

So the minimum uncertainty in the velocity is

$$\Delta v_x = \frac{\Delta p_x}{m} = \frac{1.05 \times 10^{-34}\,\text{kg m s}^{-1}}{0.1\,\text{kg}} = 1.05 \times 10^{-33}\,\text{m s}^{-1}$$

The quantum nature of the universe implies that the apple in the crate is moving with a velocity on the order of 10^{-33} m s^{-1}. This cannot be measured by any instrument; indeed, it would take the apple $\sim 10^{19}$ years to move an atomic distance of 0.1 nm.

3.5 TUNNELING PHENOMENON: QUANTUM LEAK

To understand the tunneling phenomenon, let us examine the thrilling events experienced by the roller coaster shown in Figure 3.16a. Consider what the roller coaster can do when released from rest at a height A. The conservation of energy means that the carriage can reach B and at most C, but certainly not beyond C and definitely not D and E. Classically, there is no possible way the carriage will reach E at the other side of the potential barrier D. An extra energy corresponding to the height difference, $D - A$, is needed. Anyone standing at E will be quite safe. Ignoring frictional losses, the roller coaster will go back and forth between A and C.

Now, consider an analogous event on an atomic scale. An electron moves with an energy E in a region $x < 0$ where the potential energy PE is zero; therefore, E is solely kinetic energy. The electron then encounters a potential barrier of "height" V_o, which

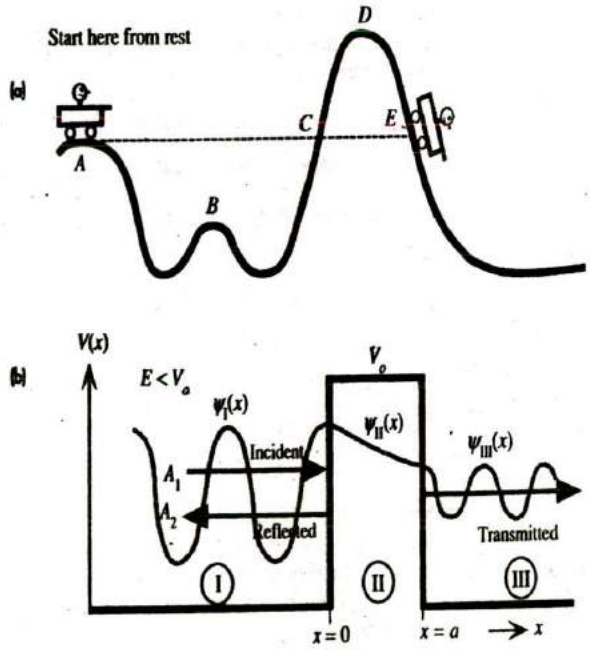


Figure 3.16

(a) The roller coaster released from A can at most make it to C, but not to E. Its PE at A is less than the PE at D. When the car is at the bottom, its energy is totally KE. CD is the energy barrier that prevents the car from making it to E. In quantum theory, on the other hand, there is a chance that the car could tunnel (leak) through the potential energy barrier between C and E and emerge on the other side of the hill at E.

(b) The wavefunction for the electron incident on a potential energy barrier (V_o). The incident and reflected waves interfere to give $\psi_I(x)$. There is no reflected wave in region III. In region II, the wavefunction decays with x because $E < V_o$.

is greater than E at $x = 0$. The extent (width) of the potential barrier is a. On the other side of the potential barrier, $x > a$, the PE is again zero. What will the electron do? Classically, just like the roller coaster, the electron should bounce back and thus be confined to the region $x < 0$, because its total energy E is less than V_o. In the quantum world, however, there is a distinct possibility that the electron will "tunnel" through the potential barrier and appear on the other side; it will leak through.

To show this, we need to solve the Schrödinger equation for the present choice of $V(x)$. Remember that the only way the Schrödinger equation will have the solution $\psi(x) = 0$ is if the PE is infinite, that is, $V = \infty$. Therefore, within any zero or finite PE region, there will always be a solution $\psi(x)$ and there always will be some probability of finding the electron.

We can divide the electron's space into three regions, I, II, and III, as indicated in Figure 3.16b. We can then solve the Schrödinger equation for each region, to obtain three wavefunctions $\psi_I(x)$, $\psi_{II}(x)$, and $\psi_{III}(x)$. In regions I and III, $\psi(x)$ must be traveling waves, as there is no PE (the electron is free and moving with a kinetic energy E). In zone II, however, $E - V_o$ is negative, so the general solution of the Schrödinger equation is the sum of an exponentially decaying function and an exponentially increasing function. In other words,

$$\psi_I(x) = A_1 \exp(jkx) + A_2 \exp(-jkx) \qquad [3.26a]$$

$$\psi_{II}(x) = B_1 \exp(\alpha x) + B_2 \exp(-\alpha x) \qquad [3.26b]$$

$$\psi_{III}(x) = C_1 \exp(jkx) + C_2 \exp(-jkx) \qquad [3.26c]$$

are the wavefunctions in which

$$k^2 = \frac{2mE}{\hbar^2} \qquad [3.27]$$

and

$$\alpha^2 = \frac{2m(V_o - E)}{\hbar^2} \qquad [3.28]$$

Both k^2 and α^2, and hence k and α, in Equations 3.26a to c are positive numbers. This means that $\exp(jkx)$ and $\exp(-jkx)$ represent traveling waves in opposite directions, and $\exp(-\alpha x)$ and $\exp(\alpha x)$ represent an exponential decay and rise, respectively. We see that in region I, $\psi_I(x)$ consists of the incident wave $A_1 \exp(jkx)$ in the $+x$ direction, and a reflected wave $A_2 \exp(-jkx)$, in the $-x$ direction. Furthermore, because the electron is traveling toward the right in region III, there is no reflected wave, so $C_2 = 0$.

We must now apply the boundary conditions and the normalization condition to determine the various constants A_1, A_2, B_1, B_2, and C_1. In other words, we must match the three waveforms in Equations 3.26a to c at their boundaries ($x = 0$ and $x = a$) so that they form a continuous single-valued wavefunction. With the boundary conditions enforced onto the wavefunctions $\psi_I(x)$, $\psi_{II}(x)$, and $\psi_{III}(x)$, all the constants can be determined in terms of the amplitude A_1 of the incoming wave. The relative probability that the electron will tunnel from region I through to III is defined as the **transmission**

coefficient T, and this depends very strongly on both the relative PE barrier height $(V_o - E)$ and the width a of the barrier. The final result that comes out from a tedious application of the boundary conditions is

$$T = \frac{|\psi_{III}(x)|^2}{|\psi_I(\text{incident})|^2} = \frac{C_1^2}{A_1^2} = \frac{1}{1 + D\sinh^2(\alpha a)} \quad [3.29]$$

Probability of tunneling

where

$$D = \frac{V_o^2}{4E(V_o - E)} \quad [3.30]$$

and α is the rate of decay of $\psi_{II}(x)$ as expressed in Equation 3.28. For a wide or high barrier, using $\alpha a \gg 1$ in Equation 3.29 and $\sinh(\alpha a) \approx \frac{1}{2}\exp(\alpha a)$, we can deduce

$$T = T_o \exp(-2\alpha a) \quad [3.31]$$

Probability of tunneling through

where

$$T_o = \frac{16E(V_o - E)}{V_o^2} \quad [3.32]$$

By contrast, the relative probability of reflection is determined by the ratio of the square of the amplitude of the reflected wave to that of the incident wave. This quantity is the **reflection coefficient R,** which is given by

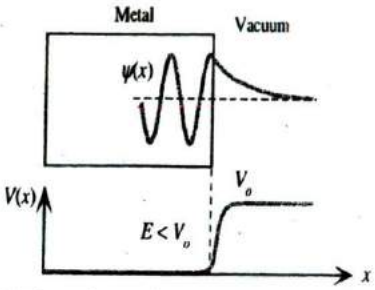
$$R = \frac{A_2^2}{A_1^2} = 1 - T \quad [3.33]$$

Reflection coefficient

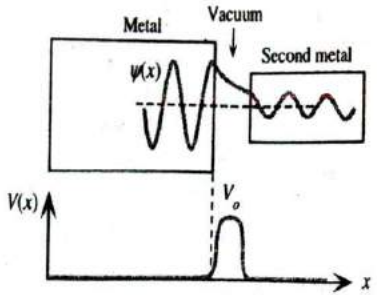
We can now summarize the entire tunneling affair as follows. When an electron encounters a potential energy barrier of height V_o greater than its energy E, there is a finite probability that it will leak through that barrier. This probability depends sensitively on the energy and width of the barrier. For a wide potential barrier, the probability of tunneling is proportional to $\exp(-2\alpha a)$, as in Equation 3.31. The wider or higher the potential barrier, the smaller the chance of the electron tunneling.

One of the most remarkable technological uses of the tunneling effect is in the scanning tunneling microscope (STM), which elegantly maps out the surfaces of solids. A conducting probe is brought so close to the surface of a solid that electrons can tunnel from the surface of the solid to the probe, as illustrated in Figure 3.17. When the probe is far removed, the wavefunction of an electron decays exponentially outside the material, by virtue of the potential energy barrier being finite (the work function is ~10 eV). When the probe is brought very close to the surface, the wavefunction penetrates into the probe and, as a result, the electron can tunnel from the material into the probe. Without an applied voltage, there will be as many electrons tunneling from the material to the probe as there are going in the opposite direction from the probe to the material, so the net current will be zero.

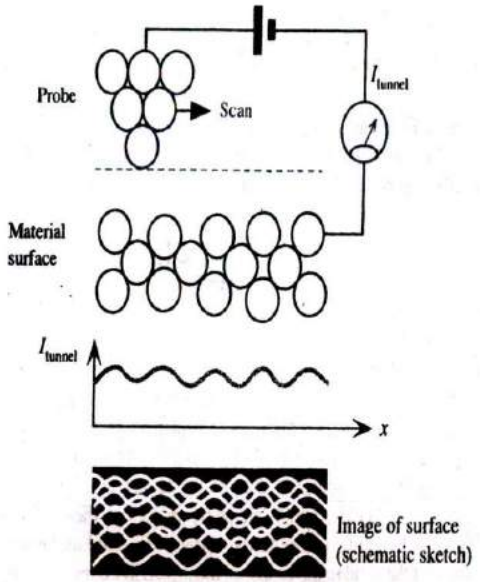
On the other hand, if a positive bias is applied to the probe with respect to the material, as shown in Figure 3.17, an electron tunneling from the material to the probe will see a lower potential barrier than one tunneling from the probe to the material. Consequently, there will be a net current from the probe to the material and this current



(a) The wavefunction decays exponentially as we move away from the surface because the PE outside the metal is V_o and the energy of the electron, $E < V_o$.



(b) If we bring a second metal close to the first metal, then the wavefunction can penetrate into the second metal. The electron can tunnel from the first metal to the second.



(c) The principle of the scanning tunneling microscope. The tunneling current depends on $\exp(-2\alpha a)$ where a is the distance of the probe from the surface of the specimen and α is a constant.

Figure 3.17

will depend very sensitively on the separation a of the probe from the surface, by virtue of Equation 3.31.

Because the tunneling current is extremely sensitive to the width of the potential barrier, the tunneling current is essentially dominated by electrons tunneling to the probe atom nearest to the surface. Thus, the probe tip has an atomic dimension. By scanning the surface of the material with the probe and recording the tunneling current

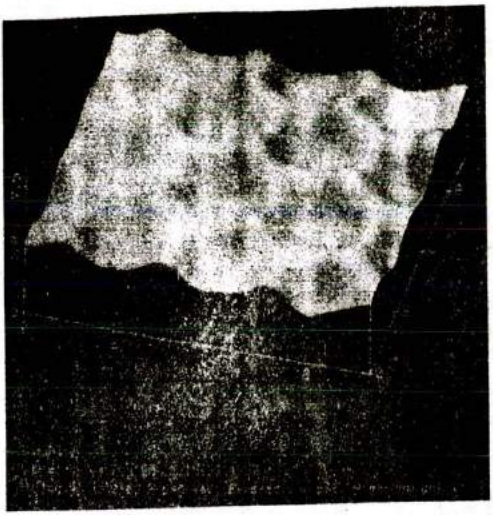


Figure 3.18 Scanning tunneling microscope (STM) image of a graphite surface where contours represent electron concentrations within the surface, and carbon rings are clearly visible. The scale is in 2 Å.
| SOURCE: Courtesy of Veeco Instruments, Metrology Division, Santa Barbara, CA.

the user can map out the surface topology of the material with a resolution comparable to the atomic dimension. The probe motion along the surface, and also perpendicular to the surface, is controlled by piezoelectric transducers to provide sufficiently small and smooth displacements. Figure 3.18 shows an STM image of a graphite surface, on which the hexagonal carbon rings can be clearly seen. Notice that the scale is 0.2 nm (2 Å). The contours in the image actually represent electron concentrations within the surface since it is the electrons that tunnel from the graphite surface to the probe tip. The astute reader will notice that not all the carbon atoms in a hexagonal ring are at the same height; three are higher and three are lower. The reason is that the exact electron concentration on the surface is also influenced by the second layer of atoms underneath the top layer. The overall effect makes the electron concentration



STM's inventors Gerd Binning (right) and Heinrich Rohrer (left), at IBM Zurich Research Laboratory with one of their early devices. They won the 1986 Nobel prize for the STM.
| SOURCE: Courtesy of IBM Zurich Research Laboratory.



An STM image of a Ni (110) surface.
| SOURCE: Courtesy of IBM.

change (alternate) from one atomic site to a neighboring site within the hexagonal rings. STM was invented by Gerd Binning and Heinrich Rohrer at the IBM Research Laboratory in Zurich, for which they were awarded the 1986 Nobel prize.[5]

EXAMPLE 3.12 **TUNNELING CONDUCTION THROUGH METAL-TO-METAL CONTACTS** Consider two copper wires separated only by their surface oxide layer (CuO). Classically, since the oxide layer is an insulator, no current should be possible through the two copper wires. Suppose that for the conduction ("free") electrons in copper, the surface oxide layer looks like a square potential energy barrier of height 10 eV. Consider an oxide layer thickness of 5 nm and evaluate the transmission coefficient for conduction electrons in copper, which have a kinetic energy of about 7 eV. What will be the transmission coefficient if the oxide barrier is 1 nm?

SOLUTION

We can calculate α from

$$\alpha = \left[\frac{2m(V_o - E)}{\hbar^2}\right]^{1/2}$$

$$= \left[\frac{2(9.1 \times 10^{-31}\,\text{kg})(10\,\text{eV} - 7\,\text{eV})(1.6 \times 10^{-19}\,\text{J/eV})}{(1.05 \times 10^{-34}\,\text{J s})^2}\right]^{1/2}$$

$$= 8.9 \times 10^{9}\,\text{m}^{-1}$$

so that

$$\alpha a = (8.9 \times 10^{9}\,\text{m}^{-1})(5 \times 10^{-9}\,\text{m}) = 44.50$$

Since this is greater than unity, we use the wide-barrier transmission coefficient in Equation 3.31.

Now,

$$T_o = \frac{16E(V_o - E)}{V_o^2} = \frac{16(7\,\text{eV})(10\,\text{eV} - 7\,\text{eV})}{(10\,\text{eV})^2} = 3.36$$

Thus,

$$T = T_o \exp(-2\alpha a)$$

$$= 3.36 \exp[-2(8.9 \times 10^{9}\,\text{m}^{-1})(5 \times 10^{-9}\,\text{m})] = 3.36 \exp(-89)$$

$$\approx 7.4 \times 10^{-39}$$

an incredibly small number.

With $a = 1$ nm,

$$T = 3.36 \exp[-2(8.9 \times 10^{9}\,\text{m}^{-1})(1 \times 10^{-9}\,\text{m})]$$

$$= 3.36 \exp(-17.8) \approx 6.2 \times 10^{-8}$$

Notice that reducing the layer thickness by five times increases the transmission probability by 10^{31}! Small changes in the barrier width lead to enormous changes in the transmission

[5] The IBM Research Laboratory in Zurich, Switzerland, received both the 1986 and the 1987 Nobel prizes. The first was for the scanning tunneling microscope by Gerd Binning and Heinrich Rohrer. The second was awarded to Georg Bednorz and Alex Müller for the discovery of high-temperature superconductors which we will examine in Chapter 8.

probability. We should note that when a voltage is applied across the two wires, the potential energy height is altered (PE = charge × voltage), which results in a large increase in the transmission probability and hence results in a current.

EXAMPLE 3.13

SIGNIFICANCE OF A SMALL $\hbar$ Estimate the probability that a roller coaster carriage that weighs 100 kg released from point A in Figure 3.16a from a height at 10 m can reach point E over a hump that is 15 m high and 10 m wide. What will this probability be in a universe where $\hbar \approx 10$ kJ s?

SOLUTION

The total energy of the carriage at height A is

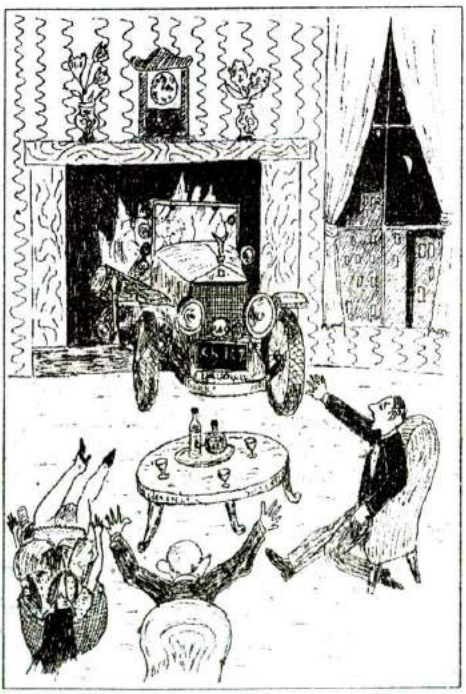
$$E = PE = mg(\text{height}) = (100\,\text{kg})(10\,\text{m s}^{-2})(10\,\text{m}) = 10^4\,\text{J}$$

Suppose that as a first approximation, we can approximate the hump as a square hill of height 15 m and width 10 m. The PE required to reach the peak would be

$$V_o = mg(\text{height}) = (100\,\text{kg})(10\,\text{m s}^{-2})(15\,\text{m}) = 1.5 \times 10^4\,\text{J}$$

Applying this, we have

$$\alpha^2 = \frac{2m(V_o - E)}{\hbar^2} = \frac{2(100\,\text{kg})(1.5 \times 10^4\,\text{J} - 10^4\,\text{J})}{(1.05 \times 10^{-34}\,\text{J s})^2} = 9.07 \times 10^{73}\,\text{m}^{-2}$$



"Just like the good old ghost of the middle ages." In a world where $\hbar$ is of the order of unity, one can expect tunneling surprises.

SOURCE: George Gamow, *Mr. Tompkins in Paperback*, Cambridge, England, University Press, 1965, p. 96. Used with permission.

and so

$$\alpha = 9.52 \times 10^{36}\,\text{m}^{-1}$$

With $a = 10$ m, we have $\alpha a \gg 1$, so we can use the wide-barrier tunneling equation,

$$T = T_o \exp(-2\alpha a)$$

where

$$T_o = \frac{16[E(V_o - E)]}{V_o^2} = 3.56$$

Thus,

$$T = 3.56 \exp[-2(9.52 \times 10^{36}\,\text{m}^{-1})(10\,\text{m})] = 3.56 \exp(-1.9 \times 10^{38})$$

which is a fantastically small number, indicating that it is impossible for the carriage to tunnel through the hump.

Suppose that $\hbar \approx 10$ kJ s. Then

$$\alpha^2 = \frac{2m(V_o - E)}{\hbar^2} = \frac{2(100\,\text{kg})(1.5 \times 10^4\,\text{J} - 10^4\,\text{J})}{(10^4\,\text{J s})^2} = 0.01\,\text{m}^{-2}$$

so that $\alpha = 0.1\,\text{m}^{-1}$. Clearly, $\alpha a = 1$, so we must use

$$T = [1 + D \sinh^2(\alpha a)]^{-1}$$

where

$$D = \frac{V_o^2}{[4E(V_o - E)]} = 1.125$$

Thus,

$$T = [1 + 1.125 \sinh^2(1)]^{-1} = 0.39$$

Thus, after three goes, the carriage would tunnel to the other side (giving the person standing at E the shock of his life).

3.6 POTENTIAL BOX: THREE QUANTUM NUMBERS

To examine the properties of a particle confined to a region of space, we take a three-dimensional space with a volume marked by a, b, c along the x, y, z axes. The *PE* is zero ($V = 0$) inside the space and is infinite on the outside, as illustrated in Figure 3.19. This is a three-dimensional potential energy well. The electron essentially lives in the "box." What will the behavior of the electron be in this box? In this case we need to solve the three-dimensional version of the Schrödinger equation,[6] which is

Schrödinger equation in three dimensions

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} + \frac{2m}{\hbar^2}(E - V)\psi = 0 \qquad [3.34]$$

[6] The term $\partial\psi/\partial x$ simply means differentiating $\psi(x, y, z)$ with respect to x while keeping y and z constant, just like $d\psi/dx$ in one dimension.

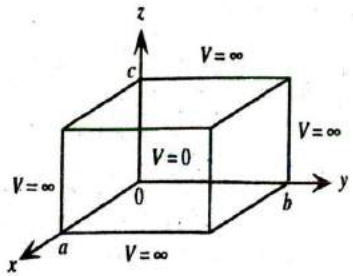


Figure 3.19 Electron confined in three dimensions by a three-dimensional infinite *PE* box.

Everywhere inside the box, $V = 0$, but outside, $V = \infty$. The electron cannot escape from the box.

with $V = 0$ in $0 < x < a, 0 < y < b$, and $0 < z < c$, and V infinite outside. We can try to solve this by separating the variables via $\psi(x, y, z) = \psi_x(x)\,\psi_y(y)\,\psi_z(z)$. Substituting this back into Equation 3.34, we can obtain three ordinary differential equations, each just like the one for the one-dimensional potential well. Having found $\psi_x(x)$, $\psi_y(y)$, and $\psi_z(z)$ we know that the total wavefunction is simply the product,

$$\psi(x, y, z) = A \sin(k_x x) \sin(k_y y) \sin(k_z z) \qquad [3.35]$$

where k_x, k_y, k_z, and A are constants to be determined. We can then apply the boundary conditions at $x = a$, $y = b$, and $z = c$ to determine the constants k_x, k_y, and k_z in the same way we found k for the one-dimensional potential well. If $\psi(x, y, z) = 0$ at $x = a$, then k_x will be quantized via

$$k_x a = n_1 \pi$$

where n_1 is a quantum number, $n_1 = 1, 2, 3, \ldots$. Similarly, if $\psi(x, y, z) = 0$ at $y = b$ and $z = c$, then k_y and k_z will be quantized, so that, overall, we will have

$$k_x = \frac{n_1 \pi}{a} \qquad k_y = \frac{n_2 \pi}{b} \qquad k_z = \frac{n_3 \pi}{c} \qquad [3.36]$$

where n_1, n_2, and n_3 are quantum numbers, each of which can be any integer except zero.

We notice immediately that in three dimensions, we have three quantum numbers n_1, n_2, and n_3 associated with $\psi_x(x)$, $\psi_y(y)$, and $\psi_z(z)$. The eigenfunctions of the electron, denoted by the quantum numbers n_1, n_2, and n_3, are now given by

Electron wavefunction in infinite PE well

$$\psi_{n_1 n_2 n_3}(x, y, z) = A \sin\left(\frac{n_1 \pi x}{a}\right) \sin\left(\frac{n_2 \pi y}{b}\right) \sin\left(\frac{n_3 \pi z}{c}\right) \qquad [3.37]$$

Notice that these consist of the products of infinite one-dimensional *PE* well-type wavefunctions, one for each dimension, and each has its own quantum number n. Each possible eigenfunction can be labeled a **state** for the electron. Thus, ψ_{111} and ψ_{121} are two possible states.

To find the constant A in Equation 3.37, we need to use the normalization condition that $|\psi_{n_1 n_2 n_3}(x, y, z)|^2$ integrated over the volume of the box must be unity,

since the electron is somewhere in the box. The result for a square box is $A = (2/a)^{3/2}$.

We can find the energy of the electron by substituting the wavefunction in Equation 3.35 into the Schrödinger Equation 3.34. The energy as a function of k_x, k_y, k_z is then found to be

$$E = E(k_x, k_y, k_z) = \frac{\hbar^2}{2m}\left(k_x^2 + k_y^2 + k_z^2\right)$$

which is quantized by virtue of k_x, k_y, and k_z being quantized. We can write this energy in terms of n_1^2, n_2^2, and n_3^2 by using Equation 3.36, as follows:

$$E_{n_1 n_2 n_3} = \frac{h^2}{8m}\left(\frac{n_1^2}{a^2} + \frac{n_2^2}{b^2} + \frac{n_3^2}{c^2}\right)$$

For a square box for which $a = b = c$, the energy is

Electron energy in infinite PE box

$$E_{n_1 n_2 n_3} = \frac{h^2\left(n_1^2 + n_2^2 + n_3^2\right)}{8ma^2} = \frac{h^2 N^2}{8ma^2} \qquad [3.38]$$

where $N^2 = (n_1^2 + n_2^2 + n_3^2)$, which can only have certain integer values. It is apparent that the energy now depends on three quantum numbers. Our conclusion is that in three dimensions, we have three quantum numbers, each one arising from boundary conditions along one of the coordinates. They quantize the energy of the electron via Equation 3.38 and its momentum in a particular direction, such as, $p_x = \pm\hbar k_x = \pm(hn_1/2a)$, though the average momentum is zero.

The lowest energy for the electron is obviously equal to E_{111}, not zero. The next energy level corresponds to E_{211}, which is the same as E_{121} and E_{112}, so there are three states (*i.e.*, ψ_{211}, ψ_{121}, ψ_{112}) for this energy. The number of states that have the same energy is termed the **degeneracy** of that energy level. The second energy level E_{211} is thus **three-fold degenerate.**

EXAMPLE 3.14 **NUMBER OF STATES WITH THE SAME ENERGY** How many states (eigenfunctions) are there at energy level E_{443} for a square potential energy box?

SOLUTION

This energy level corresponds to $n_1 = 4$, $n_2 = 4$, and $n_3 = 3$, but the energy depends on

$$N^2 = n_1^2 + n_2^2 + n_3^2 = 4^2 + 4^2 + 3^2 = 41$$

via Equation 3.38. As long as $N^2 = 41$ for any choice of (n_1, n_2, n_3), not just (4, 4, 3), the energy will be the same.

The value $N^2 = 41$ can be obtained from (4, 4, 3), (4, 3, 4), and (3, 4, 4) as well as (6, 2, 1), (6, 1, 2), (2, 6, 1), (2, 1, 6), (1, 6, 2), and (1, 2, 6). There are thus three states from (4, 4, 3) combinations and six from (6, 2, 1) combinations, giving nine possible states, each with a distinct wavefunction, $\psi_{n_1 n_2 n_3}$. However, all these $\psi_{n_1 n_2 n_3}$ for the electron have the same energy E_{443}.

3.7 HYDROGENIC ATOM

3.7.1 ELECTRON WAVEFUNCTIONS

Consider the behavior of the electron in a hydrogenic (hydrogen-like) atom, which has a nuclear charge of $+Ze$, as depicted in Figure 3.20. For the hydrogen atom, $Z = 1$, whereas for an ionized helium atom He^+, $Z = 2$. For a doubly ionized lithium atom Li^{++}, $Z = 3$, and so on. The electron is attracted by a positive nuclear charge and therefore has a Coulombic *PE*,

$$V(r) = \frac{-Ze^2}{4\pi\varepsilon_0 r} \qquad [3.39]$$

Electron PE in hydrogenic atom

Since force $F = -dV/dr$, Equation 3.39 is simply a statement of Coulomb's force between the positive charge $+Ze$ of the nucleus and the negative charge $-e$ of the electron. The task of finding $\psi(x, y, z)$ and the energy E of the electron now involves putting $V(r)$ from Equation 3.39 into the Schrödinger equation with $r = \sqrt{x^2 + y^2 + z^2}$ and solving it.

Fortunately, the problem has a spherical symmetry, and we can solve the Schrödinger equation by transforming it into the r, θ, ϕ coordinates shown in Figure 3.20. Even then, obtaining a solution is not easy. We must then ensure that the solution for $\psi(r, \theta, \phi)$ satisfies all the boundary conditions, as well as being single-valued and continuous with a continuous derivative. For example, when we go 2π around the ϕ coordinate, $\psi(r, \theta, \phi)$ should come back to its original value, or $\psi(r, \theta, \phi) = \psi(r, \theta, \phi + 2\pi)$, as is apparent from an examination of Figure 3.20. Along the radial

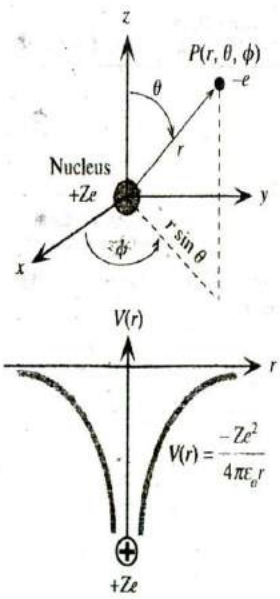


Figure 3.20 The electron in the hydrogenic atom is attracted by a central force that is always directed toward the positive nucleus.
Spherical coordinates centered at the nucleus are used to describe the position of the electron. The *PE* of the electron depends only on r.

coordinate, we need $\psi(r, \theta, \phi) \to 0$ as $r \to \infty$; otherwise, the total probability will diverge when $|\psi(r, \theta, \phi)|^2$ is integrated over all space. In an analogy with the three-dimensional potential well, there should be three quantum numbers to characterize the wavefunction, energy, and momentum of the electron. The three quantum numbers are called the **principal, orbital angular momentum,** and **magnetic quantum numbers** and are respectively denoted by n, ℓ, and m_ℓ. Unlike the three-dimensional potential well, however, not all the quantum numbers run as independent positive integers.

The solution to the Schrödinger equation $\psi(r, \theta, \phi)$ depends on three variables, r, θ, ϕ. The wavefunction $\psi(r, \theta, \phi)$ can be written as the product of two functions

$$\psi(r, \theta, \phi) = R(r)\, Y(\theta, \phi)$$

where $R(r)$ is a radial function depending only on r, and $Y(\theta, \phi)$ is called the **spherical harmonic,** which expresses the angular dependence of the wavefunction. These functions are characterized by the quantum numbers n, ℓ, m_ℓ. The radial part $R(r)$ depends on n and ℓ, whereas the spherical harmonic depends on ℓ and m_ℓ, so

$$\psi(r, \theta, \phi) = \psi_{n,\ell,m_\ell}(r, \theta, \phi) = R_{n,\ell}(r)\, Y_{\ell,m_\ell}(\theta, \phi) \qquad [3.40]$$

By solving the Schrödinger equation, these functions have already been evaluated. It turns out that we can only assign certain values to the quantum numbers n, ℓ, and m_ℓ to obtain acceptable solutions, that is, $\psi_{n,\ell,m_\ell}(r, \theta, \phi)$ that are well behaved: single-valued and with ψ and the gradient of ψ continuous. We can summarize the allowed values of n, ℓ, m_ℓ as follows:

Principal quantum number	$n = 1, 2, 3, \ldots$
Orbital angular momentum quantum number	$\ell = 0, 1, 2, \ldots, (n-1) < n$
Magnetic quantum number	$m_\ell = -\ell, -(\ell-1), \ldots, 0, \ldots, (\ell-1), \ell$ or $\lvert m_\ell \rvert \leq \ell$

The ℓ values carry a special notation inherited from spectroscopic terms. The first four ℓ values are designated by the first letters of the terms *sharp, principal, diffuse,* and *fundamental,* whereas the higher ℓ values follow from f onwards, as $g, h, i,$ etc. For example, any state ψ_{n,ℓ,m_ℓ} that has $\ell = 0$ is called an s state, whereas that which has $\ell = 1$ is termed a p state. We can also use n as a prefix to ℓ to identify n. Thus ψ_{n,ℓ,m_ℓ} with $n = 2$ and $\ell = 0$ corresponds to the $2s$ state. The notation for identifying the ℓ value and labeling a state is summarized in Table 3.1.

Table 3.1 Labeling of various $n\ell$ possibilities

	ℓ				
n	0	1	2	3	4
1	1*s*				
2	2*s*	2*p*			
3	3*s*	3*p*	3*d*		
4	4*s*	4*p*	4*d*	4*f*	
5	5*s*	5*p*	5*d*	5*f*	5*g*

Table 3.2 The radial and spherical harmonic parts of the wavefunction in the hydrogen atom ($a_o = 0.0529$ nm)

n	ℓ	$R(r)$	m_ℓ	$Y(\theta, \phi)$	
1	0	$\left(\frac{1}{a_o}\right)^{3/2} 2\exp\left(-\frac{r}{a_o}\right)$	0	$\frac{1}{2\sqrt{\pi}}$	
2	0	$\left(\frac{1}{2a_o}\right)^{3/2}\left(2-\frac{r}{a_o}\right)\exp\left(-\frac{r}{2a_o}\right)$	0	$\frac{1}{2\sqrt{\pi}}$	
2	1	$\left(\frac{1}{2a_o}\right)^{3/2}\left(\frac{r}{\sqrt{3}a_o}\right)\exp\left(-\frac{r}{2a_o}\right)$	0	$\frac{1}{2}\sqrt{\frac{3}{\pi}}\cos\theta$	$\propto \sin\theta\cos\phi$
			1	$\frac{1}{2}\sqrt{\frac{3}{2\pi}}\sin\theta e^{i\phi}$	$\propto \sin\theta\sin\phi$
			−1	$\frac{1}{2}\sqrt{\frac{3}{2\pi}}\sin\theta e^{-i\phi}$	Correspond to $m_\ell = -1$ and $+1$.

Table 3.2 summarizes the functional forms of $R_{n,\ell}(r)$ and $Y_{\ell,m_\ell}(\theta, \phi)$. For $\ell = 0$ (the s states), the angular dependence of $Y_{0,0}(\theta, \phi)$ is constant, which means that $\psi(r, \theta, \phi)$ is spherically symmetrical about the nucleus. For the $\ell = 1$ and higher states, there is a strong directionality to the wavefunctions with respect to each other. The radial part $R_{n,\ell}(r)$ is sketched in Figure 3.21a for two choices of n and ℓ. Notice that $R_{n,\ell}(r)$ is largest at $r = 0$, when $\ell = 0$. However, this does not mean that the electron will be mainly at $r = 0$, because the probability of finding the electron at a distance r actually depends on $r^2|R_{n,\ell}(r)|^2$, which vanishes as $r \to 0$.

Let us examine the probability of finding the electron at a distance r within a thin spherical shell of radius r and thickness δr (assumed to be very small). The directional dependence of the probability will be determined by the function $Y_{\ell,m_\ell}(\theta, \phi)$. We can average this over all directions (all angles θ and ϕ) to obtain $\overline{Y_{\ell,m_\ell}(\theta, \phi)}$, which turns out to be simply $1/4\pi$. The volume of the spherical shell is $\delta V = 4\pi r^2 \delta r$. The probability of finding the electron in this shell is then

$$|(\overline{Y_{\ell,m_\ell}(\theta, \phi)})(R_{n,\ell}(r))|^2 \times (4\pi r^2 \delta r)$$

If $\delta P(r)$ represents the probability that the electron is in this spherical shell of thickness δr, then

$$\delta P(r) = |R_{n,\ell}(r)|^2 r^2 \delta r \qquad [3.41]$$

The **radial probability density** $P_{n,\ell}(r)$ is defined as the probability per unit radial distance, that is, dP/dr which from Equation 3.41 is $|R_{n,\ell}(r)|^2 r^2$. The latter vanishes at the nucleus and peaks at certain locations, as shown in Figure 3.21b. This behavior implies that the probability of finding the electron within a thin spherical shell close to the nucleus also disappears. For $n = 1$, and $\ell = 0$, for example, the maximum probability is at $r = a_o = 0.0529$ nm, which is called the **Bohr radius.** Therefore, if the electron is in the $1s$ state, it spends most of its time at a distance a_o. Notice that the probability distribution does not depend on m_ℓ, but only on n and ℓ.

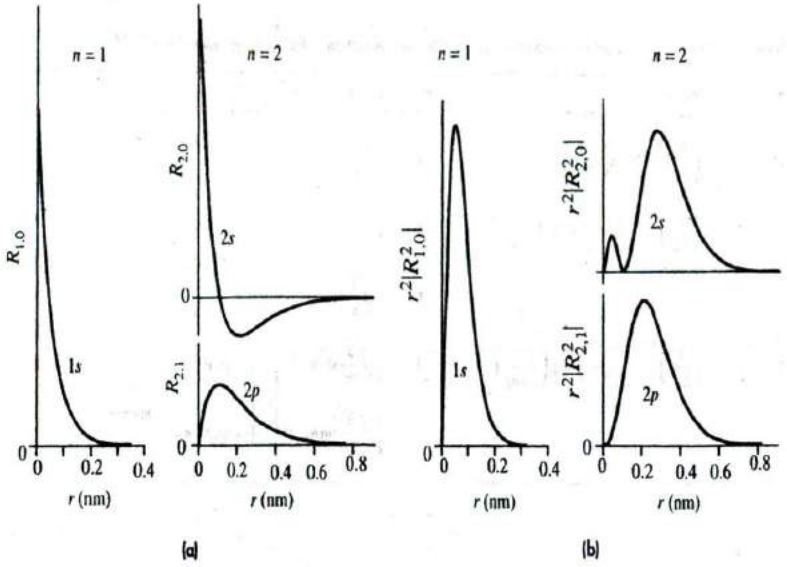


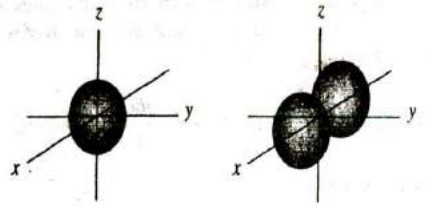
Figure 3.21

(a) Radial wavefunctions of the electron in a hydrogenic atom for various n and ℓ values.
(b) $r^2|R^2_{n,\ell}|$ gives the radial probability density. Vertical axis scales are linear in arbitrary units.

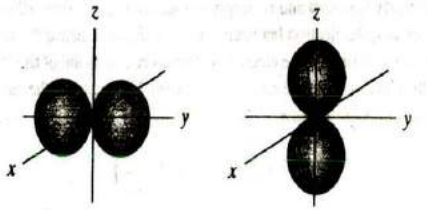
Table 3.2 summarizes the nature of the functions $R_{n,\ell}(r)$ and $Y_{\ell,m_\ell}(\theta,\phi)$ for various n, ℓ, m_ℓ values. Each possible wavefunction $\psi_{n,\ell,m_\ell}(r,\theta,\phi)$ with a particular choice of n, ℓ, m_ℓ constitutes a **quantum state** for the electron. The function $\psi_{n,\ell,m_\ell}(r,\theta,\phi)$ basically describes the behavior of the electron in the atom in probabilistic terms, as distinct from a well-defined line orbit for the electron, as one might expect from classical mechanics. For this reason, $\psi_{n,\ell,m_\ell}(r,\theta,\phi)$ is often referred to as an **orbital,** in contrast to the classical theory, which assigns an orbit to the electron.

Figure 3.22a shows the polar plots of $Y_{\ell,m_\ell}(\theta,\phi)$ for s and p orbitals. The radial distance from the origin in the polar plot represents the magnitude of $Y_{\ell,m_\ell}(\theta,\phi)$, which depends on the angles θ and ϕ. The polar plots of the probability distribution $|Y_{\ell,m_\ell}(\theta,\phi)|^2$ are shown in Figure 3.22b. Although for the s states, $Y_{1,0}(\theta,\phi)$ is spherically symmetric, resulting in a spherically symmetrical probability distribution around the nucleus, this is not so for $\ell = 1$ and higher states.

For example, each of the p states has a distinctly directional character, as illustrated in the polar plots in Figure 3.22. The angular dependence of $|\psi_{2,1,0}(r,\theta,\phi)|$, for which $m_\ell = 0$, is such that most of the probability is oriented along the z axis. This wavefunction is referred to as the $2p_z$ orbital. The two wavefunctions for $m_\ell = \pm 1$ are often represented by $\psi_{2p_x}(r,\theta,\phi)$ and $\psi_{2p_y}(r,\theta,\phi)$, or more simply, $2p_x$ and $2p_y$ orbitals, which do not possess a specific m_ℓ individually, but together represent the two $m_\ell = \pm 1$ wavefunctions. The angular dependence of $2p_x$ and $2p_y$ are essentially along the x and y directions. Thus, the three orbitals for $m_\ell = 0, \pm 1$ are all oriented perpendicular to each other, as depicted in Figure 3.22.

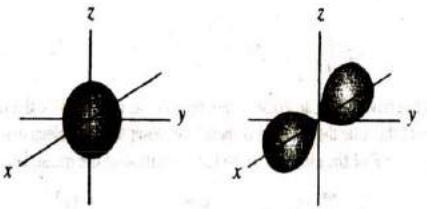


Y for a 1s orbital Y for a $2p_x$ orbital

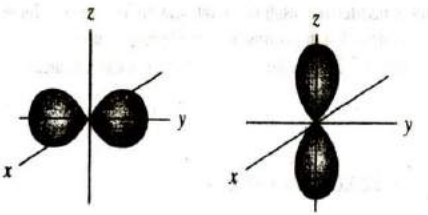


Y for a $2p_y$ orbital Y for a $2p_z$ orbital ($m_\ell = 0$)

(a)



$|Y|^2$ for a 1s orbital $|Y|^2$ for a $2p_x$ orbital



$|Y|^2$ for a $2p_y$ orbital $|Y|^2$ for a $2p_z$ orbital ($m_\ell = 0$)

(b)

Figure 3.22

(a) The polar plots of $Y_{n\ell}(\theta, \phi)$ for 1s and 2p states.

(b) The angular dependence of the probability distribution, which is proportional to $|Y_{n,\ell}(\theta, \phi)|^2$.

It should be noted that the probability distributions in Figures 3.21b and 3.22b do not depend on time. As previously mentioned, under steady-state conditions, the magnitude of the total wavefunction is

$$|\Psi(r, \theta, \phi, t)| = \left|\psi(r, \theta, \phi) \exp\left(-\frac{jEt}{\hbar}\right)\right| = |\psi(r, \theta, \phi)|$$

which is independent of time.

EXAMPLE 3.15 **PROBABILITY DENSITY FUNCTION** The quantity $|R_{n,\ell}(r)|^2 r^2$ in Equation 3.41 is called the **radial probability density function** and is simply written as $P_{n,\ell}(r)$. Thus, $dP(r) = P_{n,\ell}(r)\, dr$ is the probability of finding the electron between r and $r + dr$. We can use $P_{n,\ell}(r)$ to conveniently calculate the probability of finding the electron within a certain region of the atom, or to find the mean distance of the electron from the nucleus, and so on. For example, the electron in the 1*s* orbital has the wavefunction shown for $n = 1$, $\ell = 0$ in Table 3.2, which decays exponentially,

$$R_{n,\ell}(r) = 2a_o^{-3/2} \exp\left(-\frac{r}{a_o}\right)$$

The *total* probability of finding the electron inside the Bohr radius a_o can be found by summing (integrating) $P_{n,\ell}\, dr$ from $r = 0$ to $r = a_o$,

$$P_{\text{total}}(r < a_o) = \int_0^{a_o} P_{n,\ell}(r)\, dr = \int_0^{a_o} |R_{n,\ell}(r)|^2 r^2\, dr$$

$$= \int_0^{a_o} 4a_o^{-3} \exp\left(-\frac{2r}{a_o}\right) r^2\, dr = 0.32 \quad \text{or} \quad 32\%$$

The integration is not trivial but can nonetheless be done as indicated by the result 0.32 above. Thirty-two percent of the time the electron is therefore closer to the nucleus than the Bohr radius.

The mean distance $\bar{r}$ of the electron, from the definition of the mean, becomes

Average distance of electron from nucleus

$$\bar{r} = \int_0^{\infty} r P_{n,\ell}(r)\, dr = \frac{a_o n^2}{Z}\left[\frac{3}{2} - \frac{\ell(\ell+1)}{2n^2}\right] \qquad [3.42]$$

where we have simply inserted the result of the integration for various orbitals. (Again we take the mathematics as granted.) For the 1*s* orbital, in the hydrogen atom, $Z = 1$, $n = 1$, and $\ell = 0$, so $\bar{r} = \frac{3}{2}a_o$, further than the Bohr radius. Notice that the mean distance $\bar{r}$ of the electron increases as n^2.

3.7.2 QUANTIZED ELECTRON ENERGY

Once the wavefunctions $\psi_{n,\ell,m_\ell}(r, \theta, \phi)$ have been found, they can be substituted into the Schrödinger equation to find the possible energies of the electron. These turn out to depend only on the principal quantum number n. The energy is given by

Electron energy in hydrogenic atom

$$E_n = -\frac{me^4 Z^2}{8\varepsilon_o^2 h^2 n^2} \qquad [3.43a]$$

or

$$E_n = -\frac{Z^2 E_I}{n^2} = -\frac{Z^2(13.6\text{ eV})}{n^2} \qquad [3.43b]$$

where

$$E_I = \frac{me^4}{8\varepsilon_o^2 h^2} = 2.18 \times 10^{-18}\text{ J} \quad \text{or} \quad 13.6\text{ eV} \qquad [3.43c]$$

Ionization energy of hydrogen

This corresponds to the energy required to remove the electron in the hydrogen atom ($Z = 1$) from the lowest energy level E_1 (at $n = 1$) to infinity; hence, it represents the **ionization energy.** The energy E_n in Equation 3.43b is negative with respect to that for the electron completely isolated from the nucleus (at $r = \infty$, therefore $V = 0$). Thus, when the electron is in the vicinity of the nucleus, $+Ze$, it has a lower energy, which is a favorable situation (hence, formation of the hydrogenic atom is energetically favorable). In general, the energy required to remove an electron from the nth shell to $n = \infty$ (where the electron is free) is called the **ionization energy for the nth shell,** which from Equation 4.43b is simply $|E_n|$ or $(13.6\text{ eV})Z^2/n^2$.

Since the energy is quantized, the lowest energy of the electron corresponds to $n = 1$, which is -13.6 eV. The next higher energy value it can have is $E_2 = -3.40$ eV when $n = 2$, and so on, as sketched in Figure 3.23. Normally, the electron will take up a state corresponding to $n = 1$, because this has the lowest energy, called the **ground energy.** Its wavefunction corresponds to $\psi_{100}(r, \theta, \phi)$, which has a probability peak at $r = a_o$ and no angular dependence, as indicated in Figures 3.21 and 3.22.

The electron can only become excited to the next energy level if it is supplied by the right amount of energy $E_2 - E_1$. A photon of energy $h\nu = E_2 - E_1$ can readily supply this energy when it strikes the electron. The electron then gets excited to the state with $n = 2$ by absorbing the photon, and its wavefunction changes to $\psi_{210}(r, \theta, \phi)$, which has the maximum probability at $r = 4a_o$. The electron thus spends most of its time in this excited state, at $r = 4a_o$. It can return from the excited state at E_2 to the ground state at E_1 by emitting a photon of energy $h\nu = E_2 - E_1$.

By virtue of the quantization of energy, we see that the emission of light from excited atoms can only have certain wavelengths: those corresponding to transitions from higher quantum-number states to lower ones. In fact, in spectroscopic analysis, these wavelengths can be used to identify the elements, since each element has its unique set of emission and absorption wavelengths arising from a unique set of energy levels. Figure 3.24 illustrates the origin of the emission and absorption spectra of atoms, which are a direct consequence of the quantization of the energy.

The electrons in atoms can also be excited by other means, for example, by collisions with other atoms as a result of heating a gas. Figure 3.25 depicts how collisions with other atoms can excite an electron to higher energies. If an impinging atom has sufficient kinetic energy, it can impart just the right energy to excite the electron to a higher energy level. Since the total energy must be conserved, the incoming atom will lose some of its kinetic energy in the process. The excited electron can later return to

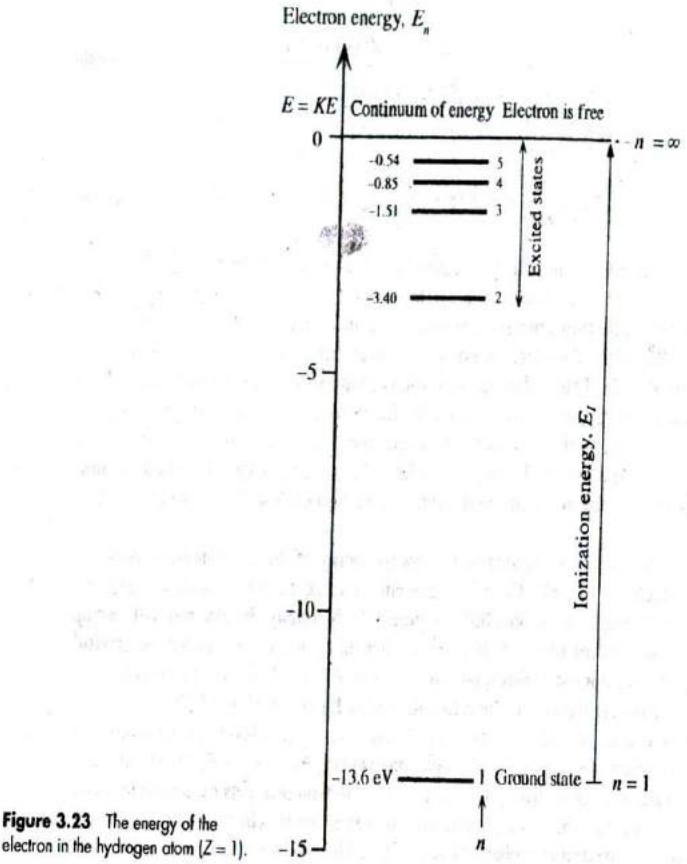


Figure 3.23 The energy of the electron in the hydrogen atom ($Z = 1$).

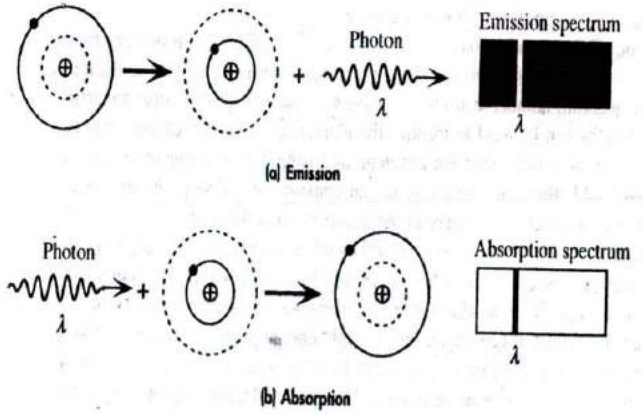


Figure 3.24 The physical origin of spectra.

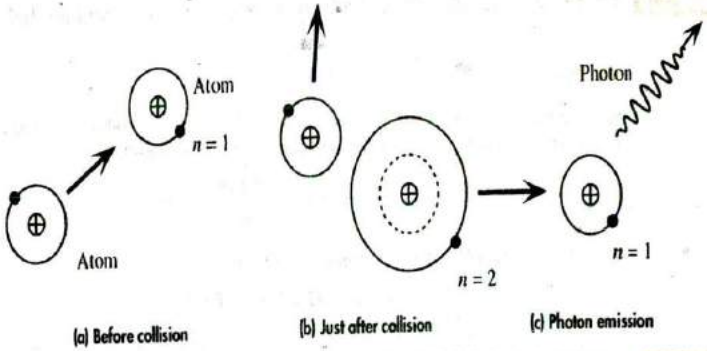


Figure 3.25 An atom can become excited by a collision with another atom. When it returns to its ground energy state, the atom emits a photon.

its ground state by emitting a photon. Excitation by atomic collisions is the process by which we obtain light from an electrical discharge in gases, a quantum phenomenon we experience every day as we read a neon sign. Indeed, this is exactly how the Ne atoms in the common laboratory HeNe laser are excited, via atomic collisions between Ne and He atoms.

Since the principal quantum number determines the energy of the electron and also the position of maximum probability, as we noticed in Figure 3.21, various n values define electron **shells**, within which we can most likely find the electron. These shells are customarily labeled $K, L, M, N, \ldots$, corresponding to $n = 1, 2, 3, \ldots$. For each n value, there are a number of ℓ values that determine the spatial distribution of the electron. For a given n, each ℓ value constitutes a **subshell**. For example, we often talk about $3s$, $3p$, $3d$ subshells within the M shell. From the radial dependence of the electron's wavefunction $\psi_{n,\ell,m_\ell}(r, \theta, \phi)$, shown in Figure 3.21, we see that for higher values of n, which correspond to more energetic states, the mean distance of the electron from the nucleus increases. In fact, we observe from Figure 3.21 that an orbital with $\ell = n - 1$ (*e.g.*, $1s$, $2p$) exhibits a single maximum in its radial probability distribution, and this maximum rapidly moves farther away from the nucleus as n increases. By examining the electron wavefunctions, we can show that the location of the maxima for these $\ell = n - 1$ states are at

Maximum probability for $\ell = n - 1$

$$r_{max} = \frac{n^2 a_o}{Z} \quad \text{for} \quad \ell = n - 1 \qquad [3.44]$$

where a_o is the radius of the ground state (0.0529 nm). The maximum probability radius r_{max} in Equation 3.44 is the Bohr radius. Note that r_{max} in Equation 3.44 is for $\ell = n - 1$ states only. For other ℓ values, there are multiple maxima, and we must think in terms of the average position of the electron from the nucleus. When we evaluate the average position from $\psi_{n,\ell,m_\ell}(r, \theta, \phi)$, we see that it depends on both n and ℓ; strongly on n and weakly on ℓ.

EXAMPLE 3.16 **THE IONIZATION ENERGY OF He^+** What is the energy required to further ionize He^+ ions to He^{++}?

SOLUTION

He^+ is a hydrogenic atom with one electron attracted by a nucleus with a $+2e$ charge. Thus $Z = 2$. The energy of the electron in a hydrogenic atom (in eV) is given by

$$E_n(\text{eV}) = -\frac{Z^2\,13.6}{n^2}$$

Since $Z = 2$, the energy required to ionize He^+ further is

$$|E_1| = |-(2^2)13.6| = 54.4\,\text{eV}$$

EXAMPLE 3.17 **IONIZATION ENERGY AND EFFECTIVE Z** The Li atom has a nucleus with a $+3e$ positive charge, which is surrounded by a full $1s$ orbital with two electrons, and a single valence electron in the outer $2s$ orbital as shown in Figure 3.26a. Intuitively we expect the valence electron to see the nuclear $+3e$ charge shielded by the two $1s$ electrons, that is, a net charge of $+1e$. It seems that we should be able to predict the ionization energy of the $2s$ electron by using the hydrogenic atom model and by taking $Z = 1$ and $n = 2$ as indicated in Figure 3.26b. However, according to quantum mechanics, the $2s$ electron has a probability distribution that has two peaks as shown in Figure 3.21; a major peak outside the $1s$ orbital, and a small peak around the $1s$ orbital. Thus, although the $2s$ electron spends a substantial time outside the $1s$ orbital, it does nonetheless penetrate the $1s$ shell and get close to the nucleus. Instead of experiencing a net $+1e$ of nuclear charge, it now experiences an effective nuclear charge that is greater than $+1e$, which we can represent as $+Z_{\text{effective}}e$, where we have used an *effective* Z. Thus, the ionization energy from Equation 3.43 is

Ionization and effective nuclear charge

$$E_{1,s} = \frac{Z^2_{\text{effective}}(13.6\,\text{eV})}{n^2} \qquad [3.45]$$

The experimental ionization energy of Li is 5.39 eV which corresponds to creating a Li^+ ion and an isolated electron. Calculate the effective nuclear charge seen by the $2s$ electron.

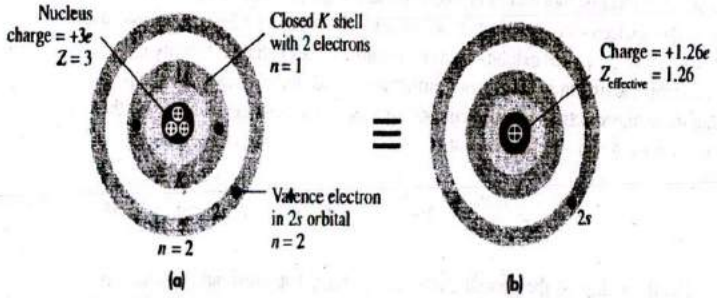


Figure 3.26

(a) The Li atom has a nucleus with charge $+3e$; two electrons in the K shell, which is closed; and one electron in the $2s$ orbital.

(b) A simple view of (a) would be one electron in the $2s$ orbital that sees a single positive charge, $Z = 1$.

SOLUTION

The most outer electron in the Li atom is in the $2s$ orbital, which is the electron that is removed in the ionization process. For this $2s$ electron, $n = 2$, and hence from Equation 3.45

$$5.39\text{ eV} = \frac{Z^2_{\text{effective}}(13.6\text{ eV})}{(2)^2}$$

Solving, we find $Z_{\text{effective}} = 1.26$. If we simply use $Z = 1$ in Equation 3.45, we would find $E_{1s} = 3.4$ eV, too small compared with the experimental value because, according to its probability distribution, the electron spends some time close to the nucleus, and hence increases its binding energy (stronger attraction). Variables Z and $Z_{\text{effective}}$ should not be confused. Z is the integer number of protons in the nucleus of the simple hydrogenic atom that are attracting the electron, as in H, He^+, or Li^{++}. $Z_{\text{effective}}$ is a convenient way of describing what the outer electron experiences in an atom because we would like to continue to use the simple expression for E_{1s}, Equation 3.45, which was originally derived for a hydrogenic atom.

3.7.3 ORBITAL ANGULAR MOMENTUM AND SPACE QUANTIZATION

The electron in the atom has an orbital angular momentum L. The electron is attracted to the nucleus by a central force, just like the Earth is attracted by the central gravitational force of the sun and thus possesses an orbital angular momentum. It is well known that in classical mechanics, under the action of a central force, both the total energy ($KE + PE$) and the orbital angular momentum (L) of an orbiting object are conserved. In quantum mechanics, the orbital angular momentum of the electron, like its energy, is also quantized, but by the quantum number ℓ. The magnitude of L is given by

Orbital angular momentum

$$L = \hbar[\ell(\ell + 1)]^{1/2} \qquad [3.46]$$

where $\ell = 0, 1, 2, \ldots < n$. Thus, for an electron in the ground state, $n = 1$ and $\ell = 0$, the angular momentum is zero, which is surprising since we always think of the electron as orbiting the nucleus. In the ground state, the spherical harmonic is a constant, independent of the angles θ and ϕ, so the electron has a spherically symmetrical probability distribution that depends only on r.

The quantum numbers n and ℓ quantize the energy and the magnitude of the orbital angular momentum. What is the significance of m_ℓ? In the presence of an external magnetic field B_z, taken arbitrarily in the z direction, the component of the angular momentum along the z axis, L_z, is also quantized and is given by

Orbital angular momentum along B_z

$$L_z = m_\ell \hbar \qquad [3.47]$$

Therefore, the quantum number m_ℓ quantizes the component of the angular momentum along the direction of an external magnetic field B_z, which for reference purposes is taken along z, as illustrated in Figure 3.27. Therefore, m_ℓ, is appropriately called the **magnetic quantum number.** For any given ℓ, quantum mechanics requires that m_ℓ must have values in the range $-\ell, -(\ell - 1), \ldots, -1, 0, 1, \ldots, (\ell - 1), \ell$. We see that $|m_\ell| \leq \ell$. Moreover, m_ℓ can be negative, since L_z can be negative or positive, depending on the orientation of the angular momentum vector $\mathbf{L}$. Since $|m_\ell| \leq \ell$, $\mathbf{L}$ can never align with the magnetic field along z; instead, it makes an angle with B_z, an

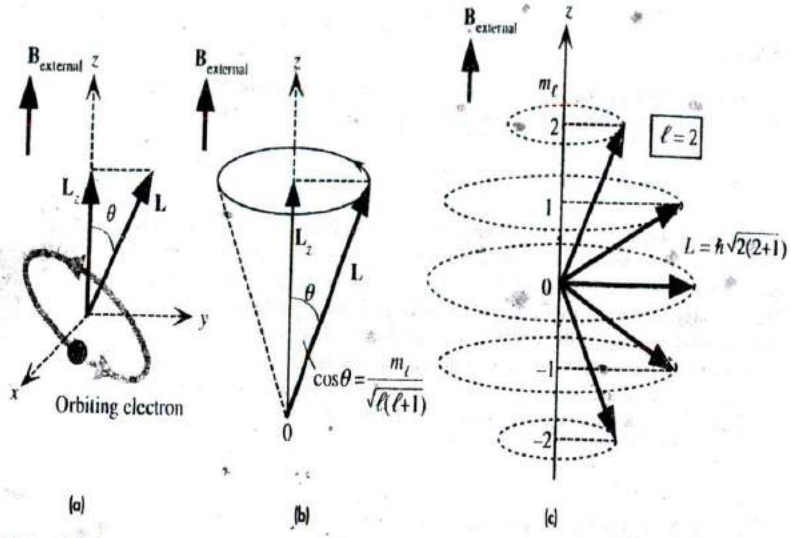


Figure 3.27

(a) The electron has an orbital angular momentum, which has a quantized component L along an external magnetic field $B_{external}$.

(b) The orbital angular momentum vector L rotates about the z axis. Its component L_z is quantized; therefore, the L orientation, which is the angle θ, is also quantized. L traces out a cone.

(c) According to quantum mechanics, only certain orientations (θ) for L are allowed, as determined by ℓ and m_ℓ.

angle that is determined by ℓ and m_ℓ. We say that **L is space quantized.** Space quantization is illustrated in Figure 3.27 for $\ell = 2$.

Since the energy of the electron does not depend on either ℓ or m_ℓ we can have a number of possible states for a given energy. For example, when the energy is E_2, then $n = 2$, which means that $\ell = 0$ or 1. For $\ell = 1$, we have $m_\ell = -1, 0, 1$, so there are a total of three different orbitals for the electron.

Since the electron has a quantized orbital angular momentum, when an electron interacts with a photon, the electron must obey the law of the conservation of angular momentum, much as an ice skater does sudden fast spins by pulling in her arms. All experiments indicate that the photon has an intrinsic angular momentum with a constant magnitude given by $\hbar$. Therefore, when a photon of energy $h\nu = E_2 - E_1$ is absorbed, the angular momentum of the electron must change. This means that following photon absorption or emission, both the principal quantum number n and the orbital angular momentum quantum number ℓ must change.

The rules that govern which transitions are allowed from one state to another as a consequence of photon absorption or emission are called **selection rules.** As a result of photon absorption or emission, we must have

Selection rules for EM radiation

$$\Delta\ell = \pm 1 \quad \text{and} \quad \Delta m_\ell = 0, \pm 1 \qquad [3.48]$$

As an example, consider the excitation of the electron in the hydrogen atom from the ground energy E_1 to a higher energy level E_2. The photon energy $h\nu$ must be

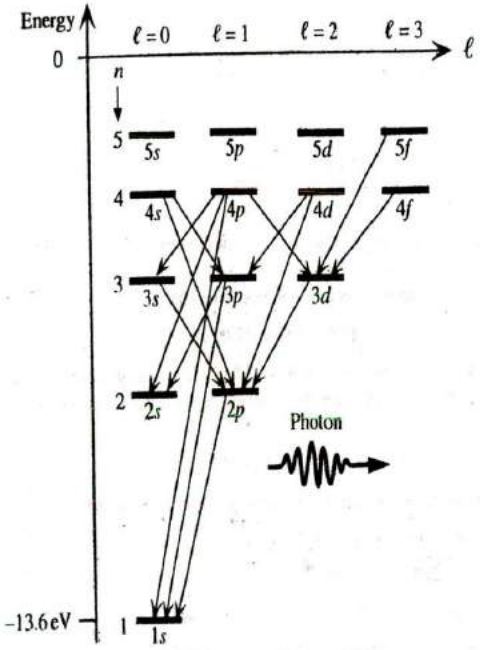


Figure 3.28 An illustration of the allowed photon emission processes. Photon emission involves $\Delta\ell = \pm 1$.

exactly $E_2 - E_1$. The wavefunction of the $1s$ ground state is $\psi_{1,0,0}$, whereas there are four wavefunctions at E_2: one $2s$ state, $\psi_{2,0,0}$; and three $2p$ states, $\phi_{2,1,-1}$, $\psi_{2,1,0}$, and $\psi_{2,1,1}$. The excited electron cannot jump into the $2s$ state, because $\Delta\ell$ must be ± 1, so it enters a $2p$ state corresponding to one of the orbitals $\psi_{2,1,-1}$, $\psi_{2,1,0}$, or $\psi_{2,1,1}$. Various allowed transitions for photon emission in the hydrogen atom are indicated in Figure 3.28.

EXAMPLE 3.18

EXCITATION BY ELECTRON-ATOM COLLISIONS IN A GAS DISCHARGE TUBE A projectile electron with a velocity 2.1×10^6 m s^{-1} collides with a hydrogen atom in a gas discharge tube. Find the nth energy level to which the electron in the hydrogen atom gets excited. Calculate the possible wavelengths of radiation that will be emitted from the excited H atom as the electron returns to its ground state.

SOLUTION

The energy of the electron in the hydrogen atom is given by $E_n\,(\text{eV}) = -13.6/n^2$. The electron must be excited from its ground state $E_1 = -13.6$ eV to a quantized energy level $-(13.6/n^2)$ eV. The change in the energy is $\Delta E = (-13.6/n^2) - (-13.6)$ eV. This must be supplied by the incoming projectile electron, which has an energy of

$$E = \frac{1}{2}mv^2 = \frac{1}{2}(9.1 \times 10^{-31}\,\text{kg})(2.1 \times 10^6\,\text{m s}^{-1})^2$$

$$= 2.01 \times 10^{-18}\,\text{J} \quad \text{or} \quad 12.5\,\text{eV}$$

Therefore,

$$12.5\,\text{eV} = 13.6\,\text{eV} - \left[\frac{(13.6\,\text{eV})}{n^2}\right]$$

Solving this for n, we find

$$n^2 = \frac{13.6}{(13.6 - 12.5)} = 12.36$$

so $n = 3.51$. But n can only be an integer; thus, the electron gets excited to the level $n = 3$ where its energy is $E_3 = -13.6/3^2 = -1.51$ eV.

The energy of the incoming electron after the collision is less by

$$(E_3 - E_1) = 13.6 - 1.51 = 12.09\,\text{eV}$$

Since the initial energy of the incoming electron was 12.5 eV, it leaves the collision with a kinetic energy of $12.5 - 12.09 = 0.41$ eV. From the E_3 level, the electron can undergo a transition from $n = 3$ to $n = 1$,

$$\Delta E_{31} = -1.51\,\text{eV} - (-13.6\,\text{eV}) = 12.09\,\text{eV}$$

The emitted radiation will have a wavelength λ given by $hc/\lambda = \Delta E$, so that

$$\lambda_{31} = \frac{hc}{\Delta E_{31}} = \frac{(6.626 \times 10^{-34}\,\text{J s})(3 \times 10^{8}\,\text{m s}^{-1})}{12.09 \times 1.6 \times 10^{-19}\,\text{J}}$$

$$= 1.026 \times 10^{-7}\,\text{m} \quad \text{or} \quad 102.6\,\text{nm} \quad \text{(in the ultraviolet region)}$$

Another possibility is the transition from $n = 3$ to $n = 2$, for which

$$\Delta E_{32} = -1.51\,\text{eV} - (-3.40\,\text{eV}) = 1.89\,\text{eV}$$

This will give a wavelength

$$\lambda_{32} = \frac{hc}{\Delta E_{32}} = 656\,\text{nm}$$

which is in the red region of the visible spectrum. For the transition from $n = 2$ to $n = 1$,

$$\Delta E_{21} = -3.40\,\text{eV} - (-13.6\,\text{eV}) = 10.2\,\text{eV}$$

which results in the emission of a photon of wavelength $\lambda_{21} = hc/\Delta E_{21} = 121.5$ nm. Note that each transition obeys $\Delta \ell = \pm 1$.

EXAMPLE 3.19 **THE FRAUNHOFER LINES IN THE SUN'S SPECTRUM** The light from the sun includes extremely sharp "dark lines" at certain wavelengths, superimposed on a bright continuum at all other wavelengths, as discovered by Josef von Fraunhofer in 1829. One of these dark lines occurs in the orange range and another in the blue. Fraunhofer measured their wavelengths to be 6563 Å and 4861 Å, respectively. With the aid of Figure 3.23, show that these are spectral lines from the hydrogen atom spectrum. (They are called the H_α and H_β Fraunhofer lines. Such lines provided us with the first clues to the chemical composition of the sun.)

SOLUTION

The energy of the electron in a hydrogenic atom is

$$E_n = -\frac{Z^2 E_I}{n^2}$$

where $E_I = me^4/(8\varepsilon_0^2 h^2)$. Photon emission resulting from a transition from quantum number n_2 to n_1 has an energy

$$\Delta E = E_{n_2} - E_{n_1} = -Z^2 E_I \left(\frac{1}{n_2^2} - \frac{1}{n_1^2}\right)$$

Emitted wavelengths for transitions in hydrogenic atom

From $h\nu = hc/\lambda = \Delta E$, we have

$$\frac{1}{\lambda} = \left(\frac{E_I}{hc}\right) Z^2 \left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right) = R_\infty Z^2 \left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

where $R_\infty = E_I/hc = 1.0974 \times 10^7\ \text{m}^{-1}$. The equation for λ is called the **Balmer–Rydberg formula**, and R_∞ is called the **Rydberg constant.** We apply the Balmer–Rydberg formula with $n_1 = 2$ and $n_2 = 3$ to obtain

$$\frac{1}{\lambda} = (1.0974 \times 10^7\ \text{m}^{-1})(1^2)\left(\frac{1}{2^2} - \frac{1}{3^2}\right) = 1.524 \times 10^6\ \text{m}^{-1}$$

to get $\lambda = 6561$ Å. We can also apply the Balmer–Rydberg formula with $n_1 = 2$ and $n_2 = 4$ to get $\lambda = 4860$ Å.

EXAMPLE 3.20

GIANT ATOMS IN SPACE Radiotelescopic studies by B. Höglund and P.-G. Mezger (*Science* vol. 150, p. 339, 1965) detected a 5009 MHz electromagnetic radiation in space. Show that this radiation comes from excited hydrogen atoms as they undergo transitions from $n = 110$ to 109. What is the size of such an excited hydrogen atom?

SOLUTION

Since the energy of the electron is $E_n = -(Z^2 E_I/n^2)$, the energy of the emitted photon in the transition from n_2 to n_1 is

$$h\nu = E_{n_2} - E_{n_1} = Z^2 E_I \left(n_1^{-2} - n_2^{-2}\right)$$

With $n_2 = 110$, $n_1 = 109$, and $Z = 1$, the frequency is

$$\nu = \frac{Z^2 E_I \left(n_1^{-2} - n_2^{-2}\right)}{h}$$

$$= \frac{[(1.6 \times 10^{-19} \times 13.6)][(109^{-2} - 110^{-2})]}{(6.626 \times 10^{-34})}$$

$$= 5 \times 10^9\ \text{s}^{-1} \quad \text{or} \quad 5000\ \text{MHz}$$

The size of the atom from Equation 3.44 is on the order of

$$2r_{max} = 2n^2 a_o = 2(110^2)(52.918 \times 10^{-12}\ \text{m}) = 1.28 \times 10^{-6}\ \text{m} \quad \text{or} \quad 1.28\ \mu\text{m}$$

A giant atom!

3.7.4 ELECTRON SPIN AND INTRINSIC ANGULAR MOMENTUM S

One aspect of electron behavior does not come from the simple Schrödinger equation. That is the spin of the electron about its own axis, which is analogous to the 24-hour

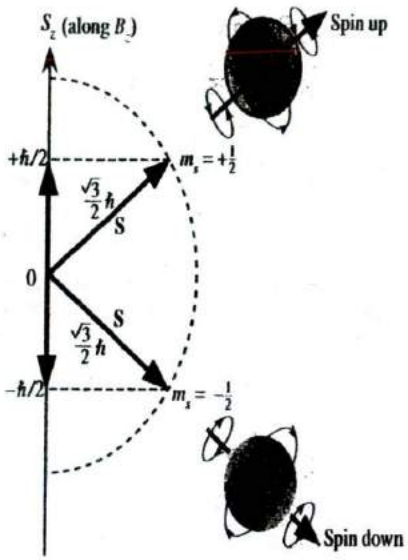


Figure 3.29 Spin angular momentum exhibits space quantization. Its magnitude along z is quantized, so the angle of **S** to the z axis is also quantized.

spin of Earth around its axis.[7] Earth has an orbital angular momentum due to its motion around the sun, and an intrinsic or spin angular momentum due to its rotation about its own axis. Similarly, the electron has a **spin** or **intrinsic angular momentum,** denoted by **S**. In classical mechanics, in the absence of external torques, spin angular momentum is conserved. In quantum mechanics, this spin angular momentum is quantized, in a manner similar to that of orbital angular momentum. The magnitude of the spin has been found to be constant, with a quantized component S_z in the z direction along a magnetic field:

Electron spin

$$S = \hbar[s(s+1)]^{1/2} \qquad s = \frac{1}{2} \tag{3.49}$$

Spin along magnetic field

$$S_z = m_s\hbar \qquad m_s = \pm\frac{1}{2} \tag{3.50}$$

where, in an analogy with ℓ and m_ℓ, we use the quantum numbers s and m_s, which are called the **spin and spin magnetic quantum numbers.** Contrary to our past experience with quantum numbers, s and m_s are not integers, but are $\frac{1}{2}$ and $\pm\frac{1}{2}$, respectively. The existence of electron spin was put forward by Goudsmit and Uhlenbeck in 1925 and derived by Dirac from relativistic quantum theory, which is beyond the scope of this book. Figure 3.29 illustrates the spin angular momentum of the electron and the two possibilities for S_z. When $S_z = +\frac{1}{2}\hbar$, using classical orbital motion as an analogy, we

[7] Do not take the meaning of "spin" too literally, as in classical mechanics. Remember that the electron is assumed to have wave-like properties, which can have no classical spin.

Table 3.3 The four quantum numbers for the hydrogenic atom

n	Principal quantum number	$n = 1, 2, 3, \ldots$	Quantizes the electron energy
ℓ	Orbital angular momentum quantum number	$\ell = 0, 1, 2, \ldots (n-1)$	Quantizes the magnitude of orbital angular momentum L
m_ℓ	Magnetic quantum number	$m_\ell = 0, \pm 1, \pm 2, \ldots, \pm \ell$	Quantizes the orbital angular momentum component along a magnetic field B_z
m_s	Spin magnetic quantum number	$m_s = \pm \frac{1}{2}$	Quantizes the spin angular momentum component along a magnetic field B_z

can label the spin of the electron as being in the clockwise direction, so $S_z = -\frac{1}{2}\hbar$ can be labeled as a counterclockwise spin. However, no such true clockwise or counterclockwise spinning of the electron can in reality[8] be identified. When $S_z = +\frac{1}{2}\hbar$, we could just as easily label the electron spin as "up," and call it "down" when $S_z = -\frac{1}{2}\hbar$. This terminology is used henceforth in this book.

Since the magnitude of the electron spin is constant, which is a remarkable fact, and is determined by $s = \frac{1}{2}$, we need not mention it further. It can simply be regarded as a fundamental property of the electron, in much the same way as its mass and charge. We do, however, need to specify whether $m_s = +\frac{1}{2}$ or $-\frac{1}{2}$, since each of these selections gives the electron a different behavior. We therefore need four quantum numbers to specify what the electron is doing. Each state of the electron needs the spin magnetic quantum number m_s, in addition to n, ℓ, and m_ℓ. For each orbital $\psi_{n,\ell,m_\ell}(r, \theta, \phi)$, we therefore have two possibilities: $m_s = \pm\frac{1}{2}$. The quantum numbers n, ℓ, and m_ℓ determine the spatial extent of the electron by specifying the form of $\psi_{n,\ell,m_\ell}(r, \theta, \phi)$, whereas m_s determines the "direction" of the electron's spin. A full description of the behavior of the electron must therefore include all four quantum numbers n, ℓ, m_ℓ, and m_s.

An **electronic state** is a wavefunction that defines both the spatial (ψ_{n,ℓ,m_ℓ}) and spin (m_s) properties of an electron. Frequently, an electronic state is simply denoted ψ_{n,ℓ,m_ℓ,m_s}, which adds the spin quantum number to the orbital wavefunction.

The quantum numbers are extremely important, because they quantize the various properties of the electron: its total energy, orbital angular momentum, and the orbital and spin angular momenta along a magnetic field. Their significance is summarized in Table 3.3.

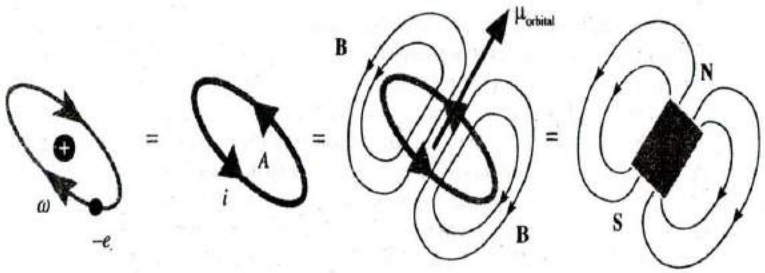
The spin angular momentum **S**, like the orbital angular momentum, is **space quantized.** $S_z = \pm(\frac{1}{2}\hbar)$ is smaller than $S = \hbar\sqrt{3}/2$, which means that S can never line up with z, or a magnetic field, and the angle θ between **S** and the z axis can only have two values corresponding to $m_\ell = +\frac{1}{2}$ and $-\frac{1}{2}$, which means that $\cos\theta = S_z/S = \pm 1/\sqrt{3}$. Classically, S_z of a spinning object, or the orientation of **S** to the z-axis, can be any value inasmuch as classical spin has no space quantization.

[8] The explanation in terms of spin and its two possible orientational directions ("clockwise" and "counterclockwise") serve as mental aids in visualizing a quantum mechanical phenomenon. One question, however, is, "If the electron is a wave, what is spinning?"

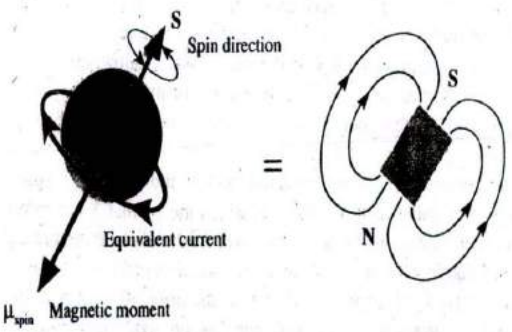
3.7.5 MAGNETIC DIPOLE MOMENT OF THE ELECTRON

Consider the electron orbiting the nucleus with an angular frequency ω as illustrated in Figure 3.30a. The orbiting electron is equivalent to a current loop. The equivalent current I due to the orbital motion of the electron is given by the charge flowing per unit time, I = charge/period $= -e(\omega/2\pi)$. The negative sign indicates that current I flows in the opposite direction to the electron motion. The magnetic field around the current loop is similar to that of a permanent magnet as depicted in Figure 3.30a. The magnetic moment is defined as $\mu = IA$, the product of the current and the area enclosed by the current loop. It is a vector normal to the surface A in a direction determined by the corkscrew rule applied to the circulation of the current I. If r is the radius of the orbit (current loop), then the magnetic moment is

$$\mu = IA = \left(-\frac{e\omega}{2\pi}\right)(\pi r^2) = -\frac{e\omega r^2}{2}$$



(a) The orbiting electron is equivalent to a current loop that behaves like a bar magnet.



(b) The spinning electron can be imagined to be equivalent to a current loop as shown. This current loop behaves like a bar magnet, just as in the orbital case.

Figure 3.30

Consider now the orbital angular momentum L, which is the linear momentum p multiplied by the radius r, or

$$L = pr = m_e vr = m_e \omega r^2$$

Using this, we can substitute for ωr^2 in $\mu = -e\omega r^2/2$ to obtain

$$\mu = -\frac{e}{2m_e} L$$

In vector notation, using the subscript "orbital" to identify the origin of the magnetic moment,

Orbital magnetic moment

$$\boldsymbol{\mu}_{\text{orbital}} = -\frac{e}{2m_e}\mathbf{L} \qquad [3.51]$$

This means that the orbital magnetic moment $\boldsymbol{\mu}_{\text{orbital}}$ is in the opposite direction to that of the orbital angular momentum $\mathbf{L}$ and is related to it by a constant $(e/2m_e)$.

Similarly, the spin angular momentum of the electron $\mathbf{S}$ leads to a **spin magnetic moment** $\boldsymbol{\mu}_{\text{spin}}$, which is in the opposite direction to $\mathbf{S}$ and given by

Spin magnetic moment

$$\boldsymbol{\mu}_{\text{spin}} = -\frac{e}{m_e}\mathbf{S} \qquad [3.52]$$

which is shown in Figure 3.30b. Notice that there is no factor of 2 in the denominator. We see that, as a consequence of the orbital motion and also of spin, the electron has two distinct magnetic moments. These moments act on each other, just like two magnets interact with each other. The result is a coupling of the orbital and the spin angular momenta $\mathbf{L}$ and $\mathbf{S}$ and their precession about the total angular momentum $\mathbf{J} = \mathbf{L} + \mathbf{S}$, which is discussed in Section 3.7.6.

Since both $\mathbf{L}$ and $\mathbf{S}$ are quantized, so are the orbital and spin magnetic moments $\boldsymbol{\mu}_{\text{orbital}}$ and $\boldsymbol{\mu}_{\text{spin}}$. In the presence of an external magnetic field $\mathbf{B}$, the electron has an additional energy term that arises from the interaction of these magnetic moments with $\mathbf{B}$. We know from electromagnetism that a magnetic dipole (equivalent to a magnet) placed in a magnetic field $\mathbf{B}$ will have a potential energy PE. (A free magnet will rotate to align with the magnetic field, as in a compass, and thereby reduce the PE.) The *potential energy* E_{BL} due to μ_{orbital} and B interacting is given by

Potential energy of a magnetic moment

$$E_{BL} = -\mu_{\text{orbital}} B \cos\theta$$

where θ is the angle between μ_{orbital} and B. The potential energy E_{BL} is minimum when μ_{orbital} (the magnet) and B are parallel, $\theta = 0$. We know that, by definition, the z axis is always along an external field $\mathbf{B}$, and L_z is the component of L along z (along B), and is quantized, so that $L_z = L\cos\theta = m_\ell \hbar$. We can substitute for μ_{orbital} to find

Potential energy of orbital angular momentum in B

$$E_{BL} = \left(\frac{e}{2m_e}\right) LB\cos\theta = \left(\frac{e}{2m_e}\right) L_z B = \left(\frac{e\hbar}{2m_e}\right) m_\ell B$$

which depends on m_ℓ, and it is minimum for the smallest m_ℓ. Since $m_\ell = -\ell, \ldots, 0, \ldots, +\ell$, negative and positive values through zero, the electron's energy splits into a number of levels determined by m_ℓ. Similarly, the spin magnetic moment $\boldsymbol{\mu}_{\text{spin}}$ and

Potential energy of orbital angular momentum in B

B interact to give the electron a potential energy E_{SL},

$$E_{SL} = \left(\frac{e\hbar}{m_e}\right) m_s B$$

which depends on m_s. Since $m_s = \pm\frac{1}{2}$, E_{SL} has only two values, positive ($m_s = +\frac{1}{2}$) and negative ($m_s = -\frac{1}{2}$), which add and subtract from the electron's energy depending on whether the spin is up or down. Thus, in an external magnetic field, the electron's spin splits the energy level into two levels. The separation ΔE_{SL} of the split levels is $(e\hbar/m_e)B$, which is 0.12 meV T^{-1}, very small compared with the energy E_n in the absence of the field. It should also be apparent that a single wavelength emission λ_0 corresponding to a particular transition from $E_{n'}$ to E_n will now be split into a number of closely spaced wavelengths around λ_0. Although the separation ΔE_{SL} is small, it is still more than sufficient even at moderate fields to be easily detected and used in various applications. As it turns out, spin splitting of the energy in a field can be fruitfully used to study the electronic structures of not only atoms and molecules, but also various defects in semiconductors in what is called *electron spin resonance*.

EXAMPLE 3.21 **STERN–GERLACH EXPERIMENT AND SPIN** The Stern–Gerlach experiment is quite famous for demonstrating the spin of the electron and its space quantization. A neutral silver atom has one outer valence electron in a 4*s* orbital and looks much like the hydrogenic atom. (We can simply ignore the inner filled subshells in the Ag atom.). The 4*s* electron has no orbital angular momentum. Because of the *spin* of this one outer 4*s* electron, the whole Ag atom has a spin magnetic moment μ_{spin}. When Otto Stern and Walther Gerlach (1921–1922) passed a beam of Ag atoms through a nonuniform magnetic field, they found that the narrow beam split into two distinct beams as depicted in Figure 3.31a. The interpretation of the experiment was that the Ag atom's magnetic moment along the field direction can have only two values, hence the split beam. This observation agrees with the quantum mechanical fact that in a field along z, $\mu_{spin,z} = -(e/m_e)m_s\hbar$ where $m_s = +\frac{1}{2}$ or $-\frac{1}{2}$; that is, the electron's spin can have only two values parallel to the field, or in other words, the electron spin is *space quantized*.

In the Stern–Gerlach experiment, the nonuniform magnetic field is generated by using a big magnet with shaped poles as in Figure 3.31a. The N-pole is sharp and the S-pole is wide, so the magnetic field lines get closer toward the N-pole and hence the magnetic field increases towards the N-pole. (This is much like a sharp point having a large electric field.) Whenever a magnetic moment, which we take to be a simple bar magnet, is in a nonuniform field, its poles experience different forces, say $\mathbf{F}_{large}$ and $\mathbf{F}_{small}$, and hence the magnet, overall, experiences a net force. The direction of the net force depends on the orientation of the magnet with respect to the z axis as illustrated in Figure 3.31b for two differently oriented magnets representing magnetic moments labeled as 1 and 2. The S-pole of magnet 1 is in the high field region and experiences a bigger pull ($\mathbf{F}_{large}$) from the big magnet's N-pole than the small force ($\mathbf{F}_{small}$) pulling the N-pole of 1 to the big magnet's S-pole. Hence magnet 1 is pulled toward the N-pole and is deflected up. The overall force on a magnetic moment is the difference between $\mathbf{F}_{large}$ and $\mathbf{F}_{small}$, and its direction here is determined by the force on whichever pole is in the high field region. Magnet 2 on the other hand has its N-pole in the high field region, and hence is pushed away from the big magnet's N-pole and is deflected down. If the magnet is at right angles to the z axis ($\theta = \pi/2$), it would experience no net force as both of its poles would be in the same field. This magnetic moment would pass through undeflected.

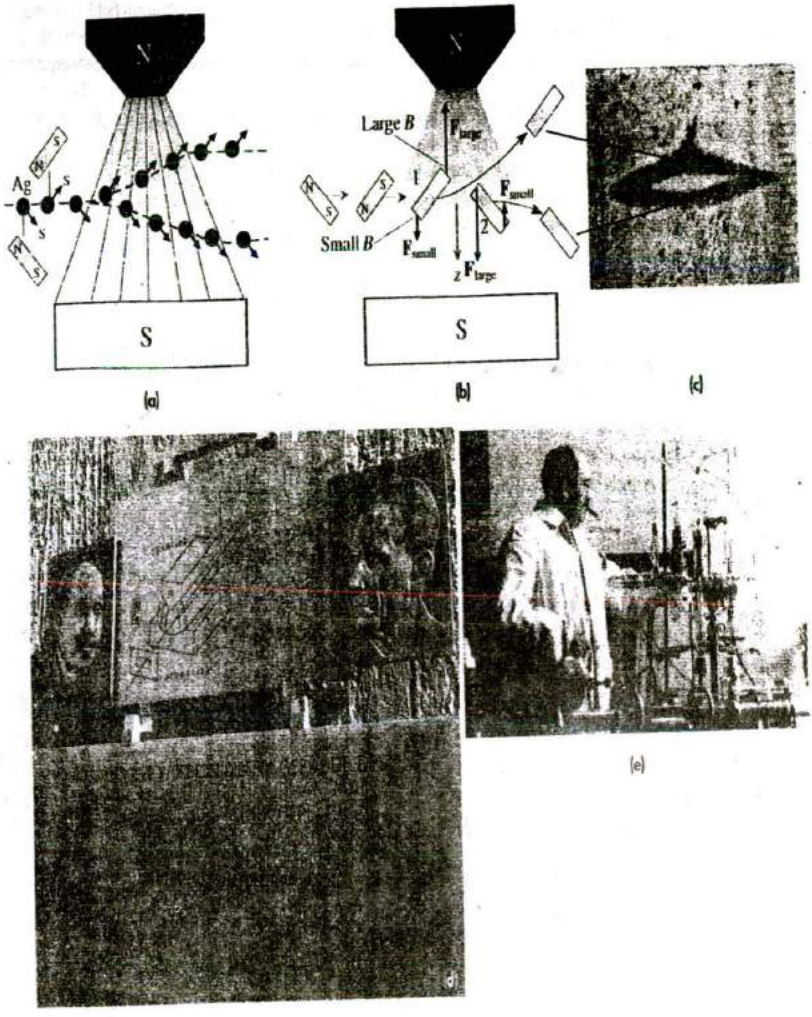


Figure 3.31

(a) Schematic illustration of the Stern–Gerlach experiment. A stream of Ag atoms passing through a nonuniform magnetic field splits into two.

(b) Explanation of the Stern–Gerlach experiment.

(c) Actual experimental result recorded on a photographic plate by Stern and Gerlach (O. Stern and W. Gerlach, *Zeitschr. fur. Physik*, **9**, 349, 1922.) When the field is turned off, there is only a single line on the photographic plate. Their experiment is somewhat different than the simple sketches in (a) and (b) as shown in (d).

(d) Stern–Gerlach memorial plaque at the University of Frankfurt. The drawing shows the original Stern–Gerlach experiment in which the Ag atom beam is passed along the long-length of the external magnet to increase the time spent in the nonuniform field, and hence increase the splitting.

(e) The photo on the lower right is Otto Stern (1888–1969), standing and enjoying a cigar while carrying out an experiment. Otto Stern won the Nobel prize in 1943 for development of the molecular beam technique.

SOURCES: (d) Courtesy of Horst Schmidt-Böcking from B. Friedrich and D. Herschbach, "Stern and Gerlach: How a Bad Cigar Helped Reorient Atomic Physics," *Physics Today*, December 2003, pp. 53–59. (e) AIP Emilio Segrè Visual Archives, Segrè Collection.

When we pass a stream of classical magnetic moments through a nonuniform field, there will be all possible orientations of the magnetic moment, from $-\pi$ to $+\pi$, with the field because there is no space quantization. Classically, the Ag atoms passing through a nonuniform field would be deflected through a distribution of angles and would not split into two distinct beams. The actual result of Stern and Gerlach's experiment is shown in Figure 3.31c, which is their photographic recording of a flat line-beam of Ag atoms passing through a long nonuniform field. In the absence of the field, the image is a simple horizontal line, the cross section of the beam. With the field turned on, the line splits into two. The edges of the line do not experience splitting because the field is very weak in the edge region. In the actual experiment, as shown in Figure 3.31c, an Ag atomic beam is passed along the long-length of the external magnet to increase the time spent in the nonuniform field, and hence increase the splitting. The physics remains the same.

3.7.6 TOTAL ANGULAR MOMENTUM J

The orbital angular momentum **L** and the spin angular momentum **S** add to give the electron a total angular momentum $\mathbf{J} = \mathbf{L} + \mathbf{S}$, as illustrated in Figure 3.32. There are a number of possibilities for the total angular momentum **J**, based on the relative orientations of **L** and **S**. For example, for a given **L**, we can add **S** either in parallel or antiparallel, as depicted in Figure 3.32a and b, respectively.

Since in classical physics the total angular momentum of a body (not experiencing an external torque) must be conserved, we can expect J (the magnitude of **J**) to be quantized. This turns out to be true. The magnitude of **J** and its z component along an external magnetic field are quantized via

Total angular momentum

$$J = \hbar[j(j+1)]^{1/2} \qquad [3.53]$$

$$J_z = m_j \hbar \qquad [3.54]$$

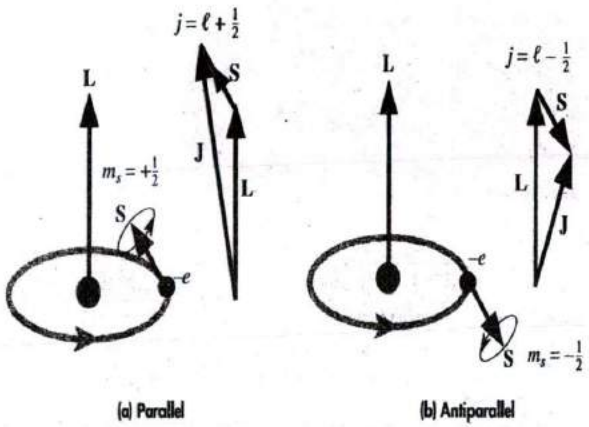


Figure 3.32 Orbital angular momentum vector **L** and spin angular momentum vector **S** can add either in parallel as in (a) or antiparallel, as in (b).
The total angular momentum vector $\mathbf{J} = \mathbf{L} + \mathbf{S}$, has a magnitude $J = \sqrt{[j(j+1)]}$, where in (a) $j = \ell + \frac{1}{2}$ and in (b) $j = \ell - \frac{1}{2}$.

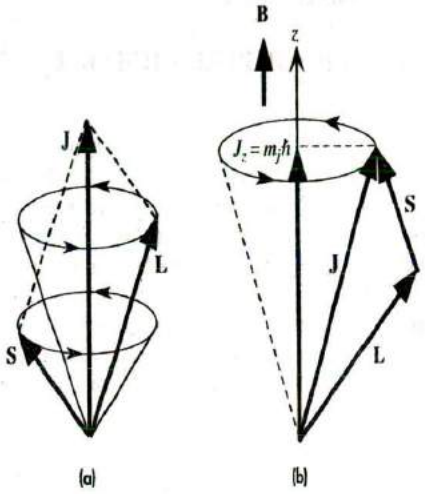


Figure 3.33
(a) The angular momentum vectors **L** and **S** precess around their resultant total angular momentum vector **J**. (b) The total angular momentum vector is space quantized. Vector **J** precesses about the z axis, along which its component must be $m_j\hbar$.

where both j and m_j are quantum numbers[9] like ℓ and m_ℓ, but j and m_j can have fractional values. A rigorous theory of quantum mechanics shows that when $\ell > s$, the quantum numbers for the total angular momentum are given by $j = \ell + s$ and $\ell - s$ and $m_j = \pm j, \pm(j-1)$. For example, for an electron in a p orbital, where $\ell = 1$, we have $j = \frac{3}{2}$ and $\frac{1}{2}$, and $m_j = \frac{3}{2}, \frac{1}{2}, -\frac{1}{2}$, and $-\frac{3}{2}$. However, when $\ell = 0$ (as for all s orbitals), we have $j = s = \frac{1}{2}$ and $m_j = m_s = \pm\frac{1}{2}$, which are the only possibilities. We note from Equations 3.53 and 3.54 that $|J_z| < J$ and both are quantized, which means that **J** is space quantized; its orientation (or angle) with respect to the z axis is determined by j and m_j.

The spinning electron actually experiences a magnetic field $\mathbf{B}_{\text{int}}$ due to its orbital motion around the nucleus. If we were sitting on the electron, then in our reference frame, the positively charged nucleus would be orbiting around us, which would be equivalent to a current loop. At the center of this current loop, there would be an "internal" magnetic field $\mathbf{B}_{\text{int}}$, which would act on the magnetic moment of the spinning electron to produce a torque. Since **L** and **S** add to give **J**, and since the latter quantity is space quantized (or conserved), then as a result of the internal torque on the electron, we must have **L** and **S** synchronously precessing about **J**, as illustrated in Figure 3.33a. If there is an external magnetic field **B** taken to be along z, this torque will act on the net magnetic moment due to **J** to cause this quantity to precess about **B**, as depicted in Figure 3.33b. Remember that the component along the z axis must be quantized and equal to $m_j\hbar$, so the torque can only cause precession. To understand the precession of the electron's angular momentum about the magnetic field **B**, think of a spinning top that precesses about the gravitational field of Earth.

[9] The quantum number j as used here should not be confused with j for $\sqrt{-1}$.

3.8 THE HELIUM ATOM AND THE PERIODIC TABLE

3.8.1 He ATOM AND PAULI EXCLUSION PRINCIPLE

In the He atom, there are two electrons in the presence of a nucleus of charge $+2e$, as depicted in Figure 3.34. (Obviously, in higher-atomic-number elements, there will be Z electrons around a nucleus of charge $+Ze$.) The *PE* of an electron in the He atom consists of two interactions. The first is due to the Coulombic attraction between itself and the positive nucleus; the second is due to the mutual repulsion between the two electrons. The *PE* function V of any one of the electrons, for example, that labeled as 1, therefore depends on both its distance from the nucleus r_1 and the separation of the two electrons r_{12}. The *PE* of electron 1 thus depends on the locations of both the electrons, or

PE of one electron in He atom

$$V(r_1, r_{12}) = -\frac{2e^2}{4\pi\varepsilon_0 r_1} + \frac{e^2}{4\pi\varepsilon_0 r_{12}} \qquad [3.55]$$

When we use this *PE* in the Schrödinger equation for a single electron, we find the wavefunction and energy of one of the electrons in the He atom. We thus obtain the **one-electron wavefunction** and the **energy of one electron** within a many-electron atom.

One immediate and obvious result is that the energy of an electron now depends not only on n but also on ℓ, because the electron–electron potential energy term (the second term in Equation 3.55, which contains r_{12}) depends on the relative orientations of the electron orbitals, which change r_{12}. We therefore denote the electron energy by $E_{n,\ell}$. The dependence on ℓ is weaker than on n, as shown in Figure 3.35. As n and ℓ increase, $E_{n,\ell}$ also increases. Notice, however, that the energy of a $4s$ state is lower than that of a $3d$ state, and the same pattern also occurs at $4s$ and $5s$.

One of the most important theorems in quantum physics is the **Pauli exclusion principle,** which is based on experimental observations. This principle states that *no two electrons within a given system (e.g., an atom) may have all four identical quantum numbers,* n, ℓ, m_ℓ, and m_s. Each set of values for n, ℓ, m_ℓ, and m_s represents a possible electronic state, that is, a wavefunction denoted by ψ_{n,ℓ,m_ℓ,m_s}, that the electron may (or may not) acquire. For example, an electron with the quantum numbers given by 2, 1, 1, $\frac{1}{2}$ will have a definite wavefunction $\psi_{n,\ell,m_\ell,m_s} = \psi_{2,1,1,1/2}$, and it is said to be

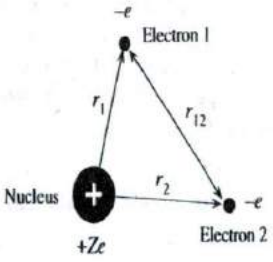


Figure 3.34 A helium-like atom.
The nucleus has a charge of $+Ze$, where $Z = 2$ for He. If one electron is removed, we have the He^+ ion, which is equivalent to the hydrogenic atom with $Z = 2$.

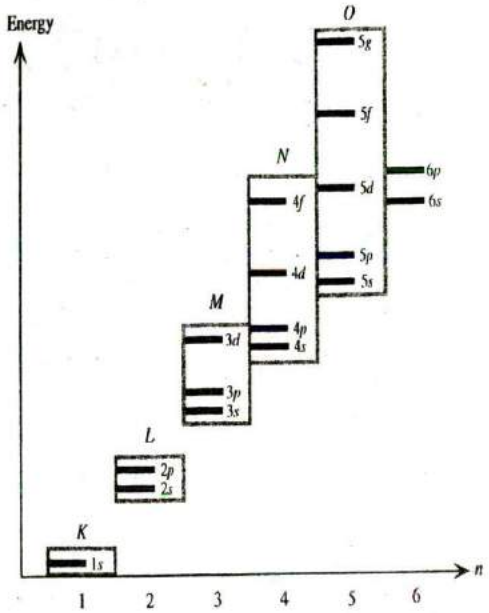


Figure 3.35 Energy of various one-electron states. The energy depends on both n and ℓ.

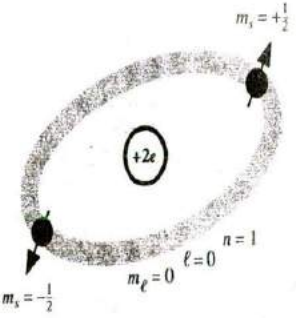


Figure 3.36 Paired spins in an orbital.

in the state $2p$, $m_\ell = 1$ and spin up. Its energy will be E_{2p}. The Pauli exclusion principle requires that no other electron be in this same state.

The orbital motion of an electron is determined by n, ℓ, and m_ℓ, whereas m_s determines the spin direction (up or down). Suppose two electrons are in the same orbital state, with identical n, ℓ, m_ℓ. By the Pauli exclusion principle, they would have to spin in opposite directions, as shown in Figure 3.36. One would have to spin "up" and the other "down." In this case we say that the electrons are **spin paired.** Two electrons can thus have the same orbitals (occupy the same region of space) if they pair their spins. However, the Pauli exclusion principle prevents a third electron from entering this orbital, since m_s can only have two values.

Using the Pauli exclusion principle, we can determine the electronic structure of many-electron atoms. For simplicity, we will use a box to represent an orbital state defined by a set of n, ℓ, m_ℓ values. Each box can take two electrons at most, with their spins paired. When we put an electron into a box, we are essentially assigning a wavefunction to that electron; that is, we are defining its orbital n, ℓ, m_ℓ. We use an arrow to show whether the electron is spinning up or down. As depicted in Figure 3.37, we arrange all the boxes to correspond to the electronic subshells. As an example, consider boron, which has five electrons. The first electron enters the $1s$ orbital at the lowest energy. The second also enters this orbital by spinning in the opposite direction. The third goes into the $n = 2$ orbital. The lowest energy there is in the s orbitals corresponding to $\ell = 0$ and $m_\ell = 0$. The fourth electron can also enter the $2s$

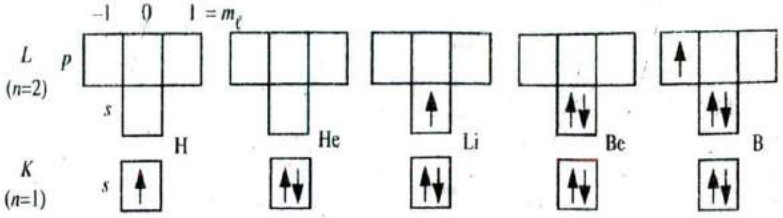


Figure 3.37 Electronic configurations for the first five elements. Each box represents an orbital $\psi(n, \ell, m_\ell)$.

orbital, provided that it spins in the opposite direction. Similarly, the fifth must go into another orbital, and the next nearest low-energy orbitals are those having $\ell = 1$ (p states) and $m_\ell = -1, 0, +1$. The final electronic structure of the B atom is shown in Figure 3.37.

We see that because the electron energy depends on n and ℓ, there are a number of states for a given energy $E_{n,\ell}$. Each of these states corresponds to different sets of m_ℓ and m_s. For example, the energy $E_{2,1}$ (or E_{2p}) corresponding to $n = 2$, $\ell = 1$ has six possible states, arising from $m_\ell = -1, 0, 1$ and $m_s = +\frac{1}{2}, -\frac{1}{2}$. Each m_ℓ state can have an electron spinning up or down, $m_s = +\frac{1}{2}$ or $m_s = -\frac{1}{2}$, respectively.

EXAMPLE 3.22 **THE NUMBER OF STATES AT AN ENERGY LEVEL** Enumerate and identify the states corresponding to the energy level E_{3d}, or $n = 3$, $\ell = 2$.

SOLUTION

When $n = 3$ and $\ell = 2$, m_ℓ and m_s can have these following values: $m_\ell = -2, -1, 0, 1, 2$, and $m_s = +\frac{1}{2}, -\frac{1}{2}$. This means there are 10 combinations. The possible wavefunctions (electron states) are

- $\psi_{3,2,2,1/2}$; $\psi_{3,2,1,1/2}$; $\psi_{3,2,0,1/2}$; $\psi_{3,2,-1,1/2}$; $\psi_{3,2,-2,1/2}$, all of which have spins up ($m_s = +\frac{1}{2}$)
- $\psi_{3,2,2,-1/2}$; $\psi_{3,2,1,-1/2}$; $\psi_{3,2,0,-1/2}$; $\psi_{3,2,-1,-1/2}$; $\psi_{3,2,-2,-1/2}$, all of which have spins down ($m_s = -\frac{1}{2}$)

3.8.2 HUND'S RULE

In the many-electron atom, the electrons take up the lowest-energy orbitals and obey the Pauli exclusion principle. However, the Pauli exclusion principle does not determine how any two electrons distribute themselves among the many states of a given n and ℓ. For example, there are six $2p$ states corresponding to $m_\ell = -1, 0, +1$, with each m_ℓ having $m_s = \pm\frac{1}{2}$. The two electrons could pair their spins and enter a given m_ℓ state, or they could align their spins (same m_s) and enter different m_ℓ states. An experimental

fact deducted from spectroscopic studies shows that *electrons in the same n, ℓ orbitals prefer their spins to be parallel* (same m_s). This is known as **Hund's rule.**

The origin of Hund's rule can be readily understood. If electrons enter the same m_ℓ state by pairing their spins (different m_s), their quantum numbers n, ℓ, m_ℓ will be the same and they will both occupy the same region of space (same ψ_{n,ℓ,m_ℓ} orbital). They will then experience a large Coulombic repulsion and will have a large Coulombic potential energy. On the other hand, if they parallel their spins (same m_s), they will each have a different m_ℓ and will therefore occupy different regions of space (different ψ_{n,ℓ,m_ℓ} orbitals), thereby reducing their Coulombic repulsion.

The oxygen atom has eight electrons and its electronic structure is shown in Figure 3.38. The first two electrons enter the 1*s* box (orbital). The next two enter the 2*s* box. But *p* states can accommodate six electrons, so the remaining four electrons have a choice. Hund's rule forces three of the four electrons to enter the boxes corresponding to $m_\ell = -1, 0, +1$, all with their spins parallel. The last electron can go into any of the 2*p* boxes, but it has no choice for spin. It must pair its spin with the electron already in the box. Thus, the oxygen atom has two unpaired electrons in half-occupied orbitals, as indicated in Figure 3.38. Since these two unpaired electrons spin in the same direction, they give the O atom a net angular momentum. An angular momentum due to charge rotation (*i.e.*, spin) gives rise to a magnetic moment $\boldsymbol{\mu}$. If there is an external magnetic field present, then $\boldsymbol{\mu}$ experiences a force given by $\boldsymbol{\mu} \cdot d\mathbf{B}/dx$. Oxygen atoms will therefore be deflected by a nonuniform magnetic field, as experimentally observed.

Following the Pauli exclusion principle and Hund's rule, it is not difficult to build the electronic structure of various elements in the Periodic Table. There are only a few instances of unusual behavior in the energy levels of the electronic states. The 4*s* state happens to be energetically lower than the 3*d* states, so the 4*s* state fills up first. Similarly, the 5*s* state is at a lower energy than the 4*d* states. These features are summarized in the energy diagram of Figure 3.35. There is a neat shorthand way of writing the electronic structure of any atom. To each $n\ell$ state, we attach a superscript to represent the number of electrons in those $n\ell$ states. For example, for oxygen, we write $1s^2 2s^2 2p^4$, or simply $[\text{He}]2s^2 2p^4$, since $1s^2$ is a full (closed) shell corresponding to He.

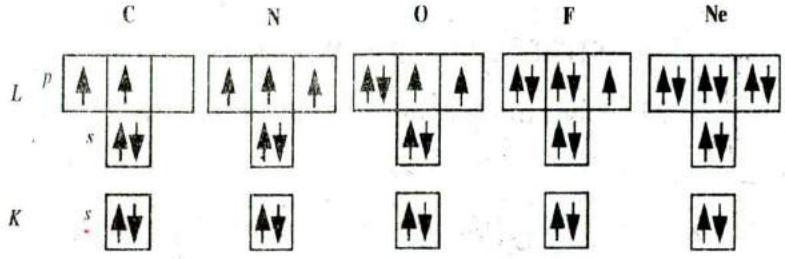


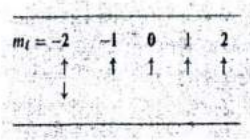
Figure 3.38 Electronic configurations for C, N, O, F, and Ne atoms.
Notice that in C, N, and O, Hund's rule forces electrons to align their spins. For the Ne atom, all the *K* and *L* orbitals are full.

EXAMPLE 3.23 **HUND'S RULE** The Fe atom has the electronic structure $[Ar]3d^64s^2$. Show that the Fe atom has four unpaired electrons and therefore a net angular momentum and a magnetic moment due to spin.

SOLUTION

In a closed subshell, for example, $2p$ subshell with six states given by $m_\ell = -1, 0, +1$ and $m_s = \pm\frac{1}{2}$, all m_ℓ and m_s values have been taken up by electrons, so each m_ℓ orbital is occupied and has paired electrons. Each positive m_ℓ (or m_s) value assigned to an electron is canceled by the negative m_ℓ (or m_s) value assigned to another electron in the subshell. Therefore, *there is no net angular momentum from a closed subshell.* Only unfilled subshells contribute to the overall angular momentum. Thus, only the six electrons in the $3d$ subshell need be considered.

There are five d orbitals, corresponding to $m_\ell = -2, -1, 0, 1, 2$. Five of the six electrons obey Hund's rule and align their spins, with each taking one of the m_ℓ values.



The sixth must take the same m_ℓ as another electron. This is only possible if they pair their spins. Consequently, there are four electrons with unpaired spins in the Fe atom, which gives the Fe atom a net angular momentum. The Fe atom therefore possesses a magnetic moment as a result of four electrons having their charges spinning in the same direction.

Many *isolated* atoms possess unpaired spins and hence also possess a magnetic moment. For example, the isolated Ag atom has one outer $5s$ electron with an unpaired spin and hence it is magnetic; it can be deflected in a magnetic field. The silver crystal, however, is nonmagnetic. In the crystal, the $5s$ electrons become detached to form the electron gas (metallic bonding) where they pair their spins, and the silver crystal has no net magnetic moment. The iron crystal is magnetic because the constituent Fe atoms retain at least two of the unpaired electron spins which then all align in the same direction to give the crystal an overall magnetic moment; iron is a magnetic metal.[10]

3.9 STIMULATED EMISSION AND LASERS

3.9.1 STIMULATED EMISSION AND PHOTON AMPLIFICATION

An electron can be excited from an energy level E_1 to a higher energy level E_2 by the absorption of a photon of energy $h\nu = E_2 - E_1$, as show in Figure 3.39a. When an electron at a higher energy level transits down in energy to an unoccupied energy level,

[10] This qualitative explanation is discussed in Chapter 8.

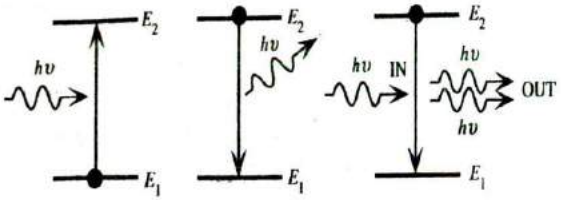


Figure 3.39 Absorption, spontaneous emission, and stimulated emission.

it emits a photon. There are essentially two possibilities for the emission process. The electron can spontaneously undergo the downward transition by itself, or it can be induced to do so by another photon.

In **spontaneous emission,** the electron falls in energy from level E_2 to E_1 and emits a photon of energy $h\nu = E_2 - E_1$, as indicated in Figure 3.39b. The transition is only spontaneous if the state with energy E_1 is not already occupied by another electron. In classical physics, when a charge accelerates and decelerates, as in an oscillatory motion, with a frequency ν, it emits an electromagnetic radiation also of frequency ν. The emission process during the transition of the electron from E_2 to E_1 appears as if the electron is oscillating with a frequency ν.

In **stimulated emission,** an incoming photon of energy $h\nu = E_2 - E_1$ stimulates the emission process by inducing the electron at E_2 to transit down to E_1. The emitted photon is in phase with the incoming photon, it is going in the same direction, and it has the same frequency, since it must also have the energy $E_2 - E_1$, as shown in Figure 3.39c. To get a feel for what is happening during stimulated emission, imagine the electric field of the incoming photon coupling to the electron and thereby driving it with the same frequency as the photon. The forced oscillation of the electron at a frequency $\nu = (E_2 - E_1)/h$ causes the electron to emit electromagnetic radiation, for which the electric field is totally in phase with that of the stimulating photon. When the incoming photon leaves the site, the electron can return to E_1, because it has emitted a photon of energy $h\nu = E_2 - E_1$.

Stimulated emission is the basis for photon amplification, since one incoming photon results in two outgoing photons, which are in phase. It is possible to achieve a practical light amplifying device based on this phenomenon. From Figure 3.39c, we see that to obtain stimulated emission, the incoming photon should not be absorbed by another electron at E_1. When we are considering using a collection of atoms to amplify light, we must therefore require that the majority of the atoms be at the energy level E_2. If this were not the case, the incoming photons would be absorbed by the atoms at E_1. When there are more atoms at E_2 than at E_1, we have what is called a **population inversion.** It should be apparent that with two energy levels, we can never achieve a population at E_2 greater than that at E_1, because, in the steady state, the incoming photon flux will cause as many upward excitations as downward stimulated emissions.

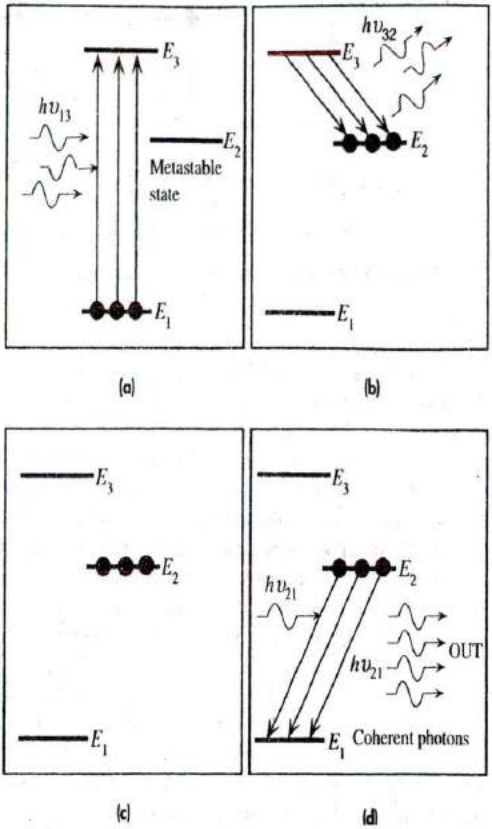
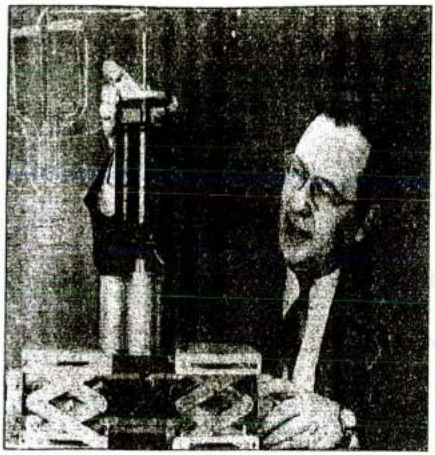


Figure 3.40 The principle of the LASER.
(a) Atoms in the ground state are pumped up to energy level E_3 by incoming photons of energy $h\nu_{13} = E_3 - E_1$.
(b) Atoms at E_3 rapidly decay to the metastable state at energy level E_2 by emitting photons or emitting lattice vibrations: $h\nu_{32} = E_3 - E_2$.
(c) Since the states at E_2 are metastable, they quickly become populated, and there is a population inversion between E_2 and E_1.
(d) A random photon of energy $h\nu_{21} = E_2 - E_1$ can initiate stimulated emission. Photons from this stimulated emission can themselves further stimulate emissions, leading to an avalanche of stimulated emissions and coherent photons being emitted.

Let us consider the three-energy-level system shown in Figure 3.40. Suppose an external excitation causes the atoms[11] in this system to become excited to energy level E_3. This is called the **pump energy level,** and the process of exciting the atoms to E_3 is called **pumping.** In the present case, **optical pumping** is used, although this is not the only means of taking the atoms to E_3. Suppose further that the atoms in E_3 decay rapidly to energy level E_2, which happens to correspond to a state that does not rapidly and spontaneously decay to a lower energy state. In other words, the state at E_2 is a **long-lived state.**[12] Quite often, the long-lived states are referred to as **metastable states.** Since the atoms cannot decay rapidly from E_2 to E_1, they accumulate at this energy level, causing a population inversion between E_2 and E_1 as pumping takes more and more atoms to E_3 and hence to E_2.

[11] An atom is in an excited state when one (or more) of its electrons is excited from the ground energy to a higher energy level. The ground state of an atom has all the electrons in their lowest energy states consistent with the Pauli exclusion principle and Hund's rule.

[12] We will not examine what causes certain states to be long lived; we will simply accept that these states do not decay rapidly and spontaneously to lower energy states.



Arthur L. Schawlow in 1961 with a ruby laser built by his Stanford group. The solid state laser was a dark ruby crystal containing Cr^{3+} ions. Lasing is obtained by stimulated emission from the Cr^{3+} ions. Arthur Schawlow won the Nobel prize in Physics in 1981 for his contribution to the development of laser spectroscopy.

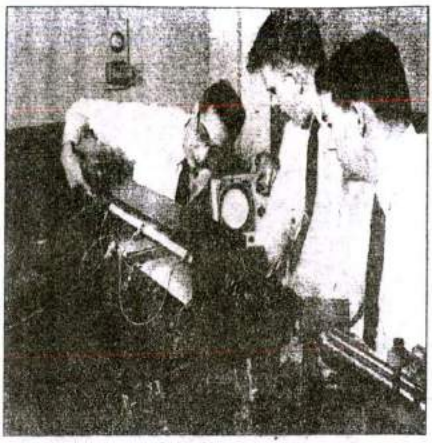
SOURCE: Stanford University, courtesy of AIP Emilio Segrè Visual Archives.

When one atom at E_2 decays spontaneously, it emits a photon, which can go on to a neighboring atom and cause that to execute stimulated emission. The photons from the latter can then go on to the next atom at E_2 and cause that atom to emit by stimulated emission, and so on. The result is an avalanche effect of stimulated emission processes with all the photons in phase, so the light output is a large collection of coherent photons. This is the principle of the ruby laser in which the energy levels E_1, E_2, and E_3 are those of the Cr^{+3} ion in the Al_2O_3 crystal. At the end of the avalanche of stimulated emission processes, the atoms at E_2 will have returned to E_1 and can be pumped again to repeat the stimulated emission cycle again. The emission from E_2 to E_1 is called the **lasing emission.**

The system we have just described for photon amplification is a **LASER,** an acronym for light amplification by stimulated emission of radiation. In the ruby laser, pumping is achieved by using a xenon flashlight. The lasing atoms are chromium ions (Cr^{3+}) in a crystal of alumina Al_2O_3 (sapphire). The ends of the ruby crystal are silvered to reflect the stimulated radiation back and forth so that its intensity builds up, in much the same way we build up voltage oscillations in an electric oscillator circuit. One of the mirrors is partially silvered to allow some of this radiation to be tapped out. What comes out is a highly coherent radiation with a high intensity. The coherency and the well-defined wavelength of this radiation are what make it distinctly different from a random stream of different-wavelength photons emitted from a tungsten bulb.

3.9.2 HELIUM–NEON LASER

With the helium–neon (HeNe) laser, the actual operation is not simple, since we need to know such things as the energy states of the whole atom. We will therefore only consider the lasing emission at 632.8 nm, which gives the well-known red color to the laser light. The actual stimulated emission occurs from the Ne atoms; He atoms are used to excite the Ne atoms by atomic collisions.



Ali Javan and his associates William Bennett Jr. and Donald Herriott at Bell Labs were first to successfully demonstrate a continuous wave (cw) helium–neon laser operation (1960).
I SOURCE: Courtesy of Bell Labs, Lucent Technologies

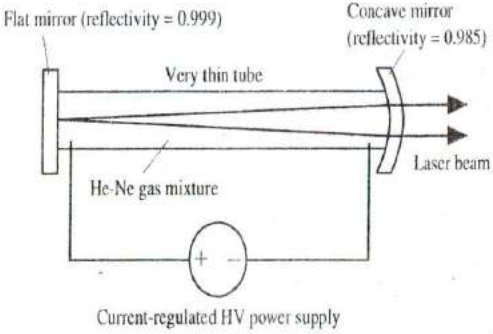


Figure 3.41 Schematic illustration of the HeNe laser.



A modern stabilized HeNe laser.
I SOURCE: Courtesy of Melles Griot.

Ne is an inert gas with a ground state ($1s^2 2s^2 2p^6$), which is represented as ($2p^6$) when the inner closed $1s$ and $2s$ subshells are ignored. If one of the electrons from the $2p$ orbital is excited to a $5s$ orbital, the excited configuration ($2p^5 5s^1$) is a state of the Ne atom that has higher energy. Similarly, He is an inert gas with the ground-state configuration of ($1s^2$). The state of He when one electron is excited to a $2s$ orbital can be represented as ($1s^1 2s^1$), which has higher energy.

The HeNe laser consists of a gaseous mixture of He and Ne atoms in a gas discharge tube, as shown schematically in Figure 3.41. The ends of the tube are mirrored to reflect the stimulated radiation and to build up the intensity within the cavity. If sufficient dc high voltage is used, electrical discharge is obtained within the tube, causing the He atoms to become excited by collisions with the drifting electrons. Thus,

$$\mathrm{He} + e^- \rightarrow \mathrm{He}^* + e^-$$

where He^* is an excited He atom.

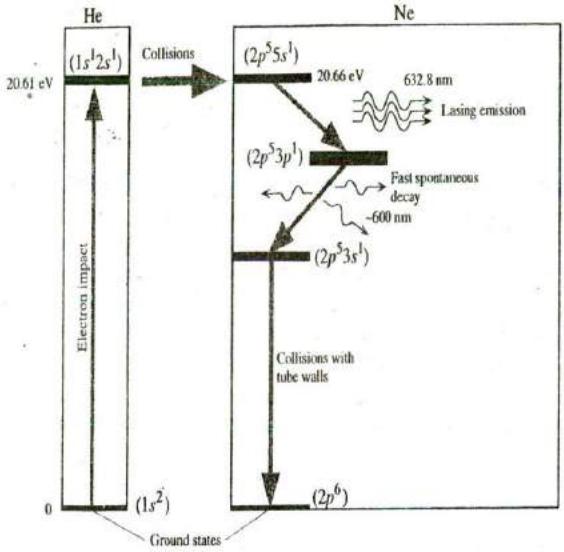


Figure 3.42 The principle of operation of the HeNe laser. Important HeNe laser energy levels (for 632.8 nm emission).

The excitation of the He atom by an electron collision puts the second electron in He into a $2s$ state, so the excited He atom, He*, has the configuration $(1s^12s^1)$. This atom is metastable (long lasting) with respect to the $(1s^2)$ state, as shown schematically in Figure 3.42. He* cannot spontaneously emit a photon and decay down to the $(1s^2)$ ground state because $\Delta\ell$ must be ± 1. Thus, a large number of He* atoms build up during the electrical discharge.

When an excited He atom collides with a Ne atom, it transfers its energy to the Ne atom by resonance energy exchange. This happens because, by good fortune, Ne has an empty energy level, corresponding to the $(2p^55s^1)$ configuration, which matches that of $(1s^12s^1)$ of He*. The collision process excites the Ne atom and de-excites He* down to its ground energy, that is,

$$\text{He}^* + \text{Ne} \rightarrow \text{He} + \text{Ne}^*$$

With many He*–Ne collisions in the gaseous discharge, we end up with a large number of Ne* atoms and a population inversion between the $(2p^55s^1)$ and $(2p^53p^1)$ states of the Ne atom, as indicated in Figure 3.42. The spontaneous emission of a photon from one Ne* atom falling from $5s$ to $3p$ gives rise to an avalanche of stimulated emission processes, which leads to a lasing emission with a wavelength of 632.8 nm, in the red.

There are a few interesting facts about the HeNe laser, some of which are quite subtle. First, the $(2p^55s^1)$ and $(2p^53p^1)$ electronic configurations of the Ne atom actually have a spread of energies. For example for Ne$(2p^55s^1)$, there are four closely spaced energy levels. Similarly, for Ne$(2p^53p^1)$, there are 10 closely separated energies. We can therefore achieve population inversion with respect to a number of energy levels. As a result, the lasing emissions from the HeNe laser contain a variety of wavelengths. The two

lasing emissions in the visible spectrum, at 632.8 nm and 543 nm, can be used to build a red or green HeNe laser. Further, we should note that the energy of the Ne($2p^5 4p^1$) state (not shown) is above that of Ne($2p^5 3p^1$) but below that of Ne($2p^5 5s^1$). Consequently, there will also be stimulated transitions from Ne($2p^5 5s^1$) to Ne($2p^5 4p^1$), and hence a lasing emission at a wavelength of ~3.39 μm infrared. To suppress lasing emissions at the unwanted wavelengths (*e.g.*, the infrared) and to obtain lasing only at the wavelength of interest, we can make the reflecting mirrors wavelength selective. This way the optical cavity builds up optical oscillations at the selected wavelength.

From ($2p^5 3p^1$) energy levels, the Ne atoms decay rapidly to the ($2p^5 3s^1$) energy levels by spontaneous emission. Most of the Ne atoms with the ($2p^5 3s^1$) configuration, however, cannot simply return to the ground state $2p^6$, because the return of the electron in 3*s* requires that its spin be flipped to close the 2*p* subshell. An electromagnetic radiation cannot change the electron spin. Thus, the Ne($2p^5 3s^1$) energy levels are metastable. The only possible means of returning to the ground state (and for the next repumping act) is collisions with the walls of the laser tube. Therefore, we cannot increase the power obtainable from a HeNe laser simply by increasing the laser tube diameter, because that will accumulate more Ne atoms at the metastable ($2p^5 3s^1$) states.

A typical HeNe laser, illustrated in Figure 3.41, consists of a narrow glass tube that contains the He and Ne gas mixture (typically, the He to Ne ratio is 10:1). The lasing emission intensity increases with tube length, since more Ne atoms are then used in stimulated emission. The intensity decreases with increasing tube diameter, since Ne atoms in the ($2p^5 3s^1$) states can only return to the ground state by collisions with the walls of the tube. The ends of the tube are generally sealed with a flat mirror (99.9 percent reflecting) at one end and, for easy alignment, a concave mirror (98.5 percent reflecting) at the other end, to obtain an optical cavity within the tube. The outer surface of the concave mirror is ground to behave like a convergent lens, to compensate for the divergence in the beam arising from reflections from the concave mirror. The output radiation from the tube is typically a beam of diameter 0.5–2 mm and a divergence of 1 milliradians at a power of a few milliwatts. In high-power HeNe lasers, the mirrors are external to the tube. In addition, Brewster windows are fused at the ends of the laser tube, to allow only polarized light to be transmitted and amplified within the cavity, so that the output radiation is polarized (that is, has electric field oscillations in one plane).

EXAMPLE 3.24 **EFFICIENCY OF THE HeNe LASER** A typical low-power 2.5 mW HeNe laser tube operates at a dc voltage of 2 kV and carries a current of 5 mA. What is the efficiency of the laser?

SOLUTION

From the definition of efficiency,

$$\text{Efficiency} = \frac{\text{Output power}}{\text{Input power}}$$

$$= \frac{(2.5 \times 10^{-3}\,\text{W})}{(5 \times 10^{-3}\,\text{A})(2000\,\text{V})} = 0.00025 \quad \text{or} \quad 0.025\%$$

3.9.3 LASER OUTPUT SPECTRUM

The output radiation from a laser is not actually at one single well-defined wavelength corresponding to the lasing transition. Instead, the output covers a spectrum of wavelengths with a central peak. This is not a simple consequence of the Heisenberg uncertainty principle (which does broaden the output). Predominantly, it is a result of the broadening of the emitted spectrum by the **Doppler effect.** We recall from the kinetic molecular theory that gas atoms are in random motion, with an average translational kinetic energy of $\frac{3}{2}kT$. Suppose that these gas atoms emit radiation of frequency ν_0 which we label as the source frequency. Then, due to the Doppler effect, when a gas atom moves toward an observer, the latter detects a higher frequency ν_2, given by

Doppler effect

$$\nu_2 = \nu_0\left(1 + \frac{v_x}{c}\right)$$

where v_x is the relative velocity of the atom with respect to the observer and c is the speed of light. When the atom moves away, the observer detects a smaller frequency, which corresponds to

Doppler effect

$$\nu_1 = \nu_0\left(1 - \frac{v_x}{c}\right)$$

Since the atoms are in random motion, the observer will detect a range of frequencies, due to this Doppler effect. As a result, the frequency or wavelength of the output radiation from a gas laser will have a "linewidth" of $\Delta\nu = \nu_2 - \nu_1$, called a Doppler-broadened **linewidth** of a laser radiation. Other mechanisms also broaden the output spectrum, but we will ignore these at present.

The reflections from the laser end mirrors give rise to traveling waves in opposite directions within the cavity. Since the waves are in phase, they interfere constructively, to set up a standing wave—in other words, stationary oscillations. Some of the energy in this wave is tapped by the 99 percent reflecting mirror to get an output, in much the same way that we tap the energy from an oscillating field in an LC circuit by attaching an antenna to it.

Only standing waves with certain wavelengths can be maintained within the optical cavity, just as only certain acoustic wavelengths can be obtained from musical instruments. Any standing wave in the cavity must have a half-wavelength $\lambda/2$ that fits into the cavity length L, or

Laser cavity modes

$$n\left(\frac{\lambda}{2}\right) = L \qquad [3.56]$$

where n is an integer called the **mode number** of the standing wave. Each possible standing wave within the laser tube (cavity) satisfying Equation 3.56 is called a **cavity mode.** The laser output thus has a broad spectrum with peaks at certain wavelengths corresponding to various cavity modes existing within the Doppler-broadened emission curve. Figure 3.43 shows the expected output from a typical gas laser. At wavelengths satisfying Equation 3.56, that is, representing certain cavity modes, we have intensity spikes in the output. The net envelope of the output

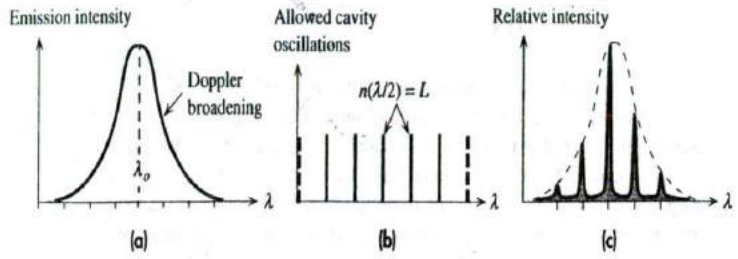


Figure 3.43

(a) Doppler-broadened emission versus wavelength characteristics of the lasing medium.

(b) Allowed oscillations and their wavelengths within the optical cavity.

(c) The output spectrum is determined by satisfying (a) and (b) simultaneously.

radiation is a Gaussian distribution, which is essentially due to the Doppler-broadened linewidth.

Even though we can try to get as parallel a beam as possible by lining the mirrors up perfectly, we will still be faced with diffraction effects at the output. When the output laser beam hits the end of the laser tube, it becomes diffracted, so the emerging beam is necessarily divergent. Simple diffraction theory can readily predict the divergence angle.

EXAMPLE 3.25 **DOPPLER-BROADENED LINEWIDTH** Calculate the Doppler-broadened linewidths $\Delta\nu$ and $\Delta\lambda$ for the HeNe laser transition $\lambda = 632.8$ nm, if the gas discharge temperature is about 127 °C. The atomic mass of Ne is 20.2 g mol^{-1}.

SOLUTION

Due to the Doppler effect, the laser radiation from gas lasers is broadened around a central frequency ν_o, which corresponds to the source frequency. Higher frequencies detected will be due to radiations emitted from atoms moving toward the observer, and lower frequencies detected will be the result of emissions from atoms moving away from the observer. Therefore, the width of the observed frequencies will be approximately

Doppler-broadened frequency width

$$\Delta\nu = \nu_o\left(1 + \frac{v_x}{c}\right) - \nu_o\left(1 - \frac{v_x}{c}\right) = \frac{2\nu_o v_x}{c}$$

From $\lambda = c/\nu$, we obtain the following by differentiation:

$$\frac{d\lambda}{d\nu} = -\frac{c}{\nu^2} = -\frac{\lambda}{\nu} = -\frac{\lambda^2}{c}$$

We need to know v_x, which is given by kinetic theory as $v_x^2 = kT/m$. For the HeNe laser, the Ne atoms lase, so

$$m = \frac{20.2 \times 10^{-3}\ \text{kg mol}^{-1}}{6.023 \times 10^{23}\ \text{mol}^{-1}} = 3.35 \times 10^{-26}\ \text{kg}$$

Thus

$$v_x = \left[\frac{(1.38 \times 10^{-23}\ \text{J K}^{-1})(127 + 273\ \text{K})}{(3.35 \times 10^{-26}\ \text{kg})}\right]^{1/2} = 406\ \text{m s}^{-1}$$

The central frequency is

$$\nu_o = \frac{c}{\lambda_o} = \frac{3 \times 10^8 \text{ m s}^{-1}}{632.8 \times 10^{-9} \text{ m}} = 4.74 \times 10^{14} \text{ s}^{-1}$$

The frequency linewidth is

$$\Delta\nu = \frac{(2\nu_o v_x)}{c} = \frac{2(4.74 \times 10^{14} \text{ s}^{-1})(406 \text{ m s}^{-1})}{3 \times 10^8 \text{ m s}^{-1}} = 1.283 \text{ GHz}$$

To get $\Delta\lambda$, we use $d\lambda/d\nu = -\lambda/\nu$, so that

$$\Delta\lambda = \Delta\nu \left| -\frac{\lambda_o}{\nu_o} \right| = \frac{(1.283 \times 10^9 \text{ Hz})(632.8 \times 10^{-9} \text{ m})}{4.74 \times 10^{14} \text{ s}^{-1}}$$

$$= 1.71 \times 10^{-12} \text{ m} \quad \text{or} \quad 0.0017 \text{ nm}$$

ADDITIONAL TOPICS

3.10 OPTICAL FIBER AMPLIFIERS

A light signal that is traveling along an optical fiber communications link over a long distance suffers marked attenuation. It becomes necessary to regenerate the light signal at certain intervals for long-haul communications over several thousand kilometers. Instead of regenerating the optical signal by photodetection, conversion to an electrical signal, amplification, and then conversion back from electrical to light energy by a laser diode, it becomes practical to amplify the signal directly by using an optical amplifier. The photons in an optical signal have a wavelength of 1550 nm, and optical amplifiers have to amplify signal photons at this wavelength.

One practical **optical amplifier** is based on the **erbium (Er^{3+} ion) doped fiber amplifier (EDFA).**[13] The core region of an optical fiber is doped with Er^{3+} ions. The host fiber core material is a glass based on SiO_3–GeO_2 and perhaps some other glass-forming oxides such as Al_2O_3. It is easily fused to a long-distance optical fiber by a technique called splicing.

When the Er^{3+} ion is implanted in the host glass material, it has the energy levels indicated in Figure 3.44 where E_1 corresponds to the lowest energy possible consistent with the Pauli exclusion principle and Hund's rule. One of the convenient energy levels for optically pumping the Er^{3+} ion is at E_3, approximately 1.27 eV above the ground energy level. The Er^{3+} ions are optically pumped, usually from a laser diode, to excite them to E_3. The wavelength for this pumping is about 980 nm. The Er^{3+} ions decay rapidly from E_3 to a **long-lived** energy level at E_2 which has a long lifetime of ~10 ms (very long on the atomic scale). The decay E_3 to E_2 involves energy losses by radiationless transitions (generation of crystal vibrations) and are very rapid. Thus, more and more Er^{3+} ions accumulate at E_2 which is 0.80 eV above the ground energy. The accumulation of Er^{3+} ions at E_2 leads to a population inversion between E_2 and E_1. Signal photons at 1550 nm have an energy of 0.80 eV, or $E_2 - E_1$, and give rise to *stimulated transitions* of Er^{3+} ions from E_2 to E_1. Any Er^{3+} ions left at E_1, however, will

[13] EDFA was first reported in 1987 by E. Desurvire, J. R. Simpson, and P. C. Becker and, within a short period, AT&T began deploying EDFA repeaters in long-haul fiber communications in 1994.

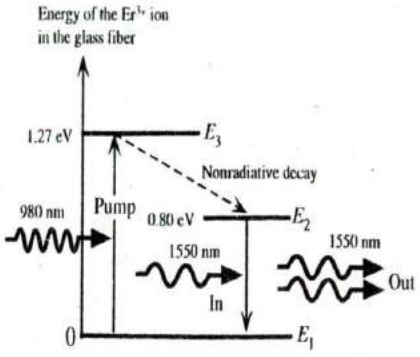


Figure 3.44 Energy diagram for the Er^{3+} ion in the glass fiber medium and light amplification by stimulated emission from E_2 to E_1.

Dashed arrows indicate radiationless transitions (energy emission by lattice vibrations).

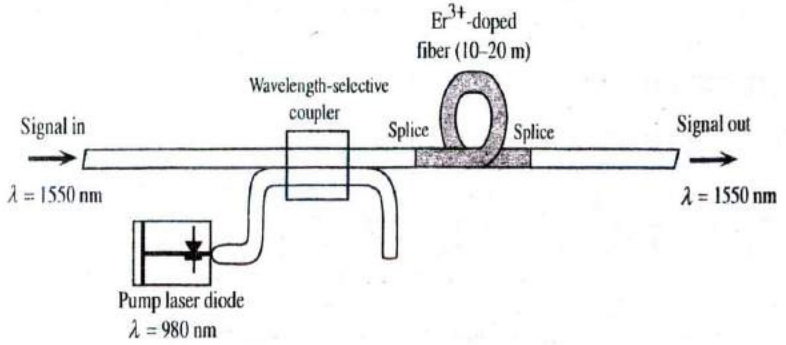


Figure 3.45 A simplified schematic illustration of an EDFA (optical amplifier).

The erbium-ion doped fiber is pumped by feeding the light from a laser pump diode, through a coupler, into the erbium-ion doped fiber.

absorb the incoming 1550 nm photons to reach E_2. To achieve light amplification we must therefore have stimulated emission exceeding absorption. This is only possible if there are more Er^{3+} ions at the E_2 level than at the E_1 level, that is, if we have population inversion. With sufficient optical pumping, population inversion is readily achieved.

In practice the erbium-doped fiber is inserted into the fiber communications line by splicing as shown in the simplified schematic diagram in Figure 3.45 and it is pumped from a laser diode through a coupling fiber arrangement which allows only the pumping wavelength to be coupled.

CD Selected Topics and Solved Problems

Selected Topics	Solved Problems
Compton Scattering	Modern Physics: Photoelectric Experiment, Ionization Energy
Stimulated Emission and Laser Principles	He–Ne Laser Problem
Stimulated Emission and Optical Amplifiers	
Time-Dependent Schrödinger Equation	

DEFINING TERMS

Angular momentum L about a point O is defined as $\mathbf{L} = \mathbf{p} \times \mathbf{r}$, where $\mathbf{p}$ is the linear momentum and $\mathbf{r}$ is the position vector of the body from O. For a circular orbit around O, the angular momentum is orbital and $L = pr = mvr$.

Bragg diffraction law describes the diffraction of an X-ray beam by a crystal in which the interplanar separation d of a given set of atomic planes causing the X-ray diffraction is related to the diffraction angle 2θ and the wavelength λ of the X-rays through $2d \sin\theta = n\lambda$, where n is an integer, usually unity.

Complementarity principle suggests that the wave model and the particle model are complementary models in that one model alone cannot be used to explain all the observations in nature. For example, the electron diffraction phenomenon is best explained by the wave model, whereas in the Compton experiment, the electron is treated as a particle; that is, it is deflected by an impinging photon that imparts an additional momentum to the electron.

Compton effect is the scattering of a high-energy photon by a "free" electron. The effect is experimentally observed when an X-ray beam is scattered from a target that contains many conduction ("free") electrons, such as a metal or graphite.

De Broglie relationship relates the wave-like properties (*e.g.*, wavelength λ) of matter to its particle-like properties (*e.g.*, momentum p) via $\lambda = h/p$.

Diffraction is the bending of waves as a result of the interaction of the waves with an object of size comparable to the wavelength. If the object has a regular pattern, periodicity, an incident beam of waves can be bent (diffracted) in certain well-defined directions that depend on the periodicity, which is used in the X-ray diffraction study of crystals.

Doppler effect is the change in the measured frequency of a wave due to the motion of the source relative to the observer. In the case of electromagnetic radiation, if v is the relative velocity of the source object toward the observer and ν_o is the source frequency, then the measured electromagnetic wave frequency is $\nu = \nu_o[1 + (v/c)]$ for $(v/c) \ll 1$.

Energy density ρ_E is the amount of energy per unit volume. In a region where the electric field is $\mathcal{E}$, the energy stored per unit volume is $\frac{1}{2}\varepsilon_0\mathcal{E}^2$.

Flux is a term used to describe the rate of flow through a unit area. If ΔN is the number of particles flowing through an area A in time Δt, then particle flux Γ is defined as $\Gamma = \Delta N/(A\Delta t)$. If an amount of energy ΔE flows through an area A in time Δt, energy flux is $\Gamma_E = \Delta E/(A\Delta t)$, which defines the intensity (I) of an electromagnetic wave.

Flux in radiometry is the flow of radiation (electromagnetic wave) energy per unit time in watts. It is simply the radiation power that is flowing. In contrast, the photon or particle flux refers to the number of photons or particles flowing per unit time per unit area. **Radiant flux emitted** by a source refers to the radiation power in watts that is emitted. Flux in radiometry normally has either *radiant* or *luminous* as an adjective, *e.g.*, radiant flux, luminous flux.

Ground state is the state of the electron with the lowest energy.

Heisenberg's uncertainty principle states that the uncertainty Δx in the position of a particle and the uncertainty Δp_x in its momentum in the x direction obey $(\Delta x)(\Delta p_x) \gtrsim \hbar$. This is a consequence of the wave nature of matter and has nothing to do with the precision of measurement. If ΔE is the uncertainty in the energy of a particle during a time Δt, then according to the uncertainty principle, $(\Delta E)(\Delta t) \gtrsim \hbar$. To measure the energy of a particle without any uncertainty means that we would need an infinitely long time $\Delta t \to \infty$.

Hund's rule states that electrons in a given subshell $n\ell$ try to occupy separate orbitals (different m_ℓ) and keep their spins parallel (same m_s). In doing so, they achieve a lower energy than pairing their spins (different m_s) and occupying the same orbital (same m_ℓ).

Intensity (I) is the flow of energy per unit area per unit time. It is equal to an energy flux.

LASER (light amplification by stimulated emission of radiation) is a device within which photon multiplication by stimulated emission produces an

output radiation that is nearly monochromatic and coherent (vis-à vis an incoherent stream of photons from a tungsten light bulb). Furthermore, the output beam has very little divergence.

Luminous flux or power Φ_v is a measure of flow of "visual energy" per unit time that takes into account the wavelength dependence of the efficiency of the human eye, that is, whether the energy that is flowing is perceptible to the human eye. It is a measure of "brightness." One lumen of luminous flux is obtained from a 1.58 mW light source emitting a single wavelength of 555 nm (green).

Magnetic quantum number m_ℓ specifies the component of the orbital angular momentum L_z in the direction of a magnetic field along z so that $L_z = \pm\hbar m_\ell$, where m_ℓ can be a negative or positive integer from $-\ell$ to $+\ell$ including 0, that is, $-\ell, -(\ell-1), \ldots, 0, \ldots, (\ell-1), \ell$. The orbital ψ of the electron depends on m_ℓ, as well as on n and ℓ. The m_ℓ, however, generally determines the angular variation of ψ.

Orbital is a region of space in an atom or molecule where an electron with a given energy may be found. Two electrons with opposite spins can occupy the same orbital. An orbit is a well-defined path for an electron, but it cannot be used to describe the whereabouts of the electron in an atom or molecule, because the electron has a probability distribution. The wavefunction $\psi_{n\ell m_\ell}(r, \theta, \phi)$ is often referred to as an orbital that represents the spatial distribution of the electron, since $|\psi_{n\ell m_\ell}(r, \theta, \phi)|^2$ is the probability of finding the electron per unit volume at (r, θ, ϕ).

Orbital (angular momentum) quantum number specifies the magnitude of the orbital angular momentum of the electron via $L = h\sqrt{[\ell(\ell+1)]}$, where ℓ is the orbital quantum number with values $0, 1, 2, 3, \ldots, n-1$. The ℓ values 0, 1, 2, 3 are labeled the s, p, d, f states.

Orbital wavefunction describes the spatial dependence of the electron, not its spin. It is $\psi(r, \theta, \phi)$, which depends on n, ℓ, and m_ℓ, with the spin dependence m_s excluded. Generally, $\psi(r, \theta, \phi)$ is simply called an orbital.

Pauli exclusion principle requires that no two electrons in a given system may have the same set of quantum numbers, n, ℓ, m_ℓ, m_s. In other words, no two electrons can occupy a given state $\psi(n, \ell, m_\ell, m_s)$. Equivalently, up to two electrons with opposite spins can occupy a given orbital $\psi(n, \ell, m_\ell)$.

Photoelectric effect is the emission of electrons from a metal upon illumination with a frequency of light above a critical value which depends on the material. The kinetic energy of the emitted electron is independent of the light intensity and dependent on the light frequency ν, via $KE = h\nu - \Phi$ where h is Planck's constant and Φ is a material-related constant called the **work function.**

Photon is a quantum of energy $h\nu$ (where h is Planck's constant and ν is the frequency) associated with electromagnetic radiation. A photon has a zero rest mass and a momentum p given by the de Broglie relationship $p = h/\lambda$, where λ the wavelength. A photon does have a "moving mass" of $h\nu/c^2$, so it experiences gravitational attraction from other masses. For example, light from a star gets deflected as it passes by the sun.

Population inversion is the phenomenon of having more atoms occupy an excited energy level E_2, higher than a lower energy level, E_1, which means that the normal equilibrium distribution is reversed; that is, $N(E_2) > N(E_1)$. Population inversion occurs temporarily as a result of the excitation of a medium (pumping). If left on its own, the medium will eventually return to its equilibrium population distribution, with more atoms at E_1 than at E_2. For gas atoms, this means $N(E_2)/N(E_1) \approx \exp[-(E_2 - E_1)/kT]$.

Principal quantum number n is an integer quantum number with values $1, 2, 3, \ldots$ that characterizes the total energy of an electron in an atom. The energy increases with n. With the other quantum numbers ℓ and m_ℓ, n determines the orbital of the electron in an atom, or $\psi_{n\ell m_\ell}(r, \theta, \phi)$. The values $n = 1, 2, 3, 4, \ldots$ are labeled the $K, L, M, N, \ldots$ shells, within each of which there may be subshells based on $\ell = 0, 1, 2, \ldots (n-1)$ and corresponding to the $s, p, d, \ldots$ states.

Pumping means exciting atoms from their ground states to higher energy states.

Radiant is a common adjective used to imply the involvement of radiation, that is, electromagnetic waves, in the noun that it qualifies; *e.g., radiant energy* is the energy transmitted by radiation.

Radiant power is radiation energy flowing, or emitted from a source, per unit time, which is also known as **optical power** even if the wavelength is not within the visible spectrum. **Radiant flux** signifies radiant power flow in radiometry, measured in watts.

Radiation normally signifies a traveling electromagnetic wave that is carrying energy. Due to the particle-like behavior of waves, radiation can also mean a *stream of photons*.

Schrödinger equation is a fundamental equation in nature, the solution of which describes the wave-like behavior of a particle. The equation cannot be derived from a more fundamental law. Its validity is based on its ability to predict any known physical phenomena. The solution requires as input the potential energy function $V(x, y, z, t)$ of the particle and the boundary and initial conditions. The *PE* function $V(x, y, z, t)$ describes the interaction of the particle with its environment. The time-independent Schrödinger equation describes the wave behavior of a particle under steady-state conditions, that is, when the *PE* is time-independent $V(x, y, z)$. If E is the total energy and $\nabla^2 = (\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2)$, then

$$\nabla^2\psi + \left(\frac{2m}{\hbar^2}\right)[E - V(x, y, z)]\psi = 0$$

The solution of the time-independent Schrödinger equation gives the wavefunction $\psi(x, y, z)$ of the electron and its energy E. The interpretation of the wavefunction $\psi(x, y, z)$ is that $|\psi(x, y, z)|^2$ is the probability of finding the electron per unit volume at point x, y, z.

Selection rules determine what values of ℓ and m_ℓ are allowed for an electron transition involving the emission and absorption of electromagnetic radiation, that is, a photon. In summary, $\Delta\ell = \pm 1$ and $\Delta m_\ell = 0, \pm 1$. The spin number m_s of the electron remains unchanged. Within an atom, the transition of the electron from one state $\psi(n, \ell, m_\ell, m_s)$ to another $\psi(n', \ell', m'_\ell, m'_s)$, due to collisions with other atoms or electrons, does not necessarily obey the selection rules.

Spin of an electron S is its intrinsic angular momentum (analogous to the spin of Earth around its own axis), which is space quantized to have two possibilities. The magnitude of the electron's spin is a constant, $\hbar\sqrt{3}/2$, but its component along a magnetic field in the z direction is $m_s\hbar$, where m_s is the **spin magnetic quantum number,** which is $+\frac{1}{2}$ or $-\frac{1}{2}$.

Spontaneous emission is the phenomenon in which a photon is emitted when an electron in a high energy state $\psi(n, \ell, m_\ell, m_s)$ with energy E_2 spontaneously falls down to a lower, unoccupied energy state $\psi(n', \ell', m'_\ell, m'_s)$ with energy E_1. The photon energy is $h\nu = (E_2 - E_1)$. Since the emitted photon has an angular momentum, the orbital quantum number ℓ of the electron must change, that is $\Delta\ell = \ell' - \ell = \pm 1$.

State is a possible wavefunction for the electron that defines its spatial (orbital) and spin properties. For example, $\psi(n, \ell, m_\ell, m_s)$ is a state of the electron. From the Schrödinger equation, each state corresponds to a certain electron energy E. We use the terms state of energy E, or *energy state*. There is generally more than one state ψ with the same energy E.

Stimulated emission is the phenomenon in which an incoming photon of energy $h\nu = E_2 - E_1$ interacts with an electron in a high-energy state $\psi(n, \ell, m_\ell, m_s)$ at E_2, and induces that electron to oscillate down to a lower, unoccupied energy state, $\psi(n', \ell', m'_\ell, m'_s)$ at E_1. The photon emitted by stimulation has the same energy and phase as the incoming photon, and it moves in the same direction. Consequently, stimulated emission results in two coherent photons with the same energy, traveling in the same direction. The stimulated emission process must obey the selection rule $\Delta\ell = \ell' - \ell = \pm 1$, just as spontaneous emission must.

Tunneling is the penetration of an electron through a potential energy barrier by virtue of the electron's wave-like behavior. In classical mechanics, if the energy E of the electron is less than the *PE* barrier V_o, the electron cannot cross the barrier. In quantum mechanics, there is a distinct probability that the electron will "tunnel" through the barrier to appear on the other side. The probability of tunneling depends very strongly on the height and width of the *PE* barrier.

Wave is a periodically occurring disturbance, such as the displacement of atoms from their equilibrium positions in a solid carrying sound waves, or a periodic variation in a measurable quantity, such as the electric

field $\mathcal{E}(x, t)$ in a medium or space. In a traveling wave, energy is transferred from one location to another by the oscillations. For example, $\mathcal{E}_y(x, t) = \mathcal{E}_0 \sin(kx - \omega t)$ is a traveling wave in the x direction, where $k = 2\pi/\lambda$ and $\omega = 2\pi\nu$. The electric field in the y direction varies periodically along x, with a period λ called the wavelength. The field also varies with time, with a period $1/\nu$, where ν is the frequency. The wave propagates along the x direction with a velocity of propagation c. Electromagnetic waves are transverse waves in which the electric and magnetic fields $\mathcal{E}_y(x, t)$ and $B_z(x, t)$ are at right angles to each other, as well as to the direction of propagation x. A traveling wave in the electric field must be accompanied by a similar traveling wave in the associated magnetic field $B_z(x, t) = B_{z0} \sin(kx - \omega t)$. Typical wave-like properties are interference and diffraction.

Wave equation is a general partial differential equation in classical physics, of the form

$$v^2 \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial t^2} = 0$$

the solution of which describes the space and time dependence of the displacement $u(x, t)$ from equilibrium or zero, given the boundary conditions. The parameter v in the wave equation is the propagation velocity of the wave. In the case of electromagnetic waves in a vacuum, the wave equation describes the variation of the electric (or magnetic) field $\mathcal{E}(x, t)$ with space and time, $(c^2 \partial^2 \mathcal{E}/\partial x^2) - (\partial^2 \mathcal{E}/\partial t^2) = 0$, where c is the speed of light.

Wavefunction $\Psi(x, y, z, t)$ is a probability-based function used to describe the wave-like properties of a particle. It is obtained by solving the Schrödinger equation, which in turn requires a knowledge of the *PE* of the particle and the boundary and initial conditions. The term $|\Psi(x, y, z, t)|^2$ is the probability per unit volume of finding the electron at (x, y, z) at time t. In other words, $|\Psi(x, y, z, t)|^2\, dx\, dy\, dz$ is the probability of finding the electron in the small volume $dx\, dy\, dz$ at (x, y, z) at time t. Under steady-state conditions, the wavefunction can be separated into a space-dependent component and a time-dependent component, *i.e.*, $\Psi(x, y, z, t) = \psi(x, y, z) \exp(-jEt/\hbar)$, where E is the energy of the particle and $\hbar = h/2\pi$. The spatial component $\psi(x, y, z)$ satisfies the time-independent Schrödinger equation.

Wavenumber (or wavevector) k is the number of waves per 2π of length, that is, $k = 2\pi/\lambda$.

Work function is the minimum energy required to remove an electron from inside a metal to vacuum.

X-rays are electromagnetic waves of wavelength typically in the range 10 pm–1 nm, which is shorter than ultraviolet light wavelengths. X-rays can be diffracted by crystals due to their wave-like properties.

QUESTIONS AND PROBLEMS

3.1 Photons and photon flux

a. Consider a 1 kW AM radio transmitter at 700 kHz. Calculate the number of photons emitted from the antenna per second.

b. The average intensity of sunlight on Earth's surface is about 1 kW m^{-2}. The maximum intensity is at a wavelength around 800 nm. Assuming that all the photons have an 800 nm wavelength, calculate the number of photons arriving on Earth's surface per unit time per unit area. What is the magnitude of the electric field in the sunlight?

c. Suppose that a solar cell device can convert each sunlight photon into an electron, which can then give rise to an external current. What is the maximum current that can be supplied per unit area (m^2) of this solar cell device?

3.2 Yellow, cyan, magenta, and white *Three primary colors*, red, green, and blue (RGB), can be added together in various proportions to generate any color on various displays and light emitting devices in what is known as the *additive theory of color*. For example, yellow can be generated from adding red and green, cyan from blue and green, and magenta from red and blue.

a. A device engineer wants to use three light emitting diodes (LEDs) to generate various colors in an LED-based color display that is still in the research stage. His three LEDs have wavelengths of 660 nm for red, 563 nm for green, and 450 nm for blue. He simply wishes to generate the yellow and cyan by mixing equal optical powers from these LEDs; *optical power*, or *radiant power*, is defined as the radiation energy emitted per unit time. What are the numbers of red and blue photons needed (to the nearest integer) to generate yellow and cyan, respectively, for every 100 green photons?

b. An equi-energy white light is generated by mixing red, green, and blue light in equal optical powers. Suppose that the wavelengths are 700 nm for red, 546 nm for green, and 436 nm for blue (which is one set of possible standard primary colors). Suppose that the optical power in each primary color is 0.1 W. Calculate the *total photon flux* (photons per second) needed from each primary color.

c. There are bright white LEDs on the market that generate the white light by mixing yellow (a combination of red and green) with blue emissions. The inexpensive types use a single blue LED to generate a strong blue radiation, some of which is absorbed by a phosphor in front of the LED which then emits yellow light. The yellow and the blue passing through the phosphor mix and make up the white light. In one type of white LED, the blue and yellow wavelengths are 450 nm and 564 nm, respectively. White light can be generated by setting the optical (radiative) power ratio of yellow to blue light emerging from the LED to be about 1.74. What is the ratio of the number of blue to yellow photons needed? (Sometimes the mix is not perfect and the white LED light tends to have a noticeable slight blue tint.) If the total optical power output from the white LED is 100 mW, calculate the blue and yellow total photon fluxes (photons per second).

3.3 **Brightness of laser pointers** The brightness of a light source depends not only on the radiation (optical) power emitted by the source but also on its wavelength because the human eye perceives each wavelength with a different efficiency. The visual "brightness" of a source as observed by an average daylight-adapted eye is proportional to the radiation power emitted, called the *radiant flux* Φ_e, and the efficiency of the eye to detect the spectrum of the emitted radiation. While the eye can see a red color source, it cannot see an infrared source and the brightness of the infrared source would be zero. The **luminous flux** Φ_v is a measure of *brightness*, in lumens (lm), and is defined by

Luminous flux, brightness

$$\Phi_v = \Phi_e \times (633 \text{ lm W}^{-1}) \times \eta_{\text{eye}} \qquad [3.57]$$

where Φ_e is the radiant flux or the radiation power emitted (in watts) and $\eta_{\text{eye}} = \eta_{\text{eye}}(\lambda)$ is the *relative luminous efficiency* (or the relative sensitivity) of an average light-adapted eye which depends on the wavelength; η_{eye} is a Gaussian looking function with a peak of unity at 555 nm. (See Figure 3.46 for η_{eye} vs. λ.) One lumen of luminous flux, or brightness, is obtained from a 1.58 mW light source emitting at a single wavelength of 555 nm (green). A typical 60 W incandescent lamp provides roughly 900 lm. When we buy a light bulb, we are buying lumens. Consider one 5 mW red 650 nm laser pointer, and another weaker 2 mW green 532 nm laser: $\eta_{\text{eye}}(650 \text{ nm}) = 0.11$ and $\eta_{\text{eye}}(532 \text{ nm}) = 0.86$. Find the luminous flux (brightness) of each laser pointer. Which is brighter? Calculate the number of photons emitted per unit time, the total *photon flux*, by each laser.

3.4 **Human eye** Photons passing through the pupil are focused by the lens onto the retina of the eye and are detected by two types of photosensitive cells, called *rods* and *cones*, as visualized in Figure 3.46. Rods are highly sensitive photoreceptors with a peak response at a wavelength of about 507 nm (green-cyan). They do not register color and are responsible for our vision under dimmed light conditions, termed **scotopic vision.** Cones are responsible for our color perception and daytime vision, called **photopic vision.** These three types of cone photoreceptors are sensitive to blue, green, and red at wavelengths, respectively, of 430 nm, 535 nm, and 575 nm. All three cones have an overall peak response at 555 nm (green), which represents the peak response of an average daylight-adapted eye or in our photopic vision.

a. Calculate the photon energy (in eV) for the peak responsivity for each of the photoreceptors in the eye (one rod and three cones).

b. Various experiments (the most well known being by Hecht et al., *J. Opt. Soc. America*, **38**, 196, 1942) have tested the threshold sensitivity of the dark-adapted eye and have estimated that visual

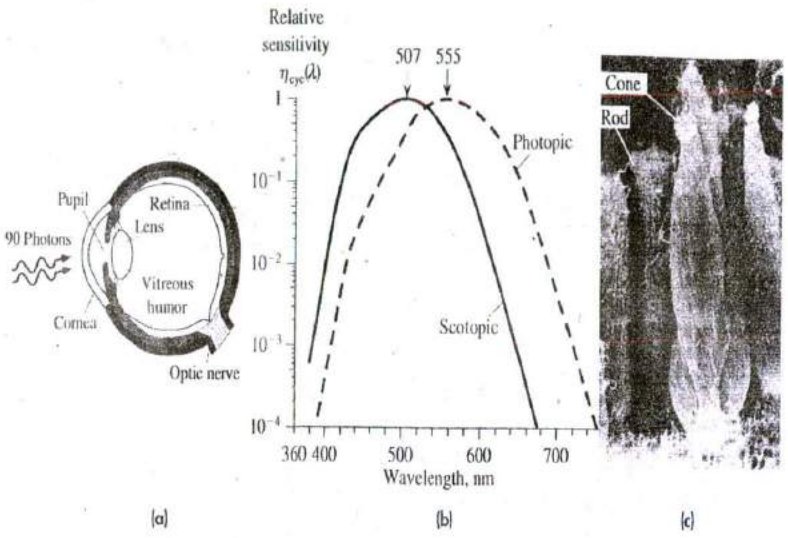


Figure 3.46

(a) The retina in the eye has photoreceptors that can sense the incident photons on them and hence provide necessary visual perception signals. It has been estimated that for minimum visual perception there must be roughly 90 photons falling on the cornea of the eye.

(b) The wavelength dependence of the relative efficiency $\eta_{eye}(\lambda)$ of the eye is different for daylight vision, or *photopic* vision (involves mainly cones), and for vision under dimmed light, or *scotopic* vision, which represents the dark-adapted eye, and involves rods.

(c) SEM photo of rods and cones in the retina.

| SOURCE: Dr. Frank Werblin, University of California, Berkeley.

perception requires a minimum of roughly 90 photons to be incident onto the cornea in front of the eye's pupil and within 1/10 second. Taking 90 incident photons every 100 ms as the threshold sensitivity, calculate the total photon flux (photons per second), total energy in eV (within 100 ms), and the optical power that is needed for threshold visual perception.

c. Not all photons incident on the eye make it to the actual photoreceptors in the retina. It has been estimated that only 1 in 10 photons arriving at the eye's cornea actually make it to rod photoreceptors, due to various reflections and absorptions in the eye and other loss mechanisms. Thus, only nine photons make it to photoreceptors on the retina.[14] It is estimated that the nine test photons fall randomly onto a circular area of about 0.0025 mm^2. What is the estimated threshold intensity for visual perception? If there are 150,000 rods mm^{-2} in this area of the eye, estimate the number of rods in this test spot. If there are a large number of rods, more than 100 in this spot, then it is likely that no single rod receives more than one photon since the nine photons arrive randomly. Thus, a rod must be able to *sense* a single photon, but it takes nine excited rods, somehow summed up by the visual system, to generate the visual sensation. Do you agree with the latter conclusion?

d. It is estimated that at least 200,000 photons per second must be incident on the eye to generate a color sensation by exciting the cones. Assuming that this occurs at the peak sensitivity at 555 nm,

[14] Sometimes one comes across a statement that the eye can detect a single photon. While a rod photoreceptor can indeed sense a single photon (or, put differently, a photon can activate a single rod), the overall human visual perception needs roughly nine photons at around 507 nm to consciously register a visual sensation.

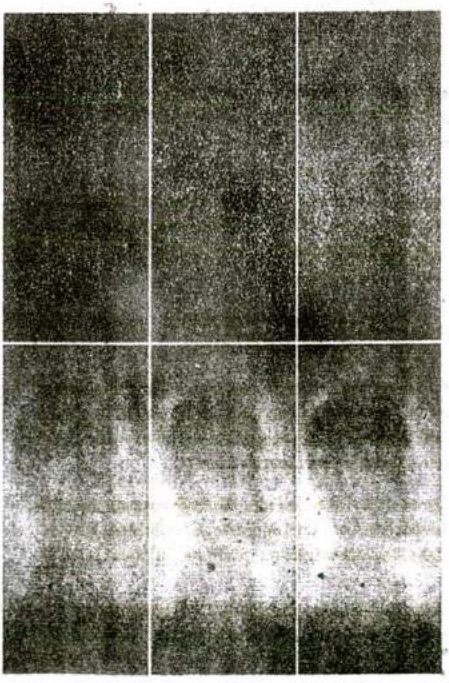
and that as in part (b) only about 10 percent of the photons make it to the retina, estimate the threshold optical power stimulating the cones in the retina.

3.5 **X-ray photons** In *chest radiology*, a patient's chest is exposed to X-rays, and the X-rays passing through the patient are recorded on a photographic film to generate an X-ray image of the chest for medical diagnosis. The average wavelength of X-rays in chest radiology is about 0.2 Å (0.02 nm). Numerous measurements indicate that the patient, on average, is exposed to a total radiation energy per unit area of roughly 0.1 μJ cm^{-2} for one chest X-ray image. Find the photon energy used in chest radiology, and the average number of photons incident on the patient per unit area (per cm^2).

***3.6** **X-rays, exposure, and roentgens** X-rays are widely used in many applications such as medical imaging, security scans, X-ray diffraction studies of crystals, and for examining defects such as cracks in objects and structures. X-rays are highly energetic photons that can easily penetrate and pass though various objects. Different materials attenuate X-rays differently, so when X-rays are passed through an object, the emerging X-rays can be recorded on a photographic film, or be captured by a modern flat panel X-ray image detector, to generate an X-ray image of the interior of the object; this is called **radiography**. X-rays also cause ionization in a medium and hence are known as *ionization radiation*. The amount of exposure (denoted by X) to X-rays, ionizing radiation, is measured in terms of the ionizing effects of the X-ray photons. One **roentgen** (1 R) is defined as an X-ray exposure that ionizes 1 cm^3 of air to generate 0.33 nC of charge in this volume at standard temperature and pressure (STP). When a body is exposed to X-rays, it will receive a certain amount of radiation energy per unit area, called **energy fluence** Ψ_E, that is, so many joules per cm^2, that depends on the exposure X. If X in roentgens is the exposure, then the energy fluence is given by

$$\Psi_E = \left[\frac{8.73 \times 10^{-6}}{\mu_{en,air}/\rho_{air}}\right] X \quad \text{J cm}^{-2} \qquad [3.58]$$

Fluence and roentgens



X-ray image of an American one-cent coin captured using an X-ray a-Se HARP camera. The first image at the top left is obtained under extremely low exposure, and the subsequent images are obtained with increasing exposure of approximately one order of magnitude between each image. The slight attenuation of the X-ray photons by Lincoln provides the image. The image sequence clearly shows the discrete nature of X-rays, and hence their description in terms of photons.

SOURCE: Courtesy of Dylan Hunt and John Rowlands, Sunnybrook Hospital, University of Toronto.

where Ψ_E is in J cm^{-2}, and $\mu_{en,air}/\rho_{air}$ is the *mass energy absorption coefficient* of air in cm^2 g^{-1} at the photon energy E_{ph} of interest; the $\mu_{en,air}/\rho_{air}$ values are listed in radiological tables. For example, for 1 R of exposure, $X = 1$, $E_{ph} = 20$ keV, and $\mu_{en,air}/\rho_{air} = 0.539$ cm^2 g^{-1}. Equation 3.58 gives $\Psi_E = 1.62 \times 10^{-5}$ J cm^{-2} incident on the object.

a. In mammography (X-ray imaging of the breasts for breast cancer), the average photon energy is about 20 keV, and the X-ray mean exposure is 12 mR. At $E_{ph} = 20$ keV, $\mu_{en,air}/\rho_{air} = 0.539$ cm^2 g^{-1}. Find the mean energy incident per unit area in μJ cm^{-2}, and the mean number of X-ray photons incident per unit area (photons cm^{-2}), called **photon fluence** Φ.

b. In chest radiography, the average photon energy is about 60 keV, and the X-ray mean exposure is 300 μR. At $E_{ph} = 60$ keV, $\mu_{en,air}/\rho_{air} = 0.0304$ cm^2 g^{-1}. Find the mean energy incident per unit area in μJ cm^{-2}, and the mean number of X-ray photons incident per unit area.

c. A modern flat panel *X-ray image detector* is a large area image sensor that has numerous arrays of tiny pixels (millions) all tiled together to make one large continuous image sensor. Each pixel is an independent X-ray detector and converts the X-rays it receives to an electrical signal. Each tiny detector is responsible for capturing a small pixel of the whole image. (Typically, the image resolution is determined by the detector pixel size.) Each pixel in a particular experimental chest radiology X-ray sensor is 150 μm × 150 μm. If the mean exposure is 300 μR, what is the number of photons received by each pixel detector? If each pixel is required to have at least 10 photons for an acceptable signal-to-noise ratio, what is the minimum exposure required in μR?

3.7 **Photoelectric effect** A photoelectric experiment indicates that violet light of wavelength 420 nm is the longest wavelength radiation that can cause the photoemission of electrons from a particular multi-alkali photocathode surface.

a. What is the work function of the photocathode surface, in eV?

b. If a UV radiation of wavelength 300 nm is incident upon the photocathode surface, what will be the maximum kinetic energy of the photoemitted electrons, in eV?

c. Given that the UV light of wavelength 300 nm has an intensity of 20 mW cm^{-2}, if the emitted electrons are collected by applying a positive bias to the opposite electrode, what will be the photoelectric current density in mA cm^{-2}?

3.8 **Photoelectric effect and quantum efficiency** Cesium metal is to be used as the photocathode material in a photoemissive electron tube because electrons are relatively easily removed from a cesium surface. The work function of a clean cesium surface is 1.9 eV.

a. What is the longest wavelength of radiation which can result in photoemission?

b. If blue radiation of wavelength 450 nm is incident onto the Cs photocathode, what will be the kinetic energy of the photoemitted electrons in eV? What should be the voltage required on the opposite electrode to extinguish the external photocurrent?

c. **Quantum efficiency** (QE) of a photocathode is defined by,

Quantum efficiency definition

$$\text{Quantum efficiency} = \frac{\text{Number of photoemitted electrons}}{\text{Number of incident photons}} \qquad [3.59]$$

QE is 100 percent if each incident photon ejects one electron. Suppose that blue light of wavelength 450 nm with an intensity of 30 mW cm^{-2} is incident on a Cs photocathode that is a circular disk of diameter 6 mm. If the emitted electrons are collected by applying a positive bias voltage to the anode, and the photocathode has a QE of 25 percent, what will be the photoelectric current?

3.9 **Photoelectric effect** A multi-alkali metal alloy is to be used as the photocathode material in a photoemissive electron tube. The work function of the metal is 1.6 eV, and the photocathode area is 0.5 cm^2. Suppose that blue light of wavelength 420 nm with an intensity of 50 mW cm^{-2} is incident on the photocathode.

a. If the photoemitted electrons are collected by applying a positive bias to the anode, what will be the photoelectric current density assuming that the quantum efficiency η is 15 percent? *Quantum efficiency*

as a percentage is the number of photoemitted electrons per 100 absorbed photons and is defined in Equation 3.59. What is the kinetic energy of a photoemitted electron at 420 nm?

b. What should be the voltage and its polarity to extinguish the current?

c. What should be the intensity of an incident red light beam of wavelength 600 nm that would give the same photocurrent if the quantum efficiency is 5 percent at this wavelength? (Normally the quantum efficiency depends on the wavelength.)

***3.10 Planck's law and photon energy distribution of radiation** Planck's law, stated in Equation 3.9, provides the spectral distribution of the black body radiation intensity in terms of wavelength through I_λ, intensity per unit wavelength. Suppose that we wish to find the distribution in terms of frequency ν or photon energy $h\nu$. Frequency $\nu = c/\lambda$ and the wavelength range λ to $\lambda + d\lambda$ corresponds to a frequency range ν to $\nu + d\nu$. ($d\lambda$ and $d\nu$ have opposite signs since ν increases as λ decreases.) The intensity $I_\lambda\, d\lambda$ in λ to $\lambda + d\lambda$ must be the same as the intensity in ν to $\nu + d\nu$, which we can write as $I_\nu\, d\nu$ where I_ν is the radiation intensity per unit frequency. Thus,

$$I_\nu = I_\lambda \left| \frac{d\lambda}{d\nu} \right|$$

The magnitude sign is needed because $\lambda = c/\nu$ results in a negative $d\lambda/d\nu$, and I_ν must be positive by definition. We can simply substitute $\lambda = c/\nu$ for λ in I_λ and obtain I_λ as a function of ν, and then find $|d\lambda/d\nu|$ to find I_ν from the preceding expression.

a. Show that

$$I_\nu = \frac{2\pi(h\nu)^3}{c^2h^2[\exp(h\nu/kT) - 1]} \qquad [3.60]$$

Black body photon energy distribution

Equation 3.60 is written to highlight that it is a function of the *photon energy* $h\nu$, which is in joules in Equation 3.60 but can be converted to eV by dividing by 1.6×10^{-19} J eV^{-1}.

b. If we integrate I_ν over all photon energies (numerically on a calculator or a computer from 0 to say 6 eV), we would obtain the total intensity at a temperature T. Find the total intensity I_T emitted at $T = 2600$ K (a typical incandescent light bulb filament temperature) and at 6000 K (roughly representing the sun's spectrum). Plot $y = I_\nu/I_T$ versus the photon energy in eV. What are the photon energies for the peaks in the distributions? Calculate the corresponding wavelength for each using $\lambda = c/\nu$ and then compare these wavelengths with those predicted by Wien's law, $\lambda_{max}T \approx 2.89 \times 10^{-3}$ m K.

3.11 Wien's law The maximum in the intensity distribution of black body radiation depends on the temperature. Substitute $x = hc/(\lambda kT)$ in Planck's law and plot I_λ versus x and find λ_{max} which corresponds to the peak of the distribution, and hence derive Wien's law. Find the peak intensity wavelength λ_{max} for a 40 W light bulb given that its filament operates at roughly 2400 °C.

3.12 Diffraction by X-rays and an electron beam Diffraction studies on a polycrystalline Al sample using X-rays gives the smallest diffraction angle (2θ) of 29.5° corresponding to diffraction from the (111) planes. The lattice parameter a of Al (FCC), is 0.405 nm. If we wish to obtain the same diffraction pattern (same angle) using an electron beam, what should be the voltage needed to accelerate the electron beam? Note that the interplanar separation d for planes (h, k, ℓ) and the lattice parameter a for cubic crystals are related by $d = a/(h^2 + k^2 + \ell^2)^{1/2}$.

3.13 Heisenberg's uncertainty principle Show that if the uncertainty in the position of a particle is on the order of its de Broglie wavelength, then the uncertainty in its momentum is about the same as the momentum value itself.

3.14 Heisenberg's uncertainty principle An excited electron in an Na atom emits radiation at a wavelength 589 nm and returns to the ground state. If the mean time for the transition is about 20 ns, calculate the inherent width in the emission line. What is the length of the photon emitted?

3.15 Tunneling

a. Consider the phenomenon of tunneling through a potential energy barrier of height V_o and width a, as shown in Figure 3.16. What is the probability that the electron will be reflected? Given the transmission coefficient T, can you find the reflection coefficient R? What happens to R as a or V_o or both become very large?

b. For a wide barrier ($\alpha a \gg 1$), show that T_o can at most be 4 and that $T_o = 4$ when $E = \frac{1}{2}V_o$.

3.16 Electron impact excitation

a. A projectile electron of kinetic energy 12.2 eV collides with a hydrogen atom in a gas discharge tube. Find the nth energy level to which the electron in the hydrogen atom gets excited.

b. Calculate the possible wavelengths of radiation (in nm) that will be emitted from the excited H atom in part (a) as the electron returns to its ground state. Which one of these wavelengths will be in the visible spectrum?

c. In neon street lighting tubes, gaseous discharge in the Ne tube involves electrons accelerated by the electric field impacting Ne atoms and exciting some of them to the $2p^5 3p^1$ states, as shown in Figure 3.42. What is the wavelength of emission? Can the Ne atom fall from the $2p^5 3p^1$ state to the ground state by spontaneous emission?

3.17 Line spectra of hydrogenic atoms Spectra of hydrogen-like atoms are classified in terms of electron transitions to a common lower energy level.

a. All transitions from energy levels $n = 2, 3, \ldots$ to $n = 1$ (the K shell) are labeled K lines and constitute the **Lyman series.** The spectral line corresponding to the smallest energy difference ($n = 2$ to $n = 1$) is labeled the K_α line, next is labeled K_β, and so on. The transition from $n = \infty$ to $n = 1$ has the largest energy difference and defines the greatest photon energy (shortest wavelength) in the K series; hence it is called the absorption edge K_{ae}. What is the range of wavelengths for the K lines? What is K_{ae}? Where are these lines with respect to the visible spectrum?

b. All transitions from energy levels $n = 3, 4, \ldots$ to $n = 2$ (L shell) are labeled L lines and constitute the **Balmer series.** What is the range of wavelengths for the L lines (*i.e.*, L_α and L_{ae})? Are these in the visible range?

c. All transitions from energy levels $n = 4, 5, \ldots$ to $n = 3$ (M shell) are labeled M lines and constitute the **Paschen series.** What is the range of wavelengths for the M lines? Are these in the visible range?

d. How would you expect the spectral lines to depend on the atomic number Z?

3.18 Ionization energy and effective Z

a. Consider the singly ionized Li ion, Li^+, which has lost its $2s$ electron. If the energy required to ionize one of the $1s$ electrons in Li^+ is 18.9 eV, calculate the effective nuclear charge seen by a $1s$ electron, that is, $Z_{effective}$ in the hydrogenic atom ionization energy expression in Equation 3.45; $E_{1,n} = (Z_{effective}/n)^2(13.6\text{ eV})$.

b. The B atom has a total of five electrons, two in the 1s orbital, two in the $2s$, and one in the $2p$. The experimental ionization energy of B is 8.30 eV. Calculate $Z_{effective}$.

c. The experimental ionization energy of Na is 3.49 eV. Calculate the effective nuclear charge seen by the $3s$ valence electron.

d. The chemical tables typically list the first, second, and third ionization energies E_1, E_2, E_3, respectively, and so on. Consider Al. E_1 represents the energy required to remove the first electron from neutral Al; E_2, the second electron from Al^+; E_3, the third electron from Al^{2+} to generate Al^{3+}. For Al, experimentally, $E_1 = 6.0$ eV, $E_2 = 18.8$ eV, and $E_3 = 28.4$ eV. For each case find the $Z_{effective}$ seen by the electron that is removed.

3.19 Atomic and ionic radii The maximum in the radial probability distribution of an electron in a hydrogen-like atom is given by Equation 3.44, that is, $r_{max} = (n^2 a_o)/Z$, for $\ell = n - 1$. The average distance $\bar{r}$ of an electron from the nucleus can be calculated by using the definition of an average and the probability

distribution function $P_{n,\ell}(r)$, that is,

$$\bar{r} = \int_0^\infty r P_{n,\ell}(r)\, dr = \frac{a_0 n^2}{Z}\left[\frac{3}{2} - \frac{\ell(\ell+1)}{2n^2}\right]$$

Average distance of electron from nucleus

in which the right-hand side represents the result of the integration (which has been done by physicists). Calculate r_{max} and $\bar{r}$ for the $2p$ valence electron in the B atom. Which value is closer to the radius of the B atom, 0.085 nm, given in the Period Table? Consider only the outermost electrons, and calculate $r_{average}$ for Li, Li^+, Be^{2+}, and B, and compare with the experimental values of the atomic or ionic sizes: 0.15 nm for Li; 0.070 nm for Li^+, 0.035 nm for Be^{2+}, and 0.085 nm for B.

***3.20 X-rays and the Moseley relation** X-rays are photons with wavelengths in the range 0.01–10 nm, with typical energies in the range 100 eV to 100 keV. When an electron transition occurs in an atom from the L to the K shell, the emitted radiation is generally in the X-ray spectrum. For all atoms with atomic number $Z > 2$, the K shell is full. Suppose that one of the electrons in the K shell has been knocked out by an energetic projectile electron impacting the atom (the projectile electron would have been accelerated by a large voltage difference). The resulting vacancy in the K shell can then be filled by an electron in the L shell transiting down and emitting a photon. The emission resulting from the L to K shell transition is labeled the K_α line. The table shows the K_α line data obtained for various materials.

	Material								
	Mg	Al	S	Ca	Cr	Fe	Cu	Rb	W
Z	12	13	16	20	24	26	29	37	74
K_α line (nm)	0.987	0.834	0.537	0.335	0.229	0.194	0.154	0.093	0.021

a. If ν is the frequency of emission, plot $\nu^{1/2}$ against the atomic number Z of the element.

b. H. G. Moseley, while still a graduate student of E. Rutherford in 1913, found the empirical relationship

$$\nu^{1/2} = B(Z - C)$$

Moseley relation

where B and C are constants. What are B and C from the plot? Can you give a simple explanation as to why K_α absorption should follow this relationship?



Henry G. J. Moseley (1887–1915), around 1910, carrying out experiments at Balliol-Trinity Laboratory at Oxford.

SOURCE: University of Oxford Museum of Science, courtesy AIP Emilio Segre Visual Archives.

3.21 **The He atom** Suppose that for the He atom, zero energy is taken to be the two electrons stationary at infinity (and infinitely apart) from the nucleus (He^{++}). Estimate the energy (in eV) of the electrons in the He atom by neglecting the electron–electron repulsion, that is, neglecting the potential energy due to the mutual Coulombic repulsion between the electrons. How does this compare with the experimental value of −79 eV? How strong is the electron–electron repulsion energy?

3.22 **Excitation energy of He** In the HeNe laser, an energetic electron is accelerated by the applied field impacts and excites the He from its ground state, $1s^2$, to an excited state He^*, $1s^12s^1$, which has one of the electrons in the $2s$ orbital. The ground energy of the He atom is −79 eV with respect to both electrons isolated at infinity, which defines the zero energy. Consider the $1s^12s^1$ state. If we neglect the electron–electron interactions, we can calculate the energy of the $1s$ and $2s$ electrons using the energy for a hydrogenic atom, $E_n = -(Z^2/n^2)(13.6\text{ eV})$. We can then add the electron–electron interaction energy by assuming that the $1s$ and $2s$ electrons are effectively separated by $3a_o$, which is the difference, $4a_o - 1a_o$, between the $1s$ and $2s$ Bohr radii. Calculate the overall energy of He^* and hence the excitation energy from He to He^*. The experimental value is about 20.6 eV.

3.23 **Electron affinity** The fluorine atom has the electronic configuration $[He]2s^22p^5$. The F atom can actually capture an electron to become a F^- ion, and release energy, which is listed as its *electron affinity*, 328 kJ mol^{-1}. We will assume that the two $1s$ electrons in the closed K shell (very close to the nucleus) and the two electrons in the $2s$ orbitals will shield four positive charges and thereby expose $+9e - 4e = +5e$ for the $2p$ orbital. Suppose that we try to calculate the energy of the F^- ion by simply assuming that the additional electron is attracted by an effective positive charge, $+e(5 - Z_{2p})$ or $+eZ_{\text{effective}}$, where Z_{2p} is the overall shielding effect of the five electrons in the $2p$ orbital, so that the tenth electron we have added sees an effective charge of $+eZ_{\text{effective}}$. Calculate Z_{2p} and $Z_{\text{effective}}$. The F atom does not enjoy losing an electron. The ionization energy of the F atom is 1681 kJ mol^{-1}. What is the $Z_{\text{effective}}$ that is experienced by a $2p$ electron? (Note: 1 kJ mol^{-1} = 0.01036 eV/atom.)

***3.24** **Electron spin resonance (ESR)** It is customary to write the spin magnetic moment of an electron as



$$\mu_{\text{spin}} = -\frac{ge}{2m_e}S$$

where S is the spin angular momentum, and g is a numerical factor, called the **g factor,** which is 2 for a free electron. Consider the interaction of an electron's spin with an external magnetic field. Show that the additional potential energy E_{BS} is given by



$$E_{BS} = \beta g m_s B$$

where $\beta = e\hbar/2m_e$ is called the **Bohr magneton.** Frequently **electron spin resonance** is used to examine various defects and impurities in semiconductors. A defect such as a dangling bond, for example, will have a single unpaired electron in an orbital and thus will possess a spin magnetic moment. A strong magnetic field is applied to the specimen to split the energy level E_1 of the unpaired spin to two levels $E_1 - E_{BS}$ and $E_1 + E_{BS}$, separated by ΔE_{BS}. The electron occupies the lower level $E_1 - E_{BS}$. Electromagnetic waves (usually in the microwave range) of known frequency ν, and hence of known photon energy $h\nu$, are passed through the specimen. The magnetic field B is varied until the EM waves are absorbed by the specimen, which corresponds to the excitation of the electron at each defect from $E_1 - E_{BS}$ to $E_1 + E_{BS}$, that is, $h\nu = \Delta E_{BS}$ at a certain field B. This maximum absorption condition is called **electron spin resonance,** as the electron's spin is made to resonate with the EM wave. If $B = 2$ T, calculate the frequency of the EM waves needed for ESR, taking $g = 2$. Note: For many molecules, and impurities and defects in crystals, g is not exactly 2, because the electron is in a different environment in each case. The experimentally measured value of g can be used to characterize molecules, impurities, and defects.

3.25 **Spin–orbit coupling** An electron in an atom will experience an internal magnetic field B_{int} because, from the electron's reference frame, it is the positive nucleus that is orbiting the electron. The electron will "see" the nucleus, take as charge $+e$, circling around it, which is equivalent to a current $I = +ef$ where f is the electron's frequency of rotation around the nucleus. The current I generates the internal

magnetic field B_{int} at the electron. From electromagnetism texts, B_{int} is given by

$$B_{int} = \frac{\mu_o I}{2r}$$

where r is the radius of the electron's orbit and μ_o is the absolute permeability. Show that

Internal magnetic field at electron (orbit) in an atom

$$B_{int} = \frac{\mu_o e}{4\pi m_e r^3} L$$

Consider the hydrogen atom with $Z = 1$, $2p$ orbital, $n = 2$, $\ell = 1$, and take $r \approx n^2 a_o$. Calculate B_{int}.

The electron's spin magnetic moment μ_{spin} will couple with this internal field, which means that the electron will now possess a magnetic potential energy E_{SL} that is due to the coupling of the *spin* with the *orbital motion*, called **spin-orbit coupling.** E_{SL} will be either negative or positive, with only two values, depending on whether the electron's spin magnetic moment is along or opposite B_{int}. Take z along B_{int} so that $E_{SL} = -B_{int}\mu_{spin,z}$, where $\mu_{spin,z}$ is μ_{spin} along z, and then show that the energy E_2 of the $2p$ orbital splits into two closely separated levels whose separation is

Spin–orbit coupling potential energy

$$\Delta E_{SL} = \left(\frac{e\hbar}{m_e}\right) B_{int}$$

Calculate ΔE_{SL} in eV and compare it with $E_2(n = 2)$and the separation $\Delta E = E_2 - E_1$. (The exact calculation of E_{SL} is much more complicated, but the calculated value here is sufficiently close to be useful.) What is the effect of E_{SL} on the observed emission spectrum from the H-atom transition from $2p$ to $1s$? What is the separation of the two wavelengths? The observation is called **fine structure splitting.**

3.26 **Hund's rule** For each of the following atoms and ions, sketch the electronic structure, using a box for an orbital wavefunction, and an arrow (up or down for the spin) for an electron.

a. Aluminum, $[Ne]3s^2p^1$
b. Silicon, $[Ne]3s^2p^2$
c. Phosphorus, $[Ne]3s^2p^3$
d. Sulfur, $[Ne]3s^2p^4$
e. Chlorine, $[Ne]3s^2p^5$
f. Titanium, $[Ar]3d^24s^2$
g. Vanadium, $[Ar]3d^34s^2$
h. Manganese, $[Ar]3d^54s^2$
i. Cobalt, $[Ar]3d^74s^2$
j. Cu^{2+}, $[Ar]3d^94s^0$

3.27 **Hund's rule** The carbon atom has the electronic structure $2s^22p^2$ in its ground state. The ground state and various possible excited states of C are shown in Figure 3.47. The following energies are known for the states *a* to *e* in Figure 3.47, not in any particular order: 0, 7.3 eV, 4.1 eV, 7.9 eV, and 1.2 eV. Using reasonable arguments match these energies to the states *a* to *e*. Use Hund's rule to establish the ground state with 0 eV. If you have to flip a spin to go from the ground to another configuration, that would cost energy. If you have to move an electron from a lower s to p or from p to a higher s, that would cost a lot of energy. Two electrons in the same orbital (obviously with paired electrons) would have substantial Coulombic repulsion energy.

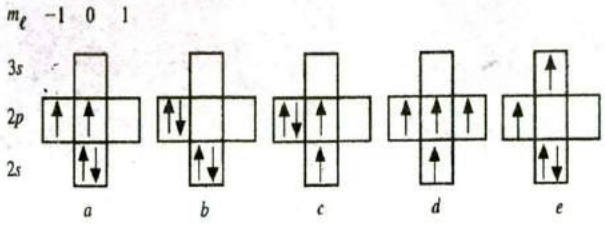


Figure 3.47
Some possible states of the carbon atom, not in any particular order.

3.28 **The HeNe laser** A particular HeNe laser operating at 632.8 nm has a tube that is 40 cm long. The operating gas temperature is about 130 °C

a. Calculate the Doppler-broadened linewidth $\Delta\lambda$ in the output spectrum.

b. What are the n values that satisfy the resonant cavity condition? How many modes are therefore allowed?

c. Calculate the frequency separation and the wavelength separation of the laser modes. How do these change as the tube warms up during operation? Taking the linear expansion coefficient to be 10^{-6} K^{-1}, estimate the change in the mode frequency separation.

3.29 **Er^{3+}-doped fiber amplifier** When the Er^{3+} ion in the Er^{3+}-doped fiber amplifer (EDFA) is pumped with 980 nm of radiation, the Er^{3+} ions absorb energy from the pump signal and become excited to E_3 (Figure 3.44). Later the Er^{3+} ions at E_2 are stimulated to add energy (coherent photons) to the signal at 1550 nm. What is the wasted energy (in eV) from the pump to the signal at each photon amplification step? (This energy is lost as heat in the glass medium.) An Er-doped fiber amplifier is 10 m long, and the radius of the core is 5 μm. The Er concentration in the core is 10^{18} cm^{-3}. The nominal power gain of the amplifier is 100 (or 20 dB). The pump wavelength is 980 nm, and the signal wavelength is 1550 nm. If the output power from the amplifier is 100 mW and assuming the signal and pump are confined to the core, what is the minimum intensity of the pump signal? How much power is wasted in this EDFA? (The pump must provide enough photons to pump the Er^{3+} ions needed to generate the additional output photons over that of input photons. The concentration of Er^{3+} ions in the fiber is given for information only.)



Wolfgan Pauli (1900–1958) won the Nobel prize in 1945 for his contributions to quantum mechanics. His exclusion principle was announced in 1925. "I don't mind your thinking slowly; I mind your publishing faster than you think." (Translation from German. Attributed to Pauli by H. Coblaus. From A. L. Mackay, *A Dictionary of Scientific Quotations*, Bristol: IOP Publishing, 1991, p. 191.)

SOURCE: AIP Emilio Segrè Visual Archives, Goudsmit Collection.



Arthur Holly Compton (1892–1962) won the Nobel prize in physics in 1927 for his discovery of the Compton effect with C. T. R. Wilson in 1923.

SOURCE: King Features Syndicate, Inc., New York and Argonne National Laboratory, courtesy AIP Emilio Segrè Visual Archives.

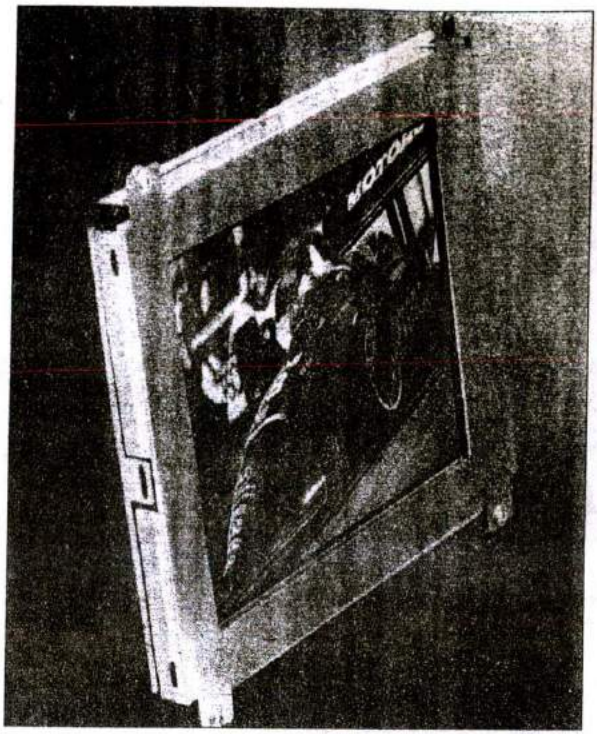


Theodore Harold Maiman was born in 1927 in Los Angeles, son of an electrical engineer. He studied engineering physics at Colorado University, while repairing electrical appliances to pay for college, and then obtained a Ph.D. from Stanford. Theodore Maiman constructed this first laser in 1960 while working at Hughes Research Laboratories. There is a vertical chromium ion-doped ruby rod in the center of a helical xenon flash tube. The ruby rod has mirrored ends. The xenon flash provides optical pumping of the chromium ions in the ruby rod. The output is a pulse of red laser light.

I SOURCE: Courtesy of HRL Laboratories, LLC, Malibu, California.


THE UNITED STATES OF AMERICA

The patent for the invention of the laser by Charles H. Townes and Arthur L. Schawlow in 1960 (Courtesy of Bell Laboratories). The laser patent was later bitterly disputed for almost three decades in "the patent wars" by Gordon Gould, an American physicist, and his designated agents. Gordon Gould eventually received the U.S. patent for optical pumping of the laser in 1977 since the original laser patent did not detail such a pumping procedure. In 1987 he also received a patent for the gas discharge laser, thereby winning his 30 year patent war. His original notebook even contained the word laser.



Motorola's prototype flat panel display based on the Fowler-Nordheim field emission principle. The display is 14 cm in diagonal and 3.5 mm thick with a viewing angle 160°. Each pixel (325 μm thick) uses field emission of electrons from microscopic sharp point sources (icebergs). Emitted electrons impinge on colored phosphors on a screen and cause light emission by cathodoluminescence. There are millions of these microscopic field emitters to constitute the image.

I SOURCE: Courtesy of Dr. Babu Chalamala, Flat Panel Display Division, Motorola.



Left: A scanning electron microscope image of an array of electron field emitters (icebergs). Center: One iceberg. Right: A cross section of a field emitter. Each iceberg is a source of electron emission arising from Fowler-Nordheim field emission; for further information see B. Chalamala, et al., *IEEE Spectrum*, April 1998, pp. 42–51.

I SOURCE: Courtesy of Dr. Babu Chalamala, Flat Panel Display Division, Motorola.