Chapter 1

Crystal Properties and Growth of Semiconductors

In studying solid state electronic devices we are interested primarily in the electrical behavior of solids. However, we shall see in later chapters that the transport of charge through a metal or a semiconductor depends not only on the properties of the electron but also on the arrangement of atoms in the solid. In the first chapter we shall discuss some of the physical properties of semiconductors compared with other solids, the atomic arrangements of various materials, and some methods of growing semiconductor crystals. Topics such as crystal structure and crystal growth technology are often the subjects of books rather than introductory chapters; thus we shall consider only a few of the more important and fundamental ideas that form the basis for understanding electronic properties of semiconductors and device fabrication.

1.1 SEMICONDUCTOR MATERIALS

Semiconductors are a group of materials having electrical conductivities intermediate between metals and insulators. It is significant that the conductivity of these materials can be varied over orders of magnitude by changes in temperature, optical excitation, and impurity content. This variability of electrical properties makes the semiconductor materials natural choices for electronic device investigations.

Semiconductor materials are found in column IV and neighboring columns of the periodic table (Table 1–1). The column IV semiconductors, silicon and germanium, are called *elemental* semiconductors because they are composed of single species of atoms. In addition to the elemental materials, compounds of column III and column V atoms, as well as certain combinations from II and VI, and from IV, make up the *compound* semiconductors.

As Table 1–1 indicates, there are numerous semiconductor materials. As we shall see, the wide variety of electronic and optical properties of these semiconductors provides the device engineer with great flexibility in the design of electronic and optoelectronic functions. The elemental semiconductor Ge was

Table 1–1. Common semiconductor materials: (a) the portion of the periodic table where semiconductors occur; (b) elemental and compound semiconductors.

(a)	II	III	IV	V	VI
		B	C	N	
		Al	Si	P	S
	Zn	Ga	Ge	As	Se
	Cd	In		Sb	Te

(b)	Elemental	IV compounds	Binary III–V compounds	Binary II–VI compounds
	Si	SiC	AlP	ZnS
	Ge	SiGe	AlAs	ZnSe
			AlSb	ZnTe
			GaN	CdS
			GaP	CdSe
			GaAs	CdTe
			GaSb	
			InP	
			InAs	
			InSb	

widely used in the early days of semiconductor development for transistors and diodes. Silicon is now used for the majority of rectifiers, transistors, and integrated circuits. However, the compounds are widely used in high-speed devices and devices requiring the emission or absorption of light. The two-element (*binary*) III–V compounds such as GaN, GaP, and GaAs are common in light-emitting diodes (LEDs). As discussed in Section 1.2.4, three-element (*ternary*) compounds such as GaAsP and four-element (*quaternary*) compounds such as InGaAsP can be grown to provide added flexibility in choosing materials properties.

Fluorescent materials such as those used in television screens usually are II–VI compound semiconductors such as ZnS. Light detectors are commonly made with InSb, CdSe, or other compounds such as PbTe and HgCdTe. Si and Ge are also widely used as infrared and nuclear radiation detectors. An important microwave device, the Gunn diode, is usually made of GaAs or InP. Semiconductor lasers are made using GaAs, AlGaAs, and other ternary and quaternary compounds.

One of the most important characteristics of a semiconductor, which distinguishes it from metals and insulators, is its *energy band gap*. This property, which we will discuss in detail in Chapter 3, determines among other things the wavelengths of light that can be absorbed or emitted by the semiconductor. For example, the band gap of GaAs is about 1.43 electron volts (eV), which corresponds to light wavelengths in the near infrared. In contrast, GaP has a band gap of about 2.3 eV, corresponding to wavelengths in

the green portion of the spectrum.[1] The band gap E_g for various semiconductor materials is listed along with other properties in Appendix III. As a result of the wide variety of semiconductor band gaps, light-emitting diodes and lasers can be constructed with wavelengths over a broad range of the infrared and visible portions of the spectrum.

The electronic and optical properties of semiconductor materials are strongly affected by impurities, which may be added in precisely controlled amounts. Such impurities are used to vary the conductivities of semiconductors over wide ranges and even to alter the nature of the conduction processes from conduction by negative charge carriers to positive charge carriers. For example, an impurity concentration of one part per million can change a sample of Si from a poor conductor to a good conductor of electric current. This process of controlled addition of impurities, called *doping*, will be discussed in detail in subsequent chapters.

To investigate these useful properties of semiconductors, it is necessary to understand the atomic arrangements in the materials. Obviously, if slight alterations in purity of the original material can produce such dramatic changes in electrical properties, then the nature and specific arrangement of atoms in each semiconductor must be of critical importance. Therefore, we begin our study of semiconductors with a brief introduction to crystal structure.

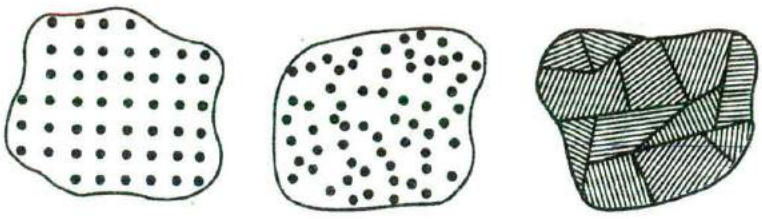
1.2 CRYSTAL LATTICES

In this section we discuss the arrangements of atoms in various solids. We shall distinguish between single crystals and other forms of materials and then investigate the periodicity of crystal lattices. Certain important crystallographic terms will be defined and illustrated in reference to crystals having a basic cubic structure. These definitions will allow us to refer to certain planes and directions within a lattice. Finally, we shall investigate the diamond lattice; this structure, with some variations, is typical of most of the semiconductor materials used in electronic devices.

1.2.1 Periodic Structures

A crystalline solid is distinguished by the fact that the atoms making up the crystal are arranged in a periodic fashion. That is, there is some basic arrangement of atoms that is repeated throughout the entire solid. Thus the crystal appears exactly the same at one point as it does at a series of other equivalent points, once the basic periodicity is discovered. However, not all solids are crystals (Fig. 1–1); some have no periodic structure at all (*amorphous* solids), and others are composed of many small regions of single-crystal material (*polycrystalline* solids). The high-resolution micrograph shown in Fig.

[1]The conversion between the energy E of a photon of light (eV) and its wavelength $\lambda(\mu m)$ is $\lambda = 1.24/E$. For GaAs, $\lambda = 1.24/1.43 = 0.87\ \mu m$.



(a) Crystalline (b) Amorphous (c) Polycrystalline

Figure 1–1
Three types of solids, classified according to atomic arrangement: (a) crystalline and (b) amorphous materials are illustrated by microscopic views of the atoms, whereas (c) polycrystalline structure is illustrated by a more macroscopic view of adjacent single-crystalline regions, such as (a).

6–33 illustrates the periodic array of atoms in the single-crystal silicon of a transistor channel compared with the amorphous SiO_2 (glass) of the oxide layer.

The periodic arrangement of atoms in a crystal is called the *lattice*. Since there are many different ways of placing atoms in a volume, the distances and orientation between atoms can take many forms. However, in every case the lattice contains a volume, called a *unit cell*, which is representative of the entire lattice and is regularly repeated throughout the crystal. As an example of such a lattice, Fig. 1–2 shows a two-dimensional arrangement of atoms with a unit cell ODEF. This cell has an atom at each corner shared with adjacent cells. Notice that we can define vectors **a** and **b** such that if the unit cell is translated by integral multiples of these vectors, a new unit cell identical to the original is found (e.g., O′D′E′F′). These vectors **a** and **b** (and **c** if the lattice is three dimensional) are called the *basis vectors* for the lattice. Points within the lattice are indistinguishable if the vector between the points is

$$\mathbf{r} = p\mathbf{a} + q\mathbf{b} + s\mathbf{c} \qquad (1\text{–}1)$$

where p, q, and s are integers.

The smallest unit cell that can be repeated to form the lattice is called a *primitive cell*. In many lattices, however, the primitive cell is not the most convenient to work with. The importance of the unit cell lies in the fact that we can analyze the crystal as a whole by investigating a representative volume. For example, from the unit cell we can find the distances between nearest atoms and next nearest atoms for calculation of the forces holding the lattice together; we can look at the fraction of the unit cell volume filled by atoms and relate the density of the solid to the atomic arrangement. But even more important for our interest in electronic devices, the properties of the periodic crystal lattice determine the allowed energies of electrons that participate in the conduction process. Thus the lattice determines not only the mechanical properties of the crystal but also its electrical properties.

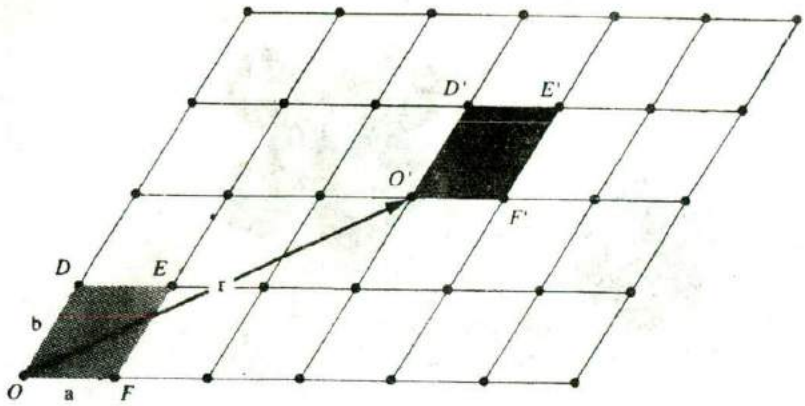
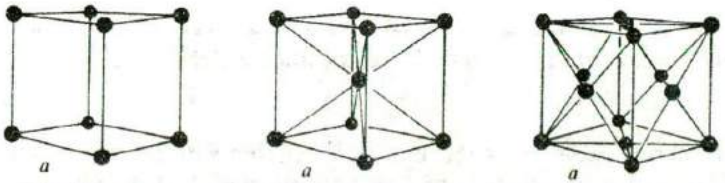


Figure 1–2
A two-dimensional lattice showing translation of a unit cell by $\mathbf{r} = 3\mathbf{a} + 2\mathbf{b}$.



Simple cubic Body-centered cubic Face-centered cubic

Figure 1–3
Unit cells for three types of cubic lattice structures.

1.2.2 Cubic Lattices

The simplest three-dimensional lattice is one in which the unit cell is a cubic volume, such as the three cells shown in Fig. 1–3. The *simple cubic* structure (abbreviated *sc*) has an atom located at each corner of the unit cell. The *body-centered cubic* (*bcc*) lattice has an additional atom at the center of the cube, and the *face-centered cubic* (*fcc*) unit cell has atoms at the eight corners and centered on the six faces.

As atoms are packed into the lattice in any of these arrangements, the distances between neighboring atoms will be determined by a balance between the forces that attract them together and other forces that hold them apart. We shall discuss the nature of these forces for particular solids in Section 3.1.1. For now, we can calculate the maximum fraction of the lattice volume that can be filled with atoms by approximating the atoms as hard spheres. For example, Fig. 1–4 illustrates the packing of spheres in a face-centered cubic cell of side a, such that the nearest neighbors touch. The dimension a for a cubic unit cell is called the *lattice constant*. For the fcc lattice the nearest neighbor distance is one-half the diagonal of a face, or $\frac{1}{2}(a\sqrt{2})$. Therefore, for the atom centered on the face to just touch the atoms at each corner of the face, the radius of the sphere must be one-half the nearest neighbor distance, or $\frac{1}{4}(a\sqrt{2})$.

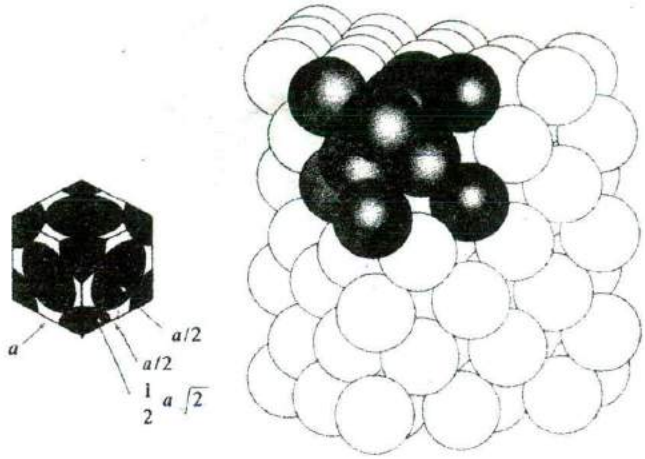


Figure 1–4
Packing of hard spheres in an fcc lattice.

EXAMPLE 1–1 Find the fraction of the fcc unit cell volume filled with hard spheres as in Fig. 1–4.

SOLUTION Each corner atom in a cubic unit cell is shared with seven neighboring cells; thus each unit cell contains $\frac{1}{8}$ of a sphere at each of the eight corners for a total of one atom. Similarly, the fcc cell contains half an atom at each of the six faces for a total of three. Thus we have

Atoms per cell = 1 (corners) + 3 (faces) = 4

Nearest neighbor distance = $\frac{1}{2}(a\sqrt{2})$

Radius of each sphere = $\frac{1}{4}(a\sqrt{2})$

Volume of each sphere = $\frac{4}{3}\pi\,[\frac{1}{4}(a\sqrt{2})]^3 = \dfrac{\pi a^3\sqrt{2}}{24}$

Maximum fraction of cell filled

$$= \frac{\text{no. of spheres} \times \text{vol. of each sphere}}{\text{total vol. of each cell}}$$

$$= \frac{4 \times (\pi a^3\sqrt{2})/24}{a^3}$$

$$= \frac{\pi\sqrt{2}}{6} = 74 \text{ percent filled}$$

Therefore, if the atoms in an fcc lattice are packed as densely as possible, with no distance between the outer edges of nearest neighbors, 74 percent of the volume is filled. This is a relatively high percentage compared with some other lattice structures (Prob. 1.14).

1.2.3 Planes and Directions

In discussing crystals it is very helpful to be able to refer to planes and directions within the lattice. The notation system generally adopted uses a set of three integers to describe the position of a plane or the direction of a vector within the lattice. The three integers describing a particular plane are found in the following way:

1. Find the intercepts of the plane with the crystal axes and express these intercepts as integral multiples of the basis vectors (the plane can be moved in and out from the origin, retaining its orientation, until such an integral intercept is discovered on each axis).
2. Take the reciprocals of the three integers found in step 1 and reduce these to the smallest set of integers h, k, and l, which have the same relationship to each other as the three reciprocals.
3. Label the plane (hkl).

EXAMPLE 1–2

The plane illustrated in Fig. 1–5 has intercepts at 2**a**, 4**b**, and 1**c** along the three crystal axes. Taking the reciprocals of these intercepts, we get $\frac{1}{2}$, $\frac{1}{4}$, and 1. These three fractions have the same relationship to each other as the integers 2, 1, and 4 (obtained by multiplying each fraction by 4). Thus the plane can be referred to as a (214) plane.

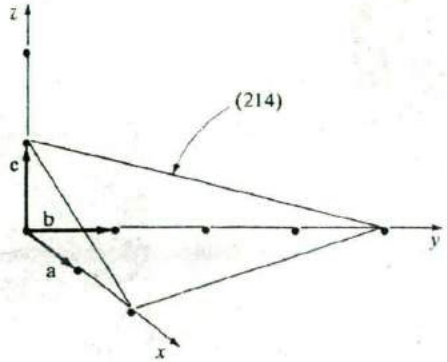


Figure 1–5
A (214) crystal plane.

The three integers h, k, and l are called the *Miller indices*; these three numbers define a set of parallel planes in the lattice. One advantage of taking the reciprocals of the intercepts is avoidance of infinities in the notation. One intercept is infinity for a plane parallel to an axis; however, the reciprocal of such an intercept is taken as zero. If a plane contains one of the axes, it is parallel to that axis and has a zero reciprocal intercept. If a plane passes through the origin, it can be translated to a parallel position for calculation of the Miller indices. If an intercept occurs on the negative branch of an axis, the minus sign is placed above the Miller index for convenience, such as $(h\bar{k}l)$.

From a crystallographic point of view, many planes in a lattice are equivalent; that is, a plane with given Miller indices can be shifted about in the lattice simply by choice of the position and orientation of the unit cell. The indices of such equivalent planes are enclosed in braces { } instead of parentheses. For example, in the cubic lattice of Fig. 1–6 all the cube faces are crystallographically equivalent in that the unit cell can be rotated in various directions and still appear the same. The six equivalent faces are collectively designated as {100}.

A direction in a lattice is expressed as a set of three integers with the same relationship as the components of a vector in that direction. The three vector components are expressed in multiples of the basis vectors, and the three integers are reduced to their smallest values while retaining the relationship among them. For example, the body diagonal in the cubic lattice (Fig. 1–7a) is composed of the components 1**a**, 1**b**, and 1**c**; therefore, this diagonal is the [111] direction. (Brackets are used for direction indices.) As in

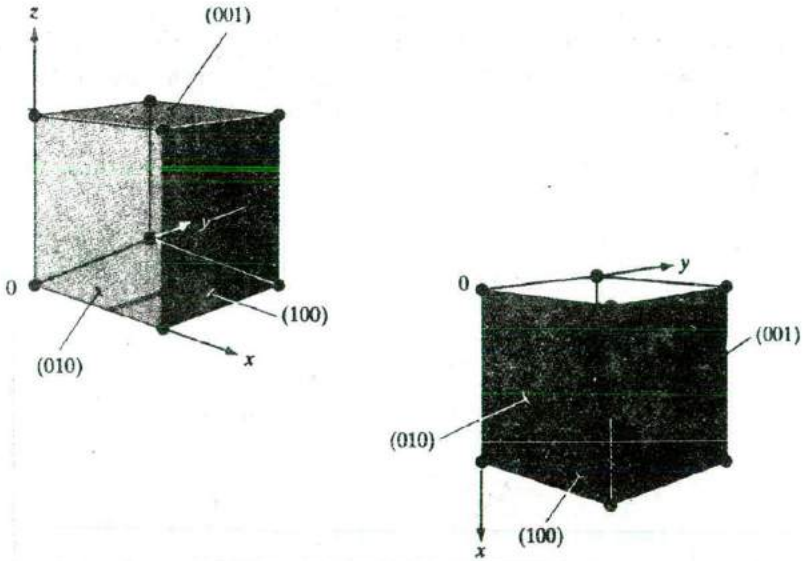


Figure 1–6
Equivalence of the cube faces ({100} planes) by rotation of the unit cell within the cubic lattice.

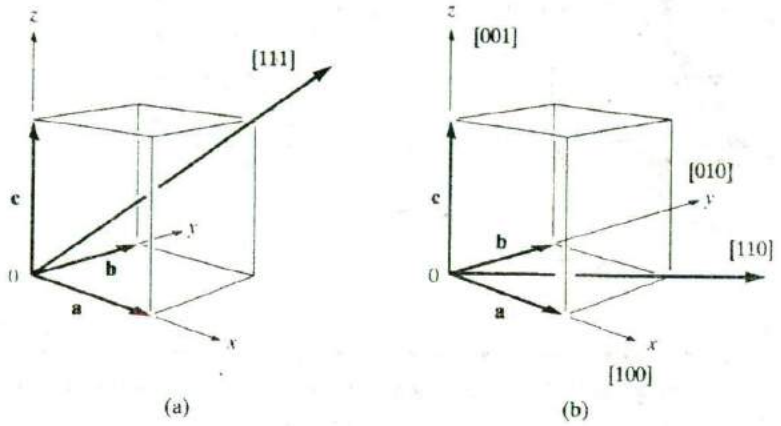


Figure 1–7
Crystal directions in the cubic lattice.

the case of planes, many directions in a lattice are equivalent, depending only on the arbitrary choice of orientation for the axes. Such equivalent direction indices are placed in angular brackets ⟨ ⟩. For example, the crystal axes in the cubic lattice [100], [010], and [001] are all equivalent and are called ⟨100⟩ directions (Fig. 1–7b).

Comparing Figs. 1–6 and 1–7, we notice that in cubic lattices a direction [*hkl*] is perpendicular to the plane (*hkl*). This is convenient in analyzing lattices with cubic unit cells, but it should be remembered that it is not necessarily true in noncubic systems.

1.2.4 The Diamond Lattice

The basic lattice structure for many important semiconductors is the *diamond* lattice, which is characteristic of Si and Ge. In many compound semiconductors, atoms are arranged in a basic diamond structure but are different on alternating sites. This is called a *zincblende* lattice and is typical of the III–V compounds. One of the simplest ways of stating the construction of the diamond lattice is the following:

> The diamond lattice can be thought of as an fcc structure with an extra atom placed at **a**/4 + **b**/4 + **c**/4 from each of the fcc atoms.

Figure 1–8a illustrates the construction of a diamond lattice from an fcc unit cell. We notice that when the vectors are drawn with components one-fourth of the cube edge in each direction, only four additional points within the same unit cell are reached. Vectors drawn from any of the other fcc atoms simply determine corresponding points in adjacent unit cells. This method of constructing the diamond lattice implies that the original fcc has associated with it a second interpenetrating fcc displaced by ¼, ¼, ¼. The two interpenetrating fcc *sublattices* can be visualized by looking down on the unit

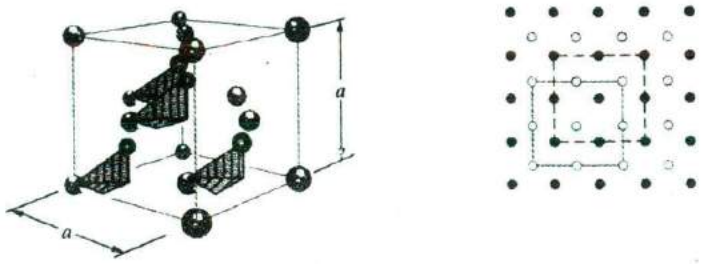


Figure 1–8
Diamond lattice structure: (a) a unit cell of the diamond lattice constructed by placing atoms $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ from each atom in an fcc; (b) top view (along any ⟨100⟩ direction) of an extended diamond lattice. The colored circles indicate one fcc sublattice and the black circles indicate the interpenetrating fcc.

cell of Fig. 1–8a from the top (or along any ⟨100⟩ direction). In the top view of Fig. 1–8b, atoms belonging to the original fcc are represented by open circles, and the interpenetrating sublattice is shaded. If the atoms are all similar, we call this structure a diamond lattice; if the atoms differ on alternating sites, it is a zincblende structure. For example, if one fcc sublattice is composed of Ga atoms and the interpenetrating sublattice is As, the zincblende structure of GaAs results. Most of the compound semiconductors have this type of lattice, although some of the II–VI compounds are arranged in a slightly different structure called the *wurtzite* lattice. We shall restrict our discussion here to the diamond and zincblende structures, since they are typical of most of the commonly used semiconductors.

EXAMPLE 1–3 Calculate the densities of Si and GaAs from the lattice constants (Appendix III), atomic weights, and Avogadro's number. Compare the results with densities given in Appendix III. The atomic weights of Si, Ga, and As are 28.1, 69.7, and 74.9, respectively.

SOLUTION For Si: $a = 5.43 \times 10^{-8}$ cm, 8 atoms/cell,

$$\frac{8}{a^3} = \frac{8}{(5.43 \times 10^{-8})^3} = 5 \times 10^{22} \text{ atoms/cm}^3$$

$$\text{density} = \frac{5 \times 10^{22}(\text{atoms/cm}^3) \times 28.1(\text{g/mole})}{6.02 \times 10^{23}(\text{atoms/mole})} = 2.33 \text{ g/cm}^3$$

For GaAs: $a = 5.65 \times 10^{-8}$ cm, 4 each Ga, As atoms/cell

$$\frac{4}{a^3} = \frac{4}{(5.65 \times 10^{-8})^3} = 2.22 \times 10^{22} \text{ atoms/cm}^3$$

$$\text{density} = \frac{2.22 \times 10^{22}(69.7 + 74.9)}{6.02 \times 10^{23}} = 5.33 \text{ g/cm}^3$$

A particularly interesting and useful feature of the III–V compounds is the ability to vary the mixture of elements on each of the two interpenetrating fcc sublattices of the zincblende crystal. For example, in the ternary compound AlGaAs, it is possible to vary the composition of the ternary alloy by choosing the fraction of Al or Ga atoms on the column III sublattice. It is common to represent the composition by assigning subscripts to the various elements. For example, $Al_xGa_{1-x}As$ refers to a ternary alloy in which the column III sublattice in the zincblende structure contains a fraction x of Al atoms and $1-x$ of Ga atoms. The composition $Al_{0.3}Ga_{0.7}As$ has 30 percent Al and 70 percent Ga on the column III sites, with the interpenetrating column V sublattice occupied entirely by As atoms. It is extremely useful to be able to grow ternary alloy crystals such as this with a given composition. For the $Al_xGa_{1-x}As$ example we can grow crystals over the entire composition range from $x = 0$ to $x = 1$, thus varying the electronic and optical properties of the material from that of GaAs ($x = 0$) to that of AlAs ($x = 1$). To vary the properties even further, it is possible to grow four-element (quaternary) compounds such as $In_xGa_{1-x}As_yP_{1-y}$ having a very wide range of properties.

It is important from an electronic point of view to notice that each atom in the diamond and zincblende structures is surrounded by four nearest neighbors (Fig. 1–9). The importance of this relationship of each atom to its neighbors will become evident in Section 3.1.1 when we discuss the bonding forces which hold the lattice together.

The fact that atoms in a crystal are arranged in certain planes is important to many of the mechanical, metallurgical, and chemical properties of the material. For example, crystals often can be cleaved along certain atomic planes, resulting in exceptionally planar surfaces. This is a familiar result in cleaved diamonds for jewelry; the facets of a diamond reveal clearly the triangular, hexagonal, and rectangular symmetries of intersecting planes in various crystallographic directions. Semiconductors with diamond and zincblende lattices have similar cleavage planes. Chemical reactions, such as etching of the crystal, often take place preferentially along certain directions. These properties serve as interesting illustrations of crystal symmetry, but in addition, each plays an important role in fabrication processes for many semiconductor devices.

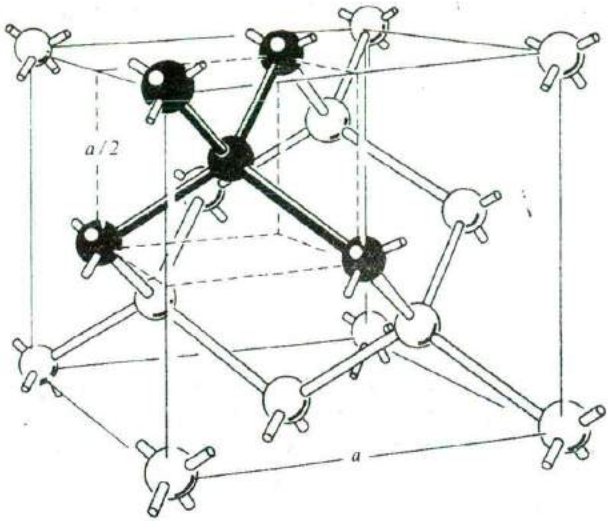


Figure 1–9 Diamond lattice unit cell, showing the four nearest neighbor structure. (From *Electrons and Holes in Semiconductors* by W. Shockley, © 1950 by Litton Educational Publishing Co., Inc.; by permission of Van Nostrand Reinhold Co., Inc.)

1.3 BULK CRYSTAL GROWTH

The progress of solid state device technology since the invention of the transistor in 1948 has depended not only on the development of device concepts but also on the improvement of materials. For example, the fact that integrated circuits can be made today is the result of a considerable breakthrough in the growth of pure, single-crystal Si in the early and mid−1950s. The requirements on the growing of device-grade semiconductor crystals are more stringent than those for any other materials. Not only must semiconductors be available in large single crystals, but also the purity must be controlled within extremely close limits. For example, Si crystals now being used in devices are grown with concentrations of most impurities of less than one part in ten billion. Such purities require careful handling and treatment of the material at each step of the manufacturing process.

1.3.1 Starting Materials

The raw feedstock for Si crystal is silicon dioxide (SiO_2). We react SiO_2 with C in the form of coke in an arc furnace at very high temperatures (~1800 °C) to reduce SiO_2 according to the following reaction:

$$SiO_2 + 2C \rightarrow Si + 2CO \qquad (1\text{–}2)$$

This forms metallurgical grade Si (MGS) which has impurities such as Fe, Al and heavy metals at levels of several hundred to several thousand parts per million (ppm). Refer back to Example 1–3 to see that 1 ppm of Si corresponds to an impurity level of $5 \times 10^{16}\text{cm}^{-3}$. While MGS is clean enough

for metallurgical applications such as using Si to make stainless steel, it is not pure enough for electronic applications; it is also not single-crystal.

The MGS is refined further to yield semiconductor-grade or electronic-grade Si (EGS), in which the levels of impurities are reduced to parts per billion or ppb (1 ppb=5×10^{13}cm^{-3}). This involves reacting the MGS with dry HCl according to the following reaction to form trichlorosilane, $SiHCl_3$, which is a liquid with a boiling point of 32 ° C.

$$Si + 3HCl \rightarrow SiHCl_3 + H_2 \qquad (1\text{–}3)$$

Along with $SiHCl_3$, chlorides of impurities such as $FeCl_3$ are formed which fortunately have boiling points that are different from that of $SiHCl_3$. This allows a technique called fractional distillation to be used, in which we heat up the mixture of $SiHCl_3$ and the impurity chlorides, and condense the vapors in different distillation towers held at appropriate temperatures. We can thereby separate pure $SiHCl_3$ from the impurities. $SiHCl_3$ is then converted to highly pure EGS by reaction with H_2,

$$2SiHCl_3 + 2H_2 \rightarrow 2Si + 6HCl \qquad (1\text{–}4)$$

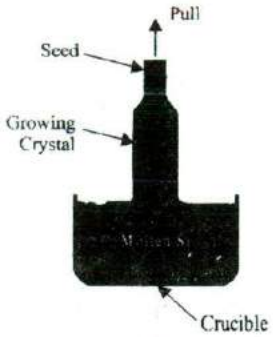
1.3.2 Growth of Single Crystal Ingots

Next, we have to convert the high purity but still polycrystalline EGS to single-crystal Si ingots or boules. This is generally done today by a process commonly called the *Czochralski* method. In order to grow single-crystal material, it is necessary to have a seed crystal which can provide a template for growth. We melt the EGS in a quartz-lined graphite crucible by resistively heating it to the melting point of Si (1412°C).

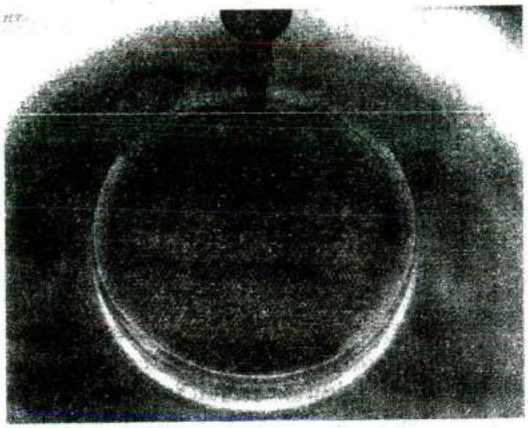
A seed crystal is lowered into the molten material and then is raised slowly, allowing the crystal to grow onto the seed (Fig. 1-10). Generally, the crystal is rotated slowly as it grows to provide a slight stirring of the melt and to average out any temperature variations that would cause inhomogeneous solidification. This technique is widely used in growing Si, Ge, and some of the compound semiconductors.

In pulling compounds such as GaAs from the melt, it is necessary to prevent volatile elements (e.g., As) from vaporizing. In one method a layer of B_2O_3, which is dense and viscous when molten, floats on the surface of the molten GaAs to prevent As evaporation. This growth method is called *liquid-encapsulated Czochralski* (*LEC*) growth.

In Czochralski crystal growth, the shape of the ingot is determined by a combination of the tendency of the cross section to assume a polygonal shape due to the crystal structure and the influence of surface tension, which encourages a circular cross section. The crystal facets are noticeable in the initial growth near the seed crystal in Fig. 1–10(b). However, the cross section of the large ingot in Fig. 1–11 is almost circular.



(a)



(b)

Figure 1–10
Pulling of a Si crystal from the melt (Czochralski method): (a) schematic diagram of the crystal growth process; (b) an 8-in. diameter, ⟨100⟩ oriented Si crystal being pulled from the melt. (Photograph courtesy of MEMC Electronics Intl.)

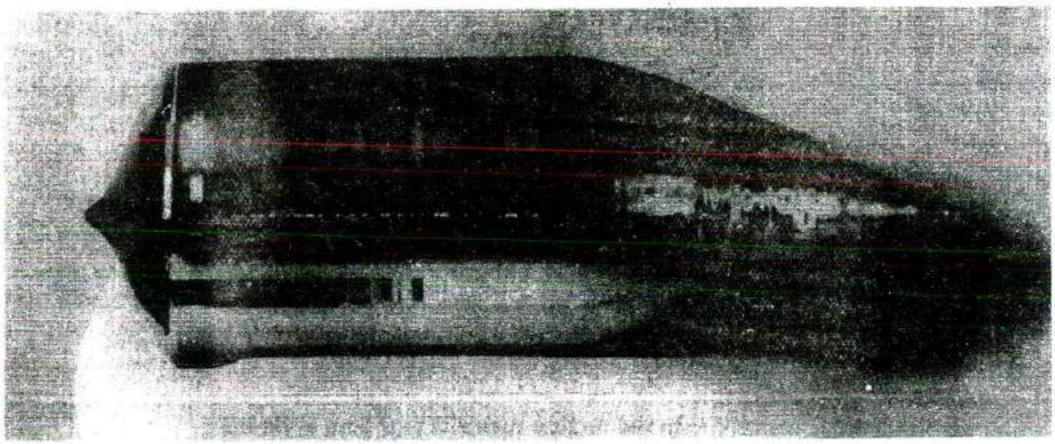


Figure 1–11
Silicon crystal grown by the Czochralski method. This large single-crystal ingot provides 300 mm (12-in.) diameter wafers when sliced using a saw. The ingot is about 1.5 m long (excluding the tapered regions), and weighs about 275 kg. (Photograph courtesy of MEMC Electronics Intl.)

In the fabrication of Si integrated circuits (Chapter 9) it is economical to use very large Si wafers, so that many IC chips can be made simultaneously. As a result, considerable research and development have gone into methods for growing very large Si crystals. For example, Fig. 1–11 illustrates a 12-inch-diameter Si ingot, 1.5 m long, weighing 275 kg.

1.3.3 Wafers

After the single-crystal ingot is grown, it is then mechanically processed to manufacture wafers. The first step involves mechanically grinding the more-

or-less cylindrical ingot into a perfect cylinder with a precisely controlled diameter. This is important because in a modern integrated circuit fabrication facility many processing tools and wafer handling robots require tight tolerances on the size of the wafers. Using X-ray crystallography, crystal planes in the ingot are identified. For reasons discussed in Section 6.4.3, most Si ingots are grown along the ⟨100⟩ direction (Fig. 1–10). For such ingots, a small notch is ground on one side of the cylinder to delineate a {110} face of the crystal. This is useful because for ⟨100⟩ Si wafers, the {110} cleavage planes are orthogonal to each other. This notch then allows the individual integrated circuit chips to be made oriented along {110} planes so that when the chips are sawed apart, there is less chance of spurious cleavage of the crystal, which could cause good chips to be lost.

Next, the Si cylinder is sawed into individual wafers about 775 μm thick, by using a diamond-tipped inner-hole blade saw, or a wire saw (Fig. 1–12a).

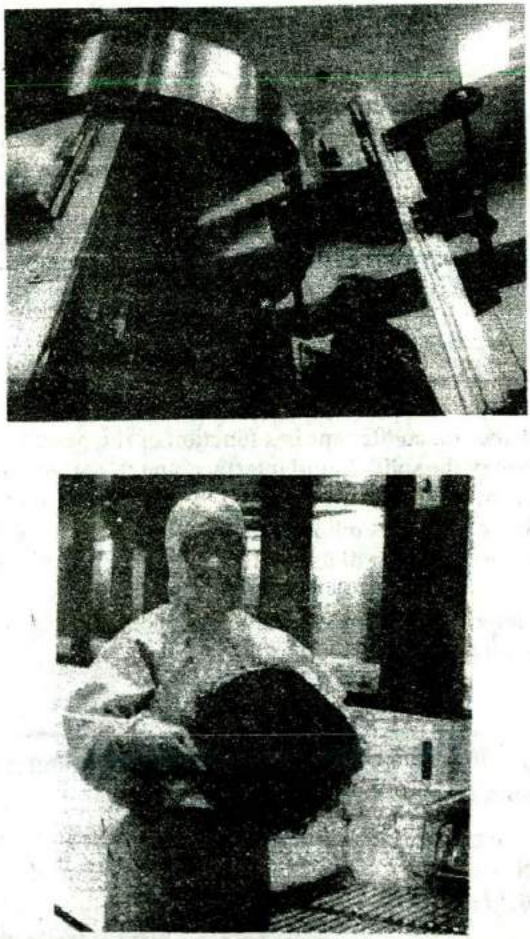


Figure 1–12
Steps involved in manufacturing Si wafers: (a) A 300 mm Si cylindrical ingot, with a notch on one side, being loaded into a wire saw to produce Si wafers; (b) a technician holding a cassette of 300 mm wafers. (Photographs courtesy of MEMC Electronics Intl.)

The resulting wafers are mechanically lapped and ground on both sides to achieve a flat surface, and to remove the mechanical damage due to sawing. Such damage would have a detrimental effect on devices. The flatness of the wafer is critical from the point of view of "depth of focus" or how sharp an image can be focussed on the wafer surface during photolithography, as discussed in Chapter 5. The Si wafers are then rounded or "chamfered" along the edges to minimize the likelihood of chipping the wafers during processing. Finally, the wafers undergo chemical-mechanical polishing using a slurry of very fine SiO_2 particles in a basic NaOH solution to give the front surface of the wafer a mirror-like finish. The wafers are now ready for integrated circuit fabrication (Fig. 1–12b). The economic value added in this process is impressive. From sand (SiO_2) costing pennies, we can obtain Si wafers costing a few hundred dollars, on which we can make hundreds of microprocessors, for example, each costing several hundred dollars.

1.3.4 Doping

As previously mentioned, there are some impurities in the molten EGS. We may also add intentional impurities or dopants to the Si melt to change its electronic properties. At the solidifying interface between the melt and the solid, there will be a certain distribution of impurities between the two phases. An important quantity that identifies this property is the *distribution coefficient* k_d, which is the ratio of the concentration of the impurity in the solid C_S to the concentration in the liquid C_L at equilibrium:

$$k_d = \frac{C_S}{C_L} \tag{1–5}$$

The distribution coefficient is a function of the material, the impurity, the temperature of the solid–liquid interface, and the growth rate. For an impurity with a distribution coefficient of one-half, the relative concentration of the impurity in the molten liquid to that in the refreezing solid is two to one. Thus the concentration of impurities in that portion of material that solidifies first is one-half the original concentration C_0. The distribution coefficient is thus important during growth from a melt. This can be illustrated by an example involving Czochralski growth:

EXAMPLE 1–4 A Si crystal is to be grown by the Czochralski method, and it is desired that the ingot contain 10^{16} phosphorus atoms/cm^3.

(a) What concentration of phosphorus atoms should the melt contain to give this impurity concentration in the crystal during the initial growth? For P in Si, $k_d = 0.35$.

(b) If the initial load of Si in the crucible is 5 kg, how many grams of phosphorus should be added? The atomic weight of phosphorus is 31.

SOLUTION

(a) Assume that $C_S = k_d C_L$ throughout the growth. Thus the initial concentration of P in the melt should be

$$\frac{10^{16}}{0.35} = 2.86 \times 10^{16}\ \text{cm}^{-3}$$

(b) The P concentration is so small that the volume of melt can be calculated from the weight of Si. From Example 1–3 the density of Si is 2.33 g/cm^3. In this example we will neglect the difference in density between solid and molten Si.

$$\frac{5000\ \text{g of Si}}{2.33\ \text{g/cm}^3} = 2146\ \text{cm}^3\ \text{of Si}$$

$$2.86 \times 10^{16}\ \text{cm}^{-3} \times 2146\ \text{cm}^3 = 6.14 \times 10^{19}\ \text{P atoms}$$

$$\frac{6.14 \times 10^{19}\ \text{atoms} \times 31\ \text{g/mole}}{6.02 \times 10^{23}\ \text{atoms/mole}} = 3.16 \times 10^{-3}\ \text{g of P}$$

Since the P concentration in the growing crystal is only about one-third of that in the melt, Si is used up more rapidly than P in the growth. Thus the melt becomes richer in P as the growth proceeds, and the crystal is doped more heavily in the latter stages of growth. This assumes that k_d is not varied; a more uniformly doped ingot can be grown by varying the pull rate (and therefore k_d) appropriately. Modern Czochralski growth systems use computer controls to vary the temperature, pull rate, and other parameters to achieve fairly uniformly doped ingots.

1.4 EPITAXIAL GROWTH

One of the most important and versatile methods of crystal growth for device applications is the growth of a thin crystal layer on a wafer of a compatible crystal. The substrate crystal may be a wafer of the same material as the grown layer or a different material with a similar lattice structure. In this process the substrate serves as the seed crystal onto which the new crystalline material grows. The growing crystal layer maintains the crystal structure and orientation of the substrate. The technique of growing an oriented single-crystal layer on a substrate wafer is called *epitaxial growth*, or *epitaxy*. As we shall see in this section, epitaxial growth can be performed at temperatures considerably below the melting point of the substrate crystal. A variety of methods are used to provide the appropriate atoms to the surface of the

growing layer. These methods include *chemical vapor deposition (CVD)*,[2] growth from a melt (*liquid-phase epitaxy, LPE*), and evaporation of the elements in a vacuum (*molecular beam epitaxy, MBE*). With this wide range of epitaxial growth techniques, it is possible to grow a variety of crystals for device applications, having properties specifically designed for the electronic or optoelectronic device being made.

1.4.1 Lattice Matching in Epitaxial Growth

When Si epitaxial layers are grown on Si substrates, there is a natural matching of the crystal lattice, and high-quality single-crystal layers result. On the other hand, it is often desirable to obtain epitaxial layers that differ somewhat from the substrate, which is known as *heteroepitaxy*. This can be accomplished easily if the lattice structure and lattice constant a match for the two materials. For example, GaAs and AlAs both have the zincblende structure, with a lattice constant of about 5.65 Å. As a result, epitaxial layers of the ternary alloy AlGaAs can be grown on GaAs substrates with little lattice mismatch. Similarly, GaAs can be grown on Ge substrates (see Appendix III).

Since AlAs and GaAs have similar lattice constants, it is also true that the ternary alloy AlGaAs has essentially the same lattice constant over the entire range of compositions from AlAs to GaAs. As a result, one can choose the composition x of the ternary compound $Al_xGa_{1-x}As$ to fit the particular device requirement, and grow this composition on a GaAs wafer. The resulting epitaxial layer will be lattice-matched to the GaAs substrate.

Figure 1–13 illustrates the energy band gap E_g as a function of lattice constant a for several III–V ternary compounds as they are varied over their composition ranges. For example, as the ternary compound InGaAs is varied by choice of composition on the column III sublattice from InAs to GaAs, the band gap changes from 0.36 to 1.43 eV while the lattice constant of the crystal varies from 6.06 Å for InAs to 5.65 Å for GaAs. Clearly, we cannot grow this ternary compound over the entire composition range on a particular binary substrate, which has a fixed lattice constant. As Fig. 1–13 illustrates, however, it is possible to grow a specific composition of InGaAs on an InP substrate. The vertical (invariant lattice constant) line from InP to the InGaAs curve shows that a midrange ternary composition (actually, $In_{0.53}Ga_{0.47}As$) can be grown lattice-matched to an InP substrate. Similarly, a ternary InGaP alloy with about 50 percent Ga and 50 percent In on the column III sublattice can be grown lattice-matched to a GaAs substrate. To achieve a broader range of alloy compositions, grown lattice-matched on particular substrates, it is helpful to

[2]The generic term *chemical vapor deposition* includes deposition of layers that may be polycrystalline or amorphous. When a CVD process results in a single-crystal epitaxial layer, a more specific term is *vapor-phase epitaxy (VPE)*.

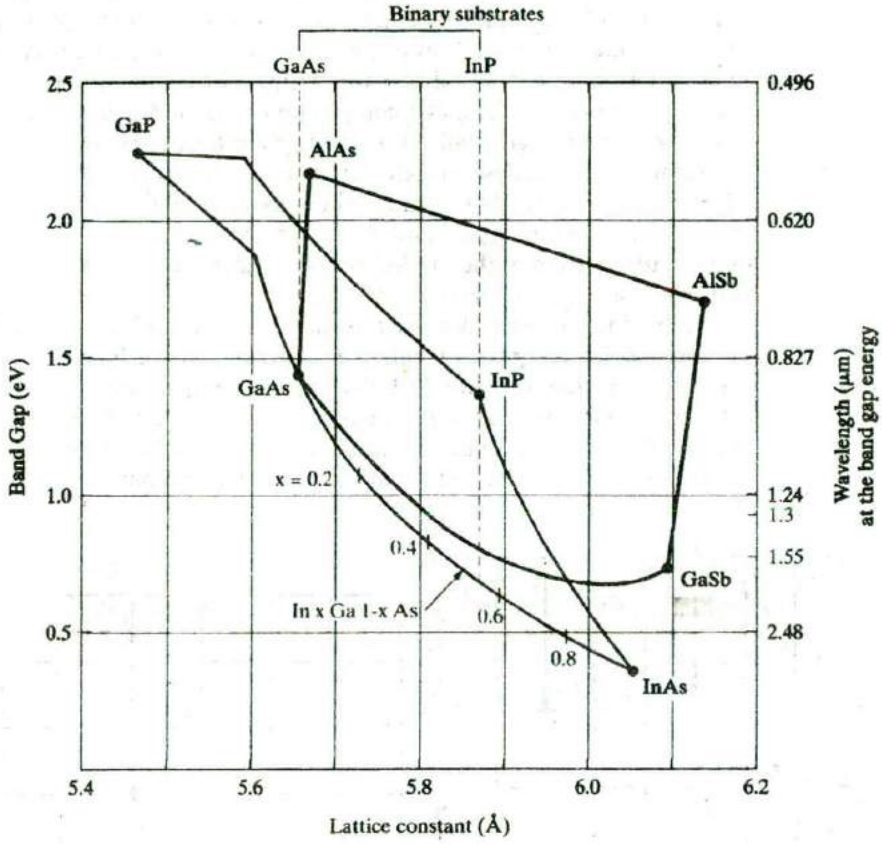


Figure 1–13
Relationship between band gap and lattice constant for alloys in the InGaAsP and AlGaAsSb systems. The dashed vertical lines show the lattice constants for the commercially available binary substrates GaAs and InP. For the marked example of $In_xGa_{1-x}As$, the ternary composition $x = 0.53$ can be grown lattice-matched on InP, since the lattice constants are the same. For quaternary alloys, the compositions on both the III and V sublattices can be varied to grow lattice-matched epitaxial layers along the dashed vertical lines between curves. For example, $In_xGa_{1-x}As_yP_{1-y}$ can be grown on InP substrates, with resulting band gaps ranging from 0.75 eV to 1.35 eV. In using this figure, assume the lattice constant a of a ternary alloy varies linearly with the composition x.

use quaternary alloys such as InGaAsP. The variation of compositions on both the column III and column V sublattices provides additional flexibility in choosing a particular band gap while providing lattice-matching to convenient binary substrates such as GaAs or InP.

In the case of GaAsP, the lattice constant is intermediate between that of GaAs and GaP, depending upon the composition. For example, GaAsP

crystals used in red LEDs have 40 percent phosphorus and 60 percent arsenic on the column V sublattice. Since such a crystal cannot be grown directly on either a GaAs or a GaP substrate, it is necessary to gradually change the lattice constant as the crystal is grown. Using a GaAs or Ge wafer as a substrate, the growth is begun at a composition near GaAs. A region ~25 μm thick is grown while gradually introducing phosphorus until the desired As/P ratio is achieved. The desired epitaxial layer (e.g., 100 μm thick) is then grown on this graded layer. By this method epitaxial growth always occurs on a crystal of similar lattice constant. Although some crystal dislocations occur due to lattice strain in the graded region, such crystals are of high quality and can be used in LEDs.

In addition to the widespread use of lattice-matched epitaxial layers, the advanced epitaxial growth techniques described in the following sections allow the growth of very thin (~100Å) layers of lattice-mismatched crystals. If the mismatch is only a few percent and the layer is thin, the epitaxial layer grows with a lattice constant in compliance with that of the seed crystal (Fig. 1–14). The resulting layer is in compression or tension along the surface plane

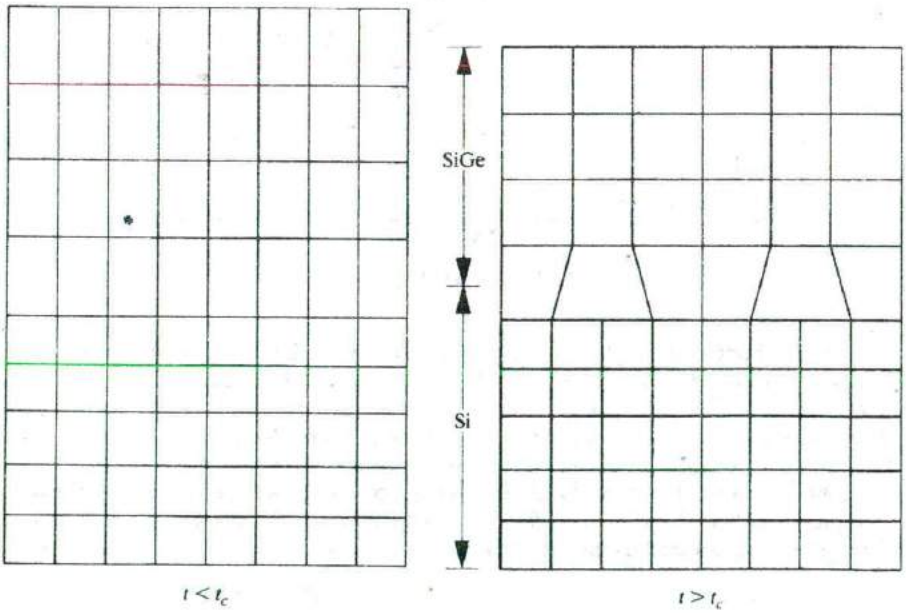


Figure 1–14
Heteroepitaxy and misfit dislocations. For example, in heteroepitaxy of a SiGe layer on Si, the lattice mismatch between SiGe and Si leads to compressive strain in the SiGe layer. The amount of strain depends on the mole fraction of Ge. (a) For layer thicknesses less than the critical layer thickness, t_c, pseudomorphic growth occurs. (b) However, above t_c, misfit dislocations form at the interface which may reduce the usefulness of the layers in device applications.

as its lattice constant adapts to the seed crystal (Fig. 1–14). Such a layer is called *pseudomorphic* because it is not lattice-matched to the substrate without strain. However, if the epitaxial layer exceeds a critical layer thickness, t_c, which depends on the lattice mismatch, the strain energy leads to formation of defects called *misfit dislocations*. Using thin alternating layers of slightly mismatched crystal layers, it is possible to grow a *strained-layer superlattice* (*SLS*) in which alternate layers are in tension and compression. The overall SLS lattice constant is an average of that of the two bulk materials.

1.4.2 Vapor-Phase Epitaxy

The advantages of low temperature and high purity epitaxial growth can be achieved by crystallization from the vapor phase. Crystalline layers can be grown onto a seed or substrate from a chemical vapor of the semiconductor material or from mixtures of chemical vapors containing the semiconductor. *Vapor-phase epitaxy* (*VPE*) is a particularly important source of semiconductor material for use in devices. Some compounds such as GaAs can be grown with better purity and crystal perfection by vapor epitaxy than by other methods. Furthermore, these techniques offer great flexibility in the actual fabrication of devices. When an epitaxial layer is grown on a substrate, it is relatively simple to obtain a sharp demarcation between the type of impurity doping in the substrate and in the grown layer. The advantages of this freedom to vary the impurity will be discussed in subsequent chapters. We point out here, however, that Si integrated-circuit devices (Chapter 9) are usually built in layers grown by VPE on Si wafers.

Epitaxial layers are generally grown on Si substrates by the controlled deposition of Si atoms onto the surface from a chemical vapor containing Si. In one method, a gas of silicon tetrachloride reacts with hydrogen gas to give Si and anhydrous HCl:

$$SiCl_4 + 2H_2 \rightleftharpoons Si + 4HCl \qquad (1\text{–}6)$$

If this reaction occurs at the surface of a heated crystal, the Si atoms released in the reaction can be deposited as an epitaxial layer. The HCl remains gaseous at the reaction temperature and does not disturb the growing crystal. As indicated, this reaction is reversible. This is very important because it implies that by adjusting the process parameters, the reaction in Eq. (1–6) can be driven to the left (providing etching of the Si rather than deposition). This etching can be used for preparing an atomically clean surface on which epitaxy can occur.

This vapor epitaxy technique requires a chamber into which the gases can be introduced and a method for heating the Si wafers. Since the chemical reactions take place in this chamber, it is called a *reaction chamber* or, more simply, a *reactor*. Hydrogen gas is passed through a heated vessel in which $SiCl_4$

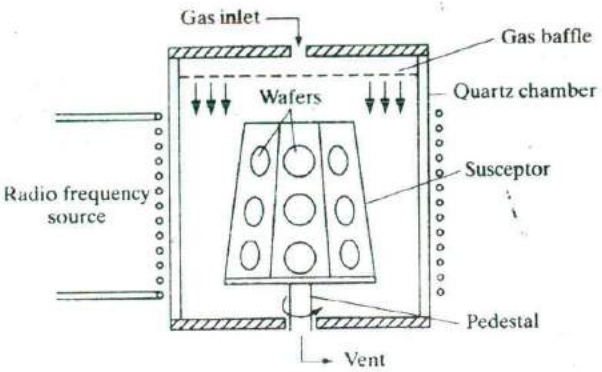


Figure 1–15
A barrel-type reactor for Si VPE. These are atmospheric pressure systems. The Si wafers are held in slots cut on the sides of a SiC-coated graphite susceptor that flares out near the base to promote gas flow patterns conducive to uniform epitaxy.

is evaporated; then the two gases are introduced into the reactor over the substrate crystal, along with other gases containing the desired doping impurities. The Si slice is placed on a graphite susceptor or some other material that can be heated to the reaction temperature with an rf heating coil or tungsten halogen lamps. This method can be adapted to grow epitaxial layers of closely controlled impurity concentration on many Si slices simultaneously (Fig. 1–15).

The reaction temperature for the hydrogen reduction of $SiCl_4$ is approximately 1150–1250°C. Other reactions may be employed at somewhat lower temperatures, including the use of dichlorosilane (SiH_2Cl_2) at 1000–1100 ° C, or the pyrolysis of silane (SiH_4) at 1000°C. Pyrolysis involves the breaking up of the silane at the reaction temperature:

$$SiH_4 \rightarrow Si + 2H_2 \qquad (1\text{–}7)$$

There are several advantages of the lower reaction temperature processes, including the fact that they reduce migration of impurities from the substrate to the growing epitaxial layer.

In some applications it is useful to grow thin Si layers on insulating substrates. For example, vapor-phase epitaxial techniques can be used to grow ~1μm Si films on sapphire and other insulators. This application of VPE is discussed in Section 9.3.2.

Vapor-phase epitaxial growth is also important in the III–V compounds, such as GaAs, GaP, and the ternary alloy GaAsP, which is widely used in the fabrication of LEDs. Substrates are held at about 800°C on a rotating wafer holder while phosphine, arsine, and gallium chloride gases are mixed and passed over the samples. The GaCl is obtained by reacting anhydrous HCl with molten Ga within the reactor. Variation of the crystal composition for GaAsP can be controlled by altering the mixture of arsine and phosphine gases.

Another useful method for epitaxial growth of compound semiconductors is called *metal-organic vapor-phase epitaxy* (*MOVPE*), or *organometallic vapor-phase epitaxy* (*OMVPE*). For example, the organometallic compound trimethylgallium can be reacted with arsine to form GaAs and methane:

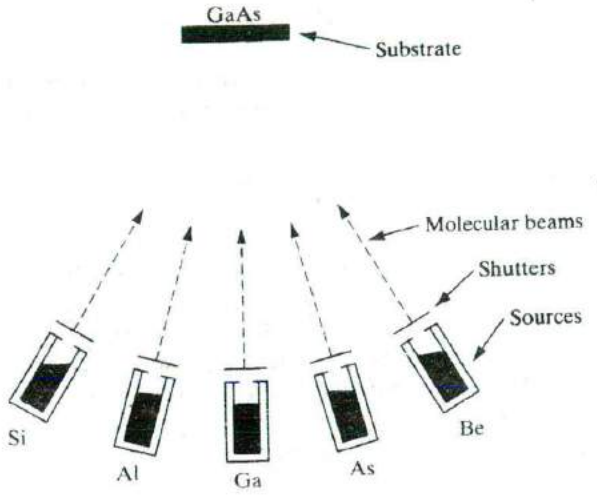
$$(CH_3)_3Ga + AsH_3 \rightarrow GaAs + 3CH_4 \qquad (1\text{-}8)$$

This reaction takes place at about 700°C, and epitaxial growth of high-quality GaAs layers can be obtained. Other compound semiconductors can also be grown by this method. For example, trimethylaluminum can be added to the gas mixture to grow AlGaAs. This growth method is widely used in the fabrication of a variety of devices, including solar cells and lasers. The convenient variability of the gas mixture allows the growth of multiple thin layers similar to those discussed below for molecular beam epitaxy.

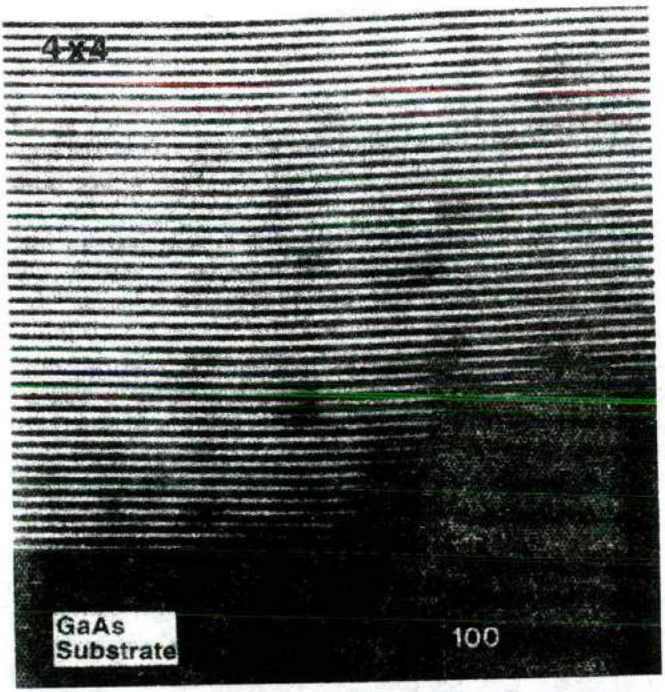
1.4.3 Molecular Beam Epitaxy

One of the most versatile techniques for growing epitaxial layers is called *molecular beam epitaxy* (*MBE*). In this method the substrate is held in a high vacuum while molecular or atomic beams of the constituents impinge upon its surface (Fig. 1–16a). For example, in the growth of AlGaAs layers on GaAs substrates, the Al, Ga, and As components, along with dopants, are heated in separate cylindrical cells. Collimated beams of these constituents escape into the vacuum and are directed onto the surface of the substrate. The rates at which these atomic beams strike the surface can be closely controlled, and growth of very high quality crystals results. The sample is held at a relatively low temperature (about 600 ° C for GaAs) in this growth procedure. Abrupt changes in doping or in crystal composition (e.g., changing from GaAs to AlGaAs) can be obtained by controlling shutters in front of the individual beams. Using slow growth rates (≤ 1 μm/h), it is possible to control the shutters to make composition changes on the scale of the lattice constant. For example, Fig. 1–16b illustrates a portion of a crystal grown with alternating layers of GaAs and AlGaAs only four monolayers thick. Because of the high vacuum and close controls involved, MBE requires a rather sophisticated setup (Fig. 1–17). However, the versatility of this growth method makes it very attractive for many applications.

As MBE has developed in recent years, it has become common to replace some of the solid sources shown in Fig. 1–16 with gaseous chemical sources. This approach, called *chemical beam epitaxy*, or *gas-source MBE*, combines many of the advantages of MBE and VPE.



(a)



(b)

Figure 1–16
Crystal growth by molecular beam epitaxy (MBE): (a) evaporation cells inside a high -vacuum chamber directing beams of Al, Ga, As, and dopants onto a GaAs substrate; (b) scanning electron micrograph of the cross section of an MBE-grown crystal having alternating layers of GaAs (dark lines) and AlGaAs (light lines). Each layer is four monolayers ($4 \times a/2 = 11.3$Å) thick. (Photograph courtesy of Bell Laboratories.)

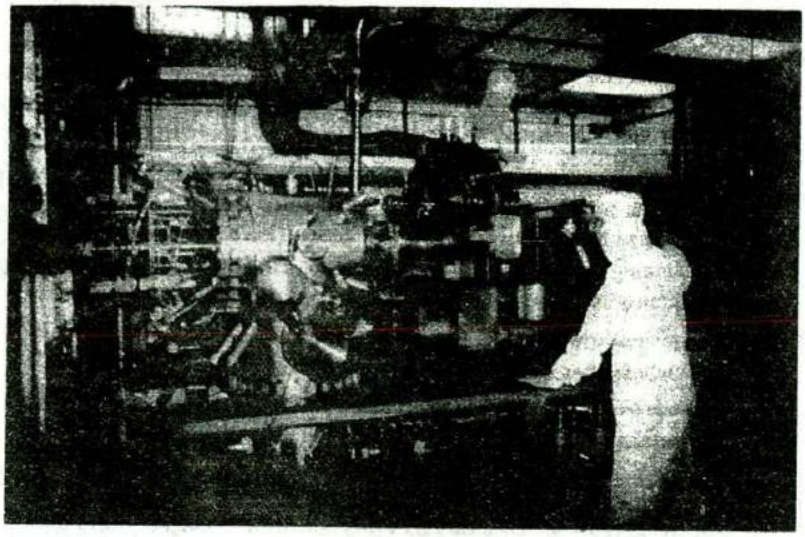


Figure 1–17 Molecular beam epitaxy facility in the Microelectronics Research Center at the University of Texas at Austin.

PROBLEMS

1.1 Using Appendix III, which of the listed semiconductors in Table 1–1 has the largest band gap? The smallest? What are the corresponding wavelengths if light is emitted at the energy E_g? Is there a noticeable pattern in the band gap energy of III–V compounds related to the column III element?

1.2 For a bcc lattice of identical atoms with a lattice constant of 5Å, calculate the maximum packing fraction and the radius of the atoms treated as hard spheres with nearest neighbors touching.

1.3 (a) Label the planes illustrated in Fig. P1–3.

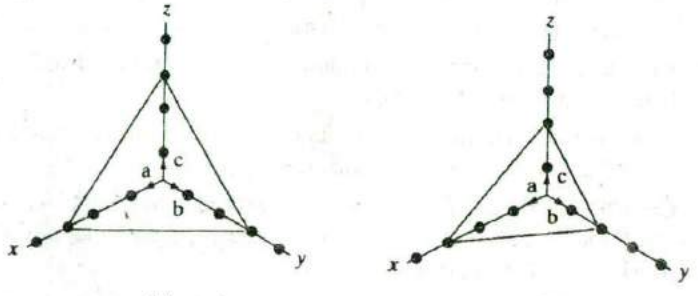


Figure P1–3

(b) Draw equivalent ⟨111⟩, ⟨100⟩, ⟨110⟩ directions in a cubic lattice; use a unit cube for illustrating each set of equivalent directions.

1.4 Calculate the volume density of Si atoms (number of atoms/cm^3) given that the lattice constant of Si is 5.43 Å. Calculate the areal density of atoms (number/cm^2) on the (110) plane. Calculate the distance between two adjacent (111) planes in Si passing through nearest-neighbor atoms.

1.5 The atomic radii of In and Sb atoms are approximately 1.44 Å and 1.36 Å, respectively. Using the hard-sphere approximation, find the lattice constant of InSb (zincblende structure), and the volume of the primitive cell. What is the atomic density on the (110) planes? (Hint: The volume of the primitive cell is $^1/_4$ the fcc unit cell volume.)

1.6 Sodium chloride (NaCl) is a cubic crystal that differs from a sc in that alternating atoms are different; each Na is surrounded by six Cl nearest neighbors and vice versa in the three-dimensional lattice. Draw a two-dimensional NaCl lattice looking down a ⟨100⟩ direction and indicate a unit cell. Remember the unit cell must be repetitive upon displacement by the basis vectors.

1.7 Sketch a view down a ⟨110⟩ direction of a diamond lattice, using Fig. 1–9 as a guide. Include lines connecting nearest neighbors.

1.8 Show by a sketch that the bcc lattice can be represented by two interpenetrating sc lattices. To simplify the sketch, show a ⟨100⟩ view of the lattice.

1.9 (a) Find the number of atoms/cm^2 on the (100) surface of a Si wafer.

(b) What is the distance (in Å) between nearest In neighbors in InP?

1.10 The ionic radii of Na^+ (atomic weight 23) and Cl^- (atomic weight 35.5) are 1.0 and 1.8 Å, respectively. Treating the ions as hard spheres, calculate the density of NaCl. Compare this with the measured density of 2.17 g/cm^3.

1.11 The atoms seen in Fig. 1–8b along a ⟨100⟩ direction of the diamond lattice are not all coplanar. Taking the top plane of colored atoms in Fig. 1–8a to be (0), the parallel plane $a/4$ down to be $(\frac{1}{4})$, the plane through the center to be $(\frac{1}{2})$, and the second plane of black atoms to be $(\frac{3}{4})$, label the plane of each atom in Fig. 1–8b.

1.12 How many atoms are found inside a unit cell of a simple cubic, body-centered cubic, and face-centered cubic crystal? How far apart in terms of lattice constant a are nearest-neighbor atoms in each case, measured from center to center?

1.13 Draw a cube such as Fig. 1–7 and show four {111} planes with different orientations. Repeat for {110} planes.

1.14 Find the maximum fractions of the unit cell volume that can be filled by hard spheres in the sc, bcc, and diamond lattices.

1.15 Calculate the densities of Ge and InP from the lattice constants (Appendix III), atomic weights, and Avogadro's number. Compare the results with the densities given in Appendix III.

1.16 Beginning with a sketch of an fcc lattice, add atoms at $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ from each fcc atom to obtain the diamond lattice. Show that only the four added atoms in Fig. 1-8a appear in the diamond unit cell.

1.17 Assuming the lattice constant varies linearly with composition x for a ternary alloy (e.g., see the variation for InGaAs in Fig. 1–13), what composition of $AlSb_xAs_{1-x}$ is lattice matched to InP? What composition of $In_xGa_{1-x}P$ is lattice-matched to GaAs? What is the band gap energy in each case?

1.18 A Si crystal is to be pulled from the melt and doped with arsenic ($k_d = 0.3$). If the Si weighs 1 kg, how many grams of arsenic should be introduced to achieve 10^{15} cm^{-3} doping during the initial growth?

READING LIST

Ashcroft, N. W., and N. D. Mermin. *Solid State Physics.* Philadelphia: W.B. Saunders, 1976.

Buckley, D. N. "The Light Fantastic: Materials and Processing Technologies for Photonics." *The Electrochemical Society Interface* 1 (Winter 1992): 41+.

Capasso, F. "Bandgap and Interface Engineering for Advanced Electronic and Photonic Devices." *Thin Solid Films* 216 (28 August 1992): 59–67.

Denbaars, S.P. "Gallium–Nitride-Based Materials for Blue to Ultraviolet Optoelectronic Devices," *Proc. IEEE* 85(11) (November 1997): 1740–1749.

Hammond, M. L. "Epitaxial Silicon Reactor Technology—A Review: I. Reactor Technology." *Solid State Technology* 31 (May 1988): 159–64.

Herman, M. A. *Molecular Beam Epitaxy: Fundamentals and Current Status.* Berlin: Springer-Verlag, 1989.

Houng, Y. M. "Chemical Beam Epitaxy." *Critical Reviews in Solid State and Materials Sciences* 17 (1992): 277–306.

Jungbluth, E. D. "Crystal Growth Methods Shape Communications Lasers." *Laser Focus World* 29 (February 1993): 61–72.

Kasper, E., and J. F. Luy. "Molecular Beam Epitaxy of Silicon Based Electronic Structures." *Microelectronics Journal* 22 (May 1992): 5–16.

Kittel, C. *Introduction to Solid State Physics*, 6th ed. New York: Wiley, 1986.

Kuphal, E. "Liquid Phase Epitaxy." *Applied Physics* A52 (June 1991): 380–409.

Levi, B. G. "What's the Shape of Things to Come in Semiconductors." *Physics Today* 45 (September 1992): 17+.

Li, S. S. *Semiconductor Physical Electronics.* New York: Plenum Press, 1993.

Liaw, H. M., and J. W. Rose. "Silicon Vapor-Phase Epitaxy." In *Epitaxial Silicon Technology*, ed. B. J. Baliga. New York: Academic Press, 1986.

Narayanamurti, V. "Artificially Structured Thin-Film Materials and Interfaces." *Science* 235 (27 February 1987): 1023+.

Schubert, E. F. *Doping in III-V Semiconductors.* Cambridge: Cambridge University Press, 1993.

Speier, P. "MOVPE for Optoelectronics." *Microelectronic Engineering* 18 (May 1992): 1–31.

Stringfellow, G. B. *Organometallic Vapor-Phase Epitaxy.* New York: Academic Press, 1989.

Swaminathan, V., and Macrander, A. T. *Material Aspects of GaAs and InP Based Structures.* Englewood Cliffs, NJ: Prentice Hall, 1991.

Tsang, W. T. "Advances in MOVPE, MBE, and CBE." *Journal of Crystal Growth* 120 (May 1992): 1–24.

Chapter 2

Atoms and Electrons

Since this book is primarily an introduction to solid state devices, it would be preferable not to delay this discussion with subjects such as atomic theory, quantum mechanics, and electron models. However, the behavior of solid state devices is directly related to these subjects. For example, it would be difficult to understand how an electron is transported through a semiconductor device without some knowledge of the electron and its interaction with the crystal lattice. Therefore, in this chapter we shall investigate some of the important properties of electrons, with special emphasis on two points: (1) the electronic structure of atoms, and (2) the interaction of atoms and electrons with excitation, such as the absorption and emission of light. By studying electron energies in an atom, we lay the foundation for understanding the influence of the lattice on electrons participating in current flow through a solid. Our discussions concerning the interaction of light with electrons form the basis for later descriptions of changes in the conductivity of a semiconductor with optical excitation, properties of light-sensitive devices, and lasers.

First, we shall investigate some of the experimental observations which led to the modern concept of the atom, and then we shall give a brief introduction to the theory of quantum mechanics. Several important concepts will emerge from this introduction: the electrons in atoms are restricted to certain energy levels by quantum rules; the electronic structure of atoms is determined from these quantum conditions; and this "quantization" defines certain allowable transitions involving absorption and emission of energy by the electrons.

2.1 INTRODUCTION TO PHYSICAL MODELS

The main effort of science is to describe what happens in nature, in as complete and concise a form as possible. In physics this effort involves observing natural phenomena, relating these observations to previously established theory, and finally establishing a physical model for the observations. The primary purpose of the model is to allow the information obtained in present observations to be used to understand new experiments. Therefore, the most useful models are expressed mathematically, so that quantitative explanations of new experiments can be made succinctly in terms of established principles. For example, we can explain the behavior of a spring-supported weight

moving up and down periodically after an initial displacement, because the differential equations describing such a simple harmonic motion have been established and are understood by students of elementary physics. But the physical model upon which these equations of motion are based arises from serious study of natural phenomena such as gravitational force, the response of bodies to accelerating forces, the relationship of kinetic and potential energy, and the properties of springs. The mass and spring problem is relatively easy to solve because each of these properties of nature is well understood.

When a new physical phenomenon is observed, it is necessary to find out how it fits into the established models and "laws" of physics. In the vast majority of cases this involves a direct extension of the mathematics of well-established models to the particular conditions of the new problem. In fact, it is not uncommon for a scientist or engineer to predict that a new phenomenon should occur before it is actually observed, simply by a careful study and extension of existing models and laws. The beauty of science is that natural phenomena are not isolated events but are related to other events by a few analytically describable laws. However, it does happen occasionally that a set of observations cannot be described in terms of existing theories. In such cases it is necessary to develop models which are based as far as possible on existing laws, but which contain new aspects arising from the new phenomena. Postulating new physical principles is a serious business, and it is done only when there is no possibility of explaining the observations with established theory. When new assumptions and models are made, their justification lies in the following question: "Does the model describe precisely the observations, and can reliable predictions be made based on the model?" The model is good or poor depending on the answer to this question.

In the 1920s it became necessary to develop a new theory to describe phenomena on the atomic scale. A long series of careful observations had been made that clearly indicated that many events involving electrons and atoms did not obey the classical laws of mechanics. It was necessary, therefore, to develop a new kind of mechanics to describe the behavior of particles on this small scale. This new approach, called *quantum mechanics*, describes atomic phenomena very well and also properly predicts the way in which electrons behave in solids—our primary interest here. Through the years, quantum mechanics has been so successful that now it stands beside the classical laws as a valid description of nature.

A special problem arises when students first encounter the theory of quantum mechanics. The problem is that quantum concepts are largely mathematical in nature and do not involve the "common sense" quality associated with classical mechanics. At first, many students find quantum concepts difficult, not so much because of the mathematics involved, but because they feel the concepts are somehow divorced from "reality." This is a reasonable reaction, since ideas which we consider to be real or intuitively satisfying are usually based on our own observation. Thus the classical laws of motion are easy to understand because we observe bodies in motion every day. On the

other hand, we observe the effects of atoms and electrons only indirectly, and naturally we have very little feeling for what is happening on the atomic scale. It is necessary, therefore, to depend on the facility of the theory to predict experimental results rather than to attempt to force classical analogues onto the nonclassical phenomena of atoms and electrons.

Our approach in this chapter will be to investigate the important experimental observations that led to the quantum theory, and then to indicate how the theory accounts for these observations. Discussions of quantum theory must necessarily be largely qualitative in such a brief presentation, and those topics that are most important to solid state theory will be emphasized here. Several good references for further individual study are given at the end of this chapter.

2.2 EXPERIMENTAL OBSERVATIONS

The experiments that led to the development of quantum theory were concerned with the nature of light and the relation of optical energy to the energies of electrons within atoms. These experiments supplied only indirect evidence of the nature of phenomena on the atomic scale; however, the cumulative results of a number of careful experiments showed clearly that a new theory was needed.

2.2.1 The Photoelectric Effect

An important observation by Planck indicated that radiation from a heated sample is emitted in discrete units of energy, called *quanta;* the energy units were described by $h\nu$, where ν is the frequency of the radiation, and h is a quantity now called Planck's constant ($h = 6.63 \times 10^{-34}$ J-s). Soon after Planck developed this hypothesis, Einstein interpreted an important experiment that clearly demonstrated the discrete nature (*quantization*) of light. This experiment involved absorption of optical energy by the electrons in a metal and the relationship between the amount of energy absorbed and the frequency of the light (Fig. 2–1). Let us suppose that monochromatic light is incident on the surface of a metal plate in a vacuum. The electrons in the metal absorb energy from the light, and some of the electrons receive enough energy to be ejected from the metal surface into the vacuum. This phenomenon is called the *photoelectric effect.* If the energy of the escaping electrons is measured, a plot can be made of the maximum energy as a function of the frequency ν of the incident light (Fig. 2–1b).

One simple way of finding the maximum energy of the ejected electrons is to place another plate above the one shown in Fig. 2–1a and then create an electric field between the two plates. The potential necessary to retard all electron flow between the plates gives the energy E_m. For a particular frequency of light incident on the sample, a maximum energy E_m is observed for the emitted electrons. The resulting plot of E_m vs. ν is linear, with a slope equal to Planck's constant. The equation of the line shown in Fig. 2–1b is

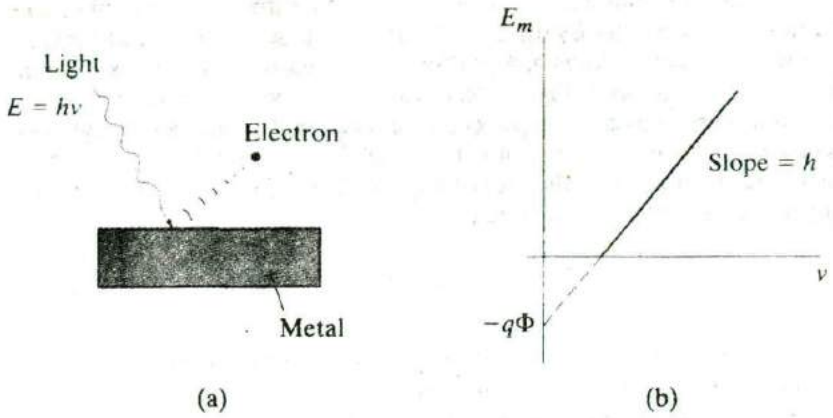


Figure 2–1
The photoelectric effect: (a) electrons are ejected from the surface of a metal when exposed to light of frequency ν in a vacuum; (b) plot of the maximum kinetic energy of ejected electrons vs. frequency of the incoming light.

$$E_m = h\nu - q\Phi \tag{2–1}$$

where q is the magnitude of the electronic charge. The quantity Φ (volts) is a characteristic of the particular metal used. When Φ is multiplied by the electronic charge, an energy (joules) is obtained which represents the minimum energy required for an electron to escape from the metal into a vacuum. The energy $q\Phi$ is called the *work function* of the metal. These results indicate that the electrons receive an energy $h\nu$ from the light and lose an amount of energy $q\Phi$ in escaping from the surface of the metal.

This experiment demonstrates clearly that Planck's hypothesis was correct—light energy is contained in discrete units rather than in a continuous distribution of energies. Other experiments also indicate that, in addition to the wave nature of light, the quantized units of light energy can be considered as localized packets of energy, called *photons*. Some experiments emphasize the wave nature of light, while other experiments reveal the discrete nature of photons. This duality is fundamental to quantum processes and does not imply an ambiguity in the theory.

2.2.2 Atomic Spectra

One of the most valuable experiments of modern physics is the analysis of absorption and emission of light by atoms. For example, an electric discharge can be created in a gas, so that the atoms begin to emit light with wavelengths characteristic of the gas. We see this effect in a neon sign, which is typically a glass tube filled with neon or a gas mixture, with electrodes for creating a discharge. If the intensity of the emitted light is measured as a function of

wavelength, one finds a series of sharp lines rather than a continuous distribution of wavelengths. By the early 1900s the characteristic spectra for several atoms were well known. A portion of the measured emission spectrum for hydrogen is shown in Fig. 2–2, in which the vertical lines represent the positions of observed emission peaks on the wavelength scale. Wavelength (λ) is usually measured in angstroms ($1\ \text{Å} = 10^{-10}$ m) and is related (in meters) to frequency by $\lambda = c/\nu$, where c is the speed of light (3×10^8 m/s). Photo energy $h\nu$ is then related to wavelength by

$$E = h\nu = \frac{hc}{\lambda} \tag{2–2}$$

The lines in Fig. 2–2 appear in several groups labeled the *Lyman*, *Balmer*, and *Paschen* series after their early investigators. Once the hydrogen spectrum was established, scientists noticed several interesting relationships among the lines. The various series in the spectrum were observed to follow certain empirical forms:

$$\text{Lyman:}\quad \nu = cR\left(\frac{1}{1^2} - \frac{1}{\mathbf{n}^2}\right),\quad \mathbf{n} = 2, 3, 4, \ldots \tag{2–3a}$$

$$\text{Balmer:}\quad \nu = cR\left(\frac{1}{2^2} - \frac{1}{\mathbf{n}^2}\right),\quad \mathbf{n} = 3, 4, 5, \ldots \tag{2–3b}$$

$$\text{Paschen:}\quad \nu = cR\left(\frac{1}{3^2} - \frac{1}{\mathbf{n}^2}\right),\quad \mathbf{n} = 4, 5, 6, \ldots \tag{2–3c}$$

where R is a constant called the Rydberg constant ($R = 109{,}678\ \text{cm}^{-1}$). If the photon energies $h\nu$ are plotted for successive values of the integer $\mathbf{n}$, we notice that each energy can be obtained by taking sums and differences of other photon energies in the spectrum (Fig. 2–3). For example, E_{42} in the Balmer series is the difference between E_{41} and E_{21} in the Lyman series. This relationship among the various series is called the *Ritz combination*

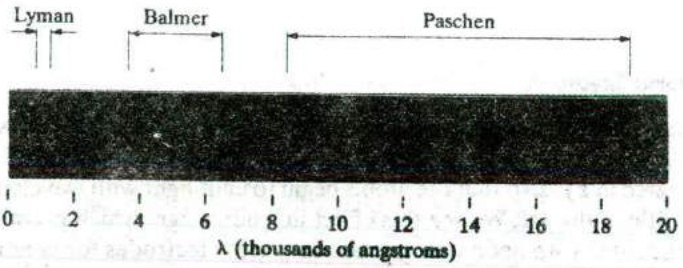


Figure 2–2
Some important lines in the emission spectrum of hydrogen.

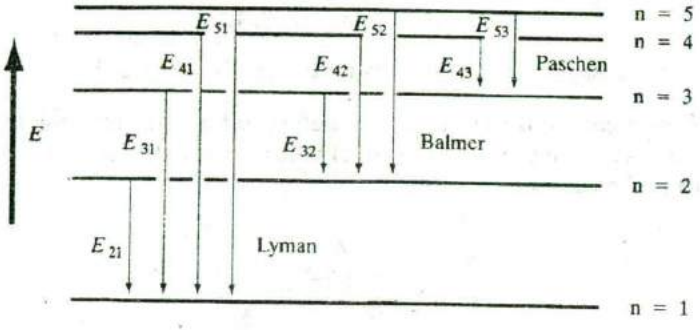


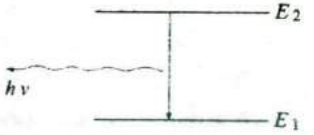
Figure 2–3 Relationships among photon energies in the hydrogen spectrum.

principle. Naturally, these empirical observations stirred a great deal of interest in constructing a comprehensive theory for the origin of the photons given off by atoms.

2.3 THE BOHR MODEL

The results of emission spectra experiments led Niels Bohr to construct a model for the hydrogen atom, based on the mathematics of planetary systems. If the electron in the hydrogen atom has a series of planetary-type orbits available to it, it can be excited to an outer orbit and then can fall to any one of the inner orbits, giving off energy corresponding to one of the lines of Fig. 2–3. To develop the model, Bohr made several postulates:

1. Electrons exist in certain stable, circular orbits about the nucleus. This assumption implies that the orbiting electron does not give off radiation as classical electromagnetic theory would normally require of a charge experiencing angular acceleration; otherwise, the electron would not be stable in the orbit but would spiral into the nucleus as it lost energy by radiation.
2. The electron may shift to an orbit of higher or lower energy, thereby gaining or losing energy equal to the difference in the energy levels (by absorption or emission of a photon of energy $h\nu$).



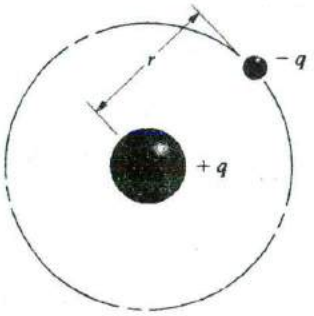
$$h\nu = E_2 - E_1 \qquad (2\text{–}4)$$

3. The angular momentum p_θ of the electron in an orbit is always an integral multiple of Planck's constant divided by 2π ($h/2\pi$ is often abbreviated $\hbar$ for convenience). This assumption,

$$p_\theta = n\hbar, \quad n = 1, 2, 3, 4, \ldots \tag{2–5}$$

is necessary to obtain the observed results of Fig. 2–3.

If we visualize the electron in a stable orbit of radius r about the proton of the hydrogen atom, we can equate the electrostatic force between the charges to the centripetal force:



$$-\frac{q^2}{Kr^2} = -\frac{mv^2}{r} \tag{2–6}$$

where $K = 4\pi\epsilon_0$ in MKS units, m is the mass of the electron, and v is its velocity. From assumption 3 we have

$$p_\theta = mvr = n\hbar \tag{2–7}$$

Since n takes on integral values, r should be denoted by r_n to indicate the nth orbit. Then Eq. (2–7) can be written

$$m^2v^2 = \frac{n^2\hbar^2}{r_n^2} \tag{2–8}$$

Substituting Eq. (2–8) in Eq. (2–6) we find that

$$\frac{q^2}{Kr_n^2} = \frac{1}{mr_n} \cdot \frac{n^2\hbar^2}{r_n^2} \tag{2–9}$$

$$r_n = \frac{Kn^2\hbar^2}{mq^2} \tag{2–10}$$

for the radius of the nth orbit of the electron. Now we must find the expression for the total energy of the electron in this orbit, so that we can calculate the energies involved in transitions between orbits.

From Eqs. (2–7) and (2–10) we have

$$v = \frac{\mathbf{n}\hbar}{mr_n} \tag{2–11}$$

$$v = \frac{\mathbf{n}\hbar q^2}{K\mathbf{n}^2\hbar^2} = \frac{q^2}{K\mathbf{n}\hbar} \tag{2–12}$$

Therefore, the kinetic energy of the electron is

$$\text{K. E.} = \frac{1}{2}mv^2 = \frac{mq^4}{2K^2\mathbf{n}^2\hbar^2} \tag{2–13}$$

The potential energy is the product of the electrostatic force and the distance between the charges:

$$\text{P. E.} = -\frac{q^2}{Kr_n} = -\frac{mq^4}{K^2\mathbf{n}^2\hbar^2} \tag{2–14}$$

Thus the total energy of the electron in the **n**th orbit is

$$E_n = \text{K. E.} + \text{P. E.} = -\frac{mq^4}{2K^2\mathbf{n}^2\hbar^2} \tag{2–15}$$

The critical test of the model is whether energy differences between orbits correspond to the observed photon energies of the hydrogen spectrum. The transitions between orbits corresponding to the Lyman, Balmer, and Paschen series are illustrated in Fig. 2–4. The energy difference between orbits $\mathbf{n}_1$ and $\mathbf{n}_2$ is given by

$$E_{n2} - E_{n1} = \frac{mq^4}{2K^2\hbar^2}\left(\frac{1}{\mathbf{n}_1^2} - \frac{1}{\mathbf{n}_2^2}\right) \tag{2–16}$$

The frequency of light given off by a transition between these orbits is

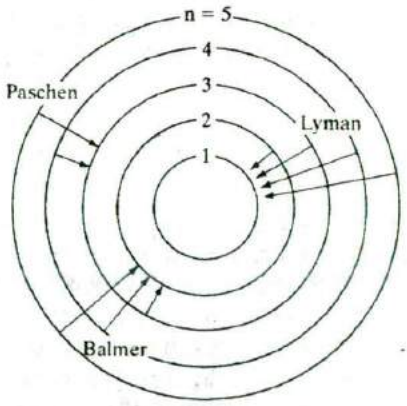


Figure 2–4
Electron orbits and transitions in the Bohr model of the hydrogen atom. Orbit spacing is not drawn to scale.

$$\nu_{21} = \left[\frac{mq^4}{2K^2\hbar^2 h}\right]\left(\frac{1}{\mathbf{n}_1^2} - \frac{1}{\mathbf{n}_2^2}\right) \tag{2–17}$$

The factor in brackets is essentially the Rydberg constant R times the speed of light c. A comparison of Eq. (2–17) with the experimental results summed up by Eq. (2–3) indicates that the Bohr theory provides a good model for electronic transitions within the hydrogen atom, as far as the early experimental evidence is concerned.

Whereas the Bohr model accurately describes the gross features of the hydrogen spectrum, it does not include many fine points. For example, experimental evidence indicates some splitting of levels in addition to the levels predicted by the theory. Also, difficulties arise in extending the model to atoms more complicated than hydrogen. Attempts were made to modify the Bohr model for more general cases, but it soon became obvious that a more comprehensive theory was needed. However, the partial success of the Bohr model was an important step toward the eventual development of the quantum theory. The concept that electrons are quantized in certain allowed energy levels, and the relationship of photon energy and transitions between levels had been established firmly by the Bohr theory.

2.4 QUANTUM MECHANICS

The principles of quantum mechanics were developed from two different points of view at about the same time (the late 1920s). One approach, developed by Heisenberg, utilizes the mathematics of matrices and is called *matrix mechanics*. Independently, Schrödinger developed an approach utilizing a wave equation, now called *wave mechanics*. These two mathematical formulations appear to be quite different. However, closer examination reveals that beyond the formalism, the basic principles of the two approaches are the same. It is possible to show, for example, that the results of matrix mechanics reduce to those of wave mechanics after mathematical manipulation. We shall concentrate here on the wave mechanics approach, since solutions to a few simple problems can be obtained with it, involving less mathematical discussion.

2.4.1 Probability and the Uncertainty Principle

It is impossible to describe with absolute precision events involving individual particles on the atomic scale. Instead, we must speak of the average values (*expectation values*) of position, momentum, and energy of a particle such as an electron. It is important to note, however, that the uncertainties revealed in quantum calculations are not based on some shortcoming of the theory. In fact, a major strength of the theory is that it describes the probabilistic nature of events involving atoms and electrons. The fact is that such quantities as the position and momentum of an electron *do not exist* apart

from a particular uncertainty. The magnitude of this inherent uncertainty is described by the *Heisenberg uncertainty principle:*[1]

> In any measurement of the position and momentum of a particle, the uncertainties in the two measured quantities will be related by

$$\boxed{(\Delta x)\,(\Delta p_x) \geq \hbar} \qquad (2\text{–}18)$$

> Similarly, the uncertainties in an energy measurement will be related to the uncertainty in the time at which the measurement was made by

$$\boxed{(\Delta E)\,(\Delta t) \geq \hbar} \qquad (2\text{–}19)$$

These limitations indicate that simultaneous measurement of position and momentum or of energy and time are inherently inaccurate to some degree. Of course, Planck's constant h is a rather small number (6.63×10^{-34} J-s), and we are not concerned with this inaccuracy in the measurement of x and p_x for a truck, for example. On the other hand, measurements of the position of an electron and its speed are seriously limited by the uncertainty principle.

One implication of the uncertainty principle is that we cannot properly speak of *the* position of an electron, for example, but must look for the "probability" of finding an electron at a certain position. Thus one of the important results of quantum mechanics is that a *probability density function* can be obtained for a particle in a certain environment, and this function can be used to find the expectation value of important quantities such as position, momentum, and energy. We are familiar with the methods for calculating discrete (single-valued) probabilities from common experience. For example, it is clear that the probability of drawing a particular card out of a random deck is $^1/_{52}$, and the probability that a tossed coin will come up heads is $^1/_2$. The techniques for making predictions when the probability varies are less familiar, however. In such cases it is common to define a probability of finding a particle within a certain volume. Given a probability density function $P(x)$ for a one-dimensional problem, the probability of finding the particle in a range from x to $x + dx$ is $P(x)dx$. Since the particle will be *somewhere*, this definition implies that

$$\int_{-\infty}^{\infty} P(x)dx = 1 \qquad (2\text{–}20)$$

if the function $P(x)$ is properly chosen. Equation (2–20) is implied by stating that the function $P(x)$ is *normalized* (i.e., the integral equals unity).

[1] This is often called the *principle of indeterminacy.*

To find the average value of a function of x, we need only multiply the value of that function in each increment dx by the probability of finding the particle in that dx and sum over all x. Thus the average value of $f(x)$ is

$$\langle f(x) \rangle = \int_{-\infty}^{\infty} f(x) P(x) dx \tag{2-21a}$$

If the probability density function is not normalized, this equation should be written

$$\langle f(x) \rangle = \frac{\int_{-\infty}^{\infty} f(x) P(x) dx}{\int_{-\infty}^{\infty} P(x) dx} \tag{2-21b}$$

2.4.2 The Schrödinger Wave Equation

There are several ways to develop the wave equation by applying quantum concepts to various classical equations of mechanics. One of the simplest approaches is to consider a few basic postulates, develop the wave equation from them, and rely on the accuracy of the results to serve as a justification of the postulates. In more advanced texts these assumptions are dealt with in more convincing detail.

Basic Postulates

1. Each particle in a physical system is described by a wave function $\Psi(x, y, z, t)$. This function and its space derivative ($\partial\Psi/\partial x + \partial\Psi/\partial y + \partial\Psi/\partial z$) are continuous, finite, and single valued.
2. In dealing with classical quantities such as energy E and momentum p, we must relate these quantities with abstract quantum mechanical operators defined in the following way:

Classical variable	Quantum operator
x	x
$f(x)$	$f(x)$
$\mathrm{p}(x)$	$\frac{\hbar}{j}\frac{\partial}{\partial x}$
E	$-\frac{\hbar}{j}\frac{\partial}{\partial t}$

and similarly for the other two directions.

3. The probability of finding a particle with wave function Ψ in the volume $dx\,dy\,dz$ is $\Psi^*\Psi\,dx\,dy\,dz$.[2] The product $\Psi^*\Psi$ is normalized according to Eq. (2–20) so that

$$\int_{-\infty}^{\infty} \Psi^*\Psi\,dx\,dy\,dz = 1$$

and the average value $\langle Q \rangle$ of any variable Q is calculated from the wave function by using the operator form Q_{op} defined in postulate 2:

$$\langle Q \rangle = \int_{-\infty}^{\infty} \Psi^* Q_{op} \Psi\,dx\,dy\,dz$$

Once we find the wave function Ψ for a particle, we can calculate its average position, energy, and momentum, within the limits of the uncertainty principle. Thus, a major part of the effort in quantum calculations involves solving for Ψ within the conditions imposed by a particular physical system. We notice from assumption 3 that the probability density function is $\Psi^*\Psi$, or $|\Psi|^2$.

The classical equation for the energy of a particle can be written:

$$\begin{array}{ccccc} \text{Kinetic energy} & + & \text{potential energy} & = & \text{total energy} \\ \dfrac{1}{2m}\mathsf{p}^2 & + & \mathsf{V} & = & E \end{array} \qquad (2\text{–}22)$$

In quantum mechanics we use the operator form for these variables (postulate 2); the operators are allowed to operate on the wave function Ψ. For a one-dimensional problem Eq. (2–22) becomes[3]

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2} + \mathsf{V}(x)\Psi(x,t) = -\frac{\hbar}{j}\frac{\partial\Psi(x,t)}{\partial t} \qquad (2\text{–}23)$$

which is the Schrödinger wave equation. In three dimensions the equation is

$$\boxed{-\frac{\hbar^2}{2m}\nabla^2\Psi + \mathsf{V}\Psi = -\frac{\hbar}{j}\frac{\partial\Psi}{\partial t}} \qquad (2\text{–}24)$$

where $\nabla^2\Psi$ is

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2}$$

[2] Ψ^* is the *complex conjugate* of Ψ, obtained by reversing the sign on each j. Thus $(e^{jx})^* = e^{-jx}$.

[3] The operational interpretation of $(\partial/\partial x)^2$ is the second derivative form $\partial^2/\partial x^2$; the square of j is -1.

The wave function Ψ in Eqs. (2–23) and (2–24) includes both space and time dependencies. It is common to calculate these dependencies separately and combine them later. Furthermore, many problems are time independent, and only the space variables are necessary. Thus we try to solve the wave equation by breaking it into two equations by the technique of separation of variables. Let $\Psi(x, t)$ be represented by the product $\psi(x)\phi(t)$. Using this product in Eq. (2–23) we obtain

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi(x)}{\partial x^2}\phi(t) + V(x)\psi(x)\phi(t) = -\frac{\hbar}{j}\psi(x)\frac{\partial\phi(t)}{\partial t} \tag{2–25}$$

Now the variables can be separated to obtain the time-dependent equation in one dimension,

$$\boxed{\frac{d\phi(t)}{dt} + \frac{jE}{\hbar}\phi(t) = 0} \tag{2–26}$$

and the time-independent equation,

$$\boxed{\frac{d^2\psi(x)}{dx^2} + \frac{2m}{\hbar^2}[E - V(x)]\psi(x) = 0} \tag{2–27}$$

We can show that the separation constant E corresponds to the energy of the particle when particular solutions are obtained, such that a wave function ψ_n corresponds to a particle energy E_n.

These equations are the basis of wave mechanics. From them we can determine the wave functions for particles in various simple systems. For calculations involving electrons, the potential term $V(x)$ usually results from an electrostatic or magnetic field.

2.4.3 Potential Well Problem

It is quite difficult to find solutions to the Schrödinger equation for most realistic potential fields. One can solve the problem with some effort for the hydrogen atom, for example, but solutions for more complicated atoms are hard to obtain. There are several important problems, however, which illustrate the theory without complicated manipulation. The simplest problem is the potential energy well with infinite boundaries. Let us assume a particle is trapped in a potential well with $V(x)$ zero except at the boundaries $x = 0$ and L, where it is infinitely large (Fig. 2–5a)

$$\begin{aligned} V(x) &= 0, \quad 0 < x < L \\ V(x) &= \infty, \quad x = 0, L \end{aligned} \tag{2–28}$$

Inside the well we set $V(x) = 0$ in Eq. (2–27)

$$\frac{d^2\psi(x)}{dx^2} + \frac{2m}{\hbar^2}E\psi(x) = 0, \quad 0 < x < L \tag{2–29}$$

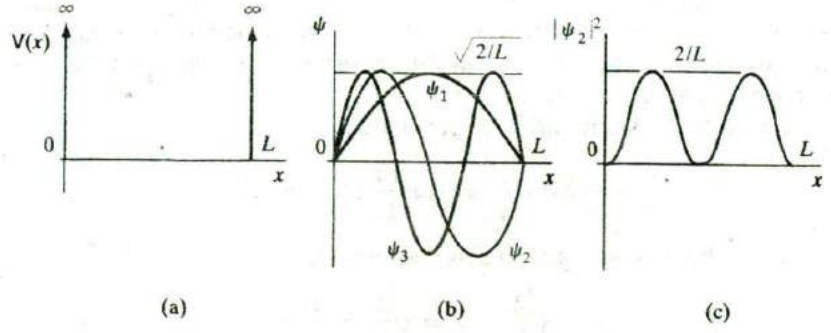


Figure 2–5
The problem of a particle in a potential well: (a) potential energy diagram; (b) wave functions in the first three quantum states; (c) probability density distribution for the second state.

This is the wave equation for a free particle; it applies to the potential well problem in the region with no potential $V(x)$.

Possible solutions to Eq. (2–29) are sin kx and cos kx, where k is $\sqrt{2mE}/\hbar$ In choosing a solution, however, we must examine the boundary conditions. The only allowable value of ψ at the walls is zero. Otherwise, there would be a nonzero $|\psi|^2$ outside the potential well, which is impossible because a particle cannot penetrate an infinite barrier. Therefore, we must choose only the sine solution and define k such that sin kx goes to zero at $x = L$.

$$\psi = A \sin kx, \quad k = \frac{\sqrt{2mE}}{\hbar} \tag{2–30}$$

The constant A is the amplitude of the wave function and will be evaluated from the normalization condition (postulate 3). If ψ is to be zero at $x = L$, then k must be some integral multiple of π/L.

$$k = \frac{n\pi}{L}, \quad n = 1, 2, 3, \ldots \tag{2–31}$$

From Eqs. (2–30) and (2–31) we can solve for the total energy E_n for each value of the integer **n**.

$$\frac{\sqrt{2mE_n}}{\hbar} = \frac{n\pi}{L} \tag{2–32}$$

$$E_n = \frac{n^2\pi^2\hbar^2}{2mL^2} \tag{2–33}$$

Thus for each allowable value of **n** the particle energy is described by Eq. (2–33). We notice that the *energy is quantized.* Only certain values of energy are allowed. The integer **n** is called a *quantum number;* the particular wave function ψ_n and corresponding energy state E_n describe the *quantum state* of the particle.

The quantized energy levels described by Eq. (2–33) appear in a variety of small-geometry structures encountered in semiconductor devices. We shall return to this potential well problem (often called the "particle in a box" problem) in later discussions.

The constant A is found from postulate 3.

$$\int_{-\infty}^{\infty} \psi^* \psi \, dx = \int_0^L A^2 \left(\sin \frac{n\pi}{L} x \right)^2 dx = A^2 \frac{L}{2} \tag{2–34}$$

Setting Eq. (2–34) equal to unity we obtain.

$$A = \sqrt{\frac{2}{L}}, \quad \psi_n = \sqrt{\frac{2}{L}} \sin \frac{n\pi}{L} x \tag{2–35}$$

The first three wave functions ψ_1, ψ_2, ψ_3, are sketched in Fig. 2–5b. The probability density function $\psi^*\psi$, or $|\psi|^2$, is sketched for ψ_2 in Fig. 2–5c.

2.4.4 Tunneling

The wave functions are relatively easy to obtain for the potential well with infinite walls, since the boundary conditions force ψ to zero at the walls. A slight modification of this problem illustrates a principle that is very important in some solid state devices—the quantum mechanical *tunneling* of an electron through a barrier of finite height and thickness. Let us consider the potential barrier of Fig. 2–6. If the barrier is not infinite, the boundary conditions do not force ψ to zero at the barrier. Instead, we must use the condition that ψ and its slope $d\psi/dx$ are continuous at each boundary of the barrier (postulate 1). Thus ψ must have a nonzero value within the barrier and also on the other side. Since ψ has a value to the right of the barrier, $\psi^*\psi$ exists

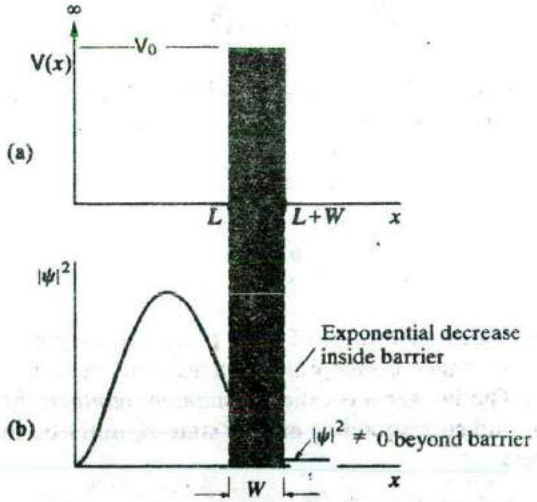


Figure 2–6 Quantum mechanical tunneling: (a) potential barrier of height V_0 and thickness W; (b) probability density for an electron with energy $E < V_0$, indicating a nonzero value of the wave function beyond the barrier.

there also, implying that there is some probability of finding the particle beyond the barrier. We notice that the particle does not go over the barrier; its total energy is assumed to be less than the barrier height V_0. The mechanism by which the particle "penetrates" the barrier is called tunneling. However, no classical analogue, including classical descriptions of tunneling through barriers, in appropriate for this effect. Quantum mechanical tunneling is intimately bound to the uncertainty principle. If the barrier is sufficiently thin, we cannot say with certainty that the particle exists only on one side. However, the wave function amplitude for the particle is reduced by the barrier as Fig. 2–6 indicates, so that by making the thickness W greater, we can reduce ψ on the right-hand side to the point that negligible tunneling occurs. Tunneling is important only over very small dimensions, but it can be of great importance in the conduction of electrons in solids, as we shall see in Chapters 5, 6 and 11.

Recently, a novel electronic device called the resonant tunneling diode was developed. This device operates by tunneling electrons through "particle in a potential well" energy levels of the type described in Section 2.4.3.

2.5 ATOMIC STRUCTURE AND THE PERIODIC TABLE

The Schrödinger equation describes accurately the interactions of particles with potential fields, such as electrons within atoms. Indeed, the modern understanding of atomic theory (the modern atomic *models*) comes from the wave equation and from Heisenberg's matrix mechanics. It should be pointed out, however, that the problem of solving the Schrödinger equation directly for complicated atoms is extremely difficult. In fact, only the hydrogen atom is generally solved directly; atoms of atomic number greater than one are usually handled by techniques involving approximations. Many atoms such as the alkali metals (Li, Na, etc.), which have a neutral core with a single electron in an outer orbit, can be treated by a rather simple extension of the hydrogen atom results. The hydrogen atom solution is also important in identifying the basic selection rules for describing allowed electron energy levels. These quantum mechanical results must coincide with the experimental spectra, and we expect the energy levels to include those predicted by the Bohr model. Without actually working through the mathematics for the hydrogen atom, in this section we shall investigate the energy level schemes dictated by the wave equation.

2.5.1 The Hydrogen Atom

Finding the wave functions for the hydrogen atom requires a solution of the Schrödinger equation in three dimensions for a coulombic potential field. Since the problem is spherically symmetric, the spherical coordinate system is used in the calculation (Fig. 2–7). The term $V(x, y, z)$ in Eq. (2–24) must be replaced by $V(r, \theta, \phi)$, representing the Coulomb potential which the electron

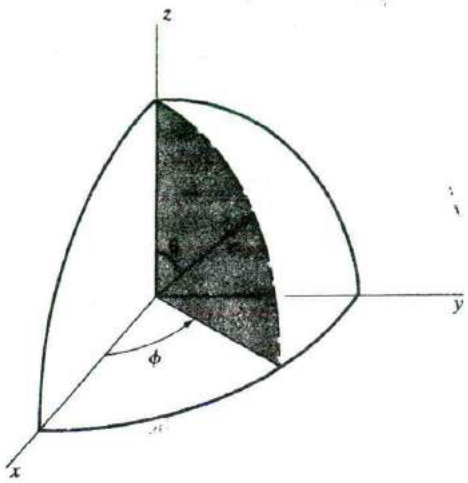


Figure 2–7
The spherical coordinate system.

experiences in the vicinity of the proton. The Coulomb potential varies only with r in spherical coordinates

$$V(r, \theta, \phi) = V(r) = -(4\pi\epsilon_0)^{-1}\frac{q^2}{r} \tag{2–36}$$

as in Eq. (2–14).

When the separation of variables is made, the time-independent equation can be written as

$$\psi(r, \theta, \phi) = R(r)\Theta(\theta)\Phi(\phi) \tag{2–37}$$

Thus the wave functions are found in three parts. Separate solutions must be obtained for the r-dependent equation, the θ-dependent equation, and the ϕ-dependent equation. After these three equations are solved, the total wave function ψ is obtained from the product.

As in the simple potential well problem, each of the three hydrogen atom equations gives a solution which is quantized. Thus we would expect a quantum number to be associated with *each* of the three parts of the wave equation. As an illustration, the ϕ-dependent equation obtained after separation of variables is

$$\frac{d^2\Phi}{d\phi^2} + \mathbf{m}^2\Phi = 0 \tag{2–38}$$

where $\mathbf{m}$ is a quantum number. The solution to this equation is

$$\Phi_m(\phi) = Ae^{j\mathbf{m}\phi} \tag{2–39}$$

where A can be evaluated by the normalization condition, as before:

$$\int_0^{2\pi} \Phi_m^*(\phi)\Phi_m(\phi)d\phi = 1 \tag{2–40}$$

$$A^2\int_0^{2\pi} e^{-jm\phi}e^{jm\phi}\,d\phi = A^2\int_0^{2\pi} d\phi = 2\pi A^2 \tag{2–41}$$

Thus the value of A is

$$A = \frac{1}{\sqrt{2\pi}} \tag{2–42}$$

and the ϕ-dependent wave function is

$$\Phi_m(\phi) = \frac{1}{\sqrt{2\pi}}e^{jm\phi} \tag{2–43}$$

Since values of ϕ repeat every 2π radians, Φ should repeat also. This occurs if **m** is an integer, including negative integers and zero. Thus the wave functions for the ϕ-dependent equation are quantized with the following selection rule for the quantum numbers:

$$\mathbf{m} = \ldots, -3, -2, -1, 0, +1, +2, +3, \ldots \tag{2–44}$$

By similar treatments, the functions $R(r)$ and $\Theta(\theta)$ can be obtained, each being quantized by its own selection rule. For the r-dependent equation, the quantum number **n** can be any positive integer (not zero), and for the θ-dependent equation the quantum number l can be zero or a positive integer. However, there are interrelationships among the equations which restrict the various quantum numbers used with a single wave function ψ_{nlm}:

$$\psi_{nlm}(r, \theta, \phi) = R_n(r)\Theta_l(\theta)\Phi_m(\phi) \tag{2–45}$$

These restrictions are summarized as follows:

$$\mathbf{n} = 1, 2, 3, \ldots \tag{2–46a}$$

$$l = 0, 1, 2, \ldots, (\mathbf{n} - 1) \tag{2–46b}$$

$$\mathbf{m} = -l, \ldots, -2, -1, 0, +1, +2, \ldots, +l \tag{2–46c}$$

In addition to the three quantum numbers arising from the three parts of the wave equation, there is an important quantization condition on the "spin" of the electron. Investigations of electron spin employ the theory of relativity as well as quantum mechanics; therefore, we shall simply state that the intrinsic angular momentum **s** of an electron with ψ_{nlm} specified is

$$\mathbf{s} = \pm\frac{\hbar}{2} \tag{2–47}$$

That is, in units of $\hbar$, the electron has a spin of $\frac{1}{2}$, and the angular momentum produced by this spin is positive or negative depending on whether

the electron is "spin up" or "spin down." The important point for our discussion is that each allowed energy state of the electron in the hydrogen atom is uniquely described by four quantum numbers: **n**, ***l***, **m** and **s**.[4]

Using these four quantum numbers, we can identify the various states which the electron can occupy in a hydrogen atom. The number **n**, called the *principal* quantum number, specifies the "orbit" of the electron in Bohr terminology. Of course, the concept of orbit is replaced by probability density functions in quantum mechanical calculations. It is common to refer to states with a given principal quantum number as belonging to a *shell* rather than an orbit.

There is considerable fine structure in the energy levels about the Bohr orbits, due to the dictates of the other three quantum conditions. For example, an electron with $\mathbf{n} = 1$ (the first Bohr orbit) can have only $\boldsymbol{l} = 0$ and $\mathbf{m} = 0$ according to Eq. (2–46), but there are two spin states allowed from Eq. (2–47). For $\mathbf{n} = 2$, $\boldsymbol{l}$ can be 0 or 1, and **m** can be $-1, 0$, or $+1$. The various allowed combinations of quantum numbers appear in the first four columns of Table 2–1. From these combinations it is apparent that the electron in a hydrogen atom can occupy any one of a large number of excited states in addition to the lowest (*ground*) state ψ_{100}. Energy differences between the various states properly account for the observed lines in the hydrogen spectrum.

2.5.2 The Periodic Table

The quantum numbers discussed in Section 2.5.1 arise from the solutions to the hydrogen atom problem. Thus the energies obtainable from the wave functions are unique to the hydrogen atom and cannot be extended to more complicated atoms without appropriate alterations. However, the quantum number selection rules are valid for more complicated structures, and we can use these rules to gain an understanding of the arrangement of atoms in the periodic table of chemical elements. Without these selection rules, it is difficult to understand why only two electrons fit into the first Bohr orbit of an atom, whereas eight electrons are allowed in the second orbit. After even the brief discussion of quantum numbers given above, we should be able to answer these questions with more insight.

Before discussing the periodic table, we must be aware of an important principle of quantum theory, the *Pauli exclusion principle*. This rule states that no two electrons in an interacting system[5] can have the same set of quantum numbers **n**, ***l***, **m**, **s**. In other words, only two electrons can have the same three quantum numbers **n**, ***l***, **m**, and those two must have opposite spin. The importance of this principle cannot be

[4] In many texts the numbers we have called **m** and **s** are referred to as $\mathbf{m}_l$ and $\mathbf{m}_s$, respectively.

[5] An interacting system is one in which electron wave functions overlap—in this case an atom with two or more electrons.

overemphasized; it is basic to the electronic structure of all atoms in the periodic table. One implication of this principle is that by listing the various combinations of quantum numbers, we can determine into which shell each electron of a complicated atom fits, and how many electrons are allowed per shell. The quantum states summarized in Table 2–1 can be used to indicate the electronic configurations for atoms in the lowest energy state.

In the first electronic shell ($\mathbf{n} = 1$), $\boldsymbol{l}$ can be only zero since the maximum value of $\boldsymbol{l}$ is always $\mathbf{n} - 1$. Similarly, $\mathbf{m}$ can be only zero since $\mathbf{m}$ runs from the negative value of $\boldsymbol{l}$ to the positive value of $\boldsymbol{l}$. Two electrons with opposite spin can fit in this ψ_{100} state; therefore, the first shell can have at most two electrons. For the helium atom (atomic number $Z = 2$) in the ground state, both electrons will be in the first Bohr orbit ($\mathbf{n} = 1$), both will have $\boldsymbol{l} = 0$ and $\mathbf{m} = 0$, and they will have opposite spin. Of course, one or both of the He atom electrons can be excited to one of the higher energy states of Table 2–1 and subsequently relax to the ground state, giving off a photon characteristic of the He spectrum.

Table 2–1 Quantum numbers to $\mathbf{n} = 3$ and allowable states for the electron in a hydorgen atom: The first four columns show the various combinations of quantum numbers allowed by the selection rules of Eq. (2–46); the last two columns indicate the number of allowed states (combinations of $\mathbf{n}$, $\boldsymbol{l}$, $\mathbf{m}$, and $\mathbf{s}$) for each $\boldsymbol{l}$ (subshell) and $\mathbf{n}$ (shell, or Bohr orbit).

n	l	m	$s/\hbar$	Allowable states in subshell	Allowable states in complete shell
1	0	0	$\pm\frac{1}{2}$	2	2
2	0	0	$\pm\frac{1}{2}$	2	8
	1	−1	$\pm\frac{1}{2}$	6	
		0	$\pm\frac{1}{2}$		
		1	$\pm\frac{1}{2}$		
3	0	0	$\pm\frac{1}{2}$	2	18
	1	−1	$\pm\frac{1}{2}$	6	
		0	$\pm\frac{1}{2}$		
		1	$\pm\frac{1}{2}$		
	2	−2	$\pm\frac{1}{2}$	10	
		−1	$\pm\frac{1}{2}$		
		0	$\pm\frac{1}{2}$		
		1	$\pm\frac{1}{2}$		
		2	$\pm\frac{1}{2}$		

As Table 2–1 indicates, there can be two electrons in the $l = 0$ subshell, six electrons when $l = 1$, and ten electrons for $l = 2$. The electronic configurations of various atoms in the periodic table can be deduced from this list of allowed states. The ground state electron structures for a number of atoms are listed in Table 2–2. There is a simple shorthand notation for electronic structures which is commonly used instead of such a table. The only new convention to remember in this notation is the naming of the l values:

$$l = 0, 1, 2, 3, 4, \ldots$$

$$s, p, d, f, g, \ldots$$

This convention was created by early spectroscopists who referred to the first four spectral groups as *s*harp, *p*rincipal, *d*iffuse, and *f*undamental. Alphabetical order is used beyond *f*. With this convention for l, we can write an electron state as follows:

$$(\mathbf{n} = 3)\ 3p^6 \quad (l = 1)$$

(6 electrons in the $3p$ subshell)

For example, the total electronic configuration for Si ($Z = 14$) in the ground state is

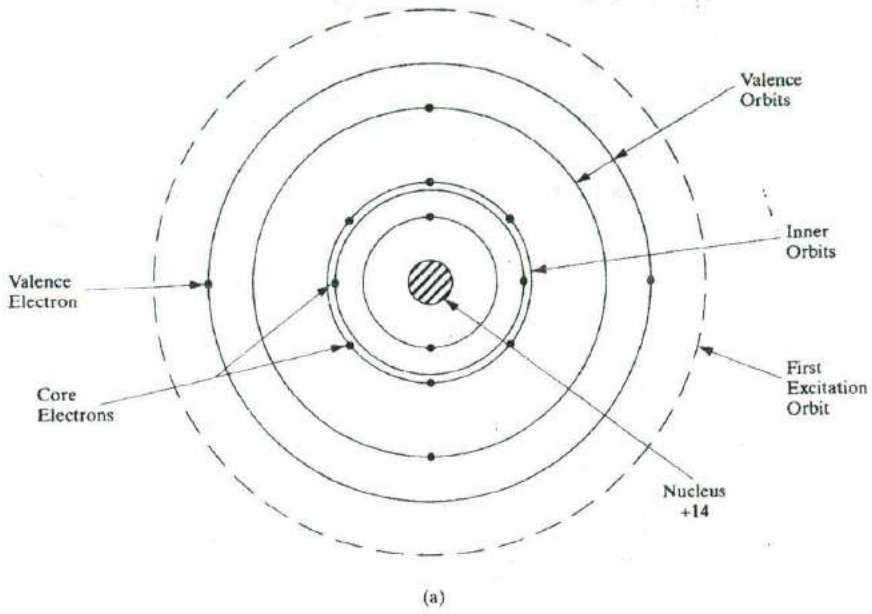
$$1s^2 2s^2 2p^6 3s^2 3p^2$$

We notice that Si has a closed Ne configuration (see Table 2–2) plus four electrons in an outer $\mathbf{n} = 3$ orbit ($3s^2 3p^2$). These are the four valence electrons of Si; two valence electrons are in an s state and two are in a p state. The Si electronic configuration can be written [Ne] $3s^2 3p^2$ for convenience, since the Ne configuration $1s^2 2s^2 2p^6$ forms a closed shell (typical of the inert elements).

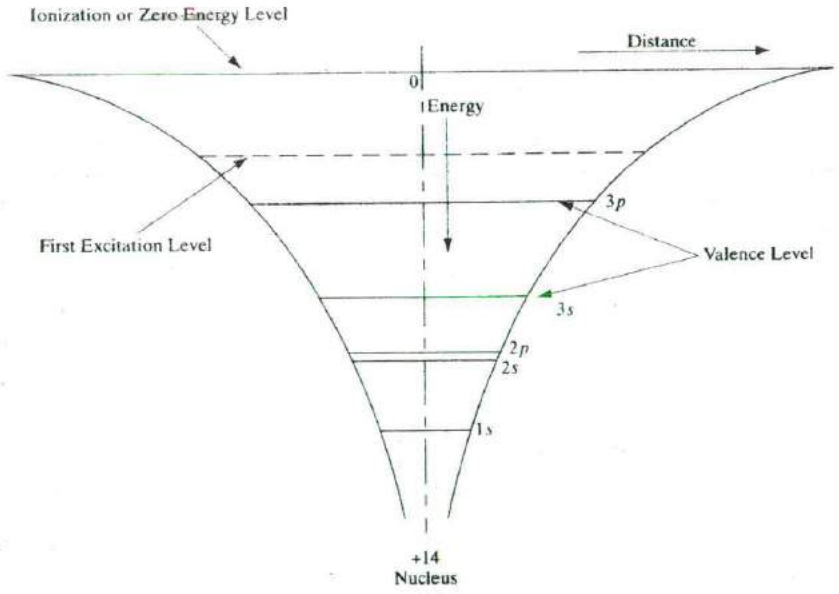
Figure 2–8a shows the orbital model of a Si atom, which has a nucleus consisting of 14 protons (with a charge of +14) and neutrons, 10 core electrons in shells $n = 1$ and 2, and 4 valence electrons in the $3s$ and $3p$ subshells. Figure 2–8b shows the energy levels of the various electrons in the coulombic potential well of the nucleus. Since unlike charges attract each other, there is an attractive potential between the negatively charged electrons and the positively charged nucleus. As indicated in Eq. (2–36), a Coulomb potential varies as $1/r$ as a function of distance from the charge, in this case the Si nucleus. The potential energy gradually goes to zero when we approach infinity. We end up getting "particle-in-a-box" states for these electrons in this potential well, as discussed in Section 2.4.3 and Eq. (2–33). Of course, in this case the shape of the potential well is not rectangular, as shown in Fig. 2–5a, but coulombic, as shown in Fig. 2–8b. Therefore, the energy levels have a form closer to those of the H atom as shown in Eq. (2–15), rather than in Eq. (2–33).

Table 2–2 Electronic configurations for atoms in the ground state.

Atomic number (Z)	Element	n = 1, l = 0: 1s	2: 0 1: 2s 2p	3: 0 1 2: 3s 3p 3d	4: 0 1: 4s 4p	Shorthand notation
		Number of electrons				
1	H	1				$1s^1$
2	He	2				$1s^2$
3	Li		1			$1s^2\ 2s^1$
4	Be		2			$1s^2\ 2s^2$
5	B		2 1			$1s^2\ 2s^2\ 2p^1$
6	C	helium core,	2 2			$1s^2\ 2s^2\ 2p^2$
7	N	2 electrons	2 3			$1s^2\ 2s^2\ 2p^3$
8	O		2 4			$1s^2\ 2s^2\ 2p^4$
9	F		2 5			$1s^2\ 2s^2\ 2p^5$
10	Ne		2 6			$1s^2\ 2s^2\ 2p^6$
11	Na			1		[Ne] $3s^1$
12	Mg			2		$3s^2$
13	Al			2 1		$3s^2\ 3p^1$
14	Si	neon core,		2 2		$3s^2\ 3p^2$
15	P	10 electrons		2 3		$3s^2\ 3p^3$
16	S			2 4		$3s^2\ 3p^4$
17	Cl			2 5		$3s^2\ 3p^5$
18	Ar			2 6		$3s^2\ 3p^6$
19	K				1	[Ar] $4s^1$
20	Ca				2	$4s^2$
21	Sc			1	2	$3d^1\ 4s^2$
22	Ti			2	2	$3d^2\ 4s^2$
23	V			3	2	$3d^3\ 4s^2$
24	Cr			5	1	$3d^5\ 4s^1$
25	Mn			5	2	$3d^5\ 4s^2$
26	Fe			6	2	$3d^6\ 4s^2$
27	Co	argon core,		7	2	$3d^7\ 4s^2$
28	Ni	18 electrons		8	2	$3d^8\ 4s^2$
29	Cu			10	1	$3d^{10}\ 4s^1$
30	Zn			10	2	$3d^{10}\ 4s^2$
31	Ga			10	2 1	$3d^{10}\ 4s^2\ 4p^1$
32	Ge			10	2 2	$3d^{10}\ 4s^2\ 4p^2$
33	As			10	2 3	$3d^{10}\ 4s^2\ 4p^3$
34	Se			10	2 4	$3d^{10}\ 4s^2\ 4p^4$
35	Br			10	2 5	$3d^{10}\ 4s^2\ 4p^5$
36	Kr			10	2 6	$3d^{10}\ 4s^2\ 4p^6$



(a)



(b)

Figure 2–8
Electronic structure and energy levels in a Si atom: (a) The orbital model of a Si atom showing the 10 core electrons ($n = 1$ and 2), and the 4 valence electrons ($n = 3$); (b) energy levels in the coulombic potential of the nucleus are also shown schematically.

If we solve the Schrödinger equation for the Si atom as we did in Section 2.5.1 for the H atom, we can get the radial and angular dependence of the wavefunctions or "orbitals" of the electrons. Let us focus on the valence shell, $\mathbf{n} = 3$, where we have two $3s$ and two $3p$ electrons. It turns out that the $3s$ orbital is spherically symmetric with no angular dependence, and is positive everywhere. It can hold 2 electrons with opposite spin according to the Pauli principle. There are 3 p-orbitals which are mutually perpendicular. These are shaped like dumb-bells with a positive lobe and a negative lobe (Fig. 2–9). The $3p$ subshell can hold up to 6 electrons, but in the case of Si has only 2. Interestingly, in a Si crystal when we bring individual atoms very close together, the s- and p-orbitals overlap so much that they lose their distinct character, and lead to four mixed sp^3 orbitals. The negative part of the p orbital cancels the s-type wavefunction, while the positive part enhances it, thereby leading to a "directed" bond in space. As shown in Fig. 2–9, these *linear combinations of atomic orbitals (LCAO)* or "hybridized" sp^3 orbitals point symmetrically in space along the 4 tetragonal directions (See Fig. 1–9). In Chapter 3 we shall see that these "directed" chemical bonds are responsible for the tetragonal diamond or zincblende lattice structure in most semiconductors. They are also very important in the understanding of energy bands, and in the conduction of charges in these semiconductors.

The column IV semiconductor Ge ($Z = 32$) has an electronic structure similar to Si, except that the four valence electrons are outside a closed $\mathbf{n} = 3$

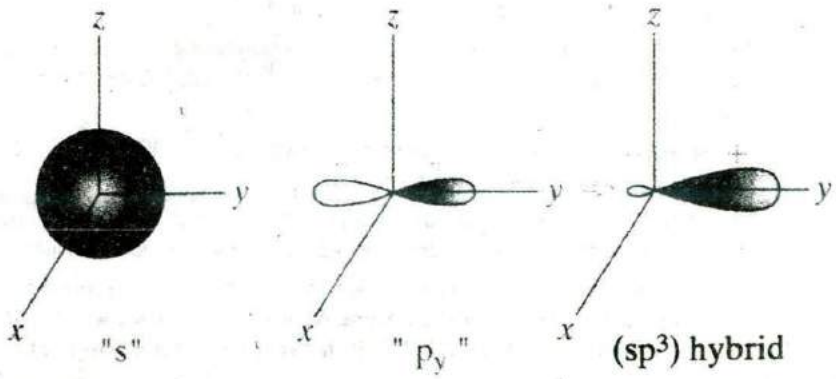


Figure 2–9 Orbitals in a Si atom: The spherically symmetric "s" type wave functions or orbitals are positive everywhere, while the three mutually perpendicular "p" type orbitals (p_x, p_y, p_z) are dumb-bell shaped and have a positive lobe and a negative lobe. The four sp^3 "hybridized" orbitals, only one of which is shown here, point symmetrically in space and lead to the diamond lattice in Si.

A-84544

shell. Thus the Ge configuration is [Ar] $3d^{10}4s^24p^2$. There are several cases in Table 2–2 that do not follow the most straight-forward choice of quantum numbers. For example, we notice that in K ($Z = 19$) and Ca ($Z = 20$) the $4s$ state is filled before the $3d$ state; in Cr ($Z = 24$) and Cu ($Z = 29$) there is a transfer of an electron back to the $3d$ state. These exceptions, required by minimum energy considerations, are discussed more fully in most atomic physics texts.

PROBLEMS

2.1 (a) Sketch a simple vacuum tube device and the associated circuitry for measuring E_m in the photoelectric effect experiment. The electrodes can be placed in a sealed glass envelope.

(b) Sketch the photocurrent I vs. retarding voltage V that you would expect to measure for a given electrode material and configuration. Make the sketch for several intensities of light at a given wavelength.

(c) The work function of platinum is 4.09 eV. What retarding potential will be required to reduce the photocurrent to zero in a photoelectric experiment with Pt electrodes if the wavelength of incident light is 2440 Å? Remember that an energy of $q\Phi$ is lost by each electron in escaping the surface.

2.2 Point A is at an electrostatic potential of +1V relative to point B in a vacuum. An electron initially at rest at B moves to A. What energy (expressed in J and eV) does the electron have at A? What is its velocity (m/s)?

2.3 (a) Show that the various lines in the hydrogen spectrum can be expressed in angstroms as

$$\lambda(\text{Å}) = \frac{911\mathbf{n}_1^2\mathbf{n}^2}{\mathbf{n}^2 - \mathbf{n}_1^2}$$

where $\mathbf{n}_1 = 1$ for the Lyman series, 2 for the Balmer series, and 3 for the Paschen series. The integer $\mathbf{n}$ is larger than $\mathbf{n}_1$.

(b) Calculate λ for the Lyman series to $\mathbf{n} = 5$, the Balmer series to $\mathbf{n} = 7$, and the Paschen series to $\mathbf{n} = 10$. Plot the results as in Fig. 2–2. What are the wavelength limits for each of the three series?

2.4 Show that the calculated Bohr expression for frequency of emitted light in the hydrogen spectrum, Eq. (2–17), corresponds to the experimental expressions, Eq. (2–3).

2.5 (a) The position of an electron is determined to within 1 Å. What is the minimum uncertainty in its momentum?

(b) An electron's energy is measured with an uncertainty of 1 eV. What is the minimum uncertainty in the time over which the measurement was made?

2.6 The de Broglie wavelength of a particle $\lambda = h/mv$ describes the wave-particle duality for small particles such as electrons. What is the de Broglie wavelength (in Å) of an electron at 100 eV? What is the wavelength for electrons at 12

keV, which is typical of electron microscopes? Comparing this to visible light, comment on the advantages of electron microscopes.

2.7 A sample of radioactive material undergoes decay such that the number of atoms $N(t)$ remaining in the unstable state at time t is related to the number N_0 at $t = 0$ by the relation $N(t) = N_0 \exp(-t/\tau)$. Show that τ is the average lifetime $\langle t \rangle$ of an atom in the unstable state before it spontaneously decays. Equation (2–21b) can be used with t substituted for x.

2.8 Given a plane wave $\psi = A exp(jk_x x)$ what is the expectation value for p_x^2 and p_z where p is momentum?

2.9 A free electron traveling in the x-direction can be described by a plane wave, with a wave function of the form $\psi_k(x) = Ae^{j\mathbf{k}_x x}$, where $\mathbf{k}_x$ is a wave vector, or propagation constant. Use postulate 3 and the momentum operator to relate the electron momentum $\langle p_x \rangle$ to $\mathbf{k}_x$.

2.10 An electron is described by a plane-wave wavefunction $\psi(x, t) = Ae^{j(10x - 7t)}$. Calculate the expectation value of the x-component of momentum, the y-component of momentum and the energy of the electron. (Give values in MKS units.)

2.11 Calculate the first three energy levels for an electron in a quantum well of width 10Å with infinite walls.

2.12 What do Li, Na, and K have in common? What do F, Cl, and Br have in common? What are the electron configurations for ionized Na and Cl?

READING LIST

Ashcroft, N. W., and N. D. Mermin. *Solid State Physics.* Philadelphia: W.B. Saunders, 1976.

Baggot, J. "Beating the Uncertainty Principle." *New Scientist* 133 (15 February 1992): 36–40.

Bate, R. T. "The Quantum-Effect Device: Tomorrow's Transistor?" *Scientific American* 258 (March 1988): 96–100.

Brehm, J. J., and W. J. Mullin. *Introduction to the Structure of Matter.* New York: Wiley, 1989.

Bube, R. H. *Electrons in Solids,* 3rd ed. Boston: Harcourt Brace Jovanovich, 1992.

Capasso, F., and S. Datta. "Quantum Electron Devices." *Physics Today* 43 (February 1990): 74–82.

Cassidy, D. C. "Heisenberg: Uncertainty and the Quantum Revolution." *Scientific American* 266 (May 1992): 106–12.

Chang, L. L., and L. Esaki. "Semiconductor Quantum Heterostructures." *Physics Today* 45 (October 1992): 36–43.

Cohen-Tannoouji, C., B. Diu, and F. Laloe. *Quantum Mechanics.* New York: Wiley, 1977.

Corcoran, E. "Diminishing Dimensions." *Scientific American* 263 (November 1990): 122–6+.

Datta, S. *Modular Series on Solid State Devices: Vol. 8. Quantum Phenomena.* Reading. MA: Addison-Wesley, 1989.

Feynman, R. P. *The Feynman Lectures on Physics, Vol. 3. Quantum Mechanics.* Reading, MA: Addison-Wesley, 1965.

Hummel, R. E. *Electronic Properties of Materials,* 2nd ed. Berlin: Springer-Verlag, 1993.

Kroemer, H. *Quantum Mechanics.* Englewood Cliffs, NJ: Prentice Hall, 1994.

Park, D. *Introduction to the Quantum Theory.* New York: McGraw-Hill, 1992.

Sakurai, J. J. *Modern Quantum Mechanics.* Reading, MA: Addison-Wesley, 1994.

Singh, J. *Semiconductor Devices.* New York: McGraw-Hill, 1994.

Sundaram, M., S. A. Chalmers, and P. F. Hopkins. "New Quantum Structures." *Science* 254 (29 November 1991): 1326–35.

Weisbuch, C., and B. Vinter. *Quantum Semiconductor Structures.* Boston: Academic Press, 1991.

Chapter 3

Energy Bands and Charge Carriers in Semiconductors

In this chapter we begin to discuss the specific mechanisms by which current flows in a solid. In examining these mechanisms we shall learn why some materials are good conductors of electric current, whereas others are poor conductors. We shall see how the conductivity of a semiconductor can be varied by changing the temperature or the number of impurities. These fundamental concepts of charge transport form the basis for later discussions of solid state device behavior.

3.1 BONDING FORCES AND ENERGY BANDS IN SOLIDS

In Chapter 2 we found that electrons are restricted to sets of discrete energy levels within atoms. Large gaps exist in the energy scale in which no energy states are available. In a similar fashion, electrons in solids are restricted to certain energies and are not allowed at other energies. The basic difference between the case of an electron in a solid and that of an electron in an isolated atom is that in the solid the electron has a *range,* or *band,* of available energies. The discrete energy levels of the isolated atom spread into bands of energies in the solid because in the solid the wave functions of electrons in neighboring atoms overlap, and an electron is not necessarily localized at a particular atom. Thus, for example, an electron in the outer orbit of one atom feels the influence of neighboring atoms, and its overall wave function is altered. Naturally, this influence affects the potential energy term and the boundary conditions in the Schrödinger equation, and we would expect to obtain different energies in the solution. Usually, the influence of neighboring atoms on the energy levels of a particular atom can be treated as a small perturbation, giving rise to shifting and splitting of energy states into energy bands.

3.1.1 Bonding Forces in Solids

The interaction of electrons in neighboring atoms of a solid serves the very important function of holding the crystal together. For example, alkali halides such as NaCl are typified by *ionic bonding.* In the NaCl lattice, each Na atom

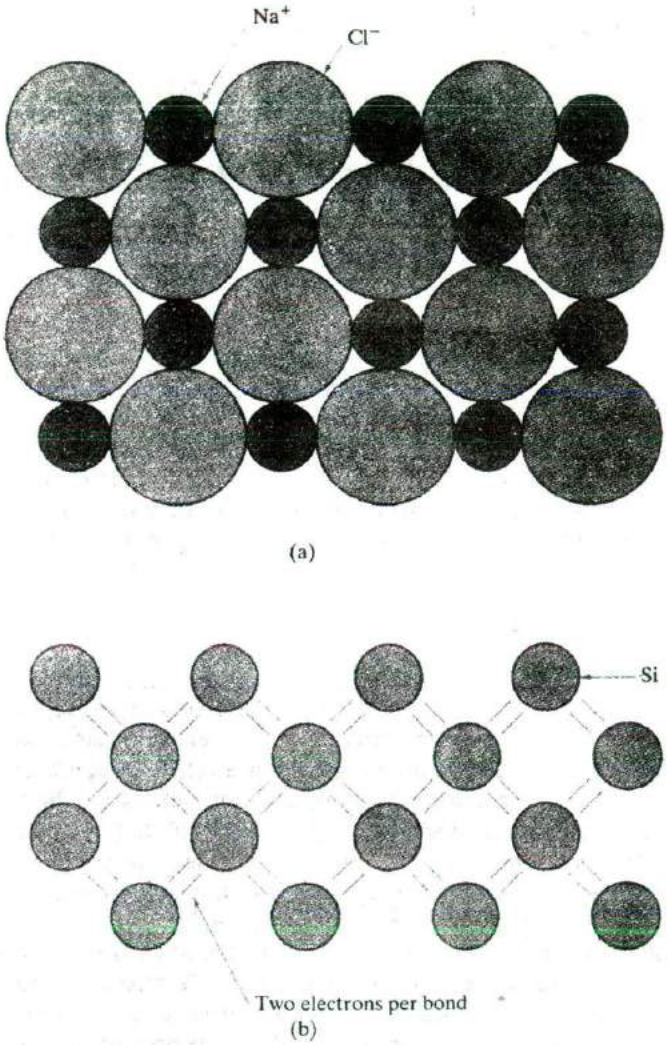


Figure 3–1
Different types of chemical bonding in solids (a) an example of ionic bonding in NaCl; (b) covalent bonding in the Si crystal, veiwed along a <100> direction (see also Figs. 1–8 and 1–9).

is surrounded by six nearest neighbor Cl atoms, and vice versa. Four of the nearest neighbors are evident in the two-dimensional representation shown in Fig. 3–1a. The electronic structure of Na ($Z = 11$) is [Ne] $3s^1$, and Cl ($Z = 17$) has the structure [Ne]$3s^2 3p^5$. In the lattice each Na atom gives up its outer $3s$ electron to a Cl atom, so that the crystal is made up of ions with the electronic structures of the inert atoms Ne and Ar (Ar has the electronic structure [Ne]$3s^2 3p^6$). However, the ions have net electric charges after the electron exchange. The Na^+ ion has a net positive charge, having lost an electron, and the Cl^- ion has a net negative charge, having gained an electron.

Each Na^+ ion exerts an electrostatic attractive force upon its six Cl^- neighbors, and vice versa. These coulombic forces pull the lattice together until a balance is reached with repulsive forces. A reasonably accurate calculation of the atomic spacing can be made by considering the ions as hard spheres being attracted together (Example 1–1).

An important observation in the NaCl structure is that all electrons are tightly bound to atoms. Once the electron exchanges have been made between the Na and Cl atoms to form the Na^+ and Cl^- ions, the outer orbits of all atoms are completely filled. Since the ions have the closed-shell configurations of the inert atoms Ne and Ar, there are no loosely bound electrons to participate in current flow; as a result, NaCl is a good insulator.

In a metal atom the outer electronic shell is only partially filled, usually by no more than three electrons. We have already noted that the alkali metals (e.g., Na) have only one electron in the outer orbit. This electron is loosely bound and is given up easily in ion formation. This accounts for the great chemical activity in the alkali metals, as well as for their high electrical conductivity. In the metal the outer electron of each alkali atom is contributed to the crystal as a whole, so that the solid is made up of ions with closed shells immersed in a sea of free electrons. The forces holding the lattice together arise from an interaction between the positive ion cores and the surrounding free electrons. This is one type of *metallic bonding.* Obviously, there are complicated differences in the bonding forces for various metals, as evidenced by the wide range of melting temperatures (234 K for Hg, 3643 K for W). However, the metals have the sea of electrons in common, and these electrons are free to move about the crystal under the influence of an electric field.

A third type of bonding is exhibited by the diamond lattice semiconductors. We recall that each atom in the Ge, Si, or C diamond lattice is surrounded by four nearest neighbors, each with four electrons in the outer orbit. In these crystals each atom shares its valence electrons with its four neighbors (Fig. 3–1b). Bonding between nearest neighbor atoms is illustrated in the diamond lattice diagram of Fig. 1–9. The bonding forces arise from a quantum mechanical interaction between the shared electrons. This is known as *covalent bonding;* each electron pair constitutes a covalent bond. In the sharing process it is no longer relevant to ask which electron belongs to a particular atom—both belong to the bond. The two electrons are indistinguishable, except that they must have opposite spin to satisfy the Pauli exclusion principle. Covalent bonding is also found in certain molecules, such as H_2.

As in the case of the ionic crystals, no free electrons are available to the lattice in the covalent diamond structure of Fig. 3–1b. By this reasoning Ge and Si should also be insulators. However, we have pictured an idealized lattice at 0 K in this figure. As we shall see in subsequent sections, an electron can be thermally or optically excited out of a covalent bond and thereby become free to participate in conduction. This is an important feature of semiconductors.

Compound semiconductors such as GaAs have mixed bonding, in which both ionic and covalent bonding forces participate. Some ionic bonding is to be expected in a crystal such as GaAs because of the difference in placement of the Ga and As atoms in the periodic table. The ionic character of the bonding becomes more important as the atoms of the compound become further separated in the periodic table, as in the II–VI compounds.

3.1.2 Energy Bands

As isolated atoms are brought together to form a solid, various interactions occur between neighboring atoms, including those described in the preceding section. The forces of attraction and repulsion between atoms will find a balance at the proper interatomic spacing for the crystal. In the process, important changes occur in the electron energy level configurations, and these changes result in the varied electrical properties of solids.

In Fig. 2–8, we showed the orbital model of a Si atom, along with the energy levels of the various electrons in the coulombic potential well of the nucleus. Let us focus on the outermost shell or valence shell, $\mathbf{n} = 3$, where two $3s$ and two $3p$ electrons interact to form the four "hybridized" sp^3 electrons when the atoms are brought close together. In Fig. 3–2, we schematically show the coulombic potential wells of two atoms close to each other, along with the wave functions of two electrons centered on the two nuclei. By solving the Schrödinger equation for such an interacting system, we find that the composite two-electron wave functions are *linear combinations* of the individual *atomic orbitals* (LCAO). The odd or anti-symmetric combination is called the anti-bonding orbital, while the even or symmetric combination is the bonding orbital. It can be seen that the bonding orbital has a higher value of the wave function (and therefore the electron probability density) than the anti-bonding state in the region between the two nuclei. This corresponds to the covalent bond between the atoms.

To determine the energy levels of the bonding and the anti-bonding states, it is important to recognize that in the region between the two nuclei the coulombic potential energy $V(r)$ is lowered (solid line in Fig. 3–2) compared to isolated atoms (dashed lines). It is easy to see why the potential energy would be lowered in this region, because an electron here would be attracted by two nuclei, rather than just one. For the bonding state the electron probability density is higher in this region of lowered potential energy than for the anti-bonding state. As a result, the original isolated atomic energy level would be split into two, a lower bonding energy level and a higher anti-bonding level. It is the lowering of the energy of the bonding state that gives rise to cohesion of the crystal. For even smaller inter-atomic spacings, the energy of the crystal goes up because of repulsion between the nuclei, and other electronic interactions. Since the probability density is given by the

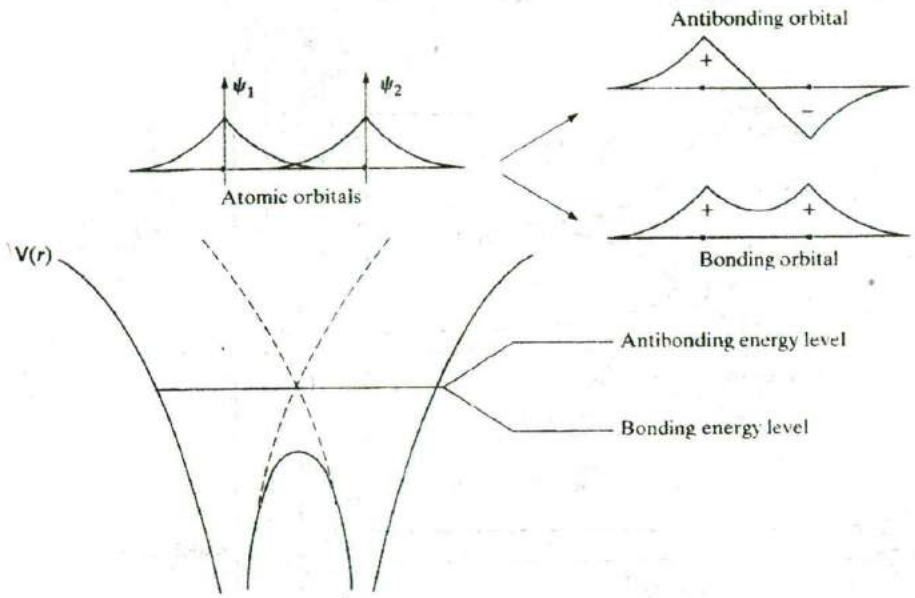


Figure 3–2
Linear combinations of atomic orbitals (LCAO): The LCAO when 2 atoms are brought together leads to 2 distinct "normal" modes—a higher energy anti-bonding orbital, and a lower energy bonding orbital. Note that the electron probability density is high in the region between the ion cores (covalent "bond"), leading to lowering of the bonding energy level and the cohesion of the crystal. If instead of 2 atoms, one brings together N atoms, there will be N distinct LCAO, and N closely-spaced energy levels in a band.

square of the wave function, if the entire wave function is multiplied by –1, it does not lead to a different LCAO. The important point to note in this discussion is that the number of distinct LCAO, and the number of distinct energy levels depends on the number of atoms that are brought together. The lowest energy level corresponds to the totally symmetric LCAO, the highest corresponds to the totally anti-symmetric case and the other combinations lead to energy levels in between.

Qualitatively, we can see that as atoms are brought together, the application of the Pauli exclusion principle becomes important. When two atoms are completely isolated from each other so that there is no interaction of electron wave functions between them, they can have identical electronic structures. As the spacing between the two atoms becomes smaller, however, electron wave functions begin to overlap. The exclusion principle dictates that no two electrons in a given interacting system may have the same quantum state; thus there must be at most one electron per level after there is a splitting of the discrete energy levels of the isolated atoms into new levels belonging to the pair rather than to individual atoms.

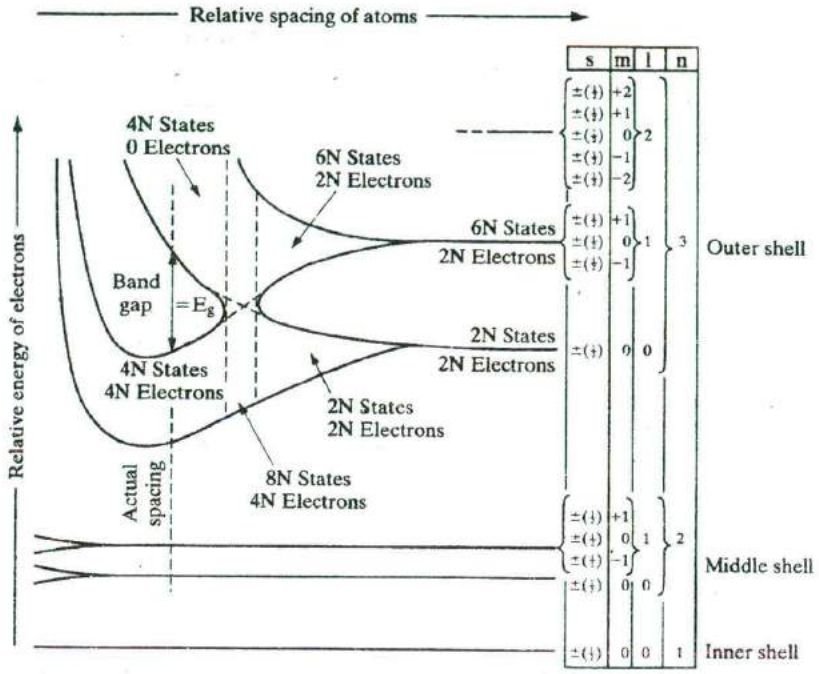


Figure 3–3
Energy levels in Si as a function of inter-atomic spacing. The core levels (**n** = 1,2) in Si are completely filled with electrons. At the actual atomic spacing of the crystal, the 2*N* electrons in the 3*s* sub-shell and the 2*N* electrons in the 3*p* sub-shell undergo sp^3 hybridization, and all end up in the lower 4*N* states (valence band), while the higher lying 4*N* states (conduction band) are empty, separated by a bandgap.

In a solid, many atoms are brought together, so that the split energy levels form essentially continuous *bands* of energies. As an example, Fig. 3–3 illustrates the imaginary formation of a silicon crystal from isolated silicon atoms. Each isolated silicon atom has an electronic structure $1s^2 2s^2 2p^6 3s^2 3p^2$ in the ground state. Each atom has available two 1*s* states, two 2*s* states, six 2*p* states, two 3*s* states, six 3*p* states, and higher states (see Tables 2–1 and 2–2). If we consider *N* atoms, there will be 2*N*, 2*N*, 6*N*, 2*N*, and 6*N* states of type 1*s*, 2*s*, 2*p*, 3*s*, and 3*p*, respectively. As the interatomic spacing decreases, these energy levels split into bands, beginning with the outer (**n** = 3) shell. As the "3*s*" and "3*p*" bands grow, they merge into a single band composed of a mixture of energy levels. This band of "3*s*–3*p*" levels contains 8*N* available states. As the distance between atoms approaches the equilibrium interatomic spacing of silicon, this band splits into two bands separated by an *energy gap* E_g. The upper band (called the *conduction band*) contains 4*N* states, as does the lower (*valence*) band. Thus, apart from the low-lying and tightly bound "core" levels, the silicon crystal has two bands of available energy levels separated

by an energy gap E_g wide, which contains no allowed energy levels for electrons to occupy. This gap is sometimes called a "forbidden band," since in a perfect crystal it contains no electron energy states.

We should pause at this point and count electrons. The lower "1*s*" band is filled with the $2N$ electrons which originally resided in the collective 1*s* states of the isolated atoms. Similarly, the 2*s* band and the 2*p* bands will have $2N$ and $6N$ electrons in them, respectively. However, there were $4N$ electrons in the original isolated $\mathbf{n} = 3$ shells ($2N$ in 3*s* states and $2N$ in 3*p* states). These $4N$ electrons must occupy states in the valence band or the conduction band in the crystal. At 0 K the electrons will occupy the lowest energy states available to them. In the case of the Si crystal, there are exactly $4N$ states in the valence band available to the $4N$ electrons. Thus at 0 K, *every* state in the valence band will be filled, while the conduction band will be completely empty of electrons. As we shall see, this arrangement of completely filled and empty energy bands has an important effect on the electrical conductivity of the solid.

3.1.3 Metals, Semiconductors, and Insulators

Every solid has its own characteristic energy band structure. This variation in band structure is responsible for the wide range of electrical characteristics observed in various materials. The silicon band structure of Fig. 3–3, for example, can give a good picture of why silicon in the diamond lattice is a good insulator. To reach such a conclusion, we must consider the properties of completely filled and completely empty energy bands in the current conduction process.

Before discussing the mechanisms of current flow in solids further, we can observe here that for electrons to experience acceleration in an applied electric field, they must be able to move into new energy states. This implies there must be empty states (allowed energy states which are not already occupied by electrons) available to the electrons. For example, if relatively few electrons reside in an otherwise empty band, ample unoccupied states are available into which the electrons can move. On the other hand, the silicon band structure is such that the valence band is completely filled with electrons at 0 K and the conduction band is empty. There can be no charge transport within the valence band, since no empty states are available into which electrons can move. There are no electrons in the conduction band, so no charge transport can take place there either. Thus silicon has a high resistivity typical of insulators.

Semiconductor materials at 0 K have basically the same structure as insulators—a filled valence band separated from an empty conduction band by a band gap containing no allowed energy states (Fig. 3–4). The difference lies in the size of the band gap E_g, which is much smaller in semiconductors than in insulators. For example, the semiconductor Si has a band gap of about 1.1 eV compared with 5 eV for diamond. The relatively small band gaps of semiconductors (Appendix III) allow for excitation of electrons from the

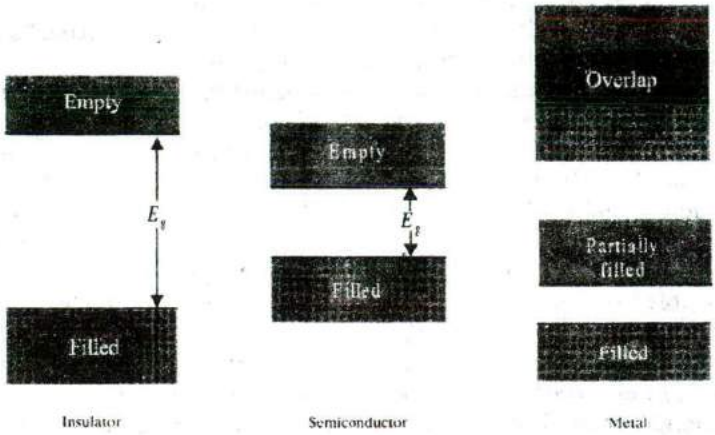


Figure 3–4
Typical band structures at 0 K.

lower (valence) band to the upper (conduction) band by reasonable amounts of thermal or optical energy. For example, at room temperature a semiconductor with a 1-eV band gap will have a significant number of electrons excited thermally across the energy gap into the conduction band, whereas an insulator with $E_g = 10$ eV will have a negligible number of such excitations. Thus an important difference between semiconductors and insulators is that the number of electrons available for conduction can be increased greatly in semiconductors by thermal or optical energy.

In metals the bands either overlap or are only partially filled. Thus electrons and empty energy states are intermixed within the bands so that electrons can move freely under the influence of an electric field. As expected from the metallic band structures of Fig. 3–4, metals have a high electrical conductivity.

3.1.4 Direct and Indirect Semiconductors

The "thought experiment" of Section 3.1.2, in which isolated atoms were brought together to form a solid, is useful in pointing out the existence of energy bands and some of their properties. Other techniques are generally used, however, when quantitative calculations are made of band structures. In a typical calculation, a single electron is assumed to travel through a perfectly periodic lattice. The wave function of the electron is assumed to be in the form of a plane wave[1] moving, for example, in the x- direction with propagation constant $\mathbf{k}$, also called a *wave vector*. The space-dependent wave function for the electron is

[1]Discussions of plane waves are available in most sophomore physics texts or in introductory electromagnetics texts.

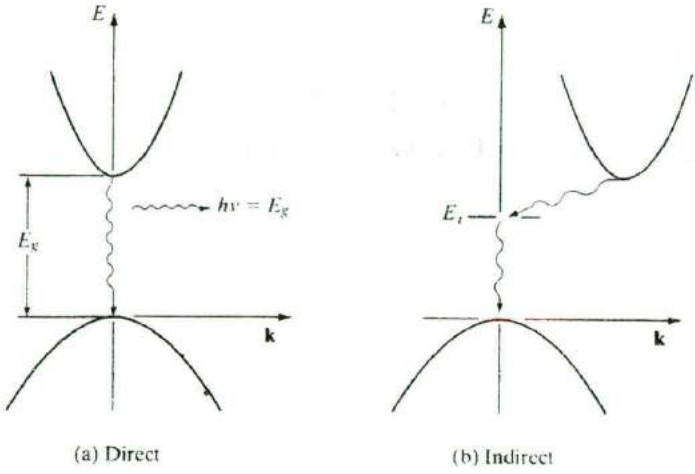


Figure 3–5
Direct and indirect electron transitions in semiconductors: (a) direct transition with accompanying photon emission; (b) indirect transition via a defect level.

$$\psi_{\mathbf{k}}(x) = U(\mathbf{k}_x, x)e^{j\mathbf{k}_x x} \tag{3–1}$$

where the function $U(\mathbf{k}_x, x)$ modulates the wave function according to the periodicity of the lattice.

In such a calculation, allowed values of energy can be plotted vs. the propagation constant $\mathbf{k}$. Since the periodicity of most lattices is different in various directions, the $(E, \mathbf{k})$ diagram must be plotted for the various crystal directions, and the full relationship between E and $\mathbf{k}$ is a complex surface which should be visualized in three dimensions.

The band structure of GaAs has a minimum in the conduction band and a maximum in the valence band for the same $\mathbf{k}$ value ($\mathbf{k} = 0$). On the other hand, Si has its valence band maximum at a different value of $\mathbf{k}$ than its conduction band minimum. Thus an electron making a smallest-energy transition from the conduction band to the valence band in GaAs can do so without a change in $\mathbf{k}$ value; on the other hand, a transition from the minimum point in the Si conduction band to the maximum point of the valence band requires some change in $\mathbf{k}$. Thus there are two classes of semiconductor energy bands; *direct* and *indirect* (Fig. 3–5). We can show that an indirect transition, involving a change in $\mathbf{k}$, requires a change of momentum for the electron.

EXAMPLE 3–1

Assuming that U is constant in Eq. (3–1) for an essentially free electron, show that the x-component of the electron momentum in the crystal is given by $\langle p_x \rangle = \hbar \mathbf{k}_x$.

SOLUTION From Eq. (3–1)

$$\psi_{\mathbf{k}}(x) = Ue^{j\mathbf{k}_x x}$$

Using Eq. (2–21b) and the momentum operator,

$$\langle p_x \rangle = \frac{\int_{-\infty}^{\infty} U^2 e^{-j\mathbf{k}_x x} \frac{\hbar}{j} \frac{\partial}{\partial x}(e^{j\mathbf{k}_x x})\, dx}{\int_{-\infty}^{\infty} U^2\, dx}$$

$$= \frac{\hbar\mathbf{k}_x \int_{-\infty}^{\infty} U^2\, dx}{\int_{-\infty}^{\infty} U^2\, dx} = \hbar\mathbf{k}_x$$

This result implies that $(E, \mathbf{k})$ diagrams such as shown in Fig. 3–5 can be considered plots of electron energy vs. momentum, with a scaling factor $\hbar$.

The direct and indirect semiconductors are identified in Appendix III. In a direct semiconductor such as GaAs, an electron in the conduction band can fall to an empty state in the valence band, giving off the energy difference E_g as a photon of light. On the other hand, an electron in the conduction band minimum of an indirect semiconductor such as Si cannot fall directly to the valence band maximum but must undergo a momentum change as well as changing its energy. For example, it may go through some defect state (E_t) within the band gap. We shall discuss such defect states in Sections 4.2.1 and 4.3.2. In an indirect transition which involves a change in $\mathbf{k}$, the energy is generally given up as heat to the lattice rather than as an emitted photon. This difference between direct and indirect band structures is very important for deciding which semiconductors can be used in devices requiring light output. For example, semiconductor light emitters and lasers (Chapter 8) generally must be made of materials capable of direct band-to-band transitions or of indirect materials with vertical transitions between defect states.

Band diagrams such as those shown in Fig. 3–5 are cumbersome to draw in analyzing devices, and do not provide a view of the variation of electron energy with distance in the sample. Therefore, in most discussions we shall use simple band pictures such as those shown in Fig. 3–4, remembering that electron transitions across the band gap may be direct or indirect.

3.1.5 Variation of Energy Bands with Alloy Composition

As III–V ternary and quaternary alloys are varied over their composition ranges (see Sections 1.2.4 and 1.4.1), their band structures change. For example, Fig. 3–6 illustrates the band structure of GaAs and AlAs, and the way in which the bands change with composition x in the ternary compound $Al_xGa_{1-x}As$.

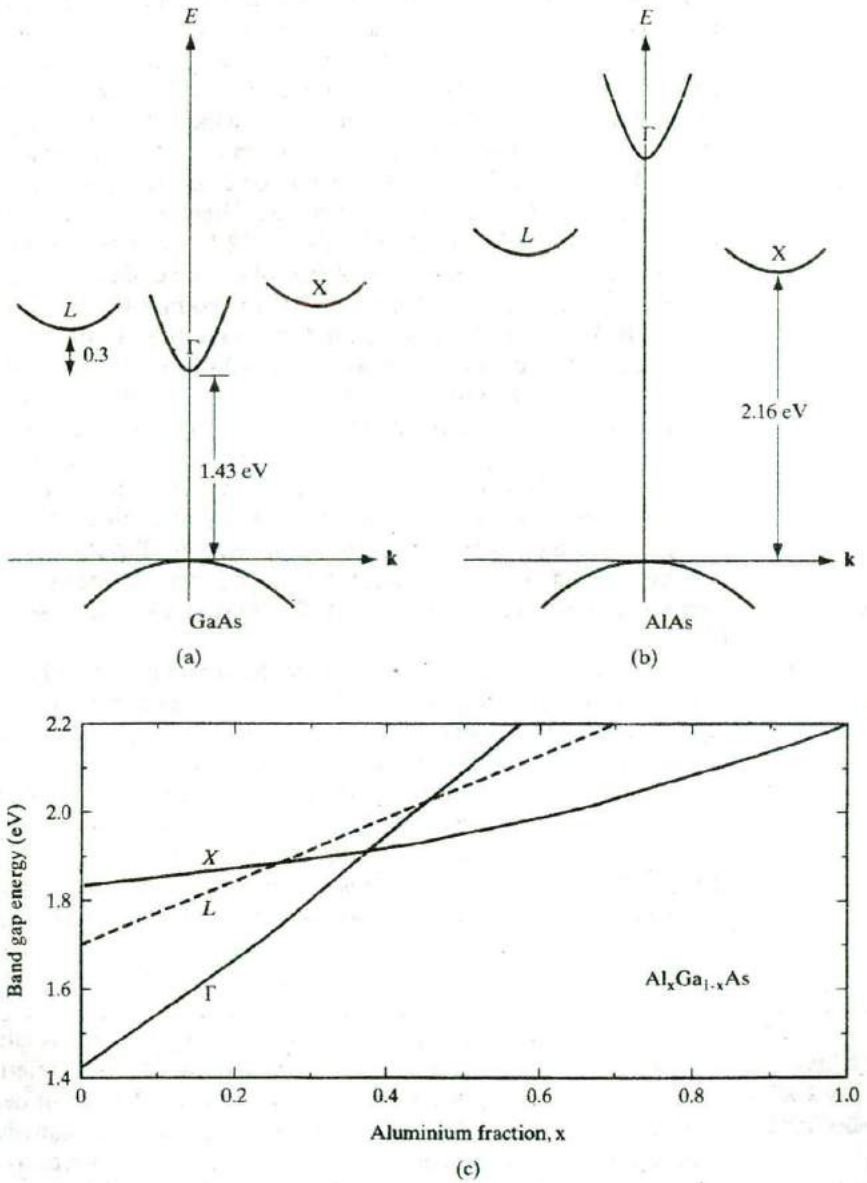


Figure 3–6
Variation of direct and indirect conduction bands in AlGaAs as a function of composition: (a) the (*E*,**k**) diagram for GaAs, showing three minima in the conduction band; (b) AlAs band diagram; (c) positions of the three conduction band minima in $Al_xGa_{1-x}As$ as *x* varies over the range of compositions from GaAs ($x = 0$) to AlAs ($x = 1$). The smallest band gap, E_g (shown in color), follows the direct Γ band to $x = 0.38$, and then follows the indirect X band.

The binary compound GaAs is a direct material, with a band gap of 1.43 eV at room temperature. For reference, we call the direct ($\mathbf{k} = 0$) conduction band minimum Γ. There are also two higher-lying indirect minima in the GaAs conduction band, but these are sufficiently far above Γ that few electrons reside there (we discuss an important exception in Chapter 10 in which high-field excitation of electrons into the indirect minima leads to the Gunn effect). We call the lowest-lying GaAs indirect minimum L and the other X. In AlAs the direct Γ minimum is much higher than the indirect X minimum, and this material is therefore indirect with a band gap of 2.16 eV at room temperature.

In the ternary alloy $Al_xGa_{1-x}As$ all of these conduction band minima move up relative to the valence band as the composition x varies from 0 (GaAs) to 1 (AlAs). However, the indirect minimum X moves up less than the others, and for compositions above about 38 percent Al this indirect minimum becomes the lowest-lying conduction band. Therefore, the ternary alloy AlGaAs is a direct semiconductor for Al compositions on the column III sublattice up to about 38 percent, and is an indirect semiconductor for higher Al mole fractions. The band gap energy E_g is shown in color on Fig. 3–6(c).

The variation of energy bands for the ternary alloy $GaAs_{1-x}P_x$ is generally similar to that of AlGaAs shown in Fig. 3–6. GaAsP is a direct semiconductor from GaAs to about $GaAs_{.55}P_{.45}$ and is indirect from this composition to GaP (see Fig. 8–11). This material is often used in visible LEDs.

Since light emission is most efficient for direct materials, in which electrons can drop from the conduction band to the valence band without changing $\mathbf{k}$ (and therefore momentum), LEDs in GaAsP are generally made in material grown with a composition less than $x = 0.45$. For example, most red LEDs in this material are made at about $x = 0.4$, where the Γ minimum is still the lowest-lying conduction band edge, and where the photon resulting from a direct transition from this band to the valence band is in the red portion of the spectrum (about 1.9 eV). The use of impurities to enhance radiative recombination in indirect material will be discussed in Section 8.2.

3.2 CHARGE CARRIERS IN SEMICONDUCTORS

The mechanism of current conduction is relatively easy to visualize in the case of a metal; the metal atoms are imbedded in a "sea" of relatively free electrons, and these electrons can move as a group under the influence of an electric field. This free electron view is oversimplified, but many important conduction properties of metals can be derived from just such a model. However, we cannot account for all of the electrical properties of semiconductors in this way. Since the semiconductor has a filled valence band and an empty conduction band at 0 K, we must consider the increase in conduction band electrons by thermal excitations across the band gap as the temperature is raised. In addition, after electrons are excited to the conduction band, the empty states left in the valence band can contribute to the conduction process

The introduction of impurities has an important effect on the energy band structure and on the availability of charge carriers. Thus there is considerable flexibility in controlling the electrical properties of semiconductors.

3.2.1 Electrons and Holes

As the temperature of a semiconductor is raised from 0 K, some electrons in the valence band receive enough thermal energy to be excited across the band gap to the conduction band. The result is a material with some electrons in an otherwise empty conduction band and some unoccupied states in an otherwise filled valence band (Fig. 3–7).[2] For convenience, an empty state in the valence band is referred to as a *hole*. If the conduction band electron and the hole are created by the excitation of a valence band electron to the conduction band, they are called an *electron–hole pair* (abbreviated EHP).

After excitation to the conduction band, an electron is surrounded by a large number of unoccupied energy states. For example, the equilibrium number of electron–hole pairs in pure Si at room temperature is only about 10^{10} EHP/cm^3, compared to the Si atom density of 5×10^{22} atoms/cm^3. Thus the few electrons in the conduction band are free to move about via the many available empty states.

The corresponding problem of charge transport in the valence band is somewhat more complicated. However, it is possible to show that the effects of current in a valence band containing holes can be accounted for by simply keeping track of the holes themselves.

In a filled band, all available energy states are occupied. For every electron moving with a given velocity, there is an equal and opposite electron motion elsewhere in the band. If we apply an electric field, the net current is zero because for every electron j moving with velocity v_j there is a corresponding electron j' with velocity $-\mathsf{v}_j$. Figure 3–8 illustrates this effect in terms of the electron energy vs. wave vector plot for the valence band. Since $\mathbf{k}$ is proportional

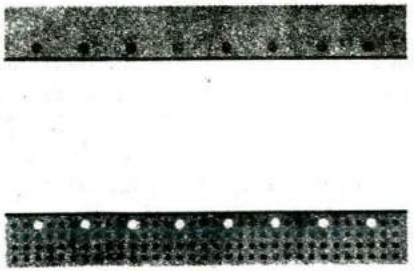


Figure 3–7 Electron-hole pairs in a semiconductor.

[2] In Fig. 3–7 and in subsequent discussions, we refer to the bottom of the conduction band as E_c and the top of the valence band as E_v.

to electron momentum, it is clear the two electrons have oppositely directed velocities. With N electrons/cm^3 in the band we express the current density using a sum over all of the electron velocities, and including the charge $-q$ on each electron. In a unit volume,

$$J = (-q)\sum_{i}^{N} \mathsf{v}_i = 0 \quad (\textit{filled band}) \tag{3–2a}$$

Now if we create a hole by removing the jth electron, the net current density in the valence band involves the sum over all velocities, minus the contribution of the electron we have removed.

$$J = (-q)\sum_{i}^{N} \mathsf{v}_i - (-q)\mathsf{v}_j \quad (j\text{th } \textit{electron missing}) \tag{3–2b}$$

But the first term is zero, from Eq. (3–2a). Thus the net current is $+q\mathsf{v}_j$. In other words, the current contribution of the hole is equivalent to that of a positively charged particle with velocity v_j, that of the missing electron. Of course, the charge transport is actually due to the motion of the new uncompensated electron (j'). Its current contribution $(-q)$ $(-\mathsf{v}_j)$ is equivalent to that of a positively charged particle with velocity $+\mathsf{v}_j$. For simplicity, it is customary to treat empty states in the valence band as charge carriers with positive charge and positive mass.

A simple analogy may help in understanding the behavior of holes. If we have two bottles, one completely filled with water and one completely empty, we can ask ourselves "Will there be any net transport of water when we tilt the bottles?" The answer is "no". In the case of the empty bottle, the answer is obvious. In the case of the completely full bottle also, there cannot be any net motion of water because there is no empty space for water to move into. Similarly, an empty conduction band completely devoid of electrons or a valence band completely full of electrons cannot give rise to a net motion of electrons, and thus to current conduction.

Next, we imagine transferring some water droplets from the full bottle into the empty bottle, leaving behind some air bubbles, and ask ourselves the same question. Now when we tilt the bottles there will be net transport of water: the water droplets will roll downhill in one bottle and the air bubbles will move uphill in the other. Similarly, a few electrons in an otherwise empty conduction band move opposite to an electric field, while holes in an otherwise filled valence band move in the direction of the field. The bubble analogy is imperfect, but it may provide a physical feel for why the charge and mass of a hole have opposite signs from those of an electron.

In all the following discussions we shall concentrate on the electrons in the conduction band and on the holes in the valence band. We can account for the current flow in a semiconductor by the motion of these two types of charge carriers. We draw valence and conduction bands on an electron energy scale E, as in Fig. 3–8. However, we should remember that in the valence

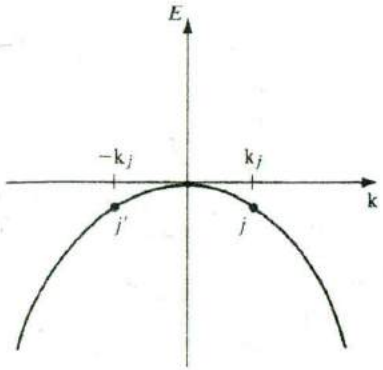


Figure 3–8
A valence band with all states filled, including states j and j', marked for discussion. The jth electron with wave vector k_j is matched by an electron at j' with the opposite wave vector $-k_j$. There is no net current in the band unless an electron is removed. For example, if the jth electron is removed, the motion of the electron at j' is no longer compensated.

band, hole energy increases oppositely to electron energy, because the two carriers have opposite charge. Thus hole energy increases downward in Fig. 3–8 and holes, seeking the lowest energy state available, are generally found at the *top* of the valence band. In contrast, conduction band electrons are found at the bottom of the conduction band.

It would be instructive to compare the $(E, \mathbf{k})$ band diagrams with the "simplified" band diagrams that are used for routine device analysis (Fig. 3–9). As discussed in Examples 3–1 and 3–2, an $(E, \mathbf{k})$ diagram is a plot of the total electron energy (potential plus kinetic) as a function of the crystal-direction–dependent electron wave vector (which is proportional to the momentum and therefore the velocity) at some point in space. Hence, the bottom of the conduction band corresponds to zero electron velocity or kinetic energy, and simply gives us the potential energy at that point in space. For holes, the top of the valence band corresponds to zero kinetic energy. For simplified band diagrams, we plot the edges of the conduction and valence bands (i.e., the potential energy) as a function of position in the device. Energies higher in the band correspond to additional kinetic energy of the electron. Also, the fact that the band edge corresponds to the electron potential energy tells us that the variation of the band edge in space is related to the electric field at different points in the semiconductor. We will show this relationship explicitly in Section 4.4.2.

In Fig. 3–9, an electron at location A sees an electric field given by the slope of the band edge (potential energy), and gains kinetic energy (at the expense of potential energy) by moving to point B. Correspondingly, in the $(E, \mathbf{k})$ diagram, the electron starts at $k = 0$, but moves to a non-zero wave vector $\mathbf{k}_B$.

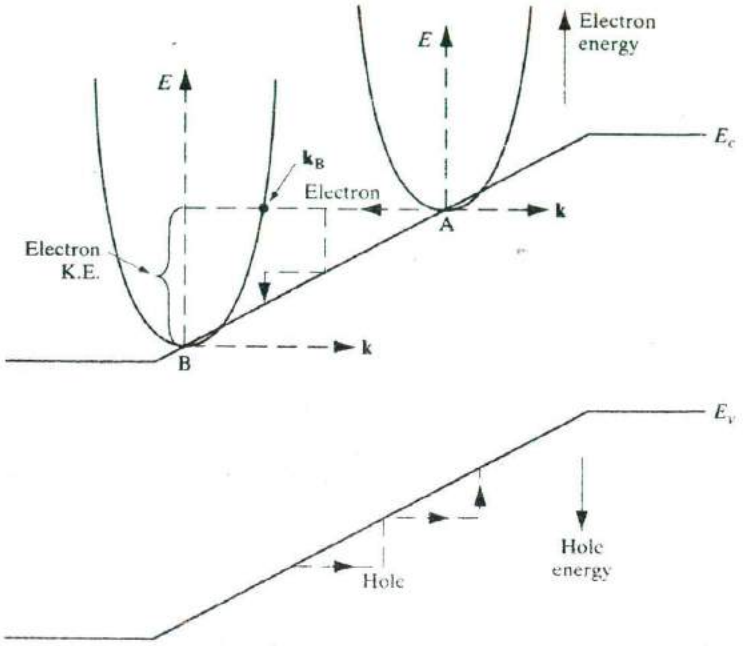


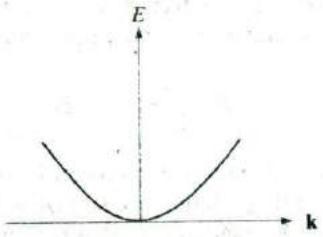
Figure 3–9 Superimposition of the (*E*,**k**) bandstructure on the *E*-versus-position simplified band diagram for a semiconductor in an electric field. Electron energies increase going up, while hole energies increase going down. Similarly, electron and hole wavevectors point in opposite directions and these charge carriers move opposite to each other, as shown.

The electron then loses kinetic energy to heat by scattering mechanisms (discussed in Section 3.4.3) and returns to the bottom of the band at B. The slopes of the (E, x) band edges at different points in space reflect the local electric fields at those points. In practice, the electron may lose its kinetic energy in stages by a series of scattering events, as shown by the dashed lines.

3.2.2 Effective Mass

The electrons in a crystal are not completely free, but instead interact with the periodic potential of the lattice. As a result, their "wave-particle" motion cannot be expected to be the same as for electrons in free space. Thus, in applying the usual equations of electrodynamics to charge carriers in a solid, we must use altered values of particle mass. In doing so, we account for most of the influences of the lattice, so that the electrons and holes can be treated as "almost free" carriers in most computations. The calculation of effective mass must take into account the shape of the energy bands in three-dimensional **k**- space, taking appropriate averages over the various energy bands.

EXAMPLE 3–2 Find the $(E, \mathbf{k})$ relationship for a free electron and relate it to the electron mass.



SOLUTION

From Example 3–1, the electron momentum is $\mathbf{p} = m\mathbf{v} = \hbar\mathbf{k}$. Then

$$E = \frac{1}{2}m\mathbf{v}^2 = \frac{1}{2}\frac{\mathbf{p}^2}{m} = \frac{\hbar^2}{2m}\mathbf{k}^2$$

Thus the electron energy is parabolic with wave vector **k**. The electron mass is inversely related to the curvature (second derivative) of the (E, **k**) relationship, since

$$\frac{d^2E}{d\mathbf{k}^2} = \frac{\hbar^2}{m}$$

Although electrons in solids are not free, most energy bands are close to parabolic at their minima (for conduction bands) or maxima (for valence bands). We can also approximate effective mass near those band extrema from the curvature of the band.

The effective mass of an electron in a band with a given (E, **k**) relationship is found in Example 3–2 to be

$$m^* = \frac{\hbar^2}{d^2E/d\mathbf{k}^2} \tag{3–3}$$

Thus the curvature of the band determines the electron effective mass. For example, in Fig. 3–6a it is clear that the electron effective mass in GaAs is much smaller in the direct Γ conduction band (strong curvature) than in the L or X minima (weaker curvature, smaller value in the denominator of the m^* expression).

A particularly interesting feature of Figs. 3–5 and 3–6 is that the curvature of $d^2E/d\mathbf{k}^2$ is positive at the conduction band minima, but is negative at the valence band maxima. Thus, the electrons near the top of the valence band have *negative effective mass*, according to Eq. (3–3). Valence band electrons with negative charge and negative mass move in an electric field in the same direction as holes with positive charge and positive mass. As discussed in Section 3.2.1, we can fully account for charge transport in the valence band by considering hole motion.

For a band centered at $\mathbf{k} = 0$ (such as the Γ band in GaAs), the $(E, \mathbf{k})$ relationship near the minimum is usually parabolic:

$$E = \frac{\hbar^2}{2m^*}\mathbf{k}^2 + E_c \tag{3–4}$$

Comparing this relation to Eq. (3–3) indicates that the effective mass m^* is constant in a parabolic band. On the other hand, many conduction bands have complex $(E, \mathbf{k})$ relationships that depend on the direction of electron transport with respect to the principal crystal directions. In this case, the effective mass is a tensor quantity. However, we can use appropriate averages over such bands in most calculations.

Figure 3–10a shows the bandstructures for Si and GaAs viewed along two major directions. While the shape is parabolic near the band edges (as indicated in Figure 3–5, and Example 3–2), there are significant non-parabolicities at higher energies. The energies are plotted along the high symmetry [111] and [100] directions in the crystal. The $\mathbf{k} = 0$ point is denoted as Γ. When we go along

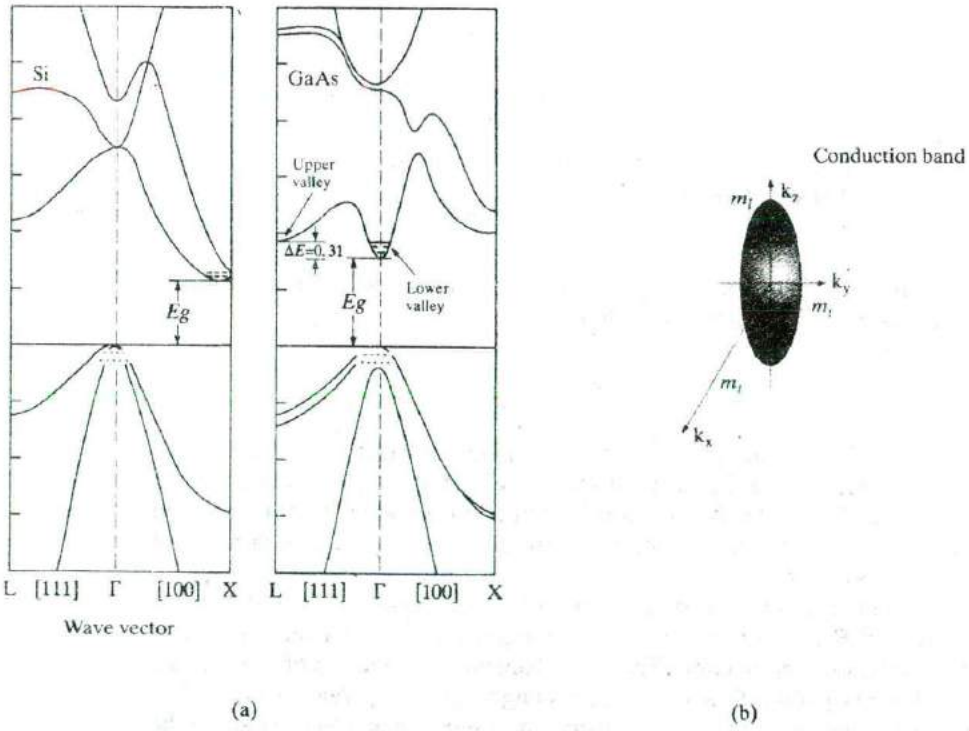


Figure 3–10
Realistic bandstructures in semiconductors: (a) Conduction and valence bands in Si and GaAs along [111] and [100]; (b) ellipsoidal constant energy surface for Si, near the 6 conduction band minima along the *X* directions. (From Chelikowsky and Cohen, Phys. Rev. B14, 556, 1976).

the [100] direction, we reach a valley near X, while we reach the L valley along the [111] direction. (Since the energies are plotted along different directions, the curves do not look symmetric.) The valence band maximum in most semiconductors is at the Γ point. It has three branches: the *heavy hole* band with the smallest curvature, a *light hole band* with a larger curvature, and a *split-off band* at a different energy. We notice that for GaAs the conduction band minimum and the valence band maximum are both at $\mathbf{k} = 0$; therefore it is direct bandgap. Silicon, on the other hand, has 6 equivalent conduction minima at X along the 6 equivalent $\langle 100 \rangle$ directions; therefore, it is indirect.

Figure 3–10b shows the constant energy surface for electrons in one of the six conduction bands for Si. The way to relate these surfaces to the bandstructures shown in Fig. 3–10a is to consider a certain value of energy, and determine all the $\mathbf{k}$ vectors in 3 dimensions for which we get this energy. We find that for Si we have 6 cigar-shaped ellipsoidal equi-energy surfaces near the conduction band minima along the six equivalent X-directions, with a longitudinal effective mass, m_l, along the major axis, and two transverse effective masses, m_t, along the minor axes. For GaAs, the conduction band is more or less spherical for low energies. On the other hand, we have warped spherical surfaces in the valence band. The importance of these surfaces will be clear in Section 3. 4.1 when we consider different types of effective masses in semiconductors.

In any calculation involving the mass of the charge carriers, we must use effective mass values for the particular material involved. In all subsequent discussions, the electron effective mass is denoted by m_n^* and the hole effective mass by m_p^*. The n subscript indicates the electron as a negative charge carrier, and the p subscript indicates the hole as a positive charge carrier.

There is nothing mysterious about the concept of an "effective" mass, m_n^*, and about the fact that it is different in different semiconductors. Indeed, the "true" mass of an electron, m, is the same in Si, Ge or GaAs—it is the same as for a free electron in vacuum. To understand why the effective mass is different from the true mass, consider Newton's second law which states that the time rate of change of momentum is the force.

$$dp/dt = d(m\mathrm{v})/dt = \text{Force} \tag{3–5a}$$

An electron in a crystal experiences a total force $\mathrm{F_{int}} + \mathrm{F_{ext}}$, where $\mathrm{F_{int}}$ is the collection of internal periodic crystal forces, and $\mathrm{F_{ext}}$ is the externally applied force. It is inefficient to solve this complicated problem involving the periodic crystal potential (which is obviously different in different semiconductors) every time we try to solve a semiconductor device problem. It is better to solve the complicated problem of carrier motion in the periodic crystal potential just once, and encapsulate that information in what is called the bandstructure, $(E, \mathbf{k})$, whose curvature gives us the effective mass, m_n^*. The electron then responds to external forces with this new m_n^*. Newton's law is then written as:

$$d(m_n^*\mathrm{v})/dt = \mathrm{F_{ext}} \tag{3–5b}$$

This is clearly an enormous simplification compared to the more detailed problem. Obviously, the periodic crystal forces depend on the details of a specific semiconductor; therefore, the effective mass is different in different materials.

Once we determine the band curvature effective mass components from the orientation-dependent $(E, \mathbf{k})$, we have to combine them appropriately for different types of calculations. We shall see in Section 3.3.2 that when we are interested in determining the numbers of carriers in the bands, we have to use a "density-of-states" effective mass by taking the geometric mean of the band curvature effective masses, and the number of equivalent band extrema. On the other hand we will find in Section 3.4.1 that in problems involving the motion of carriers, one must take the harmonic mean of the band curvature effective masses to get the "conductivity" effective mass.

3.2.3 Intrinsic Material

A perfect semiconductor crystal with no impurities or lattice defects is called an *intrinsic* semiconductor. In such material there are no charge-carriers at 0 K, since the valence band is filled with electrons and the conduction band is empty. At higher temperatures electron–hole pairs are generated as valence band electrons are excited thermally across the band gap to the conduction band. These EHPs are the only charge carriers in intrinsic material.

The generation of EHPs can be visualized in a qualitative way by considering the breaking of covalent bonds in the crystal lattice (Fig. 3–11). If one of the Si valence electrons is broken away from its position in the bonding structure such that it becomes free to move about in the lattice, a conduction electron is created and a broken bond (hole) is left behind. The energy required to break the bond is the band gap energy E_g. This model helps in visualizing the physical mechanism of EHP creation, but the energy band model is more productive for purposes of quantitative calculation. One important difficulty in the "broken bond" model is that the free electron and the hole seem deceptively localized in the lattice. Actually, the positions of the free electron and the hole are spread out over several lattice spacings and should be considered quantum mechanically by probability distributions (see Section 2.4).

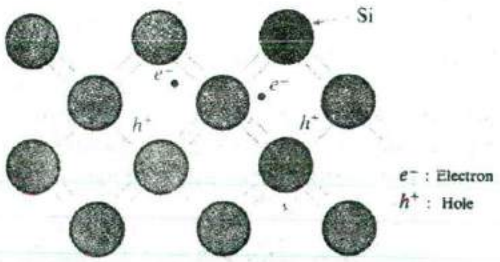


Figure 3–11 Electron–hole pairs in the covalent bonding model of the Si crystal.

Since the electrons and holes are created in pairs, the conduction band electron concentration n (electrons per cm^3) is equal to the concentration of holes in the valence band p (holes per cm^3). Each of these intrinsic carrier concentrations is commonly referred to as n_i. Thus *for intrinsic material*

$$n = p = n_i \tag{3-6}$$

At a given temperature there is a certain concentration of electron–hole pairs n_i. Obviously, if a steady state carrier concentration is maintained, there must be *recombination* of EHPs at the same rate at which they are generated. Recombination occurs when an electron in the conduction band makes a transition (direct or indirect) to an empty state (hole) in the valence band, thus annihilating the pair. If we denote the generation rate of EHPs as g_i (EHP/cm^3-s) and the recombination rate as r_i, equilibrium requires that

$$r_i = g_i \tag{3-7a}$$

Each of these rates is temperature dependent. For example, $g_i(T)$ increases when the temperature is raised, and a new carrier concentration n_i is established such that the higher recombination rate $r_i(T)$ just balances generation. At any temperature, we can predict that the rate of recombination of electrons and holes r_i is proportional to the equilibrium concentration of electrons n_0 and the concentration of holes p_0:

$$r_i = \alpha_r n_0 p_0 = \alpha_r n_i^2 = g_i \tag{3-7b}$$

The factor α_r is a constant of proportionality which depends on the particular mechanism by which recombination takes place. We shall discuss the calculation of n_i as a function of temperature in Section 3.3.3; recombination processes will be discussed in Chapter 4.

3.2.4 Extrinsic Material

In addition to the intrinsic carriers generated thermally, it is possible to create carriers in semiconductors by purposely introducing impurities into the crystal. This process, called *doping*, is the most common technique for varying the conductivity of semiconductors. By doping, a crystal can be altered so that it has a predominance of either electrons or holes. Thus there are two types of doped semiconductors, n-type (mostly electrons) and p-type (mostly holes). When a crystal is doped such that the equilibrium carrier concentrations n_0 and p_0 are different from the intrinsic carrier concentration n_i, the material is said to be *extrinsic*.

When impurities or lattice defects are introduced into an otherwise perfect crystal, additional levels are created in the energy band structure, usually within the band gap. For example, an impurity from column V of the periodic table (P, As, and Sb) introduces an energy level very near the conduction band in Ge or Si. This level is filled with electrons at 0 K, and very little thermal energy is required to excite these electrons to the conduction band (Fig. 3–12a). Thus at about 50–100 K virtually all of the electrons in the

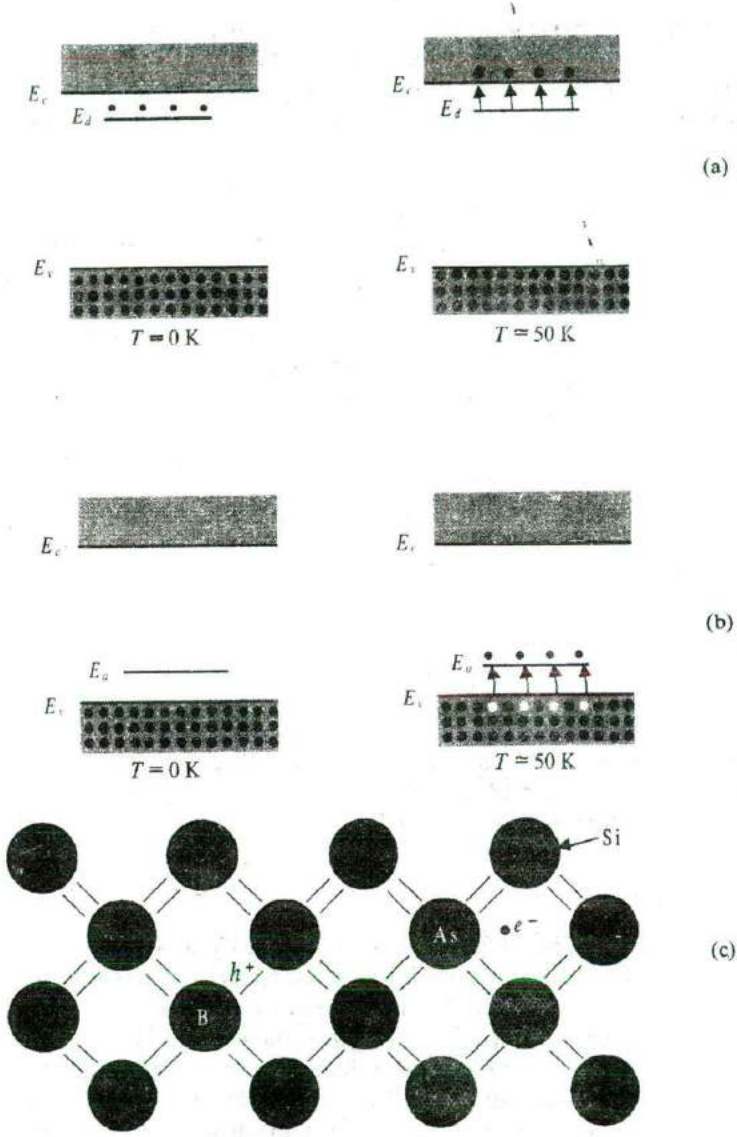


Figure 3–12
Energy band model and chemical bond model of dopants in semiconductors: (a) donation of electrons from donor level to conduction band; (b) acceptance of valence band electrons by an acceptor level, and the resulting creation of holes; (c) donor and acceptor atoms in the covalent bonding model of a Si crystal.

impurity level are "donated" to the conduction band. Such an impurity level is called a *donor* level, and the column V impurities in Ge or Si are called donor impurities. From Fig. 3–12a we note that the material doped with donor impurities can have a considerable concentration of electrons in the conduction band, even when the temperature is too low for the intrinsic EHP concentration to be appreciable. Thus semiconductors doped with a significant number of donor atoms will have $n_0 \gg (n_i, p_0)$ at room temperature. This is n-type material.

Atoms from column III (B, Al, Ga, and In) introduce impurity levels in Ge or Si near the valence band. These levels are empty of electrons at 0 K Fig. 3–12b). At low temperatures, enough thermal energy is available to excite electrons from the valence band into the impurity level, leaving behind holes in the valence band. Since this type of impurity level "accepts" electrons from the valence band, it is called an *acceptor* level, and the column III impurities are acceptor impurities in Ge and Si. As Fig. 3–12b indicates, doping with acceptor impurities can create a semiconductor with a hole concentration p_0 much greater than the conduction band electron concentration n_0 (this type is p-type material).

In the covalent bonding model, donor and acceptor atoms can be visualized as shown in Fig. 3–12c. An As atom (column V) in the Si lattice has the four necessary valence electrons to complete the covalent bonds with the neighboring Si atoms, plus one extra electron. This fifth electron does not fit into the bonding structure of the lattice and is therefore loosely bound to the As atom. A small amount of thermal energy enables this extra electron to overcome its coulombic binding to the impurity atom and be donated to the lattice as a whole. Thus it is free to participate in current conduction. This process is a qualitative model of the excitation of electrons out of a donor level and into the conduction band (Fig. 3–12a). Similarly, the column III impurity B has only three valence electrons to contribute to the covalent bonding (Fig. 3–12c), thereby leaving one bond incomplete. With a small amount of thermal energy, this incomplete bond can be transferred to other atoms as the bonding electrons exchange positions. Again, the idea of an electron "hopping" from an adjacent bond into the incomplete bond at the B site provides some physical insight into the behavior of an acceptor, but the model of Fig. 3–12b is preferable for most discussions.

We can calculate rather simply the approximate energy required to excite the fifth electron of a donor atom into the conduction band (the donor *binding energy*). Let us assume for rough calculations that the As atom of Fig. 3–12c has its four covalent bonding electrons rather tightly bound and the fifth "extra" electron loosely bound to the atom. We can approximate this situation by using the Bohr model results, considering the loosely bound electron as ranging about the tightly bound "core" electrons in a hydrogen-like orbit. From Eq. (2–15) the magnitude of the ground-state energy ($\mathbf{n} = 1$) of such an electron is

$$E = \frac{mq^4}{2K^2\hbar^2} \tag{3–8}$$

The value of K must be modified from the free-space value $4\pi\epsilon_0$ used in the hydrogen atom problem to

$$K = 4\pi\epsilon_0\,\epsilon_r \tag{3–9}$$

where ϵ_r is the relative dielectric constant of the semiconductor material. In addition, we must use the conductivity effective mass m_n^* typical of the semiconductor, discussed in more detail in Section 3.4.1.

EXAMPLE 3–3 Calculate the approximate donor binding energy for GaAs ($\epsilon_r = 13.2$, $m_n^* = 0.067m_0$).

SOLUTION From Eq. (3–8) and Appendix II we have

$$E = \frac{m_n^* q^4}{8(\epsilon_0\epsilon_r)^2\hbar^2} = \frac{0.067(9.11 \times 10^{-31})(1.6 \times 10^{-19})^4}{8(8.85 \times 10^{-12} \times 13.2)^2(6.63 \times 10^{-34})^2}$$

$$= 8.34 \times 10^{-22}\,\text{J} = 0.0052\,\text{eV}$$

Thus the energy required to excite the donor electron from the $n = 1$ state to the free state ($n = \infty$) is $\simeq 5.2$ meV. This corresponds to the energy difference $E_c - E_d$ in Fig. 3–10a and is in very close agreement with actual measured values.

Generally, the column V donor levels lie approximately 0.01 eV below the conduction band in Ge, and the column III acceptor levels lie about 0.01 eV above the valence band. In Si the usual donor and acceptor levels lie about 0.03–0.06 eV from a band edge.

In III–V compounds, column VI impurities occupying column V sites serve as donors. For example, S, Se, and Te are donors in GaAs, since they substitute for As and provide an extra electron compared with the As atom. Similarly, impurities from column II (Be, Zn, Cd) substitute for column III atoms to form acceptors in the III—V compounds. A more ambiguous case arises when a III–V material is doped with Si or Ge, from column IV. These impurities are called *amphoteric*, meaning that Si or Ge can serve as donors or acceptors depending on whether they reside on the column III or column V sublattice of the crystal. In GaAs it is common for Si impurities to occupy Ga sites. Since the Si has an extra electron compared with the Ga it replaces, it serves as a donor. However, an excess of As vacancies arising during growth or processing of the GaAs can cause Si impurities to occupy As sites, where they serve as acceptors.

The importance of doping will become obvious when we discuss electronic devices made from junctions between p-type and n-type semiconductor material. The extent to which doping controls the electronic properties of semiconductors can be illustrated here by considering changes in the sample resistance which occur with doping. In Si, for example, the intrinsic carrier concentration n_i is about 10^{10} cm^{-3} at room temperature. If we dope Si with 10^{15} As atoms/cm^3, the conduction electron concentration changes by five orders of magnitude. The resistivity of Si changes from about 2×10^5 Ω-cm to 5 Ω-cm with this doping.

When a semiconductor is doped n-type or p-type, one type of carrier dominates. In the example given above, the conduction band electrons outnumber the holes in the valence band by many orders of magnitude. We refer

Atoms from column III (B, Al, Ga, and In) introduce impurity levels in Ge or Si near the valence band. These levels are empty of electrons at 0 K Fig. 3–12b). At low temperatures, enough thermal energy is available to excite electrons from the valence band into the impurity level, leaving behind holes in the valence band. Since this type of impurity level "accepts" electrons from the valence band, it is called an *acceptor* level, and the column III impurities are acceptor impurities in Ge and Si. As Fig. 3–12b indicates, doping with acceptor impurities can create a semiconductor with a hole concentration p_0 much greater than the conduction band electron concentration n_0 (this type is p-type material).

In the covalent bonding model, donor and acceptor atoms can be visualized as shown in Fig. 3–12c. An As atom (column V) in the Si lattice has the four necessary valence electrons to complete the covalent bonds with the neighboring Si atoms, plus one extra electron. This fifth electron does not fit into the bonding structure of the lattice and is therefore loosely bound to the As atom. A small amount of thermal energy enables this extra electron to overcome its coulombic binding to the impurity atom and be donated to the lattice as a whole. Thus it is free to participate in current conduction. This process is a qualitative model of the excitation of electrons out of a donor level and into the conduction band (Fig. 3–12a). Similarly, the column III impurity B has only three valence electrons to contribute to the covalent bonding (Fig. 3–12c), thereby leaving one bond incomplete. With a small amount of thermal energy, this incomplete bond can be transferred to other atoms as the bonding electrons exchange positions. Again, the idea of an electron "hopping" from an adjacent bond into the incomplete bond at the B site provides some physical insight into the behavior of an acceptor, but the model of Fig. 3–12b is preferable for most discussions.

We can calculate rather simply the approximate energy required to excite the fifth electron of a donor atom into the conduction band (the donor *binding energy*). Let us assume for rough calculations that the As atom of Fig. 3–12c has its four covalent bonding electrons rather tightly bound and the fifth "extra" electron loosely bound to the atom. We can approximate this situation by using the Bohr model results, considering the loosely bound electron as ranging about the tightly bound "core" electrons in a hydrogen-like orbit. From Eq. (2–15) the magnitude of the ground-state energy ($\mathbf{n} = 1$) of such an electron is

$$E = \frac{mq^4}{2K^2\hbar^2} \tag{3–8}$$

The value of K must be modified from the free-space value $4\pi\epsilon_0$ used in the hydrogen atom problem to

$$K = 4\pi\epsilon_0\,\epsilon_r \tag{3–9}$$

where ϵ_r is the relative dielectric constant of the semiconductor material. In addition, we must use the conductivity effective mass m_n^* typical of the semiconductor, discussed in more detail in Section 3.4.1.

EXAMPLE 3–3 Calculate the approximate donor binding energy for GaAs ($\epsilon_r = 13.2$, $m_n^* = 0.067m_0$).

SOLUTION From Eq. (3–8) and Appendix II we have

$$E = \frac{m_n^* q^4}{8(\epsilon_0\epsilon_r)^2 h^2} = \frac{0.067(9.11 \times 10^{-31})(1.6 \times 10^{-19})^4}{8(8.85 \times 10^{-12} \times 13.2)^2(6.63 \times 10^{-34})^2}$$

$$= 8.34 \times 10^{-22}\,\text{J} = 0.0052\,\text{eV}$$

Thus the energy required to excite the donor electron from the $n = 1$ state to the free state ($n = \infty$) is $\simeq 5.2$ meV. This corresponds to the energy difference $E_c - E_d$ in Fig. 3–10a and is in very close agreement with actual measured values.

Generally, the column V donor levels lie approximately 0.01 eV below the conduction band in Ge, and the column III acceptor levels lie about 0.01 eV above the valence band. In Si the usual donor and acceptor levels lie about 0.03–0.06 eV from a band edge.

In III–V compounds, column VI impurities occupying column V sites serve as donors. For example, S, Se, and Te are donors in GaAs, since they substitute for As and provide an extra electron compared with the As atom. Similarly, impurities from column II (Be, Zn, Cd) substitute for column III atoms to form acceptors in the III—V compounds. A more ambiguous case arises when a III–V material is doped with Si or Ge, from column IV. These impurities are called *amphoteric*, meaning that Si or Ge can serve as donors or acceptors depending on whether they reside on the column III or column V sublattice of the crystal. In GaAs it is common for Si impurities to occupy Ga sites. Since the Si has an extra electron compared with the Ga it replaces, it serves as a donor. However, an excess of As vacancies arising during growth or processing of the GaAs can cause Si impurities to occupy As sites, where they serve as acceptors.

The importance of doping will become obvious when we discuss electronic devices made from junctions between p-type and n-type semiconductor material. The extent to which doping controls the electronic properties of semiconductors can be illustrated here by considering changes in the sample resistance which occur with doping. In Si, for example, the intrinsic carrier concentration n_i is about 10^{10} cm^{-3} at room temperature. If we dope Si with 10^{15} As atoms/cm^3, the conduction electron concentration changes by five orders of magnitude. The resistivity of Si changes from about 2×10^5 Ω-cm to 5 Ω-cm with this doping.

When a semiconductor is doped n-type or p-type, one type of carrier dominates. In the example given above, the conduction band electrons outnumber the holes in the valence band by many orders of magnitude. We refer

to the small number of holes in n-type material as *minority carriers* and the relatively large number of conduction band electrons as *majority carriers*. Similarly, electrons are the minority carriers in p-type material, and holes are the majority carriers.

3.2.5 Electrons and Holes in Quantum Wells

We have discussed single-valued (*discrete*) energy levels in the band gap arising from doping, and a *continuum* of allowed states in the valence and conduction bands. A third possibility is the formation of discrete levels for electrons and holes as a result of quantum-mechanical confinement.

One of the most useful applications of MBE or OMVPE growth of multi-layer compound semiconductors, as described in Section 1.4, is the fact that a continuous single crystal can be grown in which adjacent layers have different band gaps. For example, Fig. 3–13 shows the spatial variation in conduction and valence bands for a multilayer structure in which a very thin layer of GaAs is sandwiched between two layers of AlGaAs, which has a wider band gap that the GaAs. We will discuss the details of such *heterojunctions* (junctions between dissimilar materials) in Section 5.8. It is interesting to point out here, however, that a consequence of confining electrons and holes in a very thin layer is that these particles behave according to the *particle in a potential well* problem, with quantum states calculated in Section

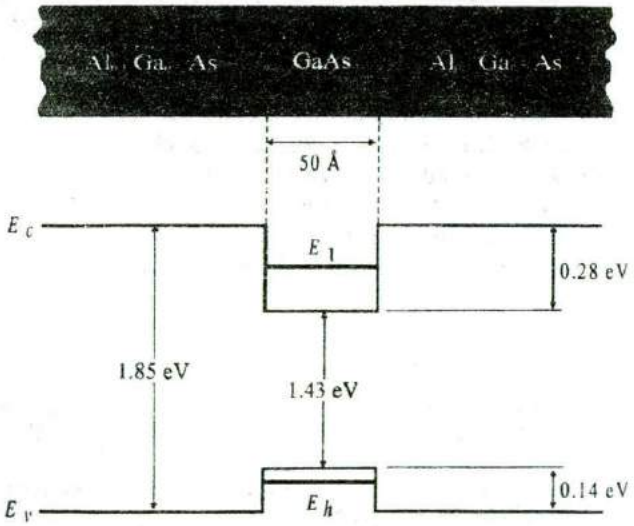


Figure 3–13
Energy band discontinuities for a thin layer of GaAs sandwiched between layers of wider band gap AlGaAs. In this case, the GaAs region is so thin that quantum states are formed in the valence and conduction bands. Electrons in the GaAs conduction band reside on "particle in a potential well" states such a E_1 shown here, rather than in the usual conduction band states. Holes in the quantum well occupy similar discrete states, such a E_h.

2.4.3. Therefore, instead of having the continuum of states normally available in the conduction band, the conduction band electrons in the narrow-gap material are confined to discrete quantum states as described by Eq. (2–33), modified for effective mass and finite barrier height. Similarly, the states in the valence band available for holes are restricted to discrete levels in the quantum well. This is one of the clearest demonstrations of the quantum mechanical results discussed in Chapter 2. From a practical device point of view, the formation of discrete quantum states in the GaAs layer of Fig. 3–13 changes the energy at which photons can be emitted. An electron on one of the discrete conduction band states (E_1 in Fig. 3–13) can make a transition to an empty discrete valence band state in the GaAs quantum well (such as E_h), giving off a photon of energy $E_g + E_1 + E_h$, greater than the GaAs band gap. Semiconductor lasers have been made in which such a quantum well is used to raise the energy of the transition from the infrared, typical of GaAs, to the red portion of the spectrum. We will see other examples of quantum wells in semiconductor devices in later chapters.

3.3 CARRIER CONCENTRATIONS

In calculating semiconductor electrical properties and analyzing device behavior, it is often necessary to know the number of charge carriers per cm^3 in the material. The majority carrier concentration is usually obvious in heavily doped material, since one majority carrier is obtained for each impurity atom (for the standard doping impurities). The concentration of minority carriers is not obvious, however, nor is the temperature dependence of the carrier concentrations.

To obtain equations for the carrier concentrations we must investigate the distribution of carriers over the available energy states. This type of distribution is not difficult to calculate, but the derivation requires some background in statistical methods. Since we are primarily concerned here with the application of these results to semiconductor materials and devices, we shall accept the distribution function as given.

3.3.1 The Fermi Level

Electrons in solids obey *Fermi–Dirac* statistics.[3] In the development of this type of statistics, one must consider the indistinguishability of the electrons,

[3]Examples of other types of statistics are *Maxwell–Boltzmann* for classical particles (e.g., gas) and *Bose–Einstein* for photons. For two discrete energy levels, E_2 and E_1 (with $E_2 > E_1$), classical gas atoms follow a Boltzmann distribution; the number n_2 of atoms in state E_2 is related to the number n_1 in E_1 at thermal equilibrium by

$$\frac{n_2}{n_1} = \frac{N_2 e^{-E_2/kT}}{N_1 e^{-E_1/kT}} = \frac{N_2}{N_1} e^{-(E_2 - E_1)/kT}$$

assuming the two levels have N_2 and N_1 number of states, respectively. The exponential term $\exp(-\Delta E/kT)$ is commonly called the *Boltzmann factor*. It appears also in the denominator of the Fermi–Dirac distribution function. We shall return to the Boltzmann distribution in Chapter 8 in discussions of the properties of lasers.

their wave nature, and the Pauli exclusion principle. The rather simple result of these statistical arguments is that the distribution of electrons over a range of allowed energy levels at thermal equilibrium is

$$f(E) = \frac{1}{1 + e^{(E-E_F)/kT}} \tag{3-10}$$

where k is Boltzmann's constant ($k = 8.62 \times 10^{-5}$ eV/K $= 1.38 \times 10^{-23}$ J/K). The function $f(E)$, the *Fermi–Dirac distribution function*, gives the probability that an available energy state at E will be occupied by an electron at absolute temperature T. The quantity E_F is called the *Fermi level*, and it represents an important quantity in the analysis of semiconductor behavior. We notice that, for an energy E equal to the Fermi level energy E_F, the occupation probability is

$$f(E_F) = [1 + e^{(E_F - E_F)/kT}]^{-1} = \frac{1}{1+1} = \frac{1}{2} \tag{3-11}$$

Thus an energy state at the Fermi level has a probability of $^1/_2$ of being occupied by an electron.

A closer examination of $f(E)$ indicates that at 0 K the distribution takes the simple rectangular form shown in Fig. 3–14. With $T = 0$ in the denominator of the exponent, $f(E)$ is $1/(1 + 0) = 1$ when the exponent is negative ($E < E_F$), and is $1/(1 + \infty) = 0$ when the exponent is positive ($E > E_F$). This rectangular distribution implies that at 0 K every available energy state up to E_F is filled with electrons, and all states above E_F are empty.

At temperatures higher than 0 K, some probability exists for states above the Fermi level to be filled. For example, at $T = T_1$ in Fig. 3–14 there is some probability $f(E)$ that states above E_F are filled, and there is a corresponding probability $[1 - f(E)]$ that states below E_F are empty. The Fermi function is symmetrical about E_F for all temperatures; that is, the probability $f(E_F + \Delta E)$ that a state ΔE above E_F is filled is the same as the probability $[1 - f(E_F - \Delta E)]$ that a state ΔE below E_F is empty. The symmetry of the

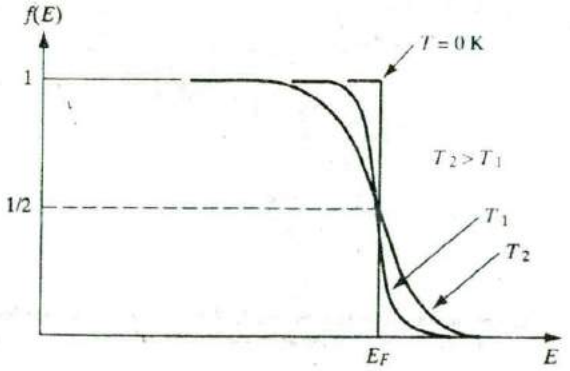


Figure 3–14
The Fermi-Dirac distribution function.

distribution of empty and filled states about E_F makes the Fermi level a natural reference point in calculations of electron and hole concentrations in semiconductors.

In applying the Fermi–Dirac distribution to semiconductors, we must recall that $f(E)$ is the probability of occupancy of an *available* state at E. Thus if there is no available state at E (e.g., in the band gap of a semiconductor), there is no possibility of finding an electron there. We can best visualize the relation between $f(E)$ and the band structure by turning the $f(E)$ vs. E diagram on its side so that the E scale corresponds to the energies of the band diagram (Fig. 3–15). For intrinsic material we know that the concentration of holes in the valence band is equal to the concentration of electrons in the conduction band. Therefore, the Fermi level E_F must lie at the middle of the band gap in intrinsic material.[4] Since $f(E)$ is symmetrical about E_F, the electron probability "tail" of $f(E)$ extending into the conduction band of Fig. 3–15a is symmetrical with the hole probability tail $[1 - f(E)]$ in the valence band. The distribution function has values within the band gap between

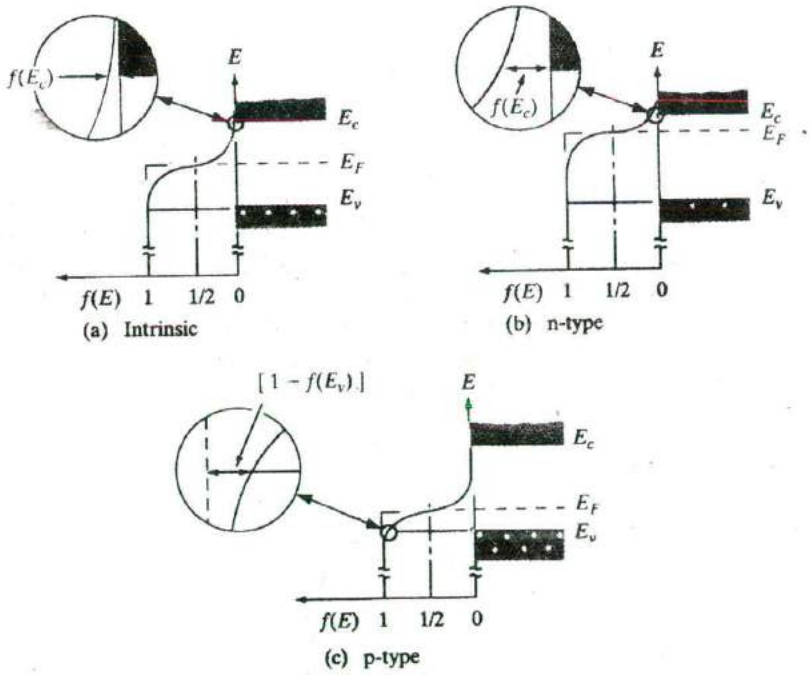


Figure 3–15
The Fermi distribution function applied to semiconductors: (a) intrinsic material; (b) n-type material; (c) p-type material.

[4]Actually the intrinsic E_F is displaced slightly from the middle of the gap, since the densities of available states in the valence and conduction bands are not equal (Section 3.3.2).

E_v and E_c but there are no energy states available, and no electron occupancy results from $f(E)$ in this range.

The tails in $f(E)$ are exaggerated in Fig. 3–15 for illustrative purposes. Actually, the probability values at E_v and E_c are quite small for intrinsic material at reasonable temperatures. For example, in Si at 300 K, $n_i = p_i \approx 10^{10}$ cm^{-3}, whereas the densities of available states at E_v and E_c are on the order of 10^{19} cm^{-3}. Thus the probability of occupancy $f(E)$ for an individual state in the conduction band and the hole probability $[1 - f(E)]$ for a state in the valence band are quite small. Because of the relatively large density of states in each band, small changes in $f(E)$ can result in significant changes in carrier concentration.

In n-type material there is a high concentration of electrons in the conduction band compared with the hole concentration in the valence band (recall Fig. 3–12a). Thus in n-type material the distribution function $f(E)$ must lie above its intrinsic position on the energy scale (Fig. 3–15b). Since $f(E)$ retains its shape for a particular temperature, the larger concentration of electrons at E_c in n-type material implies a correspondingly smaller hole concentration at E_v. We notice that the value of $f(E)$ for each energy level in the conduction band (and therefore the total electron concentration n_0) increases as E_F moves closer to E_c. Thus the energy difference $(E_c - E_F)$ gives a measure of n; we shall express this relation mathematically in the following section.

For p-type material the Fermi level lies near the valence band (Fig. 3–15c) such that the $[1 - f(E)]$ tail below E_v is larger than the $f(E)$ tail above E_c. The value of $(E_F - E_v)$ indicates how strongly p-type the material is.

It is usually inconvenient to draw $f(E)$ vs. E on every energy band diagram to indicate the electron and hole distributions. Therefore, it is common practice merely to indicate the position of E_F in band diagrams. This is sufficient information, since for a particular temperature the position of E_F implies the distributions in Fig. 3–15.

3.3.2 Electron and Hole Concentrations at Equilibrium

The Fermi distribution function can be used to calculate the concentrations of electrons and holes in a semiconductor, if the densities of available states in the valence and conduction bands are known. For example, the concentration of electrons in the conduction band is

$$n_0 = \int_{E_c}^{\infty} f(E)N(E)dE \tag{3–12}$$

where $N(E)dE$ is the density of states (cm^{-3}) in the energy range dE. The subscript 0 used with the electron and hole concentration symbols (n_0, p_0) indicates equilibrium conditions. The number of electrons per unit volume in the energy range dE is the product of the density of states and the probability of occupancy $f(E)$. Thus the total electron concentration is the integral

over the entire conduction band, as in Eq. (3–12).[5] The function $N(E)$ can be calculated by using quantum mechanics and the Pauli exclusion principle (Appendix IV).

It is shown in Appendix IV that $N(E)$ is proportional to $E^{1/2}$, so the density of states in the conduction band increases with electron energy. On the other hand, the Fermi function becomes extremely small for large energies. The result is that the product $f(E)N(E)$ decreases rapidly above E_c, and very few electrons occupy energy states far above the conduction band edge. Similarly, the probability of finding an empty state (hole) in the valence band $[1 - f(E)]$ decreases rapidly below E_v, and most holes occupy states near the top of the valence band. This effect is demonstrated in Fig. 3–16, which shows the density of available states, the Fermi function, and the resulting number of electrons and holes occupying available energy states in the conduction and valence bands at thermal equilibrium (i.e., with no excitations except thermal energy). For holes, increasing energy points down in Fig. 3–16, since the E scale refers to electron energy.

The result of the integration of Eq. (3–12) is the same as that obtained if we represent all of the distributed electron states in the conduction band by an *effective density of states* N_c located at the conduction band edge E_c. Therefore, the conduction band electron concentration is simply the effective density of states at E_c times the probability of occupancy at E_c[6]

$$n_0 = N_c f(E_c) \tag{3–13}$$

In this expression we assume the Fermi level E_F lies at least several kT below the conduction band. Then the exponential term is large compared with unity, and the Fermi function $f(E_c)$ can be simplified as

$$f(E_c) = \frac{1}{1 + e^{(E_c - E_F)/kT}} \simeq e^{-(E_c - E_F)/kT} \tag{3–14}$$

Since kT at room temperature is only 0.026 eV, this is generally a good approximation. For this condition the concentration of electrons in the conduction band is

$$\boxed{n_0 = N_c e^{-(E_c - E_F)/kT}} \tag{3–15}$$

The effective density of states N_c is shown in Appendix IV to be

$$N_c = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{3/2} \tag{3–16}$$

[5]The upper limit is actually improper in Eq. (3–12), since the conduction band does not extend to infinite energy. This is unimportant in the calculation of n_0, however, since $f(E)$ becomes negligibly small for large values of E. Most electrons occupy states near the bottom of the conduction band at equilibrium.

[6]The simple expression for n_0 obtained in Eq. (3–13) is the direct result of integrating Eq. (3–12), as in Appendix IV. Equations (3–15) and (3–19) properly include the effects of the conduction and valence bands through the density-of-states terms.

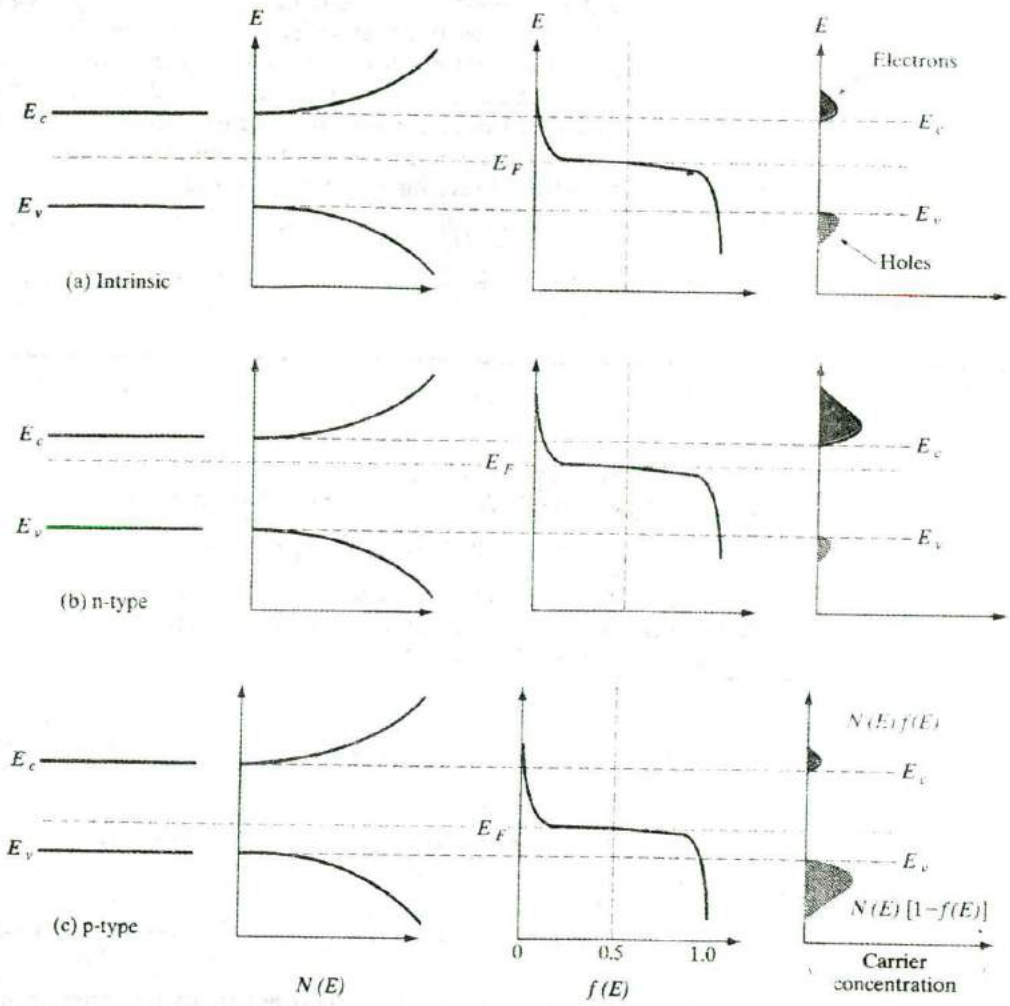


Figure 3–16
Schematic band diagram, density of states, Fermi–Dirac distribution, and the carrier concentrations for (a) intrinsic, (b) n-type, and (c) p-type semiconductors at thermal equilibrium.

Since the quantities in Eq. (3–16a) are known, values of N_c can be tabulated as a function of temperature. As Eq. (3–15) indicates, the electron concentration increases as E_F moves closer to the conduction band. This is the result we would predict from Fig. 3–15b.

In Eq. (3–16a), m_n^* is the density-of-states effective mass for electrons. To illustrate how it is obtained from the band curvature effective masses mentioned in Section 3.2.2, let us consider the 6 equivalent conduction band

minima along the X-directions for Si. Looking at the cigar-shaped equi-energy surfaces in Fig. 3–10b, we find that we have more than one band curvature to deal with in calculating effective masses. There is a longitudinal effective mass m_l along the major axis of the ellipsoid, and the transverse effective mass m_t along the two minor axes. Since we have $(m_n^*)^{3/2}$ appearing in the density-of-states expression Eq. (3–16a), by using dimensional equivalence and adding contributions from all 6 valleys, we get

$$(m_n^*)^{3/2} = 6(m_l m_t^2)^{1/2} \tag{3–16b}$$

It can be seen that this is the geometric mean of the effective masses.

EXAMPLE 3–4 Calculate the density-of-states effective mass of electrons in Si.

SOLUTION

For Si, $m_l = 0.98\ m_0$; $m_t = 0.19\ m_0$ from Appendix III.
There are six equivalent X valleys in the conduction band.

$$m_n^* = 6^{2/3}[0.98(0.19)^2]^{1/3} m_0 = 1.1\ m_0$$

Note: For GaAs, the conduction band equi-energy surfaces are spherical. So there is only one band curvature effective mass, and it is equal to the density-of-states effective mass ($= 0.067\ m_0$).

By similar arguments, the concentration of holes in the valence band is

$$p_0 = N_v[1 - f(E_v)] \tag{3–17}$$

where N_v is the effective density of states in the valence band. The probability of finding an empty state at E_v is

$$1 - f(E_v) = 1 - \frac{1}{1 + e^{(E_v - E_F)/kT}} \simeq e^{-(E_F - E_v)/kT} \tag{3–18}$$

for E_F larger than E_v by several kT. From these equations, the concentration of holes in the valence band is

$$\boxed{p_0 = N_v e^{-(E_F - E_v)/kT}} \tag{3–19}$$

The effective density of states in the valence band reduced to the band edge is

$$N_v = 2\left(\frac{2\pi m_p^* kT}{h^2}\right)^{3/2} \tag{3–20}$$

As expected from Fig. 3–15c, Eq. (3–19) predicts that the hole concentration increases as E_F moves closer to the valence band.

The electron and hole concentrations predicted by Eqs. (3–15) and (3–19) are valid whether the material is intrinsic or doped, provided thermal equilibrium is maintained. Thus *for intrinsic material, E_F lies at some intrinsic level E_i* near the middle of the band gap (Fig. 3–15a), and the intrinsic electron and hole concentrations are

$$n_i = N_c e^{-(E_c - E_i)/kT}, \quad p_i = N_v e^{-(E_i - E_v)/kT} \tag{3–21}$$

The product of n_0 and p_0 at equilibrium is a constant for a particular material and temperature, even if the doping is varied:

$$n_0 p_0 = (N_c e^{-(E_c - E_F)/kT})(N_v e^{-(E_F - E_v)/kT}) = N_c N_v e^{-(E_c - E_v)/kT} \tag{3–22a}$$
$$= N_c N_v e^{-E_g/kT}$$

$$n_i p_i = (N_c e^{-(E_c - E_i)/kT})(N_v e^{-(E_i - E_v)/kT}) = N_c N_v e^{-E_g/kT} \tag{3–22b}$$

The intrinsic electron and hole concentrations are equal (since the carriers are created in pairs), $n_i = p_i$; thus the intrinsic concentration is

$$n_i = \sqrt{N_c N_v}\, e^{-E_g/2kT} \tag{3–23}$$

The constant product of electron and hole concentrations in Eq. (3–22) can be written conveniently as

$$\boxed{n_0 p_0 = n_i^2} \tag{3–24}$$

This is an important relation, and we shall use it extensively in later calculations. The intrinsic concentration for Si at room temperature is approximately $n_i = 1.5 \times 10^{10}\ \text{cm}^{-3}$.

Comparing Eqs. (3–21) and (3–23), we note that the intrinsic level E_i is the middle of the band gap ($E_c - E_i = E_g/2$), if the effective densities of states N_c and N_v are equal. There is usually some difference in effective mass for electrons and holes, however, and N_c and N_v are slightly different as Eqs. (3–16) and (3–20) indicate. The intrinsic level E_i is displaced from the middle of the band gap, more for GaAs than for Ge or Si.

Another convenient way of writing Eqs. (3–15) and (3–19) is

$$\boxed{n_0 = n_i e^{(E_F - E_i)/kT}} \tag{3–25a}$$
$$\boxed{p_0 = n_i e^{(E_i - E_F)/kT}} \tag{3–25b}$$

obtained by the application of Eq. (3–21). This form of the equations indicates directly that the electron concentrations is n_i when E_F is at the intrinsic level E_i, and that n_0 increases exponentially as the Fermi level moves away from E_i toward the conduction band. Similarly, the hole concentration p_0 varies from n_i to larger values as E_F moves from E_i toward the valence band. Since these equations reveal the qualitative features of carrier concentration so directly, they are particularly convenient to remember.

EXAMPLE 3–5 A Si sample is doped with 10^{17} As atoms/cm^3. What is the equilibrium hole concentration p_0 at 300 K? Where is E_F relative to E_i?

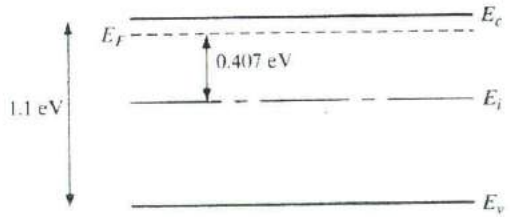
SOLUTION Since $N_d \gg n_i$, we can approximate $n_0 = N_d$ and

$$p_0 = \frac{n_i^2}{n_0} = \frac{2.25 \times 10^{20}}{10^{17}} = 2.25 \times 10^3 \text{ cm}^{-3}$$

From Eq. (3–25a), we have

$$E_F - E_i = kT \ln \frac{n_0}{n_i} = 0.0259 \ln \frac{10^{17}}{1.5 \times 10^{10}} = 0.407 \text{ eV}$$

The resulting band diagram is:



3.3.3 Temperature Dependence of Carrier Concentrations

The variation of carrier concentration with temperature is indicated by Eq. (3–25). Initially, the variation of n_0 and p_0 with T seems relatively straightforward in these relations. The problem is complicated, however, by the fact that n_i has a strong temperature dependence [Eq. (3–23)] and that E_F can also vary with temperature. Let us begin by examining the intrinsic carrier concentration. By combining Eqs. (3–23), (3–16a), and (3–20) we obtain

$$n_i(T) = 2\left(\frac{2\pi kT}{h^2}\right)^{3/2} (m_n^* m_p^*)^{3/4} e^{-E_g/2kT} \tag{3–26}$$

The exponential temperature dependence dominates $n_i(T)$, and a plot of ln n_i vs. $10^3/T$ appears linear (Fig. 3–17).[7] In this figure we neglect variations due to the $T^{3/2}$ dependence of the density-of-states function and the fact

[7]When plotting quantities such as carrier concentration, which involve a Boltzmann factor, it is common to use an inverse temperature scale. This allows terms which are exponential in $1/T$ to appear linear in the semi-logarithmic plot. When reading such graphs, remember that temperature increases from right to left.

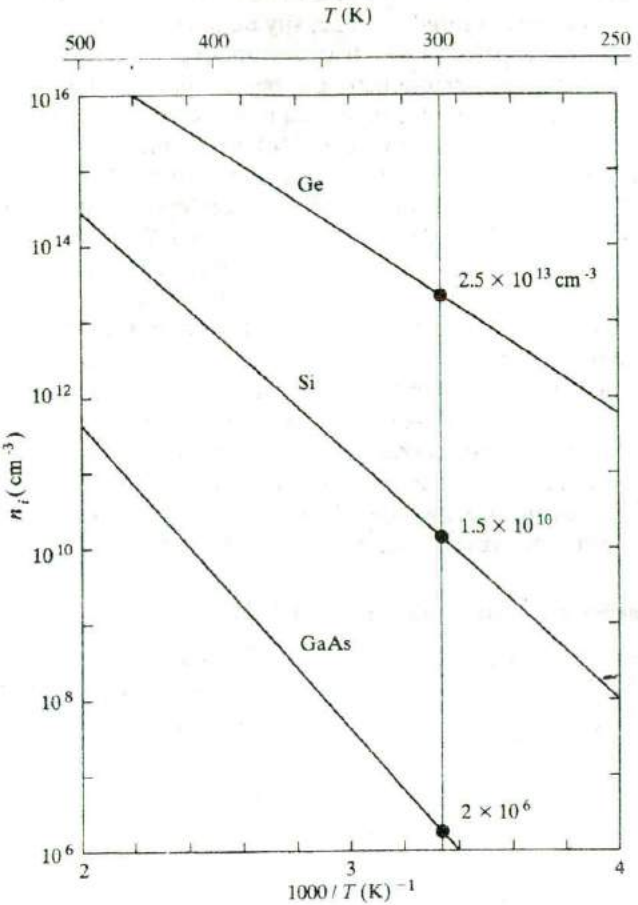


Figure 3–17
Intrinsic carrier concentration for Ge, Si, and GaAs as a function of inverse temperature. The room temperature values are marked for reference.

that E_g varies somewhat with temperature.[8] The value of n_i at any temperature is a definite number for a given semiconductor, and is known for most materials. Thus we can take n_i as given in calculating n_0 or p_0 from Eq. (3–25).[9]

With n_i and T given, the unknowns in Eq. (3–25) are the carrier concentrations and the Fermi level position relative to E_i. One of these two

[8]For Si the band gap E_g varies from about 1.11 eV at 300 K to about 1.16 eV at 0 K.

[9]Care must be taken to use consistent units in these calculations. For example, if an energy such as E_g is expressed in electron volts (eV), it should be multiplied by q (1.6×10^{-19} C) to convert to joules if k is in J/K; alternatively, E_g can be kept in eV and the value of k in eV/K can be used. At 300 K we can use $kT = 0.0259$ eV and E_g in eV.

quantities must be given if the other is to be found. If the carrier concentration is held at a certain value, as in heavily doped extrinsic material, E_F can be obtained from Eq. (3–25). The temperature dependence of electron concentration in a doped semiconductor can be visualized as shown in Fig. 3–18. In this example, Si is doped n-type with a donor concentration N_d of 10^{15} cm^{-3}. At very low temperatures (large $1/T$), negligible intrinsic EHPs exist, and the donor electrons are bound to the donor atoms. As the temperature is raised, these electrons are donated to the conduction band, and at about 100 K ($1000/T = 10$) all the donor atoms are ionized. This temperature range is called the *ionization* region. Once the donors are ionized, the conduction band electron concentration is $n_0 \simeq N_d = 10^{15}$ cm^{-3}, since one electron is obtained for each donor atom. When every available extrinsic electron has been transferred to the conduction band, n_0 is virtually constant with temperature until the concentration of intrinsic carriers n_i becomes comparable to the extrinsic concentration N_d. Finally, at higher temperatures n_i is much greater than N_d, and the intrinsic carriers dominate. In most devices it is desirable to control the carrier concentration by doping rather than by thermal EHP generation. Thus one usually dopes the material such that the extrinsic range extends beyond the highest temperature at which the device is to be used.

3.3.4 Compensation and Space Charge Neutrality

When the concept of doping was introduced, we assumed the material contained either N_d donors or N_a acceptors, so that the extrinsic majority carrier concentrations were $n_0 \simeq N_d$ or $p_0 \simeq N_a$, respectively, for the n-type or p-type

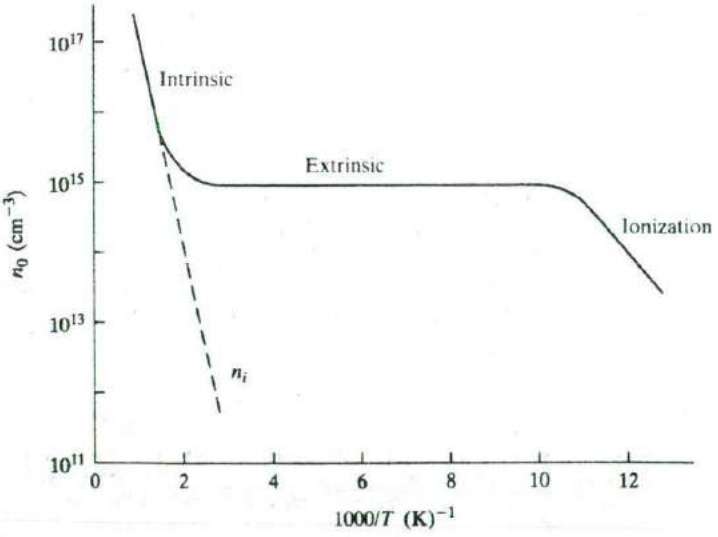


Figure 3–18 Carrier concentration vs. inverse temperature for Si doped with 10^{15} donors/cm^3.

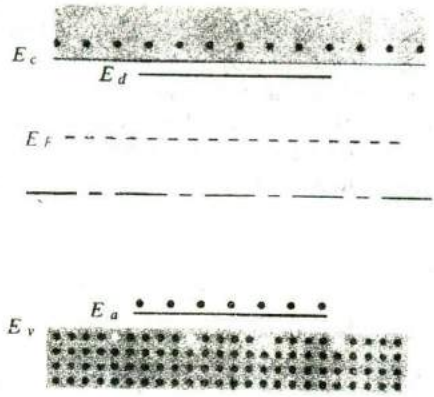


Figure 3–19
Compensation in an n-type semiconductor ($N_d > N_a$).

material. If often happens, however, that a semiconductor contains both donors and acceptors. For example, Fig. 3–19 illustrates a semiconductor for which both donors and acceptors are present, but $N_d > N_a$. The predominance of donors makes the material n-type, and the Fermi level is therefore in the upper part of the band gap. Since E_F is well above the acceptor level E_a, this level is essentially filled with electrons. However, with E_F above E_i, we cannot expect a hole concentration in the valence band commensurate with the acceptor concentration. In fact, the filling of the E_a states occurs at the expense of the donated conduction band electrons. The mechanism can be visualized as follows: Assume an acceptor state is filled with a valence band electron as described in Fig. 3–12b, with a hole resulting in the valence band. This hole is then filled by recombination with one of the conduction band electrons. Extending this logic to all the acceptor atoms, we expect the resultant concentration of electrons in the conduction band to be $N_d - N_a$ instead of the total N_d. This process is called *compensation*. By this process it is possible to begin with an *n*-type semiconductor and add acceptors until $N_a = N_d$ and no donated electrons remain in the conduction band. In such compensated material, $n_0 = n_i = p_0$ and intrinsic conduction is obtained. With further acceptor doping the semiconductor becomes p-type with a hole concentration of essentially $N_a - N_d$.

The exact relationship among the electron, hole, donor, and acceptor concentrations can be obtained by considering the requirements for *space charge neutrality*. If the material is to remain electrostatically neutral, the sum of the positive charges (holes and ionized donor atoms) must balance the sum of the negative charges (electrons and ionized acceptor atoms):

$$p_0 + N_d^+ = n_0 + N_a^- \tag{3–27}$$

Thus in Fig. 3–19 the net electron concentration in the conduction band is

$$n_0 = p_0 + (N_d^+ - N_a^-) \tag{3–28}$$

If the material is doped n-type ($n_0 \gg p_0$) and all the impurities are ionized, we can approximate Eq. (3–28) by $n_0 \simeq N_d - N_a$.

Since the intrinsic semiconductor itself is electrostatically neutral and the doping atoms we add are also neutral, the requirement of Eq. (3–27) must be maintained at equilibrium. The electron and hole concentrations and the Fermi level adjust such at Eqs. (3–27) and (3–25) are satisfied.

3.4 DRIFT OF CARRIERS IN ELECTRIC AND MAGNETIC FIELDS

Knowledge of carrier concentrations in a solid is necessary for calculating current flow in the presence of electric or magnetic fields. In addition to the values of n and p, we must be able to take into account the collisions of the charge carriers with the lattice and with the impurities. These processes will affect the ease with which electrons and holes can flow through the crystal, that is, their *mobility* within the solid. As should be expected, these collision and scattering processes depend on temperature, which affects the thermal motion of the lattice atoms and the velocity of the carriers.

3.4.1 Conductivity and Mobility

The charge carriers in a solid are in constant motion, even at thermal equilibrium. At room temperature, for example, the thermal motion of an individual electron may be visualized as random scattering from lattice vibrations, impurities, other electrons, and defects (Fig. 3–20). Since the scattering is random, there is no net motion of the group of n electrons/cm^3 over any period of time. This is not true of an individual electron, of course. The probability of the electron in Fig. 3–20 returning to its starting point after some time t is negligibly small. However, if a large number of electrons is considered (e.g., 10^{16} cm^{-3} in an n-type semiconductor), there will be no preferred direction of motion for the group of electrons and no net current flow.

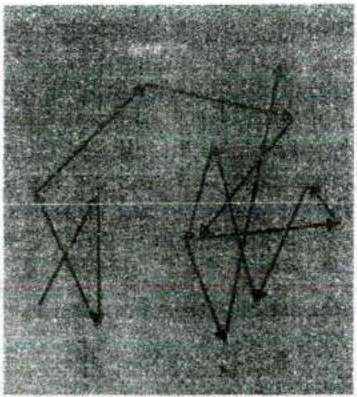


Figure 3–20 Thermal motion of an electron in a solid.

If an electric field $\mathscr{E}_x$ is applied in the x-direction, each electron experiences a net force $-q\mathscr{E}_x$ from the field. This force may be insufficient to alter appreciably the random path of an individual electron; the effect when averaged over all the electrons, however, is a net motion of the group in the $-x$-direction. If p_x is the x-component of the total momentum of the group, the force of the field on the n electrons/cm^3 is

$$-nq\mathscr{E}_x = \left.\frac{d\mathsf{p}_x}{dt}\right|_{\text{field}} \tag{3–29}$$

Initially, Eq. (3–29) seems to indicate a continuous acceleration of the electrons in the $-x$-direction. This is not the case, however, because the net acceleration of Eq. (3–29) is just balanced in steady state by the decelerations of the collision processes. Thus while the steady field $\mathscr{E}_x$ does produce a net momentum p_{-x}, the net rate of change of momentum when collisions are included must be zero in the case of steady state current flow.

To find the total rate of momentum change from collisions, we must investigate the collision probabilities more closely. If the collisions are truly random, there will be a constant probability of collision at any time for each electron. Let us consider a group of N_0 electrons at time $t = 0$ and define $N(t)$ as the number of electrons that *have not* undergone a collision by time t. The rate of decrease in $N(t)$ at any time t is proportional to the number left unscattered at t,

$$-\frac{dN(t)}{dt} = \frac{1}{\bar{t}}N(t) \tag{3–30}$$

where $\bar{t}^{-1}$ is a constant of proportionality.

The solution to Eq. (3–30) is an exponential fuction

$$N(t) = N_0 e^{-t/\bar{t}} \tag{3–31}$$

and $\bar{t}$ represents the mean time between scattering events,[10] called the *mean free time*. The probability that any electron has a collision in the time interval dt is $dt/\bar{t}$. Thus the differential change in p_x due to collisions in time dt is

$$d\mathsf{p}_x = -\mathsf{p}_x\frac{dt}{\bar{t}} \tag{3–32}$$

The rate of change of p_x due to the decelerating effect of collisions is

$$\left.\frac{d\mathsf{p}_x}{dt}\right|_{\text{collisions}} = -\frac{\mathsf{p}_x}{\bar{t}} \tag{3–33}$$

[10]Equations (3–30) and (3–31) are typical of events dominated by random processes, and the forms of these equations occur often in many branches of physics and engineering. For example, in the radioactive decay of unstable nuclear isotopes, N_0 nuclides decay exponentially with a mean lifetime $\bar{t}$. Other examples will be found in this text, including the absorption of light in a semiconductor and the recombination of excess EHPs.

The sum of acceleration and deceleration effects must be zero for steady state. Taking the sum of Eqs. (3–29) and (3–33), we have

$$-\frac{p_x}{\bar{t}} - nq\mathscr{E}_x = 0 \tag{3–34}$$

The average momentum per electron is

$$\langle p_x \rangle = \frac{p_x}{n} = -q\bar{t}\,\mathscr{E}_x \tag{3–35}$$

where the angular brackets indicate an average over the entire group of electrons. As expected for steady state, Eq. (3–35) indicates that the electrons have *on the average* a constant net velocity in the negative x-direction:

$$\langle v_x \rangle = \frac{\langle p_x \rangle}{m_n^*} = -\frac{q\bar{t}}{m_n^*}\mathscr{E}_x \tag{3–36}$$

Actually, the individual electrons move in many directions by thermal motion during a given time period, but Eq. (3–36) tells us the *net drift* of an average electron in response to the electric field. The drift speed described by Eq. (3–36) is usually much smaller than the random speed due to the thermal motion v_{th}.

The current density resulting from this net drift is just the number of electrons crossing a unit area per unit time ($n\langle v_x \rangle$) multiplied by the charge on the electron ($-q$):

$$\boxed{J_x = -qn\langle v_x \rangle} \qquad \frac{\text{ampere}}{\text{cm}^2} = \frac{\text{coulomb}}{\text{electron}} \cdot \frac{\text{electrons}}{\text{cm}^3} \cdot \frac{\text{cm}}{\text{s}} \tag{3–37}$$

Using Eq. (3–36) for the average velocity, we obtain

$$J_x = \frac{nq^2\bar{t}}{m_n^*}\mathscr{E}_x \tag{3–38}$$

Thus the current density is proportional to the electric field, as we expect from Ohm's law:

$$J_x = \sigma\mathscr{E}_x, \quad \text{where } \sigma \equiv \frac{nq^2\bar{t}}{m_n^*} \tag{3–39}$$

The conductivity $\sigma(\Omega\text{-cm})^{-1}$ can be written

$$\sigma = qn\mu_n, \quad \text{where } \mu_n \equiv \frac{q\bar{t}}{m_n^*} \tag{3–40a}$$

The quantity μ_n, called the *electron mobility*, describes the ease with which electrons drift in the material. Mobility is a very important quantity in characterizing semiconductor materials and in device development.

Here m_n^* is the conductivity effective mass for electrons, different from the density-of-states effective mass mentioned in Eq. (3–16b). While we use the density-of-states effective mass to count the number of carriers in bands, we must use the conductivity effective mass for charge transport problems. To illustrate how it is obtained from the band curvature effective masses mentioned in Section 3.2.2, once again let us consider the 6 equivalent conduction band minima along the X-directions for Si, with the band curvature longitudinal effective mass, m_l, along the major axis of the ellipsoid, and the transverse effective mass, m_t, along the two minor axes (Fig. 3–10b). Since we have $1/m_n^*$ in the mobility expression Eq. (3–40a), by using dimensional equivalence, we can write the conductivity effective mass as the harmonic mean of the band curvature effective masses.

$$\frac{1}{m_n^*} = \frac{1}{3}\left(\frac{1}{m_l} + \frac{2}{m_t}\right) \tag{3–40b}$$

EXAMPLE 3–6

Calculate the conductivity effective mass of electrons in Si.

SOLUTION

For Si, $m_l = 0.98\, m_0$; $m_t = 0.19\, m_0$ (Appendix III)
There are 6 equivalent X valleys in the conduction band.

$$1/m_n^* = 1/3(1/m_x + 1/m_y + 1/m_z) = 1/3(1/m_l + 2/m_t)$$

$$1/m_n^* = \frac{1}{3}\left(\frac{1}{0.98\, m_0} + \frac{2}{0.19\, m_0}\right)$$

$$m_n^* = 0.26\, m_0$$

Note: For GaAs, the conduction band equi-energy surfaces are spherical. So there is only one band curvature effective mass. (The density of states effective mass and the conductivity effective mass are both $0.067\, m_0$.)

The mobility defined in Eq. (3–40a) can be expressed as the average particle drift velocity per unit electric field. Comparing Eqs. (3–36) and (3–40a), we have

$$\mu_n = -\frac{\langle v_x \rangle}{\mathscr{E}_x} \tag{3–41}$$

The units of mobility are (cm/s)/(V/cm) = cm^2/V-s, as Eq. (3–41) suggests. The minus sign in the definition results in a positive value of mobility, since electrons drift opposite to the field.

The current density can be written in terms of mobility as

$$J_x = qn\mu_n \mathscr{E}_x \tag{3–42}$$

This derivation has been based on the assumption that the current is carried primarily by electrons. For hole conduction we change n to p, $-q$ to $+q$, and μ_n to μ_p, where $\mu_p = +\langle v_x \rangle / \mathscr{E}_x$ is the mobility for holes. If both electrons and holes participate, we must modify Eq. (3–42) to

$$\boxed{J_x = q(n\mu_n + p\mu_p)\mathscr{E}_x = \sigma\mathscr{E}_x} \tag{3–43}$$

Values of μ_n and μ_p are given for many of the common semiconductor materials in Appendix III. According to Eq. (3–40), the parameters determining mobility are m^* and mean free time $\bar{t}$. Effective mass is a property of the material's band structure, as described by Eq. (3–3). Thus we expect m_n^* to be small in the strongly curved Γ minimum of the GaAs conduction band (Fig. 3–6), with the result that μ_n is very high. In a more gradually curved band, a larger m^* in the denominator of Eq. (3–40) leads to a smaller value of mobility. It is reasonable to expect that lighter particles are more mobile than heavier particles (which is satisfying, since the common-sense value of effective mass is not always apparent). The other parameter determining mobility is the mean time between scattering events, $\bar{t}$. In Section 3.4.3 we shall see that this is determined primarily by temperature and impurity concentration in the semiconductor.

3.4.2 Drift and Resistance

Let us look more closely at the drift of electrons and holes. If the semiconductor bar of Fig. 3–21 contains both types of carrier, Eq. (3–43) gives the conductivity of the material. The resistance of the bar is then

$$R = \frac{\rho L}{wt} = \frac{L}{wt}\frac{1}{\sigma} \tag{3–44}$$

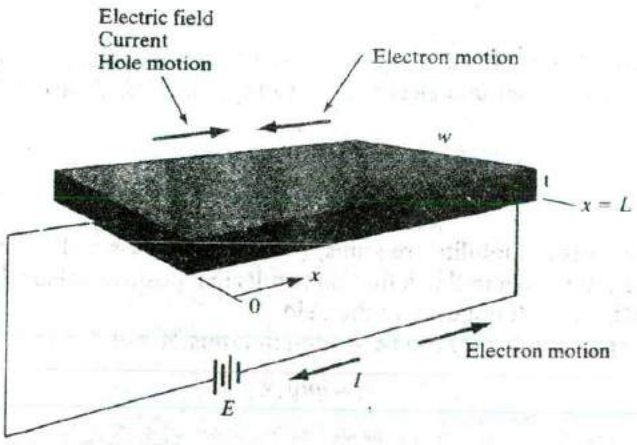


Figure 3–21 Drift of electrons and holes in a semiconductor bar.

where ρ is the resistivity (Ω-cm). The physical mechanism of carrier drift requires that the holes in the bar move as a group in the direction of the electric field and that the electrons move as a group in the opposite direction. Both the electron and the hole components of current are in the direction of the $\mathscr{E}$ field, since conventional current is positive in the direction of hole flow and opposite to the direction of electron flow. The drift current described by Eq. (3–43) is constant throughout the bar. A valid question arises, therefore, concerning the nature of the electron and hole flow at the contacts and in the external circuit. We should specify that the contacts to the bar of Fig. 3–21 are *ohmic*, meaning that they are perfect sources and sinks of both carrier types and have no special tendency to inject or collect either electrons or holes.

If we consider that current is carried around the external circuit by electrons, there is no problem in visualizing electrons flowing into the bar at one end and out at the other (always opposite to I). Thus for every electron leaving the left end ($x = 0$) of the bar in Fig. 3–21, there is a corresponding electron entering at $x = L$, so that the electron concentration in the bar remains constant at n. But what happens to the holes at the contacts? As a hole reaches the ohmic contact at $x = L$, it recombines with an electron, which must be supplied through the external circuit. As this hole disappears, a corresponding hole must appear at $x = 0$ to maintain space charge neutrality. It is reasonable to consider the source of this hole as the generation of an EHP at $x = 0$, with the hole flowing into the bar and the electron flowing into the external circuit.

EXAMPLE 3–7

Find the resistivity of intrinsic Si at 300 K.

SOLUTION

From Appendix III, $\mu_n = 1350$ and $\mu_p = 480$ cm^2/V-s for intrinsic Si. Thus, since $n_0 = p_0 = n_i$,

$$\sigma_i = q(\mu_n + \mu_p)n_i = 1.6 \times 10^{-19}(1830)(1.5 \times 10^{10})$$
$$= 4.39 \times 10^{-6}\,(\Omega - \text{cm})^{-1}$$
$$\rho_i = \sigma_i^{-1} = 2.28 \times 10^5\,\Omega - \text{cm}$$

3.4.3 Effects of Temperature and Doping on Mobility

The two basic types of scattering mechanisms that influence electron and hole mobility are *lattice scattering* and *impurity scattering*. In lattice scattering a carrier moving through the crystal is scattered by a vibration of the lattice, resulting from the temperature.[11] The frequency of such scattering events

[11]Collective vibrations of atoms in the crystal are called *phonons*. Thus lattice scattering is also known as *phonon scattering*.

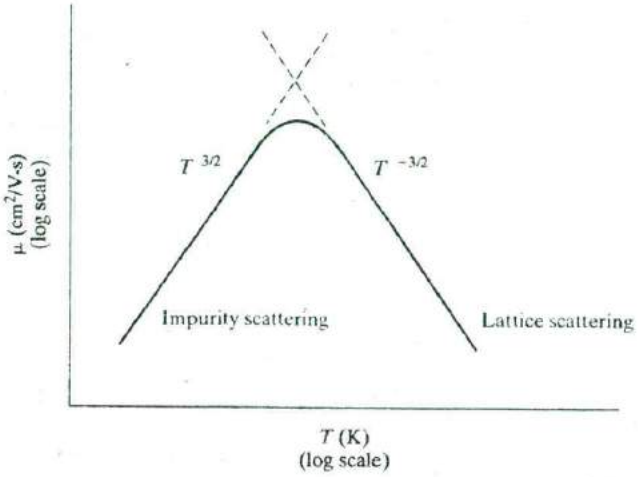


Figure 3–22
Approximate temperature dependence of mobility with both lattice and impurity scattering.

increases as the temperature increases, since the thermal agitation of the lattice becomes greater. Therefore, we should expect the mobility to decrease as the sample is heated (Fig. 3–22). On the other hand, scattering from crystal defects such as ionized impurities becomes the dominant mechanism at low temperatures. Since the atoms of the cooler lattice are less agitated, lattice scattering is less important; however, the thermal motion of the carriers is also slower. Since a slowly moving carrier is likely to be scattered more strongly by an interaction with a charged ion than is a carrier with greater momentum, impurity scattering events cause a decrease in mobility with decreasing temperature. As Fig. 3–22 indicates, the approximate temperature dependencies are $T^{-3/2}$ for lattice scattering and $T^{3/2}$ for impurity scattering. Since the scattering probability of Eq. (3–32) is inversely proportional to the mean free time and therefore to mobility, the mobilities due to two or more scattering mechanisms add inversely:

$$\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots \tag{3–45}$$

As a result, the mechanism causing the lowest mobility value dominates, as shown in Fig. 3–22.

As the concentration of impurities increases, the effects of impurity scattering are felt at higher temperatures. For example, the electron mobility μ_n of intrinsic silicon at 300 K is 1350 cm^2/(V-s). With a donor doping concentration of 10^{17} cm^{-3}, however, μ_n is 700 cm^2/(V-s). Thus the presence of the 10^{17} ionized donors/cm^3 introduces a significant amount of impurity scattering. This effect is illustrated in Fig. 3–23, which shows the variation of mobility with doping concentration at room temperature.

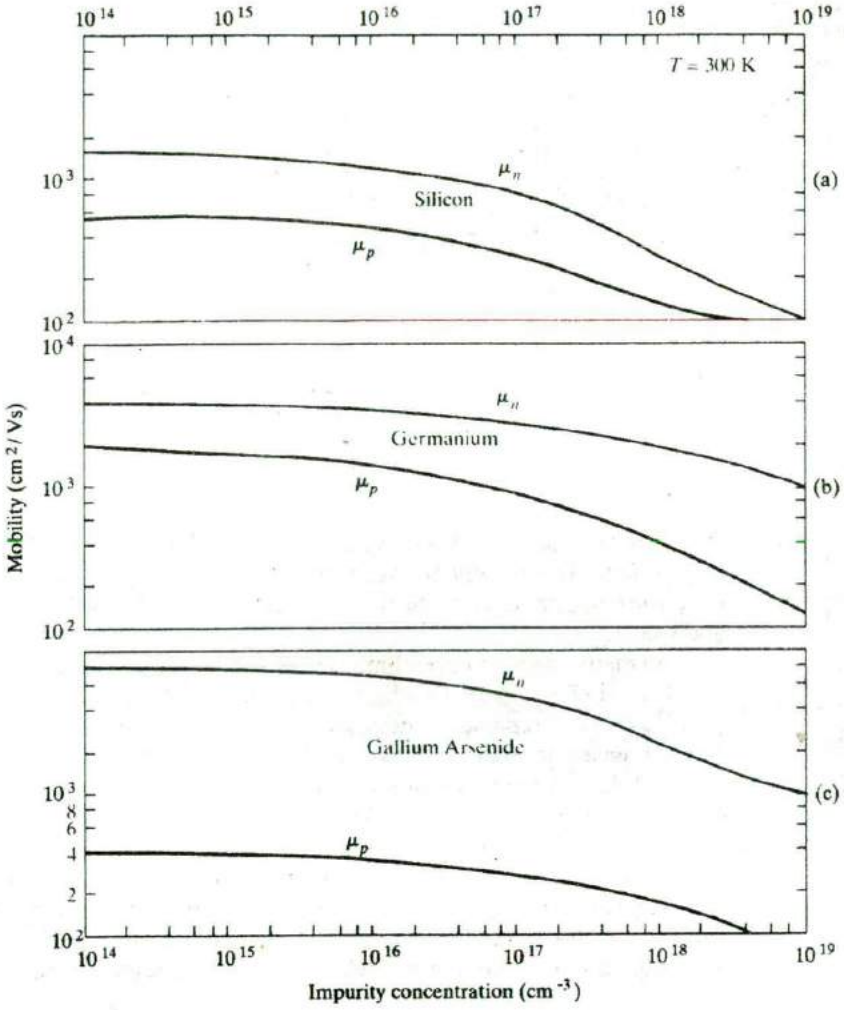


Figure 3–23
Variation of mobility with total doping impurity concentration ($N_a + N_d$) for Ge, Si, and GaAs at 300 K.

3.4.4 High-Field Effects

One assumption implied in the derivation of Eq. (3–39) was that Ohm's law is valid in the carrier drift processes. That is, it was assumed that the drift current is proportional to the electric field and that the proportionality constant (σ) is not a function of field $\mathscr{E}$. This assumption is valid over a wide range of $\mathscr{E}$. However, large electric fields ($> 10^3$V/cm) can cause the drift velocity and

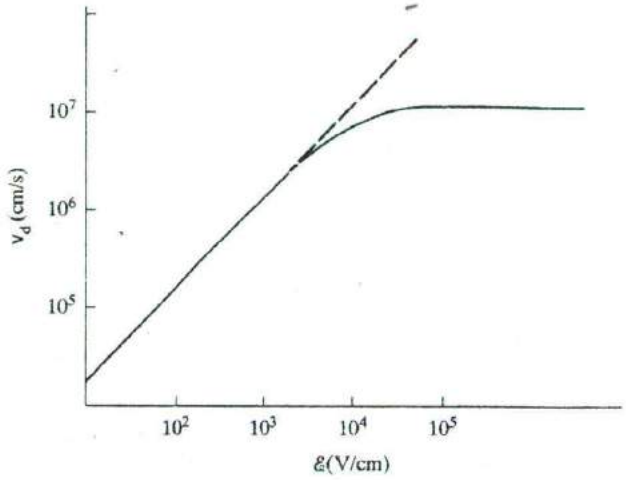


Figure 3–24 Saturation of electron drift velocity at high electric fields for Si.

therefore the current $J = -qnv_d$ to exhibit a sublinear dependence on the electric field. This dependence of σ upon $\mathscr{E}$ is an example of a *hot carrier* effect, which implies that the carrier drift velocity v_d is comparable to the thermal velocity v_{th}.

In many cases an upper limit is reached for the carrier drift velocity in a high field (Fig. 3–24). This limit occurs near the mean thermal velocity ($\simeq 10^7$ cm/s) and represents the point at which added energy imparted by the field is transferred to the lattice rather than increasing the carrier velocity. The result of this *scattering limited velocity* is a fairly constant current at high field. This behavior is typical of Si, Ge, and some other semiconductors. However, there are other important effects in some materials; for example, in Chapter 10 we shall discuss a *decrease* in electron velocity at high fields for GaAs and certain other materials, which results in negative conductivity and current instabilities in the sample. Another important high-field effect is avalanche multiplication, which we shall discuss in Section 5.4.2.

3.4.5 The Hall Effect

If a magnetic field is applied perpendicular to the direction in which holes drift in a p-type bar, the path of the holes tends to be deflected (Fig. 3–25). Using vector notation, the total force on a single hole due to the electric and magnetic fields is

$$\mathbf{F} = q(\mathscr{E} + \mathbf{v} \times \mathscr{B}) \tag{3–46}$$

In the y-direction the force is

$$F_y = q(\mathscr{E}_y - v_x \mathscr{B}_z) \tag{3–47}$$

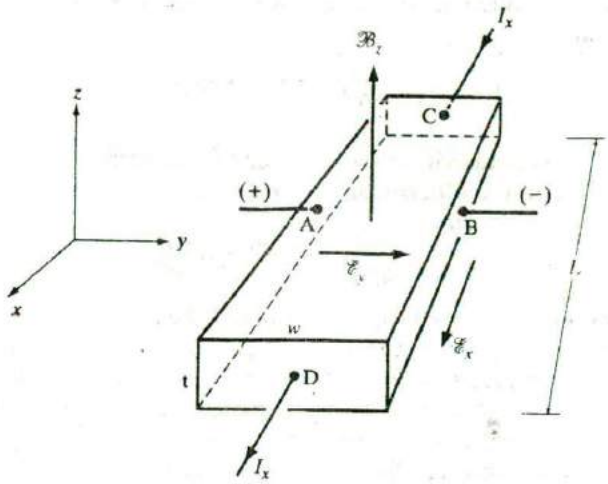


Figure 3–25
The Hall effect.

The important result of Eq. (3–47) is that unless an electric field $\mathcal{E}_y$ is established along the width of the bar, each hole will experience a net force (and therefore an acceleration) in the $-y$-direction due to the $qv_x\mathcal{B}_z$ product. Therefore, to maintain a steady state flow of holes down the length of the bar, the electric field $\mathcal{E}_y$ must just balance the product $v_x\mathcal{B}_z$;

$$\mathcal{E}_y = v_x\mathcal{B}_z \tag{3–48}$$

so that the net force F_y is zero. Physically, this electric field is set up when the magnetic field shifts the hole distribution slightly in the $-y$-direction. Once the electric field $\mathcal{E}_y$ becomes as large as $v_x\mathcal{B}_z$, no net lateral force is experienced by the holes as they drift along the bar. The establishment of the electric field $\mathcal{E}_y$ is known as the *Hall effect*, and the resulting voltage $V_{AB} = \mathcal{E}_y w$ is called the *Hall voltage*. If we use the expression derived in Eq. (3–37) for the drift velocity (using $+q$ and p_0 for holes), the field $\mathcal{E}_y$ becomes

$$\mathcal{E}_y = \frac{J_x}{qp_0}\mathcal{B}_z = R_H J_x \mathcal{B}_z, \quad R_H \equiv \frac{1}{qp_0} \tag{3–49}$$

Thus the Hall field is proportional to the product of the current density and the magnetic flux density. The proportionality constant $R_H = (qp_0)^{-1}$ is called the *Hall coefficient*. A measurement of the Hall voltage for a known current and magnetic field yields a value for the hole concentration p_0

$$p_0 = \frac{1}{qR_H} = \frac{J_x\mathcal{B}_z}{q\mathcal{E}_y} = \frac{(I_x/wt)\mathcal{B}_z}{q(V_{AB}/w)} = \frac{I_x\mathcal{B}_z}{qtV_{AB}} \tag{3–50}$$

Since all of the quantities in the right-hand side of Eq. (3–50) can be measured, the Hall effect can be used to give quite accurate values for carrier concentration.

If a measurement of resistance R is made, the sample resistivity ρ can be calculated:

$$\rho(\Omega-\text{cm}) = \frac{Rwt}{L} = \frac{V_{CD}/I_x}{L/wt} \tag{3-51}$$

Since the conductivity $\sigma = 1/\rho$ is given by $q\mu_p p_0$, the mobility is simply the ratio of the Hall coefficient and the resistivity:

$$\mu_p = \frac{\sigma}{qp_0} = \frac{1/\rho}{q(1/qR_H)} = \frac{R_H}{\rho} \tag{3-52}$$

Measurements of the Hall coefficient and the resistivity over a range of temperatures yield plots of majority carrier concentration and mobility vs. temperature. Such measurements are extremely useful in the analysis of semiconductor materials. Although the discussion here has been related to p-type material, similar results are obtained for n-type material. A negative value of q is used for electrons, and the Hall voltage V_{AB} and Hall coefficient R_H are negative. In fact, measurement of the sign of the Hall voltage is a common technique for determining if an unknown sample is p-type or n-type.

EXAMPLE 3–8 A sample of Si is doped with 10^{17} phosphorus atoms/cm^3. What would you expect to measure for its resistivity? What Hall voltage would you expect in a sample 100 μm thick if $I_x = 1$mA and $\mathcal{B}_z = 1$ kG $= 10^{-5}$ Wb/cm^2?

SOLUTION From Fig. 3–23, the mobility is 700 cm^2/(V-s). Thus the conductivity is

$$\sigma = q\mu_n n_0 = (1.6 \times 10^{-19})\,(700)\,(10^{17}) = 11.2(\Omega-\text{cm})^{-1}$$

since p_0 is negligible. The resistivity is

$$\rho = \sigma^{-1} = 0.0893\ \Omega-\text{cm}$$

The Hall coefficient is

$$R_H = -(qn_0)^{-1} = -62.5\ \text{cm}^3/\text{C}$$

from Eq. (3–49), or we could use Eq. (3–52). The Hall voltage is

$$V_{AB} = \frac{I_x \mathcal{B}_z}{t} R_H = \frac{(10^{-3}\ \text{A})(10^{-5}\ \text{Wb/cm}^2)}{10^{-2}\ \text{cm}}(-62.5\ \text{cm}^3/\text{C}) = -62.5\ \mu\text{V}$$

3.5 INVARIANCE OF THE FERMI LEVEL AT EQUILIBRIUM

In this chapter we have discussed homogeneous semiconductors, without variations in doping and without junctions between dissimilar materials. In the following chapters we will be considering cases in which nonuniform doping occurs in a given semiconductor, or junctions occur between differ-

ent semiconductors or a semiconductor and a metal. These cases are crucial to the various types of electronic and optoelectronic devices made in semiconductors. In anticipation of those discussions, an important concept should be established here regarding the demands of equilibrium. That concept can be summarized by noting that *no discontinuity or gradient can arise in the equilibrium Fermi level* E_F.

To demonstrate this assertion, let us consider two materials in intimate contact such that electrons can move between the two (Fig. 3–26). These may be, for example, dissimilar semiconductors, n- and p-type regions, a metal and a semiconductor, or simply two adjacent regions of a nonuniformly doped semiconductor. Each material is described by a Fermi–Dirac distribution function and some distribution of available energy states that electrons can occupy.

There is no current, and therefore no net charge transport, at thermal equilibrium. There is also no net transfer of energy. Therefore, for each energy E in Fig. 3–26 any transfer of electrons from material 1 to material 2 must be exactly balanced by the opposite transfer of electrons from 2 to 1. We will let the density of states at energy E in material 1 be called $N_1(E)$ and in material 2 we will call it $N_2(E)$. At energy E the rate of transfer of electrons from 1 to 2 is proportional to the number of filled states at E in material 1 times the number of empty states at E in material 2:

$$\text{rate from 1 to 2} \propto N_1(E)f_1(E) \cdot N_2(E)[1 - f_2(E)] \tag{3–53}$$

where $f(E)$ is the probability of a state being filled at E in each material, i.e., the Fermi–Dirac distribution function given by Eq. (3–10). Similarly,

$$\text{rate from 2 to 1} \propto N_2(E)f_2(E) \cdot N_1(E)[1 - f_1(E)] \tag{3–54}$$

At equilibrium these must be equal:

$$N_1(E)f_1(E) \cdot N_2(E)[1 - f_2(E)] = N_2(E)f_2(E) \cdot N_1(E)[1 - f_1(E)] \tag{3–55}$$

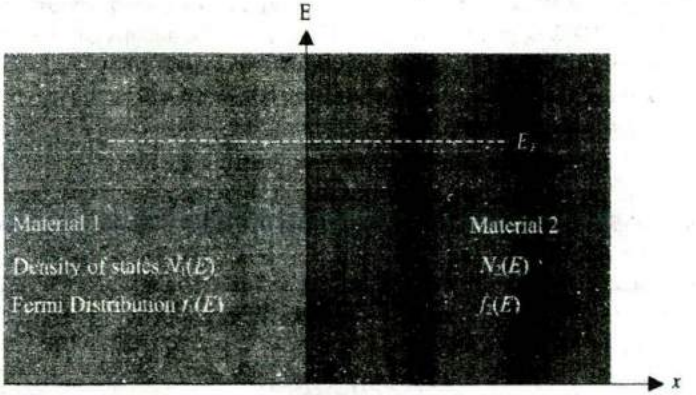


Figure 3–26
Two materials in intimate contact at equilibrium. Since the net motion of electrons is zero, the equilibrium Fermi level must be constant throughout.

Rearranging terms, we have, at energy E,

$$N_1 f_1 N_2 - N_1 f_1 N_2 f_2 = N_2 f_2 N_1 - N_2 f_2 N_1 f_1 \tag{3–56}$$

which results in

$$f_1(E) = f_2(E), \quad \text{that is,} \quad [1 + e^{(E-E_{F1})/kT}]^{-1} = [1 + e^{(E-E_{F2})/kT}]^{-1} \tag{3–57}$$

Therefore, we conclude that $E_{F1} = E_{F2}$. That is, there is no discontinuity in the equilibrium Fermi level. More generally, we can state that the Fermi level at equilibrium must be constant throughout materials in intimate contact. One way of stating this is that no gradient exists in the Fermi level at equilibrium:

$$\boxed{\frac{dE_F}{dx} = 0} \tag{3–58}$$

We will make considerable use of this result in the chapters to follow.

PROBLEMS

3.1 It was mentioned in Section 3.2 that the covalent bonding model gives a false impression of the localization of carriers. As an illustration, calculate the radius of the electron orbit around the donor in Fig. 3–12c, assuming a ground state hydrogen-like orbit in Si. Compare with the Si lattice constant. Use $m_n^* = 0.26m_0$ for Si.

3.2 Calculate values for the Fermi function $f(E)$ at 300 K and plot vs. energy in eV as in Fig. 3–14. Choose $E_F = 1$ eV and make the calculated points closer together near the Fermi level to obtain a smooth curve. Notice that $f(E)$ varies quite rapidly within a few kT of E_F.

3.3 A semiconductor such as Si has a bandstructure about the minimum along [100] described approximately by $E = E_0 - A\cos(\alpha k_x) - B\{\cos(\beta k_y) + \cos(\beta k_z)\}$. What is the density-of-states effective mass associated with the X minimum? [Hint: $\cos(2x) = 1 - 2x^2$ for small x.]

3.4 At room temperature, an unknown, intrinsic, cubic semiconductor has the following bandstructure: there are 6 X minima along the <100> directions. If $m_n^*(\Gamma) = 0.065m_0$, $m_n^*(X) = 0.30m_0$ (for each of the X minima and $m_p^* = 0.47m_0$, at what temperature is the number of electrons in the Γ minima and the X minima equal if the Γ to X energy separation is 0.35 eV, and the bandgap is 1.7 eV (m_0 = free electron mass)?

3.5 Consider n-type GaAs and assume that the total number of conduction electrons, n, is independent of temperature. The density-of-states effective mass, m_L^*, in the L valley is 15 times larger than in the Γ valley. Also the energy separation, E_s, between the Γ and L minima is 0.35 eV, and the mobility in the Γ minimum is 50 times that in L. Calculate and sketch how the conductivity varies from low $T(\ll E_s/k)$ to high $T(\gg E_s/k)$. What is the ratio of the conductivities at 1000°C and 300°C?

ent semiconductors or a semiconductor and a metal. These cases are crucial to the various types of electronic and optoelectronic devices made in semiconductors. In anticipation of those discussions, an important concept should be established here regarding the demands of equilibrium. That concept can be summarized by noting that *no discontinuity or gradient can arise in the equilibrium Fermi level* E_F.

To demonstrate this assertion, let us consider two materials in intimate contact such that electrons can move between the two (Fig. 3–26). These may be, for example, dissimilar semiconductors, n- and p-type regions, a metal and a semiconductor, or simply two adjacent regions of a nonuniformly doped semiconductor. Each material is described by a Fermi–Dirac distribution function and some distribution of available energy states that electrons can occupy.

There is no current, and therefore no net charge transport, at thermal equilibrium. There is also no net transfer of energy. Therefore, for each energy E in Fig. 3–26 any transfer of electrons from material 1 to material 2 must be exactly balanced by the opposite transfer of electrons from 2 to 1. We will let the density of states at energy E in material 1 be called $N_1(E)$ and in material 2 we will call it $N_2(E)$. At energy E the rate of transfer of electrons from 1 to 2 is proportional to the number of filled states at E in material 1 times the number of empty states at E in material 2:

$$\text{rate from 1 to 2} \propto N_1(E)f_1(E) \cdot N_2(E)[1 - f_2(E)] \tag{3–53}$$

where $f(E)$ is the probability of a state being filled at E in each material, i.e., the Fermi–Dirac distribution function given by Eq. (3–10). Similarly,

$$\text{rate from 2 to 1} \propto N_2(E)f_2(E) \cdot N_1(E)[1 - f_1(E)] \tag{3–54}$$

At equilibrium these must be equal:

$$N_1(E)f_1(E) \cdot N_2(E)[1 - f_2(E)] = N_2(E)f_2(E) \cdot N_1(E)[1 - f_1(E)] \tag{3–55}$$

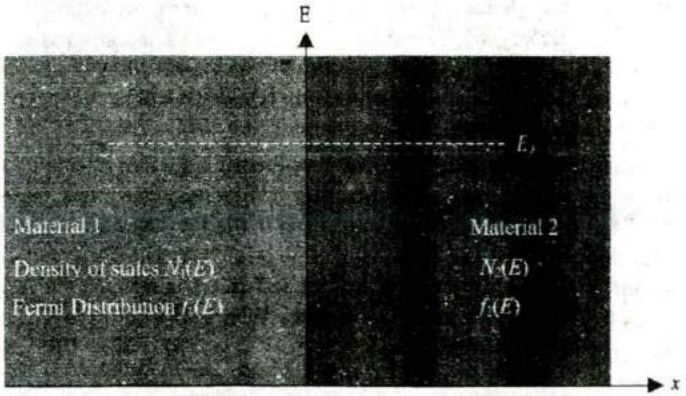


Figure 3–26
Two materials in intimate contact at equilibrium. Since the net motion of electrons is zero, the equilibrium Fermi level must be constant throughout.

Rearranging terms, we have, at energy E,

$$N_1 f_1 N_2 - N_1 f_1 N_2 f_2 = N_2 f_2 N_1 - N_2 f_2 N_1 f_1 \tag{3–56}$$

which results in

$$f_1(E) = f_2(E), \quad \text{that is,} \quad [1 + e^{(E-E_{F1})/kT}]^{-1} = [1 + e^{(E-E_{F2})/kT}]^{-1} \tag{3–57}$$

Therefore, we conclude that $E_{F1} = E_{F2}$. That is, there is no discontinuity in the equilibrium Fermi level. More generally, we can state that the Fermi level at equilibrium must be constant throughout materials in intimate contact. One way of stating this is that no gradient exists in the Fermi level at equilibrium:

$$\boxed{\frac{dE_F}{dx} = 0} \tag{3–58}$$

We will make considerable use of this result in the chapters to follow.

PROBLEMS

3.1 It was mentioned in Section 3.2 that the covalent bonding model gives a false impression of the localization of carriers. As an illustration, calculate the radius of the electron orbit around the donor in Fig. 3–12c, assuming a ground state hydrogen-like orbit in Si. Compare with the Si lattice constant. Use $m_n^* = 0.26m_0$ for Si.

3.2 Calculate values for the Fermi function $f(E)$ at 300 K and plot vs. energy in eV as in Fig. 3–14. Choose $E_F = 1$ eV and make the calculated points closer together near the Fermi level to obtain a smooth curve. Notice that $f(E)$ varies quite rapidly within a few kT of E_F.

3.3 A semiconductor such as Si has a bandstructure about the minimum along [100] described approximately by $E = E_0 - A\,cos(\alpha k_x) - B[cos(\beta k_y) + cos(\beta k_z)]$. What is the density-of-states effective mass associated with the X minimum? [Hint: $cos(2x) = 1 - 2x^2$ for small x.]

3.4 At room temperature, an unknown, intrinsic, cubic semiconductor has the following bandstructure: there are 6 X minima along the <100> directions. If $m_n^*(\Gamma) = 0.065m_0$, $m_n^*(X) = 0.30m_0$ (for each of the X minima and $m_p^* = 0.47m_0$, at what temperature is the number of electrons in the Γ minima and the X minima equal if the Γ to X energy separation is 0.35 eV, and the bandgap is 1.7 eV (m_0 = free electron mass)?

3.5 Consider n-type GaAs and assume that the total number of conduction electrons, n, is independent of temperature. The density-of-states effective mass, m_L^*, in the L valley is 15 times larger than in the Γ valley. Also the energy separation, E_s, between the Γ and L minima is 0.35 eV, and the mobility in the Γ minimum is 50 times that in L. Calculate and sketch how the conductivity varies from low $T(\ll E_s/k)$ to high $T(\gg E_s/k)$. What is the ratio of the conductivities at 1000°C and 300°C?

3.6 Calculate the band gap of Si from Eq. (3–23) and the plot of n_i vs. $1000/T$ (Fig. 3–17). *Hint:* the slope cannot be measured directly from a semilogarithmic plot; read the values from two points on the plot and take the natural logarithm as needed for the solution.

3.7 Show that Eq. (3–25) results from Eqs. (3–15) and (3–19). If $n_0 = 10^{16}$ cm^{-3}, where is the Fermi level relative to E_i in Si at 300 K?

3.8 Derive an expression relating the intrinsic level E_i to the center of the band gap $E_g/2$. Calculate the displacement of E_i from $E_g/2$ for Si at 300 K, assuming the effective mass values for electrons and holes are $1.1m_0$ and $0.56m_0$, respectively.

3.9 (a) Explain why holes are found at the *top* of the valence band, whereas electrons are found at the *bottom* of the conduction band.

(b) Explain why Si doped with 10^{14} cm^{-3} Sb is n-type at 400 K but similarly doped Ge is not.

3.10 A Si sample is doped with 6×10^{15} cm^{-3} donors and 2×10^{15} cm^{-3} acceptors. Find the position of the Fermi level with respect to E_i at 300 K. What is the value and sign of the Hall coefficient?

3.11 (a) Show that the minimum conductivity of a semiconductor sample occurs when $n_0 = n_i\sqrt{\mu_p/\mu_n}$. Hint: begin with Eq. (3–43) and apply Eq. (3–24).

(b) What is the expression for the minimum conductivity σ_{min}?

(c) Calculate σ_{min} for Si at 300 K and compare with the intrinsic conductivity.

3.12 (a) A Si bar 0.1 cm long and 100 μm^2 in cross-sectional area is doped with 10^{17} cm^{-3} phosphorus. Find the current at 300 K with 10V applied. Repeat for a Si bar 1 μm long.

(b) How long does it take an average electron to drift 1 μm in pure Si at an electric field of 100 V/cm? Repeat for 10^5 V/cm.

3.13 A perfect III–V semiconductor (relative dielectric constant = 13) is doped with column VI and column II impurities. Given that $\mu_n = 1000$ cm^2/V-s, $\mu_p = 500$ cm^2/V-s, what energy levels are introduced in the bandgap? (The mean free time = 0.1 ps for electrons and 0.4 ps for holes.)

3.14 In soldering wires to a sample such as that shown in Fig. 3–25, it is difficult to align the Hall probes *A* and *B* precisely. If *B* is displaced slightly down the length of the bar from *A*, an erroneous Hall voltage results. Show that the true Hall voltage V_H can be obtained from two measurements of V_{AB}, with the magnetic field first in the $+z$-direction and then in the $-z$-direction.

3.15 We put 11 electrons in an infinite 1-D potential well of size 100 Å. What is the Fermi level at 0 K? What is the probability of exciting a carrier to the first excited state at $T = 300$ K? Use the free electron mass in this problem.

3.16 Use Eq. (3–45) to calculate and plot the mobility vs. temperature μ(T) from 10 K to 500 K for Si doped with $N_d = 10^{14}$, 10^{16}, and 10^{18} donors cm^{-3}. Consider the mobility to be determined by impurity and phonon (lattice) scattering. Impurity scattering limited mobility can be described by

$$\mu_I = 3.29 \times 10^{15} \frac{\epsilon_r^2 T^{3/2}}{N_d^+ (m_n^*/m_0)^{1/2} \left[\ln(1+z) - \frac{z}{1+z}\right]}$$

where

$$z = 1.3 \times 10^{13} \epsilon_r T^2 (m_n^*/m_0)(N_d^+)^{-1}$$

Assume that the ionized impurity concentration N_d^+ is equal to N_d at all temperatures.

The conductivity effective mass m_n^* for Si is 0.26 m_0. Acoustic phonon (lattice) scattering limited mobility can be described by

$$\mu_{AC} = 1.18 \times 10^{-5} c_1 (m_n^*/m_0)^{-5/2} T^{-3/2} (E_{AC})^{-2}$$

where the stiffness (c_1) is given by

$c_1 = 1.9 \times 10^{12}$ dyne cm^{-2} for Si

and the conduction band acoustic deformation potential (E_{AC}) is

$E_{AC} = 9.5$ eV for Si

3.17 Rework Prob. 3.16 considering carrier freeze-out onto donors at low T. That is, consider

$$N_d^+ = \frac{N_d}{1 + \exp(E_d/kT)}$$

as the ionized impurity concentration. Consider the donor ionization energy (E_d) to be 45 meV for Si.

3.18 Hall measurements are made on a p-type semiconductor bar 500 μm wide and 20 μm thick. The Hall contacts A and B are displaced 2 μm with respect to each other in the direction of current flow of 3 mA. The voltage between A and B with a magnetic field of 10 kG (1kG = 10^{-5} Wb/cm^2) pointing out of the plane of the sample is 3.2 mV. When the magnetic field direction is reversed the voltage changes to −2.8 mV. What is the hole concentration and mobility?

3.19 For a hypothetical semiconductor, we have $\mu_n = \mu_p = 1000$ cm^2/V-s and $N_c = N_v = 10^{19}$ cm^{-3}. If the conductivity of the intrinsic semiconductor at 300 K is 4×10^{-6} (Ω-cm)$^{-1}$, what is the conductivity at 600 K?

3.20 An unknown semiconductor has $E_g = 1.1$ eV and $N_c = N_v$. It is doped with 10^{15} cm^{-3} donors where the donor level is 0.2 eV below E_C. Given that E_F is 0.25 eV below E_c, calculate n_i and the concentration of electrons and holes in the semiconductor at 300 K.

3.21 Referring to Fig. 3.25, consider a semiconductor bar with $w = 0.1$ mm, $t = 10$ μm and $L = 5$ mm. For $\mathfrak{B} = 10$ kG in the direction shown (1 kG = 10^{-5} Wb/cm^2) and a current of 1mA, we have $V_{AB} = -2$ mV, $V_{CD} = 100$ mV. Find the type, concentration and mobility of the majority carrier.

READING LIST

Ashcroft, N. W., and N. D. Mermin. *Solid State Physics.* Philadelphia: W.B. Saunders, 1976.

Blakemore, J. S. *Semiconductor Statistics.* New York: Dover Publications, 1987.

Bube, R. H. *Electrons in Solids,* 3rd ed. Boston: Harcourt Brace Jovanovich, 1992.

Burns, G. *Solid State Physics.* San Diego: Academic Press, 1985.

Capasso, F. "Bandgap and Interface Engineering for Advanced Electronic and Photonic Devices." *Thin Solid Films* 216 (28 August 1992): 59–67.

Capasso, F., and S. Datta. "Quantum Electron Devices." *Physics Today* 43 (February 1990): 74–82.

Drummund, T. J., P. L. Gourley, and T. E. Zipperian. "Quantum-Tailored Solid-State Devices." *IEEE Spectrum* 25 (June 1988): 33–37.

Hummel, R. E. *Electronic Properties of Materials,* 2nd ed. Berlin: Springer-Verlag, 1993.

Kittel, C. *Introduction to Solid State Physics,* 7th ed. New York: Wiley, 1996.

Li, S. S. *Semiconductor Physical Electronics.* New York: Plenum Press, 1993.

Muller, R. S., and T. I. Kamins. *Device Electronics for Integrated Circuits.* New York: Wiley, 1986.

Neamen, D. A. *Semiconductor Physics and Devices: Basic Principles.* Homewood, IL: Irwin, 1992.

Pierret, R. F. *Advanced Semiconductor Fundamentals.* Reading, MA: Addison-Wesley, 1987.

Schubert, E. F. *Doping in III–V Semiconductors.* Cambridge: Cambridge University Press, 1993.

Singh, J. *Physics of Semiconductors and Their Heterostructures.* New York: McGraw-Hill, 1993.

Singh, J. *Semiconductor Devices.* New York: McGraw-Hill, 1994.

Sundaram, M., S. A. Chalmers, and P. F. Hopkins. "New Quantum Structures." *Science* 254 (29 November 1991): 1326–35.

Swaminathan, V., and Macrander, A. T. *Material Aspects of GaAs and InP Based Structures.* Englewood Cliffs, NJ: Prentice Hall, 1991.

Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

Wolfe, C. M., G. E. Stillman, and N. Holonyak, Jr. *Physical Properties of Semiconductors.* Englewood Cliffs, NJ: Prentice Hall, 1989.

Chapter 4

Excess Carriers in Semiconductors

Most semiconductor devices operate by the creation of charge carriers in excess of the thermal equilibrium values. These excess carriers can be created by optical excitation or electron bombardment, or as we shall see in Chapter 5, they can be injected across a forward-biased p-n junction. However the excess carriers arise, they can dominate the conduction processes in the semiconductor material. In this chapter we shall investigate the creation of excess carriers by optical absorption and the resulting properties of photoluminescence and photoconductivity. We shall study more closely the mechanism of electron–hole pair recombination and the effects of carrier trapping. Finally, we shall discuss the diffusion of excess carriers due to a carrier gradient, which serves as a basic mechanism of current conduction along with the mechanism of drift in an electric field.

4.1 OPTICAL ABSORPTION[1]

An important technique for measuring the band gap energy of a semiconductor is the absorption of incident photons by the material. In this experiment photons of selected wavelengths are directed at the sample, and relative transmission of the various photons is observed. Since photons with energies greater than the band gap energy are absorbed while photons with energies less than the band gap are transmitted, this experiment gives an accurate measure of the band gap energy.

It is apparent that a photon with energy $h\nu \geq E_g$ can be absorbed in a semiconductor (Fig. 4–1). Since the valence band contains many electrons and the conduction band has many empty states into which the electrons may be excited, the probability of photon absorption is high. As Fig. 4–1 indicates, an electron excited to the conduction band by optical absorption may initially have more energy than is common for conduction band electrons (almost all electrons are near E_c unless the sample is very heavily doped). Thus the excited electron loses energy to the lattice in scattering events until its velocity reaches the thermal equilibrium velocity of other conduction band elec-

[1]In this context the word "optical" does not necessarily imply that the photons absorbed are in the visible part of the spectrum. Many semiconductors absorb photons in the infrared region, but this is included in the term "optical absorption."

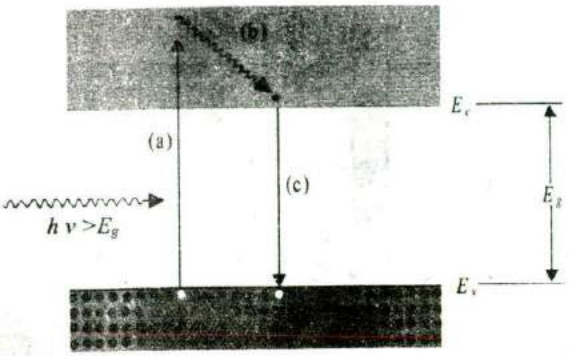


Figure 4–1
Optical absorption of a photon with $h\nu > E_g$: (a) an EHP is created during photon absorption; (b) the excited electron gives up energy to the lattice by scattering events; (c) the electron recombines with a hole in the valence band.

trons. The electron and hole created by this absorption process are *excess carriers;* since they are out of balance with their environment, they must eventually recombine. While the excess carriers exist in their respective bands, however, they are free to contribute to the conductivity of the material.

A photon with energy less than E_g is unable to excite an electron from the valence band to the conduction band. Thus in a pure semiconductor, there is negligible absorption of photons with $h\nu < E_g$. This explains why some materials are transparent in certain wavelength ranges. We are able to "see through" certain insulators, such as a good NaCl crystal, because a large energy gap containing no electron states exists in the material. If the band gap is about 2 eV wide, only long wavelengths (infrared) and the red part of the visible spectrum are transmitted; on the other hand, a band gap of about 3 eV allows infrared and the entire visible spectrum to be transmitted.

If a beam of photons with $h\nu > E_g$ falls on a semiconductor, there will be some predictable amount of absorption, determined by the properties of the material. We would expect the ratio of transmitted to incident light intensity to depend on the photon wavelength and the thickness of the sample. To calculate this dependence, let us assume that a photon beam of intensity $\mathbf{I}_0$ (photons/cm^2-s) is directed at a sample of thickness l (Fig. 4–2). The beam contains only photons of wavelength λ, selected by a monochromator. As the beam passes through the sample, its intensity at a distance x from the surface can be calculated by considering the probability of absorption within any increment dx. Since a photon which has survived to x without absorption has no memory of how far it has traveled, its probability of absorption in any dx is constant. Thus the degradation of the intensity $-d\mathbf{I}(x)/dx$ is proportional to the intensity remaining at x:

$$-\frac{d\mathbf{I}(x)}{dx} = \boldsymbol{\alpha}\mathbf{I}(x) \tag{4–1}$$

The solution to this equation is

$$\mathbf{I}(x) = \mathbf{I}_0 e^{-\boldsymbol{\alpha}x} \tag{4–2}$$

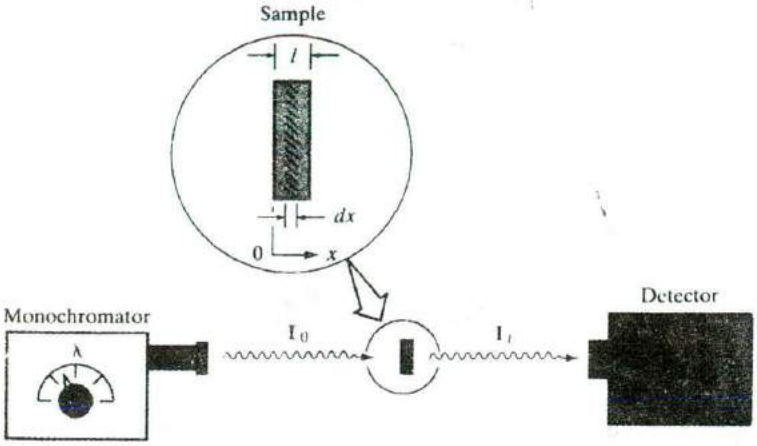


Figure 4–2 Optical absorption experiment.

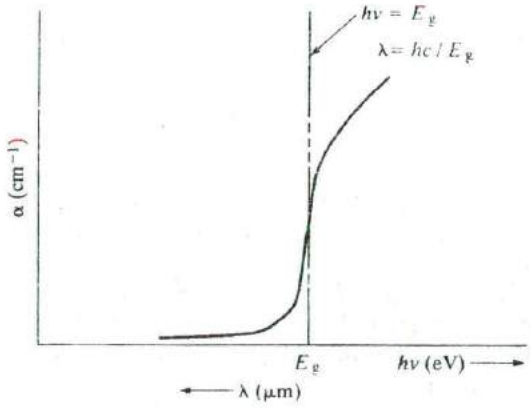


Figure 4–3 Dependence of optical absorption coefficient α for a semiconductor on the wavelength of incident light.

and the intensity of light transmitted through the sample thickness l is

$$I_t = I_0 e^{-\alpha l} \tag{4–3}$$

The coefficient α is called the *absorption coefficient* and has units of cm^{-1}. This coefficient will of course vary with the photon wavelength and with the material. In a typical plot of α vs. wavelength (Fig. 4–3), there is negligible absorption at long wavelengths ($h\nu$ small) and considerable absorption of photons with energies larger than E_g. According to Eq. (2–2), the relation between photon energy and wavelength is $E = hc/\lambda$. If E is given in electron volts and λ in micrometers, this becomes $E = 1.24/\lambda$.

Figure 4–4 indicates the band gap energies of some of the common semiconductors, relative to the visible, infrared, and ultraviolet portions of the spectrum. We observe that GaAs, Si, Ge, and InSb lie outside the vis-

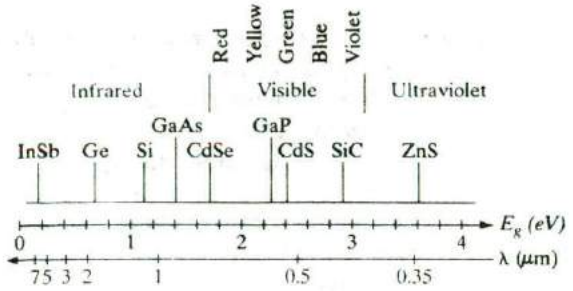


Figure 4–4
Band gaps of some common semiconductors relative to the optical spectrum.

ible region, in the infrared. Other semiconductors, such as GaP and CdS, have band gaps wide enough to pass photons in the visible range. It is important to note here that a semiconductor absorbs photons with energies equal to the band gap, *or larger.* Thus Si absorbs not only band gap light (~1 μm) but also shorter wavelengths, including those in the visible part of the spectrum.

4.2 LUMINESCENCE

When electron–hole pairs are generated in a semiconductor, or when carriers are excited into higher impurity levels from which they fall to their equilibrium states, light can be given off by the material. Many of the semiconductors are well suited for light emission, particularly the compound semiconductors with direct band gaps. The general property of light emission is called *luminescence.*[2] This overall category can be subdivided according to the excitation mechanism: If carriers are excited by photon absorption, the radiation resulting from the recombination of the excited carriers is called *photoluminescence;* if the excited carriers are created by high-energy electron bombardment of the material, the mechanism is called *cathodoluminescence;* if the excitation occurs by the introduction of current into the sample, the resulting luminescence is called *electroluminescence.* Other types of excitation are possible, but these three are the most important for device applications.

4.2.1 Photoluminescence

The simplest example of light emission from a semiconductor occurs for direct excitation and recombination of an EHP, as depicted in Fig. 3–5a. If the recombination occurs directly rather than via a defect level, band gap light is given off in the process. For steady state excitation, the recombination of EHPs occurs at the same rate at the generation, and one photon is emitted

[2]The emission processes considered here should not be confused with radiation due to incandescence which occurs in heated materials. The various luminescent mechanisms can be considered "cold" processes as compared to the "hot" process of incandescence, which increases with temperature. In fact, most luminescent processes become more efficient as the temperature is lowered.

for each photon absorbed. Direct recombination is a fast process; the mean lifetime of the EHP is usually on the order of 10^{-8} s or less. Thus the emission of photons stops within approximately 10^{-8} s after the excitation is turned off. Such fast luminescent processes are often referred to as *fluorescence*. In some materials, however, emission continues for periods up to seconds or minutes after the excitation is removed. These slow processes are called *phosphorescence*, and the materials are called *phosphors*. An example of a slow process is shown in Fig. 4–5. This material contains a defect level (perhaps due to an impurity) in the band gap which has a strong tendency to temporarily capture (*trap*) electrons from the conduction band. The events depicted in the figure are as follows; (a) An incoming photon with $h\nu_1 > E_g$ is absorbed, creating an EHP; (b) the excited electron gives up energy to the lattice by scattering until it nears the bottom of the conduction band; (c) the electron is *trapped* by the impurity level E_t and remains trapped until it can be thermally reexcited to the conduction band (d); (e) finally direct recombination occurs as the electron falls to an empty state in the valence band, giving off a photon ($h\nu_2$) of approximately the band gap energy. The delay time between excitation and recombination can be relatively long if the probability of thermal reexcitation from the trap (d) is small. Even longer delay times result if the electron is retrapped several times before recombination. If the trapping probability is greater than the probability of recombination, an electron may make several trips between the trap and the conduction band before recombination finally occurs. In such material the emission of phosphorescent light persists for a relatively long time after the excitation is removed.

The color of light emitted by a phosphor such as ZnS depends primarily on the impurities present, since many radiative transitions involve impurity levels within the band gap. This selection of colors is particularly useful in the fabrication of a color television screen.

One of the most common examples of photoluminescence is the fluorescent lamp. Typically such a lamp is composed of a glass tube filled with gas

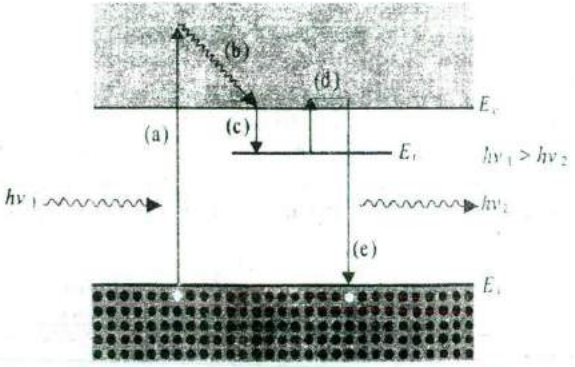


Figure 4–5
Excitation and recombination mechanisms in photoluminescence with a trapping level for electrons.

(e.g., a mixture of argon and mercury), with a fluorescent coating on the inside of the tube. When an electric discharge is induced between electrodes in the tube, the excited atoms of the gas emit photons, largely in the visible and ultra-violet regions of the spectrum. This light is absorbed by the luminescent coating, and the visible photons are emitted. The efficiency of such a lamp is considerably better than that of an incandescent bulb, and the wavelength mixture in light given off can be adjusted by proper selection of the fluorescent material.

EXAMPLE 4–1

A 0.46-μm-thick sample of GaAs is illuminated with monochromatic light of $h\nu = 2$ eV. The absorption coefficient α is 5×10^4 cm^{-1}. The power incident on the sample is 10 mW.

(a) Find the total energy absorbed by the sample per second (J/s).

(b) Find the rate of excess thermal energy given up by the electrons to the lattice before recombination (J/s).

(c) Find the number of photons per second given off from recombination events, assuming perfect quantum efficiency.

SOLUTION

(a) From Eq. (4–3),

$$\mathbf{I}_t = \mathbf{I}_0 e^{-\alpha l} = 10^{-2} \exp(-5 \times 10^4 \times 0.46 \times 10^{-4})$$
$$= 10^{-2} e^{-2.3} = 10^{-3} \text{ W}$$

Thus the absorbed power is

$$10 - 1 = 9 \text{ mW} = 9 \times 10^{-3} \text{ J/s}$$

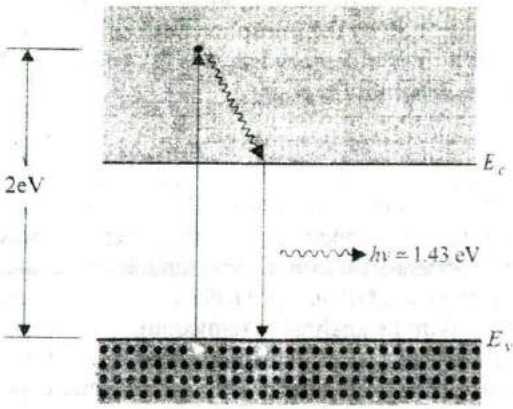


Figure 4–6 Excitation and band-to-band recombination leading to photoluminescence.

(b) The fraction of each photon energy unit which is converted to heat is

$$\frac{2 - 1.43}{2} = 0.285$$

Thus the amount of energy converted to heat per second is

$$0.285 \times 9 \times 10^{-3} = 2.57 \times 10^{-3}\ \text{J/s}$$

(c) Assuming one emitted photon for each photon absorbed (perfect quantum efficiency), we have

$$\frac{9 \times 10^{-3}\ \text{J/s}}{1.6 \times 10^{-19}\ \text{J/eV} \times 2\ \text{eV/photon}} = 2.81 \times 10^{16}\ \text{photons/s}$$

Alternative solution: Recombination radiation accounts for $9 - 2.57 = 6.43$ mW at 1.43 eV/photon.

$$\frac{6.43 \times 10^{-3}}{1.6 \times 10^{-19} \times 1.43} = 2.81 \times 10^{16}\ \text{photons/s}$$

4.2.2 Electroluminescence

There are many ways by which electrical energy can be used to generate photon emission in a solid. In LEDs an electric current causes the injection of minority carriers into regions of the crystal where they can recombine with majority carriers, resulting in the emission of recombination radiation. This important effect (*injection electroluminescence*) will be discussed in Chapter 8 in terms of p-n junction theory.

The first electroluminescent effect to be observed was the emission of photons by certain phosphors in an alternating electric field (the Destriau effect). In this device, a phosphor powder such as ZnS is held in a binder material (often a plastic) of a high dielectric constant. When an a-c electric field is applied, light is given off by the phosphor. Such cells can be useful as lighting panels, although their efficiency has thus far been too low for most applications and their reliability is poor.

4.3 CARRIER LIFETIME AND PHOTOCONDUCTIVITY

When excess electrons and holes are created in a semiconductor, there is a corresponding increase in the conductivity of the sample as indicated by Eq. (3–43). If the excess carriers arise from optical luminescence, the resulting increase in conductivity is called *photoconductivity*. This is an important effect, with useful applications in the analysis of semiconductor materials and in the operation of several types of devices. In this section we shall examine the mechanisms by which excess electrons and holes recombine and apply the recom-

bination kinetics to the analysis of photoconductive devices. However, the importance of recombination is not limited to cases in which the excess carriers are created optically. In fact, virtually every semiconductor device depends in some way on the recombination of excess electrons and holes. Therefore, the concepts developed in this section will be used extensively in the analyses of diodes, transistors, lasers, and other devices in later chapters.

4.3.1 Direct Recombination of Electrons and Holes

It was pointed out in Section 3.1.4 that electrons in the conduction band of a semiconductor may make transitions to the valence band (i.e., recombine with holes in the valence band) either directly or indirectly. In direct recombination, an excess population of electrons and holes decays by electrons falling from the conduction band to empty states (holes) in the valence band. Energy lost by an electron in making the transition is given up as a photon. Direct recombination occurs *spontaneously;* that is, the probability that an electron and a hole will recombine is constant in time. As in the case of carrier scattering, this constant probability leads us to expect an exponential solution for the decay of the excess carriers. In this case the rate of decay of electrons at any time t is proportional to the number of electrons remaining at t and the number of holes, with some constant of proportionality for recombination, α_r. The *net* rate of change in the conduction band electron concentration is the thermal generation rate $\alpha_r n_i^2$ from Eq. (3–7) minus the recombination rate

$$\frac{dn(t)}{dt} = \alpha_r n_i^2 - \alpha_r n(t)p(t) \tag{4–4}$$

Let us assume the excess electron–hole population is created at $t = 0$, for example by a short flash of light, and the initial excess electron and hole concentrations Δn and Δp are equal.[3] Then as the electrons and holes recombine in pairs, the instantaneous concentrations of excess carriers $\delta n(t)$ and $\delta p(t)$ are also equal. Thus we can write the total concentrations of Eq. (4–4) in terms of the equilibrium values n_0 and p_0 and the excess carrier concentrations $\delta n(t) = \delta p(t)$. Using Eq. (3–24) we have

$$\begin{aligned}\frac{d\delta n(t)}{dt} &= \alpha_r n_i^2 - \alpha_r[n_0 + \delta n(t)][p_0 + \delta p(t)] \\ &= -\alpha_r[(n_0 + p_0)\delta n(t) + \delta n^2(t)]\end{aligned} \tag{4–5}$$

This nonlinear equation would be difficult to solve in its present form. Fortunately, it can be simplified for the case of low-level injection. If the excess

[3]We will use $\delta n(t)$ and $\delta p(t)$ to mean instantaneous excess carrier concentrations, and Δn, Δp for their values at $t = 0$. Later we will use similar symbolism for spatial distributions, such as $\delta n(x)$ and $\Delta n(x = 0)$.

carrier concentrations are small, we can neglect the δn^2 term. Furthermore, if the material is extrinsic, we can usually neglect the term representing the equilibrium minority carriers. For example, if the material is p-type ($p_0 \gg n_0$), Eq. (4–5) becomes

$$\frac{d\delta n(t)}{dt} = -\alpha_r p_0 \delta n(t) \tag{4–6}$$

The solution to this equation is an exponential decay from the original excess carrier concentration Δn:

$$\delta n(t) = \Delta n e^{-\alpha_r p_0 t} = \Delta n e^{-t/\tau_n} \tag{4–7}$$

Excess electrons in a p-type semiconductor recombine with a decay constant $\tau_n = (\alpha_r p_0)^{-1}$, called the *recombination lifetime*. Since the calculation is made in terms of the minority carriers, τ_n is often called the *minority carrier lifetime*. The decay of excess holes in n-type material occurs with $\tau_p = (\alpha_r n_0)^{-1}$. In the case of direct recombination, the excess majority carriers decay at exactly the same rate as the minority carriers.

There is a large percentage change in the minority carrier electron concentration in Example 4–2 and a small percentage change in the majority hole concentration. Basically, the approximations of extrinsic material and low-level injection allow us to represent $n(t)$ in Eq. (4–4) by the excess concentration $\delta n(t)$ and $p(t)$ by the equilibrium value p_0. Figure 4–7 indicates that this is a good approximation for the example. A more general expression for the carrier lifetime is

$$\tau_n = \frac{1}{\alpha_r(n_0 + p_0)} \tag{4–8}$$

This expression is valid for n- or p-type material if the injection level is low.

EXAMPLE 4–2

A numerical example may be helpful in visualizing the approximations made in the analysis of direct recombination. Let us assume a sample of GaAs is doped with 10^{15} acceptors/cm^3. The intrinsic carrier concentration of GaAs is approximately 10^6 cm^{-3}; thus the minority electron concentration is $n_0 = n_i^2/p_0 = 10^{-3}$ cm^{-3}. Certainly the approximation of $p_0 \gg n_0$ is valid in this case. Now if 10^{14} EHP/cm^3 are created at $t = 0$, we can calculate the decay of these carriers in time. The approximation of $\delta n \ll p_0$ is reasonable, as Fig. 4–7 indicates. This figure shows the decay in time of the excess populations for a carrier recombination lifetime of $\tau_n = \tau_p = 10^{-8}$ s.

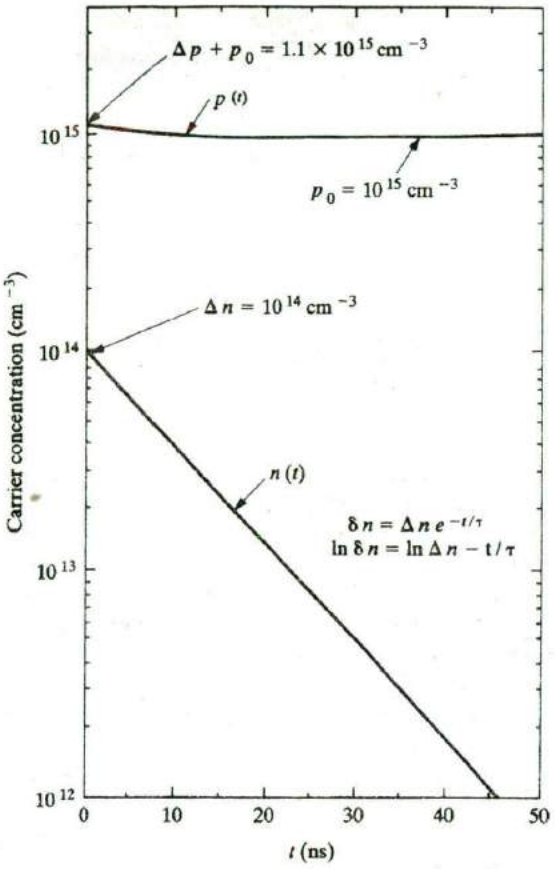


Figure 4–7
Decay of excess electrons and holes by recombination, for $\Delta n = \Delta p = 0.1 p_0$, with n_0 negligible, and $\tau = 10$ ns (Example 4–2). The exponential decay of $\delta n(t)$ is linear on this semilogarithmic graph.

4.3.2 Indirect Recombination; Trapping

In column IV semiconductors and in certain compounds, the probability of direct electron–hole recombination is very small (Appendix III). There is some band gap light given off by materials such as Si and Ge during recombination, but this radiation is very weak and may be detected only by sensitive equipment. The vast majority of the recombination events in indirect materials occur via *recombination levels* within the band gap, and the resulting energy loss by recombining electrons is usually given up to the lattice as heat rather than by the emission of photons. Any impurity or lattice defect can serve as a recombination center if it is capable of receiving a carrier of one type and subsequently capturing the opposite type of carrier, thereby annihilating the pair. For example, Fig. 4–8 illustrates a recombination level E_r which is below E_F at equilibrium and therefore is substantially filled with

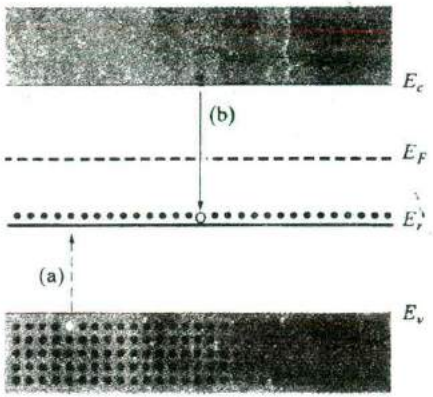


Figure 4–8
Capture processes at a recombination level: (a) hole capture at a filled recombination center; (b) electron capture at an empty center.

electrons. When excess electrons and holes are created in this material, each EHP recombines at E_r in two steps: (a) hole capture and (b) electron capture.

Since the recombination centers in Fig. 4–8 are filled at equilibrium, the first event in the recombination process is hole capture. It is important to note that this event is equivalent to an electron at E_r falling to the valence band, leaving behind an empty state in the recombination level. Thus in hole capture, energy is *given up* as heat to the lattice. Similarly, energy is given up when a conduction band electron subsequently falls to the empty state in E_r. When both of these events have occurred, the recombination center is back to its original state (filled with an electron), but an EHP is missing. Thus one EHP recombination has taken place, and the center is ready to participate in another recombination event by capturing a hole.

The carrier lifetime resulting from indirect recombination is somewhat more complicated than is the case for direct recombination, since it is necessary to account for unequal times required for capturing each type of carrier. In particular, recombination is often delayed by the tendency for a captured carrier to be thermally reexcited to its original band before capture of the opposite type of carrier can occur (Section 4.2.1). For example, if electron capture (b) does not follow immediately after hole capture (a) in Fig. 4–8, the hole may be thermally reexcited to the valence band. Energy is required for this process, which is equivalent to a valence band electron being raised to the empty state in the recombination level. This process delays the recombination, since the hole must be captured again before recombination can be completed.

When a carrier is trapped temporarily at a center and then is reexcited without recombination taking place, the process is often called *temporary trapping*. Although the nomenclature varies somewhat, it is common to refer to an impurity or defect center as a *trapping center* (or simply *trap*) if, after capture of one type of carrier, the most probable next event is reexcitation. If the most probable next event is capture of the opposite type of carrier,

the center is predominately a recombination center. The recombination can be slow or fast, depending on the average time the first carrier is held before the second carrier is captured. In general, trapping levels located deep in the band gap are slower in releasing trapped carriers than are the levels located near one of the bands. This results from the fact that more energy is required, for example, to reexcite a trapped electron from a center near the middle of the gap to the conduction band than is required to reexcite an electron from a level closer to the conduction band.

As an example of impurity levels in semiconductors, Fig. 4–9[4] shows the energy level positions of various impurities in Si. In this diagram a superscript indicates whether the impurity is positive (donor) or negative (acceptor) when ionized. Some impurities introduce multiple levels in the band gap; for example, Zn introduces a level (Zn^-) located 0.31 eV above the valence band and a second level ($Zn^=$) near the middle of the gap. Each Zn impurity atom is capable of accepting two electrons from the semiconductor, one in the lower level and then one in the upper level.

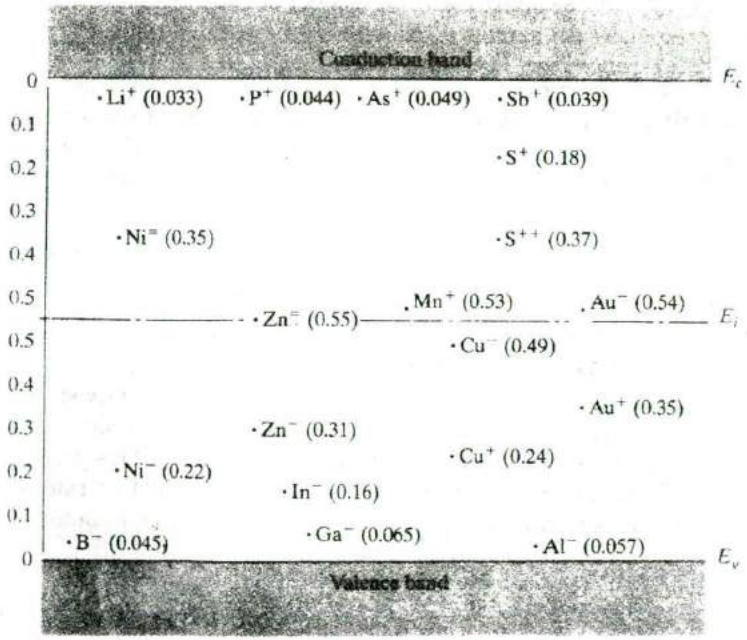


Figure 4–9
Energy levels of impurities in Si. The energies are measured from the nearest band edge (E_v or E_c); donor levels are designated by a plus sign and acceptors by a minus sign.

[4]References: S. M. Sze and J. C. Irvin, "Resistivity, Mobility, and Impurity Levels in GaAs, Ge and Si at 300 K," *Solid State Electronics*, vol. 11, pp. 599–602 (June 1968); E. Schibli and A. G. Milnes, "Deep Impurities in Silicon," *Materials Science and Engineering*, vol. 2, pp. 173–180 (1967).

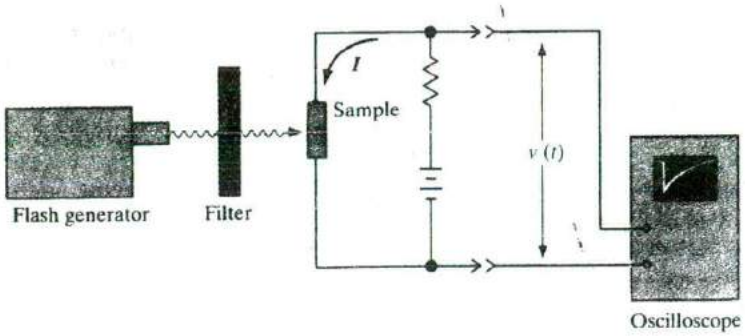


Figure 4–10 Experimental arrangement for photoconductive decay measurements, and a typical oscilloscope trace.

The effects of recombination and trapping can be measured by a *photoconductive decay* experiment. As Fig. 4–7 shows, a population of excess electrons and holes disappears with a decay constant characteristic of the particular recombination process. The conductivity of the sample during the decay is

$$\sigma(t) = q[n(t)\mu_n + p(t)\mu_p] \tag{4–9}$$

Therefore, the time dependence of the carrier concentrations can be monitored by recording the sample resistance as a function of time. A typical experimental arrangement is shown schematically in Fig. 4–10. A source of short pulses of light is required, along with an oscilloscope for displaying the sample voltage as the resistance varies. Microsecond light pulses can be obtained by periodically discharging a capacitor through a flash tube containing a gas such as xenon. For shorter pulses, special techniques such as the use of a pulsed laser must be used.

4.3.3 Steady State Carrier Generation; Quasi-Fermi Levels

In the previous discussion we emphasized the transient decay of an excess EHP population. However, the various recombination mechanisms are also important in a sample at thermal equilibrium or with a steady state EHP generation-recombination balance.[5] For example, a semiconductor at equilibrium experiences thermal generation of EHPs at a rate $g(T) = g_i$ described by Eq. (3–7). This generation is balanced by the recombination rate so that the equilibrium concentrations of carriers n_0 and p_0 are maintained:

$$g(T) = \alpha_r n_i^2 = \alpha_r n_0 p_0 \tag{4–10}$$

This equilibrium rate balance can include generation from defect centers as well as band-to-band generation.

[5]The term *equilibrium* refers to a condition of no external excitation except for temperature, and no net motion of charge (e.g., a sample at a constant temperature, in the dark, with no fields applied). *Steady state* refers to a nonequilibrium condition in which all processes are constant and are balanced by opposing processes (e.g., a sample with a constant current or a constant optical generation of EHPs just balanced by recombination).

If a steady light is shone on the sample, an optical generation rate g_{op} will be added to the thermal generation, and the carrier concentrations n and p will increase to new steady state values. We can write the balance between generation and recombination in terms of the equilibrium carrier concentrations and the departures from equilibrium δn and δp:

$$g(T) + g_{op} = \alpha_r np = \alpha_r(n_0 + \delta n)(p_0 + \delta p) \tag{4–11}$$

For steady state recombination and no trapping, $\delta n = \delta p$; thus Eq. (4–11) becomes

$$g(T) + g_{op} = \alpha_r n_0 p_0 + \alpha_r[(n_0 + p_0)\delta n + \delta n^2] \tag{4–12}$$

The term $\alpha_r n_0 p_0$ is just equal to the thermal generation rate $g(T)$. Thus, neglecting the δn^2 term for low-level excitation, we can rewrite Eq. (4–12) as

$$g_{op} = \alpha_r(n_0 + p_0)\delta n = \frac{\delta n}{\tau_n} \tag{4–13}$$

The excess carrier concentration can be written as

$$\delta n = \delta p = g_{op}\tau_n \tag{4–14}$$

More general expressions are given in Eq. (4–16), which allow for the case $\tau_p \neq \tau_n$, when trapping is present.

EXAMPLE 4–3

As a numerical example, let us assume that 10^{13} EHP/cm^3 are created optically every microsecond in a Si sample with $n_0 = 10^{14}$ cm^{-3} and $\tau_n = \tau_p = 2$ μsec. The steady state excess electron (or hole) concentration is then 2×10^{13} cm^{-3} from Eq. (4–14). While the percentage change in the majority electron concentration is small, the minority carrier concentration changes from

$$p_0 = n_i^2/n_0 = (2.25 \times 10^{20})/10^{14} = 2.25 \times 10^6 \text{ cm}^{-3} \quad (\textit{equilibrium})$$

to

$$p = 2 \times 10^{13} \text{ cm}^{-3} \quad (\textit{steady state})$$

Note that the equilibrium equation $n_0 p_0 = n_i^2$ cannot be used with the subscripts removed; that is, $np \neq n_i^2$ when excess carriers are present.

It is often desirable to refer to the steady state electron and hole concentrations in terms of Fermi levels, which can be included in band diagrams for various devices. The Fermi level E_F used in Eq. (3–25) is meaningful only when no excess carriers are present. However, we can write expressions for the steady state concentrations in the same *form* as the equilibrium expressions by defining separate *quasi-Fermi levels* F_n and F_p for electrons and holes. The resulting carrier concentration equations

$$\boxed{\begin{aligned} n &= n_i e^{(F_n - E_i)/kT} \\ p &= n_i e^{(E_i - F_p)/kT} \end{aligned}} \tag{4–15}$$

can be considered as defining relations for the quasi-Fermi levels.[6]

EXAMPLE 4–4 In Example 4–3, the steady state electron concentration is

$$n = n_0 + \delta n = 1.2 \times 10^{14} = (1.5 \times 10^{10}) e^{(F_n - E_i)/0.0259}$$

where $kT = 0.0259$ eV at room temperature. Thus the electron quasi-Fermi level position $F_n - E_i$ is found from

$$F_n - E_i = 0.0259 \ln(8 \times 10^3) = 0.233 \text{ eV}$$

and F_n lies 0.233 eV above the intrinsic level. By a similar calculation, the hole quasi-Fermi level lies 0.186 eV below E_i (Fig. 4–11). In this example, the equilibrium Fermi level is $0.0259 \ln(6.67 \times 10^3) = 0.228$ eV above the intrinsic level.

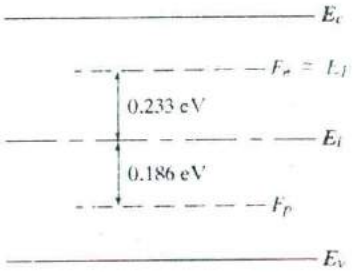


Figure 4–11 Quasi-Fermi levels F_n and F_p for a Si sample with $n_0 = 10^{14}$ cm^{-3}, $\tau_p = 2$ μs, and $g_{op} = 10^{19}$ EHP/cm^3-s (Example 4–4).

The quasi-Fermi levels of Fig. 4–11 illustrate dramatically the deviation from equilibrium caused by the optical excitation; the steady state F_n is only slightly above the equilibrium E_F, whereas F_p is greatly displaced below E_F. From the figure it is obvious that the excitation causes a large percentage change in minority carrier hole concentration and a relatively small change in the electron concentration.

In summary, the quasi-Fermi levels F_n and F_p are the steady state analogues of the equilibrium Fermi level E_F. When excess carriers are present, the deviations of F_n and F_p from E_F indicate how far the electron and hole populations are from the equilibrium values n_0 and p_0. A given concentration of excess EHPs causes a large shift in the minority carrier quasi-Fermi level

[6] In some texts the quasi-Fermi level is called *IMREF*, which is Fermi spelled backward.

compared with that for the majority carriers. The separation of the quasi-Fermi levels $F_n - F_p$ is a direct measure of the deviation from equilibrium (at equilibrium $F_n = F_p = E_F$). The concept of quasi-Fermi levels is very useful in visualizing minority and majority carrier concentrations in devices where these quantities vary with position.

4.3.4 Photoconductive Devices

There are a number of applications for devices which change their resistance when exposed to light. For example, such light detectors can be used in the home to control automatic night lights which turn on at dusk and turn off at dawn. They can also be used to measure illumination levels, as in exposure meters for cameras. Many systems include a light beam aimed at the photoconductor, which signals the presence of an object between the source and detector. Such systems are useful in moving-object counters, burglar alarms, and many other applications. Detectors are used in optical signaling systems in which information is transmitted by a light beam and is received at a photoconductive cell.

Considerations in choosing a photoconductor for a given application include the sensitive wavelength range, time response, and optical sensitivity of the material. In general, semiconductors are most sensitive to photons with energies equal to the band gap or slightly more energetic than band gap. Less energetic photons are not absorbed, and photons with $h\nu \gg E_g$ are absorbed at the surface and contribute little to the bulk conductivity. Therefore, the table of band gaps (Appendix III) indicates the photon energies to which most semiconductor photodetectors respond. For example, CdS ($E_g = 2.42$ eV) is commonly used as a photoconductor in the visible range, and narrow-gap materials such as Ge (0.67 eV) and InSb (0.18 eV) are useful in the infrared portion of the spectrum. Some photoconductors respond to excitations of carriers from impurity levels within the band gap and therefore are sensitive to photons of less than band gap energy.

The optical sensitivity of a photoconductor can be evaluated by examining the steady state excess carrier concentrations generated by an optical generation rate g_{op}. If the mean time each carrier spends in its respective band before capture is τ_n and τ_p, we have

$$\delta n = \tau_n g_{op} \quad \text{and} \quad \delta p = \tau_p g_{op} \tag{4-16}$$

and the photoconductivity change is

$$\Delta\sigma = q g_{op}(\tau_n \mu_n + \tau_p \mu_p) \tag{4-17}$$

For simple recombination, τ_n and τ_p will be equal. If trapping is present, however, one of the carriers may spend little time in its band before being trapped. From Eq. (4–17) it is obvious that for maximum photoconductive response, we want high mobilities and long lifetimes. Some semiconductors are especially good candidates for photoconductive devices

because of their high mobility; for example, InSb has an electron mobility of about 10^5 $cm^2/V\text{-}s$ and therefore is used as a sensitive infrared detector in many applications.

The time response of a photoconductive cell is limited by the recombination times, the degree of carrier trapping, and the time required for carriers to drift through the device in an electric field. Often these properties can be adjusted by proper choice of material and device geometry, but in some cases improvements in response time are made at the expense of sensitivity. For example, the drift time can be reduced by making the device short, but this substantially reduces the responsive area of the device. In addition, it is often desirable that the device have a large dark resistance, and for this reason, shortening the length may not be practical. There is usually a compromise between sensitivity, response time, dark resistance, and other requirements in choosing a device for a particular application.

4.4 DIFFUSION OF CARRIERS

When excess carriers are created nonuniformly in a semiconductor, the electron and hole concentrations vary with position in the sample. Any such spatial variation (*gradient*) in n and p calls for a net motion of the carriers from regions of high carrier concentration to regions of low carrier concentration. This type of motion is called *diffusion* and represents an important charge transport process in semiconductors. The two basic processes of current conduction are diffusion due to a carrier gradient and drift in an electric field.

4.4.1 Diffusion Processes

When a bottle of perfume is opened in one corner of a closed room, the scent is soon detected throughout the room. If there is no convection or other net motion of air, the scent spreads by diffusion. The diffusion is the natural result of the *random motion* of the individual molecules. Consider, for example, a volume of arbitrary shape with scented air molecules inside and unscented molecules outside the volume. All the molecules undergo random thermal motion and collisions with other molecules. Thus each molecule moves in an arbitrary direction until it collides with another air molecule, after which it moves in a new direction. If the motion is truly random, a molecule at the edge of the volume has equal probabilities of moving into or out of the volume on its next step (assuming the curvature of the surface is negligible on the molecular scale). Therefore, after a mean free time $\bar{t}$, half the molecules at the edge will have moved into the volume and half will have moved out of the volume. The net effect is that the volume containing scented molecules has increased. This process will continue until the molecules are uniformly distributed in the room. Only then will a given volume gain as many molecules as it loses in a given time. In other words, net diffusion will continue as long as gradients exist in the distribution of scented molecules.

Carriers in a semiconductor diffuse in a carrier gradient by random thermal motion and scattering from the lattice and impurities. For example, a pulse of excess electrons injected at $x = 0$ at time $t = 0$ will spread out in time as shown in Fig. 4–12. Initially, the excess electrons are concentrated at $x = 0$; as time passes, however, electrons diffuse to regions of low electron concentration until finally $n(x)$ is constant.

We can calculate the rate at which the electrons diffuse in a one-dimensional problem by considering an arbitrary distribution $n(x)$ such as Fig. 4–13a. Since the mean free path $\bar{l}$ between collisions is a small incremental distance, we can divide x into segments $\bar{l}$ wide, with $n(x)$ evaluated at the center of each segment (Fig. 4–13b).

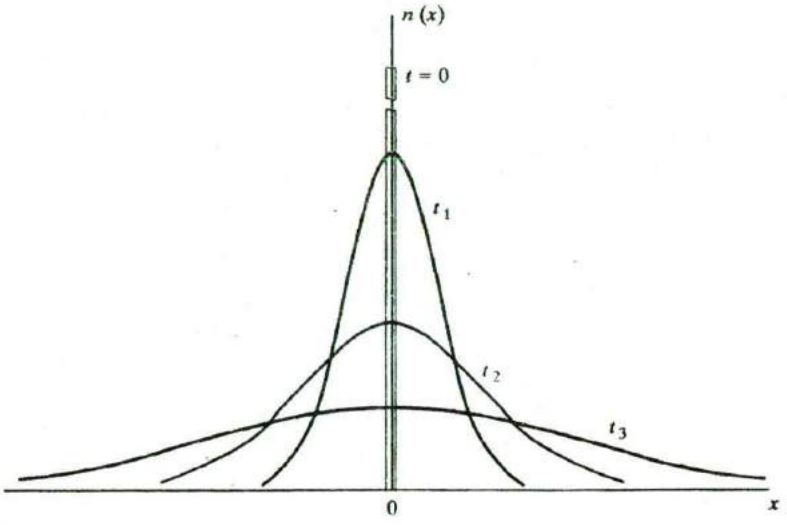


Figure 4–12 Spreading of a pulse of electrons by diffusion.

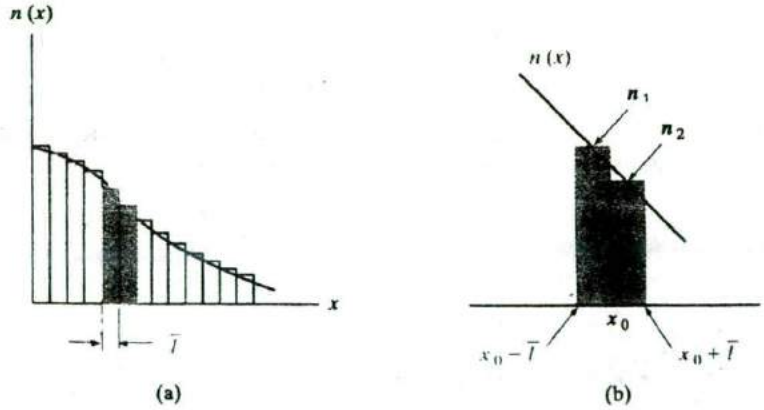


Figure 4–13 An arbitrary electron concentration gradient in one dimension: (a) division of $n(x)$ into segments of length equal to a mean free path for the electrons; (b) expanded view of two of the segments centered at x_0.

The electrons in segment (1) to the left of x_0 in Fig. 4–13b have equal chances of moving left or right, and in a mean free time $\bar{t}$ one-half of them will move into segment (2). The same is true of electrons within one mean free path of x_0 to the right; one-half of these electrons will move through x_0 from right to left in a mean free time. Therefore, the *net* number of electrons passing x_0 from left to right in one mean free time is $\frac{1}{2}(n_1\bar{l}A) - \frac{1}{2}(n_2\bar{l}A)$, where the area perpendicular to x is A. The rate of electron flow in the $+x$-direction per unit area (the electron flux density ϕ_n) is given by

$$\phi_n(x_0) = \frac{\bar{l}}{2\bar{t}}(n_1 - n_2) \tag{4–18}$$

Since the mean free path $\bar{l}$ is a small differential length, the difference in electron concentration $(n_1 - n_2)$ can be written as

$$n_1 - n_2 = \frac{n(x) - n(x + \Delta x)}{\Delta x}\bar{l} \tag{4–19}$$

where x is taken at the center of segment (1) and $\Delta x = \bar{l}$. In the limit of small Δx (i.e., small mean free path $\bar{l}$ between scattering collisions), Eq. (4–18) can be written in terms of the carrier gradient $dn(x)/dx$:

$$\phi_n(x) = \frac{\bar{l}^2}{2\bar{t}} \lim_{\Delta x \to 0} \frac{n(x) - n(x + \Delta x)}{\Delta x} = \frac{-\bar{l}^2}{2\bar{t}} \frac{dn(x)}{dx} \tag{4–20}$$

The quantity $\bar{l}^2/2\bar{t}$ is called the *electron diffusion coefficient*[7] D_n, with units cm^2/s. The minus sign in Eq. (4–20) arises from the definition of the derivative; it simply indicates that the net motion of electrons due to diffusion is in the direction of *decreasing* electron concentration. This is the result we expect, since net diffusion occurs from regions of high particle concentration to regions of low particle concentration. By identical arguments, we can show that holes in a hole concentration gradient move with a diffusion coefficient D_p. Thus

$$\phi_n(x) = -D_n \frac{dn(x)}{dx} \tag{4–21a}$$

$$\phi_p(x) = -D_p \frac{dp(x)}{dx} \tag{4–21b}$$

The diffusion current crossing a unit area (the current density) is the particle flux density multiplied by the charge of the carrier:

$$J_n(\text{diff.}) = -(-q)D_n \frac{dn(x)}{dx} = +qD_n \frac{dn(x)}{dx} \tag{4–22a}$$

[7] If motion in three dimensions were included, the diffusion would be smaller in the x-direction. Actually, the diffusion coefficient should be calculated from the true energy distributions and scattering mechanisms. Diffusion coefficients are usually determined experimentally for a particular material, as described in Section 4.4.5.

$$J_p(\text{diff.}) = -(+q)D_p\frac{dp(x)}{dx} = -qD_p\frac{dp(x)}{dx} \tag{4–22b}$$

It is important to note that electrons and holes move together in a carrier gradient [Eqs. (4–21)], but the resulting currents are in opposite directions [Eqs. (4–22)] because of the opposite charge of electrons and holes.

4.4.2 Diffusion and Drift of Carriers; Built-in Fields

If an electric field is present in addition to the carrier gradient, the current densities will each have a drift component and a diffusion component

$$J_n(x) = q\mu_n n(x)\mathcal{E}(x) + qD_n\frac{dn(x)}{dx} \tag{4–23a}$$

drift diffusion

$$J_p(x) = q\mu_p p(x)\mathcal{E}(x) - qD_p\frac{dp(x)}{dx} \tag{4–23b}$$

and the total current density is the sum of the contributions due to electrons and holes:

$$J(x) = J_n(x) + J_p(x) \tag{4–24}$$

We can best visualize the relation between the particle flow and the current of Eqs. (4–23) by considering a diagram such as shown in Fig. 4–14. In this figure an electric field is assumed to be in the x-direction, along with carrier distributions $n(x)$ and $p(x)$ which decrease with increasing x. Thus the derivatives in Eqs. (4–21) are negative, and diffusion takes place in the $+x$-direction. The resulting electron and hole diffusion currents [J_n (diff.) and J_p (diff.)] are in opposite directions, according to Eqs. (4–22). Holes drift in the direction of the electric field [ϕ_p (drift)], whereas electrons drift in the opposite direction because of their negative charge. The resulting drift current is in the $+x$-direction in each case. Note that the drift and diffusion components of the current are additive for holes when the field is in the direction of decreasing hole concentration, whereas the two components are subtractive for electrons under similar conditions. The total current may be due primarily to the flow of electrons or

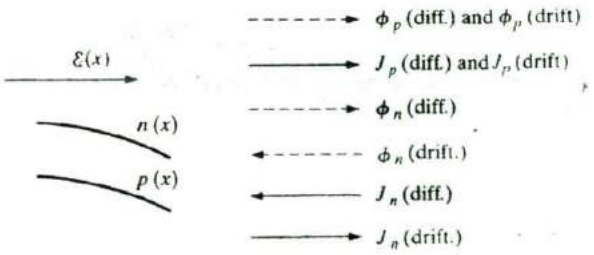


Figure 4–14
Drift and diffusion directions for electrons and holes in a carrier gradient and an electric field. Particle flow directions are indicated by dashed arrows, and the resulting currents are indicated by solid arrows.

holes, depending on the relative concentrations and the relative magnitudes and directions of electric field and carrier gradients.

An important result of Eqs. (4–23) is that minority carriers can contribute significantly to the current through diffusion. Since the drift terms are proportional to carrier concentration, minority carriers seldom provide much drift current. On the other hand, diffusion current is proportional to the *gradient* of concentration. For example, in n-type material the minority hole concentration p may be many orders of magnitude smaller than the electron concentration n, but the gradient dp/dx may be significant. As a result, minority carrier currents through diffusion can sometimes be as large as majority carrier currents.

In discussing the motion of carriers in an electric field, we should indicate the influence of the field on the energies of electrons in the band diagrams. Assuming an electric field $\mathscr{E}(x)$ in the x-direction, we can draw the energy bands as in Fig. 4–15, to include the change in potential energy of electrons in the field. Since electrons drift in a direction opposite to the field, we expect the potential energy for electrons to increase in the direction of the field, as in Fig. 4–15. The electrostatic potential $\mathscr{V}(x)$ varies in the opposite direction, since it is defined in terms of positive charges and is therefore related to the electron potential energy $E(x)$ displayed in the figure by $\mathscr{V}(x) = E(x)/(-q)$.

From the definition of electric field,

$$\mathscr{E}(x) = -\frac{d\mathscr{V}(x)}{dx} \tag{4–25}$$

we can relate $\mathscr{E}(x)$ to the electron potential energy in the band diagram by choosing some reference in the band for the electrostatic potential. We are interested only in the spatial variation $\mathscr{V}(x)$ for Eq. (4–25). Choosing E_i as a convenient reference, we can relate the electric field to this reference by

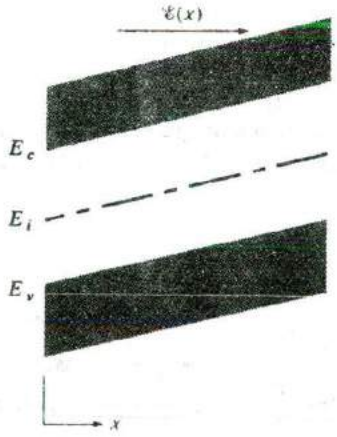


Figure 4–15
Energy band diagram of a semiconductor in an electric field $\mathscr{E}(x)$.

$$\mathscr{E}(x) = -\frac{dV(x)}{dx} = -\frac{d}{dx}\left[\frac{E_i}{(-q)}\right] = \frac{1}{q}\frac{dE_i}{dx} \tag{4–26}$$

Therefore, the variation of band energies with $\mathscr{E}(x)$ as drawn in Fig. 4–15 is correct. The direction of the slope in the bands relative to $\mathscr{E}$ is simple to remember: Since the diagram indicates electron energies, we know the slope in the bands must be such that electrons drift "downhill" in the field. Therefore, $\mathscr{E}$ points "uphill" in the band diagram.

At equilibrium, no net current flows in a semiconductor. Thus any fluctuation which would begin a diffusion current also sets up an electric field which redistributes carriers by drift. An examination of the requirements for equilibrium indicates that the diffusion coefficient and mobility must be related. Setting Eq. (4–23b) equal to zero for equilibrium, we have

$$\mathscr{E}(x) = \frac{D_p}{\mu_p}\frac{1}{p(x)}\frac{dp(x)}{dx} \tag{4–27}$$

Using Eq. (3–25b) for $p(x)$,

$$\mathscr{E}(x) = \frac{D_p}{\mu_p}\frac{1}{kT}\left(\frac{dE_i}{dx} - \frac{dE_F}{dx}\right) \tag{4–28}$$

The equilibrium Fermi level does not vary with x, and the derivative of E_i is given by Eq. (4–26). Thus Eq. (4–28) reduces to

$$\frac{D}{\mu} = \frac{kT}{q} \tag{4–29}$$

This result is obtained for either carrier type. This important equation is called the *Einstein relation.* It allows us to calculate either D or μ from a measurement of the other. Table 4–1 lists typical values of D and μ for several semiconductors at room temperature. It is clear from these values that $D/\mu \simeq 0.026$ V.

An important result of the balance of drift and diffusion at equilibrium is that *built-in* fields accompany gradients in E_i [see Eq. (4–26)]. Such gradients in the bands at equilibrium (E_F constant) can arise when the band gap varies due to changes in alloy composition. More commonly, built-in fields result from doping gradients. For example, a donor distribution $N_d(x)$

Table 4–1 Diffusion coefficient and mobility of electrons and holes for intrinsic semiconductors at 300 K. *Note:* Use Fig. 3–23 for doped semiconductors.

	D_n (cm²/s)	D_p (cm²/s)	μ_n (cm²/V-s)	μ_p (cm²/V-s)
Ge	100	50	3900	1900
Si	35	12.5	1350	480
GaAs	220	10	8500	400

causes a gradient in $n_0(x)$, which must be balanced by a built-in electric field $\mathscr{E}(x)$.

EXAMPLE 4–5

An intrinsic Si sample is doped with donors from one side such that $N_d = N_0 \exp(-ax)$. (a) Find an expression for $\mathscr{E}(x)$ at equilibrium over the range for which $N_d \gg n_i$. (b) Evaluate $\mathscr{E}(x)$ when $a = 1(\mu\text{m})^{-1}$. (c) Sketch a band diagram such as in Fig. 4–15 and indicate the direction of $\mathscr{E}$.

SOLUTION

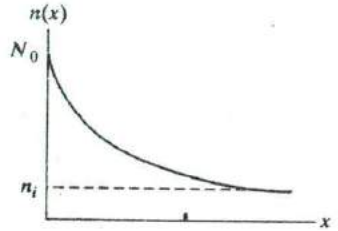
(a) From Eq. (4–23a),

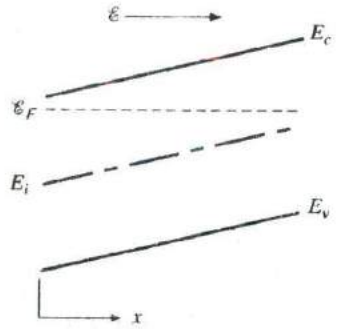
$$\mathscr{E}(x) = -\frac{D_n}{\mu_n}\frac{dn/dx}{n} = -\frac{kT}{q}\frac{N_0(-a)e^{-ax}}{N_0 e^{-ax}} = +\frac{kT}{q}a$$

We notice for this exponential impurity distribution, $\mathscr{E}(x)$ depends on a but not on N_0 or x.

(b) $\mathscr{E}(x) = 0.0259(10^4) = 259$ V/cm

(c)





4.4.3 Diffusion and Recombination; The Continuity Equation

In the discussion of diffusion of excess carriers, we have thus far neglected the important effects of recombination. These effects must be included in a description of conduction processes, however, since recombination can cause a variation in the carrier distribution. For example, consider a differential length Δx of a semiconductor sample with area A in the yz-plane (Fig. 4–16). The hole current density leaving the volume, $J_p(x + \Delta x)$, can be larger or smaller than the current density entering, $J_p(x)$, depending on the generation and recombination of carriers taking place within the volume. The net increase in hole concentration per unit time, $\partial p/\partial t$, is the difference between the hole flux per unit volume entering and leaving, minus the recombination rate. We can convert hole current density to hole particle flux density by dividing J_p by q. The current densities are already expressed per unit area; thus dividing $J_p(x)/q$

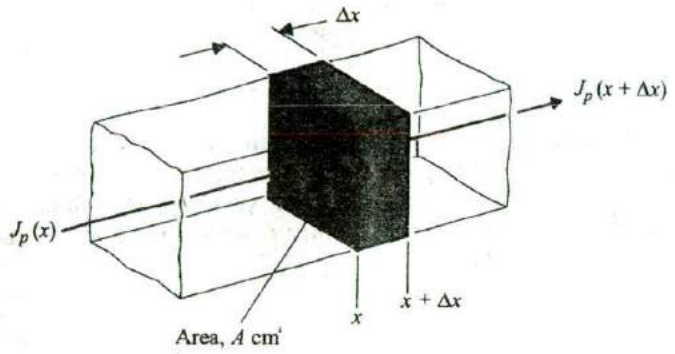


Figure 4–16 Current entering and leaving a volume ΔxA.

by Δx gives the number of carriers per unit volume entering ΔxA per unit time, and $(1/q)J_p(x + \Delta x)/\Delta x$ is the number leaving per unit volume and time:

$$\left.\frac{\partial p}{\partial t}\right|_{x \to x + \Delta x} = \frac{1}{q}\frac{J_p(x) - J_p(x + \Delta x)}{\Delta x} - \frac{\delta p}{\tau_p} \tag{4–30}$$

$$\text{Rate of hole buildup} = \text{increase of hole concentration in } \Delta xA \text{ per unit time} - \text{recombination rate}$$

As Δx approaches zero, we can write the current change in derivative form:

$$\frac{\partial p(x, t)}{\partial t} = \frac{\partial \delta p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} - \frac{\delta p}{\tau_p} \tag{4–31a}$$

The expression (4–31a) is called the *continuity equation* for holes. For electrons we can write

$$\frac{\partial \delta n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - \frac{\delta n}{\tau_n} \tag{4–31b}$$

since the electronic charge is negative.

When the current is carried strictly by diffusion (negligible drift), we can replace the currents in Eqs. (4–31) by the expressions for diffusion current; for example, for electron diffusion we have

$$J_n(\text{diff.}) = qD_n\frac{\partial \delta n}{\partial x} \tag{4–32}$$

Substituting this into Eq. (4–31b) we obtain the *diffusion equation* for electrons,

$$\boxed{\frac{\partial \delta n}{\partial t} = D_n\frac{\partial^2 \delta n}{\partial x^2} - \frac{\delta n}{\tau_n}} \tag{4–33a}$$

and similarly for holes,

$$\boxed{\frac{\partial \delta p}{\partial t} = D_p \frac{\partial^2 \delta p}{\partial x^2} - \frac{\delta p}{\tau_n}} \tag{4–33b}$$

These equations are useful in solving transient problems of diffusion with recombination. For example, a pulse of electrons in a semiconductor (Fig. 4–12) spreads out by diffusion and disappears by recombination. To solve for the electron distribution in time, $n(x,t)$, we would begin with the diffusion equation, Eq. (4–33a).

4.4.4 Steady State Carrier Injection; Diffusion Length

In many problems a steady state distribution of excess carriers is maintained, such that the time derivatives in Eqs. (4–33) are zero. In the steady state case the diffusion equations become

$$\boxed{\frac{d^2 \delta n}{dx^2} = \frac{\delta n}{D_n \tau_n} \equiv \frac{\delta n}{L_n^2}} \tag{4–34a}$$

$$\boxed{\frac{d^2 \delta p}{dx^2} = \frac{\delta p}{D_p \tau_p} \equiv \frac{\delta p}{L_p^2}} \tag{4–34b}$$

(steady state)

where $L_n \equiv \sqrt{D_n \tau_n}$ is called the electron *diffusion length* and L_p is the diffusion length for holes. We no longer need partial derivatives, since the time variation is zero for steady state.

The physical significance of the diffusion length can be understood best by an example. Let us assume that excess holes are somehow injected into a semi-infinite semiconductor bar at $x = 0$, and the steady state hole injection maintains a constant excess hole concentration at the injection point $\delta p(x = 0) = \Delta p$. The injected holes diffuse along the bar, recombining with a characteristic lifetime τ_p. In steady state we expect the distribution of excess holes to decay to zero for large values of x, because of the recombination (Fig. 4–17). For this problem we use the steady state diffusion equation for holes, Eq. (4–34b). The solution to this equation has the form

$$\delta p(x) = C_1 e^{x/L_p} + C_2 e^{-x/L_p} \tag{4–35}$$

We can evaluate C_1 and C_2 from the boundary conditions. Since recombination must reduce $\delta p(x)$ to zero for large values of x, $\delta p = 0$ at $x = \infty$ and therefore $C_1 = 0$. Similarly, the condition $\delta p = \Delta p$ at $x = 0$ gives $C_2 = \Delta p$, and the solution is

$$\boxed{\delta p(x) = \Delta p e^{-x/L_p}} \tag{4–36}$$

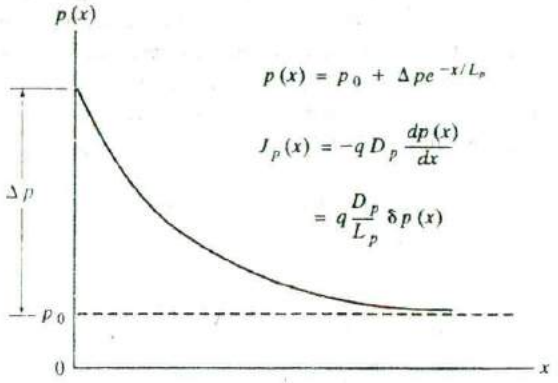


Figure 4–17 Injection of holes at $x = 0$, giving a steady state hole distribution $p(x)$ and a resulting diffusion current density $J_p(x)$.

The injected excess hole concentration dies out exponentially in x due to recombination, and the diffusion length L_p represents the distance at which the excess hole distribution is reduced to $1/e$ of its value at the point of injection. We can show that *L_p is the average distance a hole diffuses before recombining.* To calculate an average diffusion length, we must obtain an expression for the probability that an injected hole recombines in a particular interval dx. The probability that a hole injected at $x = 0$ *survives* to x without recombination is $\delta p(x)/\Delta p = \exp(-x/L_p)$, the ratio of the steady state concentrations at x and 0. On the other hand, the probability that a hole at x will *recombine* in the subsequent interval dx is

$$\frac{\delta p(x) - \delta p(x + dx)}{\delta p(x)} = \frac{-(d\delta p(x)/dx)dx}{\delta p(x)} = \frac{1}{L_p}\,dx \qquad (4\text{–}37)$$

Thus the total probability that a hole injected at $x = 0$ will recombine in a given dx is the product of the two probabilities:

$$(e^{-x/L_p})\left(\frac{1}{L_p}\,dx\right) = \frac{1}{L_p}e^{-x/L_p}\,dx \qquad (4\text{–}38)$$

Then, using the usual averaging techniques described by Eq. (2–21), the average distance a hole diffuses before recombining is

$$\langle x \rangle = \int_0^\infty x\,\frac{e^{-x/L_p}}{L_p}\,dx = L_p \qquad (4\text{–}39)$$

The steady state distribution of excess holes causes diffusion, and therefore a hole current, in the direction of decreasing concentration. From Eqs. (4–22b) and (4–36) we have

$$J_p(x) = -qD_p\frac{dp}{dx} = -qD_p\frac{d\delta p}{dx} = q\frac{D_p}{L_p}\Delta p e^{-x/L_p} = q\frac{D_p}{L_p}\delta p(x) \qquad (4\text{–}40)$$

Since $p(x) = p_0 + \delta p(x)$, the space derivative involves only the excess concentration. We notice that since $\delta p(x)$ is proportional to its derivative for an exponential distribution, the diffusion current at any x is just proportional to the excess concentration δp at that position.

Although this example seems rather restricted, its usefulness will become apparent in Chapter 5 in the discussion of p-n junctions. The injection of minority carriers across a junction often leads to exponential distributions as in Eq. (4–36), with the resulting diffusion current of Eq. (4–40).

4.4.5 The Haynes-Shockley Experiment

One of the classic semiconductor experiments is the demonstration of drift and diffusion of minority carriers, first performed by J.R Haynes and W. Shockley in 1951 at the Bell Telephone Laboratories. The experiment allows independent measurement of the minority carrier mobility μ and diffusion coefficient D. The basic principles of the Haynes–Shockley experiment are as follows: A pulse of holes is created in an n-type bar (for example) that contains an electric field (Fig. 4–18); as the pulse drifts in the field and spreads out by diffusion, the excess hole concentration is monitored at some point down the bar; the time required for the holes to drift a given distance in the field gives a measure of the mobility; and the spreading of the pulse during a given time is used to calculate the diffusion coefficient.

In Fig. 4–18 a pulse of excess carriers is created by a light flash at some point $x = 0$ in an n-type semiconductor ($n_0 \gg p_0$). We assume that the excess carriers have a negligible effect on the electron concentration but change the hole concentration significantly. The excess holes drift in the direction of the electric field and eventually reach the point $x = L$, where they are

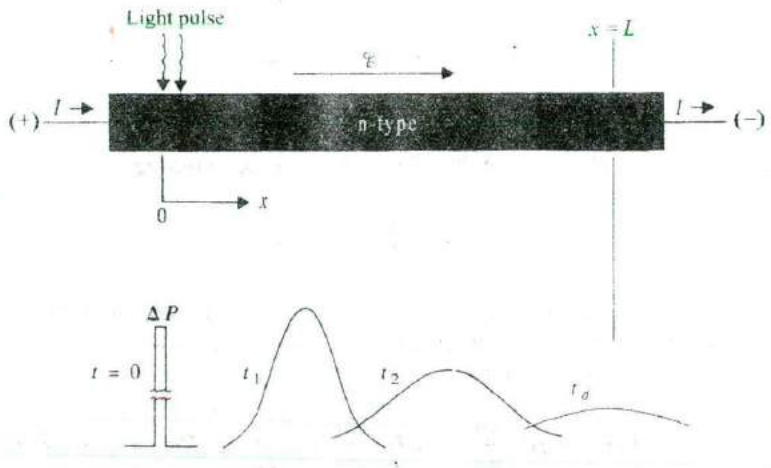


Figure 4–18 Drift and diffusion of a hole pulse in an n-type bar: (a) sample geometry; (b) position and shape of the pulse for several times during its drift down the bar.

monitored. By measuring the drift time t_d, we can calculate the drift velocity v_d and, therefore, the hole mobility:

$$v_d = \frac{L}{t_d} \tag{4–41}$$

$$\mu_p = \frac{v_d}{\mathscr{E}} \tag{4–42}$$

Thus the hole mobility can be calculated directly from a measurement of the drift time for the pulse as it moves down the bar. In contrast with the Hall effect (Section 3.4.5), which can be used with resistivity to obtain the *majority* carrier mobility, the Haynes–Shockley experiment is used to measure the *minority* carrier mobility.

As the pulse drifts in the $\mathscr{E}$ field it also spreads out by diffusion. By measuring the spread in the pulse, we can calculate D_p. To predict the distribution of holes in the pulse as a function of time, let us first reexamine the case of diffusion of a pulse *without drift, neglecting recombination* (Fig. 4–12). The equation which the hole distribution must satisfy is the time-dependent diffusion equation, Eq. (4–33b). For the case of negligible recombination (τ_p long compared with the times involved in the diffusion), we can write the diffusion equation as

$$\frac{\partial \delta p(x,t)}{\partial t} = D_p \frac{\partial^2 \delta p(x,t)}{\partial x^2} \tag{4–43}$$

The function which satisfies this equation is called a *gaussian distribution,*

$$\delta p(x,t) = \left[\frac{\Delta P}{2\sqrt{\pi D_p t}}\right] e^{-x^2/4D_p t} \tag{4–44}$$

where ΔP is the number of holes per unit area created over a negligibly small distance at $t = 0$. The factor in brackets indicates that the peak value of the pulse (at $x = 0$) decreases with time, and the exponential factor predicts the spread of the pulse in the positive and negative x-directions (Fig. 4–19). If we

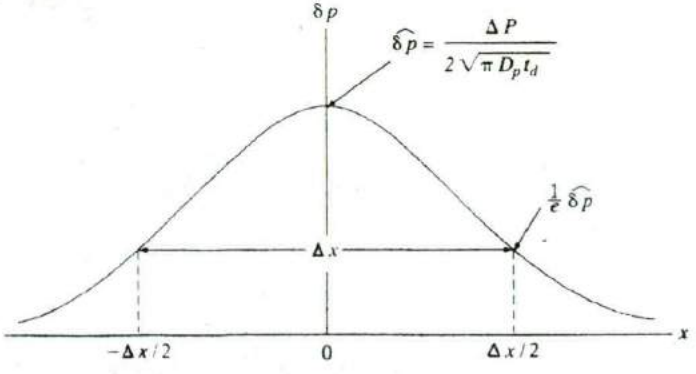


Figure 4–19
Calculation of D_p from the shape of the δp distribution after time t_d. No drift or recombination is included.

designate the peak value of the pulse as $\delta\hat{p}$ at any time (say t_d), we can use Eq. (4–44) to calculate D_p from the value of δp at some point x. The most convenient choice is the point $\Delta x/2$, at which δp is down by $1/e$ of its peak value $\delta\hat{p}$. At this point we can write

$$e^{-1}\delta\hat{p} = \delta\hat{p}e^{-(\Delta x/2)^2/4D_p t_d} \tag{4–45}$$

$$D_p = \frac{(\Delta x)^2}{16t_d} \tag{4–46}$$

Since Δx cannot be measured directly, we use an experimental setup such as Fig. 4–20, which allows us to display the pulse on an oscilloscope as the carriers pass under a detector. As we shall see in Chapter 5, a forward-biased p-n junction serves as an excellent injector of minority carriers, and a reverse-biased junction serves as a detector. The measured quantity in Fig. 4–20 is the pulse width Δt displayed on the oscilloscope in time. It is related to Δx by the drift velocity, as the pulse drifts past the detector point (2).

$$\Delta x = \Delta t v_d = \Delta t \frac{L}{t_d} \tag{4–47}$$

EXAMPLE 4–6 An n-type Ge sample is used in the Haynes–Shockley experiment shown in Fig. 4–20. The length of the sample is 1 cm, and the probes (1) and (2) are separated by 0.95 cm. The battery voltage E_0 is 2 V. A pulse arrives at point (2)

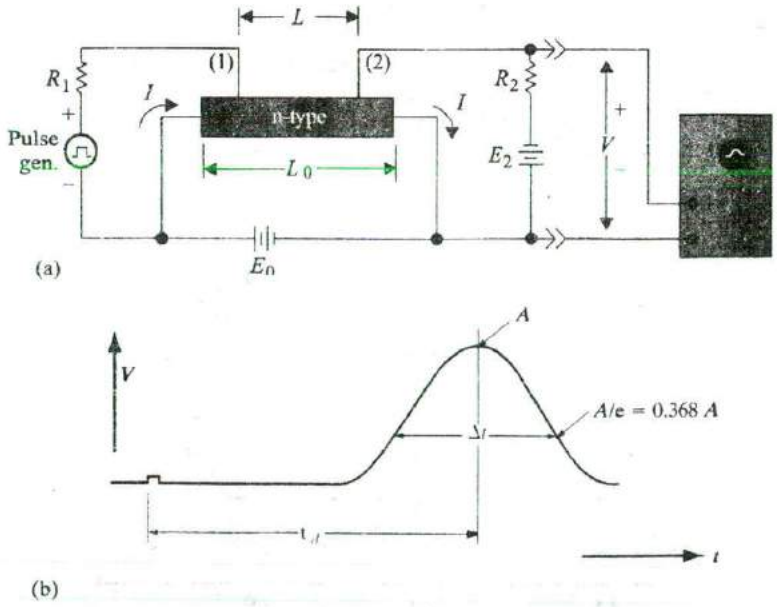


Figure 4–20 The Haynes–Shockley experiment: (a) circuit schematic; (b) typical trace on the oscilloscope screen.

0.25 ms after injection at (1); the width of the pulse Δt is 117 μs. Calculate the hole mobility and diffusion coefficient, and check the results against the Einstein relation.

SOLUTION

$$\mu_p = \frac{v_d}{\mathscr{E}} = \frac{0.95/(0.25 \times 10^{-3})}{2/1} = 1900 \text{ cm}^2/(\text{V-s})$$

$$D_p = \frac{(\Delta x)^2}{16t_d} = \frac{(\Delta t L)^2}{16t_d^3}$$

$$= \frac{(117 \times 0.95)^2 \times 10^{-12}}{16(0.25)^3 \times 10^{-9}} = 49.4 \text{ cm}^2/\text{s}$$

$$\frac{D_p}{\mu_p} = \frac{49.4}{1900} = 0.026 = \frac{kT}{q}$$

4.4.6 Gradients in the Quasi-Fermi Levels

In Section 3.5 we saw that equilibrium implies no gradient in the Fermi level E_F. In contrast, any combination of drift and diffusion implies a gradient in the steady state quasi-Fermi level.

We can use the results of Eqs. (4–23), (4–26), and (4–29) to demonstrate the power of the concept of quasi-Fermi levels in semiconductors [see Eq. (4–15)]. If we take the general case of nonequilibrium electron concentration with drift and diffusion, we must write the total electron current as

$$J_n(x) = q\mu_n n(x)\mathscr{E}(x) + qD_n \frac{dn(x)}{dx} \tag{4–48}$$

where the gradient in electron concentration is

$$\frac{dn(x)}{dx} = \frac{d}{dx}\left[n_i e^{(F_n - E_i)/kT}\right] = \frac{n(x)}{kT}\left(\frac{dF_n}{dx} - \frac{dE_i}{dx}\right) \tag{4–49}$$

Using the Einstein relation, the total electron current becomes

$$J_n(x) = q\mu_n n(x)\mathscr{E}(x) + \mu_n n(x)\left[\frac{dF_n}{dx} - \frac{dE_i}{dx}\right] \tag{4–50}$$

But Eq. (4–26) indicates that the subtractive term in the brackets is just $q\mathscr{E}(x)$, giving a direct cancellation of $q\mu_n n(x)\mathscr{E}(x)$ and leaving

$$J_n(x) = \mu_n n(x)\frac{dF_n}{dx} \tag{4–51}$$

Thus, the processes of electron drift and diffusion are summed up by the spatial variation of the quasi-Fermi level. The same derivation can be made for holes, and we can write the current due to drift and diffusion in the form of a *modified Ohm's law*

$$J_n(x) = q\mu_n n(x)\frac{d(F_n/q)}{dx} = \sigma_n(x)\frac{d(F_n/q)}{dx} \tag{4–52a}$$

$$J_p(x) = q\mu_p p(x)\frac{d(F_p/q)}{dx} = \sigma_p(x)\frac{d(F_p/q)}{dx} \tag{4–52b}$$

Therefore, any drift, diffusion, or combination of the two in a semiconductor results in currents proportional to the gradients of the two quasi-Fermi levels. Conversely, a lack of current implies constant quasi-Fermi levels.

PROBLEMS

4.1 With E_F located 0.4 eV above the valence band in a Si sample, what charge state would you expect for most Ga atoms in the sample? What would be the predominant charge state of Zn? Au? *Note:* By charge state we mean neutral, singly positive, doubly negative, etc.

4.2 A Si sample is doped with 10^{16} cm^{-3} Sb. How many Zn atoms/cm^3 must be added to exactly compensate this material ($n_0 = p_0 = n_i$)?

4.3 Construct a semilogarithmic plot such as Fig. 4–7 for GaAs doped with 2×10^{15} donors/cm^3 and having 4×10^{14} EHP/cm^3 created uniformly at $t = 0$. Assume that $\tau_n = \tau_p = 50$ ns.

4.4 Calculate the recombination coefficient α_r, for the low-level excitation described in Prob. 4.3. Assume that this value of α_r applies when the GaAs sample is uniformly exposed to a steady state optical generation rate $g_{op} = 10^{20}$ EHP/cm^3-s. Find the steady state excess carrier concentration $\Delta n = \Delta p$.

4.5 A sample is doped with donors such that $n_0 = Gx$ for $n_0 \gg n_i$, where G is a constant. Find the built-in electric field $\mathscr{E}(x)$.

4.6 A Si sample with 10^{15}/cm^3 donors is uniformly optically excited at room temperature such that 10^{19}/cm^3 electron–hole pairs are generated per second. Find the separation of the quasi-Fermi levels and the change of conductivity upon shining the light. Electron and hole lifetimes are both 10 μs. $D_p = 12$ cm^2/s.

4.7 An n-type Si sample with $N_d = 10^{15}$ cm^{-3} is steadily illuminated such that $g_{op} = 10^{21}$ EHP/cm^3-s. If $\tau_n = \tau_p = 1\,\mu$s for this excitation, calculate the separation in the quasi-Fermi levels, $(F_n - F_p)$. Draw a band diagram such as Fig. 4–11.

4.8 For a 2 cm long doped Si bar ($N_d = 10^{16}$ cm^{-3}) with a cross-sectional area = 0.05 cm^2, what is the current if we apply 10V across it? If we generate 10^{20} electron–hole pairs per second per cm^3 uniformly in the bar and the lifetime $\tau_n = \tau_p = 10^{-4}$s, what is the new current? Assume the low level α_r doesn't change for high level injection. If the voltage is then increased to 100,000 V, what is the new current? Assume $\mu_p = 500$ cm^2/V-s, but you must choose the appropriate values for electrons.

4.9 Design and sketch a photoconductor using a 5-μm-thick film of CdS, assuming that $\tau_n = \tau_p = 10^{-6}$ s and $N_d = 10^{14}$ cm^{-3}. The dark resistance (with $g_{op} = 0$) should be 10 MΩ, and the device must fit in a square 0.5 cm on a side; therefore, some

sort of folded or zigzag pattern is in order. With an excitation of $g_{op} = 10^{21}$ EHP/cm^3-s, what is the resistance change?

4.10 In a very long p-type Si bar with cross-sectional area = 0.5 cm^2 and $N_a = 10^{17}$ cm^{-3}, we inject holes such that the steady state *excess* hole concentration is 5×10^{16} cm^{-3} at $x = 0$. What is the steady state separation between F_p and E_c at $x = 1000$Å? What is the hole current there? How much is the excess stored hole charge? Assume $\mu_p = 500$ cm^2/V-s and $\tau_p = 10^{-10}$s.

4.11 Assume that a photoconductor in the shape of a bar of length L and area A has a constant voltage V applied, and it is illuminated such that g_{op} EHP/cm^3-s are generated uniformly throughout. If $\mu_n \gg \mu_p$, we can assume the optically induced change in current ΔI is dominated by the mobility μ_n and lifetime τ_n for electrons. Show that $\Delta I = qAL g_{op}\tau_n/\tau_t$ for this photoconductor, where τ_t is the transit time of electrons drifting down the length of the bar.

4.12 For the steady state minority hole distribution shown in Fig. 4–17, find the expression for the hole quasi-Fermi level position $E_i - F_p(x)$ while $p(x) \gg p_0$ (i.e., while F_p is below E_F). On a band diagram, draw the variation of $F_p(x)$. Be careful—when the minority carriers are few (e.g., when δp is n_i), F_p still has a long way to go to reach E_F.

4.13 Boron is diffused into an intrinsic Si sample, giving the acceptor distribution shown in Figure P4-13. Sketch the equilibrium band diagram and show the direction of the resulting electric field, for $N_a(x) \gg n_i$. Repeat for phosphorus, with $N_d(x) \gg n_i$.

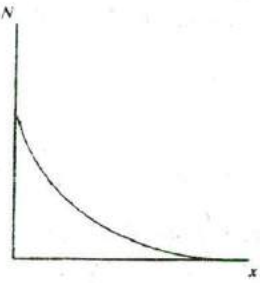


Figure P4–13

4.14 The current required to feed the hole injection at $x = 0$ in Fig. 4–17 is obtained by evaluating Eq. (4–40) at $x = 0$. The result is $I_p(x = 0) = qAD_p\Delta p/L_p$. Show that this current can be calculated by integrating the charge stored in the steady state hole distribution $\delta p(x)$ and then dividing by the average hole lifetime τ_p. Explain why this approach gives $I_p(x = 0)$.

4.15 We wish to use the Haynes–Shockley experiment to calculate the hole lifetime τ_p in an n-type sample. Assume the peak voltage of the pulse displayed on the oscilloscope screen is proportional to the hole concentration under the collector terminal at time t_d, and that the displayed pulse can be approximated as a gaussian, as in Eq. (4–44), which decays due to recombination by e^{-t/τ_p}. The electric field is varied and the following data taken: For $t_d = 200$ μs, the peak is 20 mV; for $t_d = 50$ μs, the peak is 80 mV. What is τ_p?

4.16 Consider a sample of GaAs ($n_i = 10^6$ cm^{-3} at 300 K) doped with 10^{15} donors per cm^3 illuminated with the 5145 Å line of an argon ion laser. For GaAs at 5145 Å, $\alpha = 10^4$ cm^{-1}. Calculate and plot the steady state excess electron profile $\delta n(x)$ in the region within 5 μm of the surface for photon fluxes of 10^{15}, 10^{17}, and 10^{19} photons cm^{-2} s^{-1} using low-level injection assumptions and directly solving Eq. (4–12). For this problem, assume that $\tau_n = \tau_p = 10^{-6}$ s. Neglect diffusion.

4.17 For the sample of Prob. 4–16, calculate and plot the steady state excess electron profile $\delta n(x)$ in the region within 5 μm of the surface for a photon flux of 10^{19} photons cm^{-2} s^{-1} using low-level injection assumptions and directly solving Eq. (4–12) for values of α_r of 10^{-9}, 10^{-7}, and 10^{-5} cm^3 s^{-1}.

4.18 Using the results of Prob. 4–16 obtained for a photon flux of 10^{15} photons cm^{-2} s^{-1}, calculate and plot the transient excess carrier profile, 1, 2, and 5 ns after the laser flux is interrupted, by integrating Eq. (4–5) within each depth interval, using 10^{-6} cm^3 s^{-1} for α_r. In this case, ignore carrier diffusion.

4.19 Assume an n-type semiconductor bar is illuminated over a narrow region of its length, such that $\Delta n = \Delta p$ in the illuminated zone, and excess carriers diffuse away and recombine in both directions along the bar. Assuming $\delta n = \delta p$, sketch the excess carrier distribution and, on a band diagram, sketch the quasi-Fermi levels F_n and F_p over several diffusion lengths from the illuminated zone. See the cautionary note in Prob. 4–12.

READING LIST

Ashcroft, N. W., and N. D. Mermin. *Solid State Physics.* Philadelphia: W.B. Saunders, 1976.

Bhattacharya, P. *Semiconductor Optoelectronic Devices.* Englewood Cliffs, NJ: Prentice Hall, 1994.

Blakemore, J. S. *Semiconductor Statistics.* New York: Dover Publications, 1987.

Collins, R. T., and M. A. Tischler. "Silicon Emits Light, but How?" *Laser Focus World* 28 (February 1992): 18+.

Ghandhi, S. K. *VLSI Fabrication Principles,* 2nd ed. New York: Wiley, 1994.

Gupta, S., S. L. Williamson, and J. F. Whitaker. "Epitaxial Methods Produce Robust Ultrafast Detectors." *Laser Focus World* 28 (June 1992): 97–101+.

Hummel, R. E. *Electronic Properties of Materials,* 2nd ed. Berlin: Springer-Verlag, 1993.

Li, S. S. *Semiconductor Physical Electronics.* New York: Plenum Press, 1993.

Madden M. R., and P. C. Williamson. "Photodetector Hybrids: What Are They and Who Needs Them?" *Laser Focus World* 28 (July 1992): 107–109+.

Muller, R. S., and T. I. Kamins. *Device Electronics for Integrated Circuits.* New York: Wiley, 1986.

Neamen, D. A. *Semiconductor Physics and Devices: Basic Principles.* Homewood, IL: Irwin, 1992.

Neudeck, G. W. *Modular Series on Solid State Devices: Vol II. The PN Junction Diode.* Reading, MA: Addison-Wesley, 1983.

Pankove, J. I. *Optical Processes in Semiconductors.* Englewood Cliffs, NJ: Prentice Hall, 1971.

Pierret, R. F. *Advanced Semiconductor Fundamentals.* Reading, MA: Addison-Wesley, 1987.

Singh, J. *Physics of Semiconductors and Their Heterostructures.* New York: McGraw-Hill, 1993.

Singh, J. *Semiconductor Devices.* New York: McGraw-Hill, 1994.

Swaminathan, V., and Macrander, A. T. *Material Aspects of GaAs and InP Based Structures.* Englewood Cliffs, NJ: Prentice Hall, 1991.

Sze, S. M. *Physics of Semiconductor Devices.* New York: Wiley, 1981.

Thomas, G. A. "An Electron–Hole Liquid." *Scientific American* 234 (June 1976): 28–37.

Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englev ood Cliffs, NJ: Prentice Hall, 1989.

Weisbuch, C., and B. Vinter. *Quantum Semiconductor Structures.* Boston: Academic Press, 1991.

Wolfe, C. M., G. E. Stillman, and N. Holonyak, Jr. *Physical Properties of Semiconductors.* Englewood Cliffs, NJ: Prentice Hall, 1989.