CHAPTER

4

Modern Theory of Solids

One of the great successes of modern physics has been the application of quantum mechanics or the Schrödinger equation to the behavior of molecules and solids. For example, quantum mechanics explains the nature of the bond between atoms, and its consequences. How can carbon bond with four other carbon atoms? What determines the direction and strength of a bond? An intuitively obvious outcome from quantum mechanics is that the energy of the electron is still quantized in the molecule. In addition, the application of quantum mechanics to many atoms, as in a solid, leads to energy bands within which the electron energy levels are almost continuous. The electron energy falls within possible values in a band of energies. It is nearly impossible to comprehend the principles of operation of modern solid-state electronic devices without a good grasp of the band theory of solids. Since we are dealing with a large number of electrons in the solid, we must consider a statistical way of describing their behavior, just as we use the Maxwell distribution of velocities to explain the behavior of gas atoms. An equally important question, therefore, is "What is the probability that an electron is in a state with energy E within an energy band?"

4.1 HYDROGEN MOLECULE: MOLECULAR ORBITAL THEORY OF BONDING

Consider what happens when two hydrogen atoms approach each other to form the hydrogen molecule. This is the H–H (or H_2) system. Let us examine the energy levels of the H–H system as a function of the interatomic distance R. When the atoms are infinitely separated, each atom has its own set of energy levels, labeled $1s$, $2s$, $2p$, etc. The electron energy in each atom is -13.6 eV with respect to the "free" state (electron infinitely separated from the parent nucleus). The energy of the two isolated hydrogen atoms is twice -13.6 eV.

As the atoms approach closer, the electrons interact both with each other and with the other nuclei. To obtain the wavefunctions and the new energy of the electrons, we

need to find the new potential energy function *PE* for the electrons in this new environment and then solve the Schrödinger equation with this new *PE* function. The new energy is actually *lower* than twice −13.6 eV, which means that the H_2 formation is energetically favorable.

The bond formation between two H atoms can be easily explained by describing the behavior of the electron within the molecule. We use a **molecular orbital** ψ, which depends on the interaction of individual atomic wavefunctions and is regarded as an electron wavefunction within the molecule.

In the H_2 molecule, we cannot have two sets of identical atomic ψ_{1s} orbitals, for two reasons. First, this would violate the Pauli exclusion principle, which requires that, in a given system of electrons (those within the H_2 molecule), we cannot have two sets of identical quantum numbers. When the atoms were separated, we did not have this problem, because we had two isolated systems.

Second, as the two atoms approach each other, as shown in Figure 4.1, the atomic ψ_{1s} wavefunctions overlap. This overlap produces two new wavefunctions with different energies and hence different quantum numbers. When the two atomic wavefunctions interfere, they can overlap either in phase (both positive or both negative) or out of phase

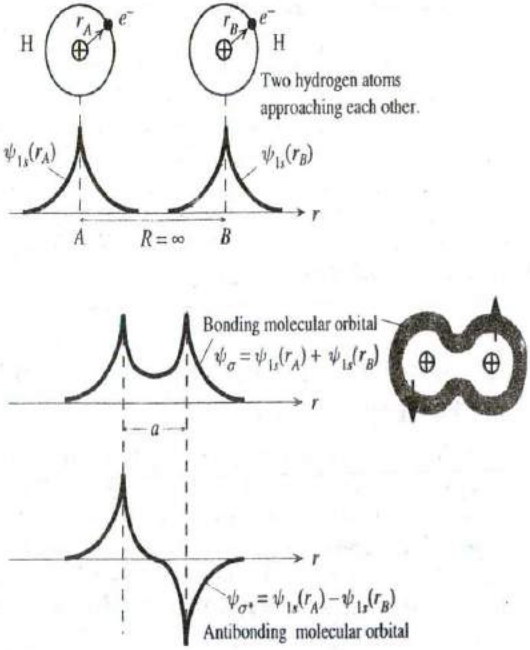


Figure 4.1 Formation of molecular orbitals, bonding, and antibonding (ψ_σ and ψ_{σ^*}) when two H atoms approach each other.
The two electrons pair their spins and occupy the bonding orbital ψ_σ.

(one positive and the other negative), as a result of which two molecular orbitals are formed. These are conventionally labeled ψ_σ and ψ_{σ^*} as illustrated in Figure 4.1. Thus, two of the molecular orbitals in the H–H system are

$$\psi_\sigma = \psi_{1s}(r_A) + \psi_{1s}(r_B) \quad [4.1]$$

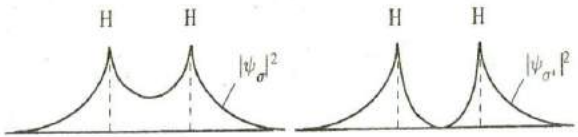
$$\psi_{\sigma^*} = \psi_{1s}(r_A) - \psi_{1s}(r_B) \quad [4.2]$$

where the two hydrogen atoms are labeled A and B, and r_A and r_B are the respective distances of the electrons from their parent nucleus. In generating two separate molecular orbitals ψ_σ and ψ_{σ^*} from a linear combination of two identical atomic orbitals ψ_{1s}, we have used the **linear combination of atomic orbitals (LCAO)** method.

The first molecular orbital ψ_σ is *symmetric* and has considerable magnitude between the nuclei, whereas the second ψ_{σ^*}, is *antisymmetric* and has a node between the nuclei. The resulting electron probability distributions $|\psi_\sigma|^2$ and $|\psi_{\sigma^*}|^2$ are shown in Figure 4.2.

In an analogy to hydrogenic wavefunctions, since ψ_{σ^*} has a node, we would expect it to have a higher energy than the ψ_σ orbital and therefore a different energy quantum number, which means that the Pauli exclusion principle is no longer violated. We can also expect that because $|\psi_\sigma|^2$ has an appreciable electron concentration between the two nuclei, the electrostatic PE, and hence the total energy for the wavefunction ψ_σ, will be lower than that for ψ_{σ^*}, as well as those for the individual atomic wavefunctions.

Of course, the true wavefunctions of the electrons in the H_2 system must be determined by solving the Schrödinger equation, but an intelligent guess is that these must look like ψ_σ and ψ_{σ^*}. We can therefore use ψ_σ and ψ_{σ^*} in the Schrödinger equation, with the correct form of the PE term V, to evaluate the energies E_σ and E_{σ^*} of ψ_σ and ψ_{σ^*}, respectively, as a function of R. The PE function V in the H–H system has positive PE contributions arising from electron–electron repulsions and proton–proton



(a) Electron probability distributions for bonding and antibonding orbitals, ψ_σ and ψ_{σ^*}.

(b) Lines representing contours of constant probability (darker lines represent greater relative probability).

Figure 4.2

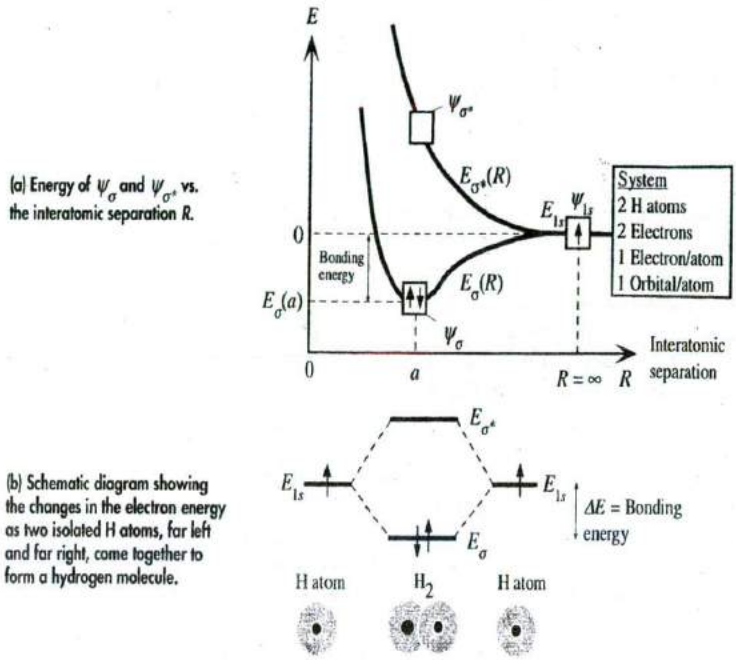
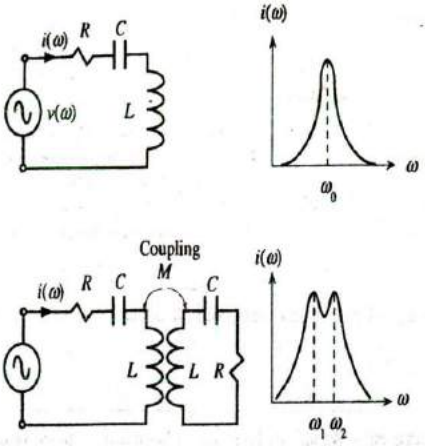


Figure 4.3 Electron energy in the system comprising two hydrogen atoms.

repulsions, but negative PE contributions arising from the attractions of the two electrons to the two protons.

The two energies, E_σ and E_{σ^*}, are widely different, with E_σ below E_{1s} and E_{σ^*} above E_{1s}, as shown schematically in Figure 4.3a. As R decreases and the two H atoms get closer, the energy of the ψ_σ orbital state passes through a minimum at $R = a$. Each orbital state can hold two electrons with spins paired, and within the two hydrogen atoms, we have two electrons. If these enter the ψ_σ orbital and pair their spins, then this new configuration is energetically more favorable than two isolated H atoms. It corresponds to the hydrogen molecule H_2. The energy difference between that of the two isolated H atoms and the E_σ minimum energy at $R = a$ is the bonding energy, as illustrated in Figure 4.3a. When the two electrons in the H_2 molecule occupy the ψ_σ orbital, their probability distribution (and hence, the negative charge distribution) is such that the negative PE, arising from the attractions of these two electrons to the two protons, is stronger in magnitude than the positive PE, arising from electron–electron repulsions and proton–proton repulsions and the kinetic energy of the two electrons. Therefore, the H_2 molecule is energetically stable.

The wavefunction ψ_σ corresponding to the lowest electron energy is called the **bonding orbital,** and ψ_{σ^*} is the **antibonding orbital.** When two atoms are brought together, the two identical atomic wavefunctions combine in two ways to generate two different molecular orbitals, each with a different energy. Effectively, then, an atomic



(a) There is one resonant frequency, ω_0, in an isolated *RLC* circuit.

(b) There are two resonant frequencies in two coupled *RLC* circuits: one below and the other above ω_0.

Figure 4.4

energy level, such as E_{1s}, splits into two, E_σ and E_{σ^*}. The splitting is due to the interaction (or overlap) between the atomic orbitals. Figure 4.3b schematically illustrates the changes in the electron energy levels as two isolated H atoms are brought together to form the H_2 molecule.

The splitting of a one-atom energy level when a molecule is formed is analogous to the splitting of the resonant frequency in an *RLC* circuit when two such circuits are brought together and coupled. Consider the *RLC* circuit shown in Figure 4.4a. The circuit is excited by an ac voltage source. The current peaks at the resonant frequency ω_0, as indicated in Figure 4.4a. When two such identical *RLC* circuits are coupled together and driven by an ac voltage source, the current develops two peaks, at frequencies ω_1 and ω_2, below and above ω_0, as illustrated in Figure 4.4b. The two peaks at ω_1 and ω_2 are due to the mutual inductance that couples the two circuits, allowing them to interact. From this analogy, we can intuitively accept the energy splitting observed in Figure 4.3a.

Consider what happens when two He atoms come together. Recall that the $1s$ orbital has paired electrons and is full. The $1s$ atomic energy level will again split into two levels, E_σ and E_{σ^*}, associated with the molecular orbitals ψ_σ and ψ_{σ^*}, as illustrated in Figure 4.5. However, in the He–He system, there are four electrons, so two occupy the ψ_σ orbital state and two go to the ψ_{σ^*} orbital state. Consequently, the system energy is not lowered by bringing the two He atoms closer. Furthermore, quantum mechanical calculations show that the antibonding energy level E_{σ^*} shifts higher than the bonding level E_σ shifts lower. By the same token, although we could put an additional electron at E_{σ^*} in H_2 to make H_2^-, we could not make H_2^{2-} by placing two electrons at E_{σ^*}.

From the He–He example, we can conclude that, as a general rule, the overlap of full atomic orbital states does not lead to bonding. In fact, full orbitals repel each other, because any overlap results in an increase in the system energy. To form a bond

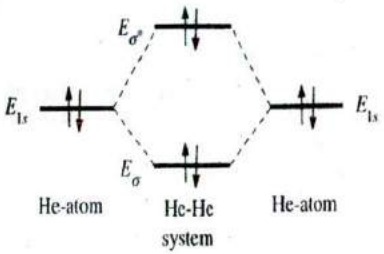


Figure 4.5 Two He atoms have four electrons. When He atoms come together, two of the electrons enter the E_σ level and two the E_{σ^*} level, so the overall energy is greater than two isolated He atoms.

between two atoms, we essentially need an overlap of half-occupied orbitals, as in the H_2 molecule.

EXAMPLE 4.1 **HYDROGEN HALIDE MOLECULE (HF)** We already know that H has a half-occupied $1s$ orbital, which can take part in bonding. Since the F atom has the electronic structure $1s^2 2s^2 p^5$, two of the p orbitals are full and one p orbital, p_x, is half full. This means that only the p_x orbital can participate in bonding. Figure 4.6 shows the electron orbitals in both H and F. When the H atom and the F atom approach each other to form an HF molecule, the ψ_{1s} orbital of H overlaps the p_x orbital of F. There are two possibilities for the overlap. First, ψ_{1s} and p_x can overlap in phase (both positive or both negative), to give a ψ_σ orbital that does not have a node between H and F, as shown in Figure 4.6. Second, they can overlap out of phase (one positive and the other negative), so that the overlap orbital ψ_{σ^*} has a node (similar to ψ_{σ^*} in Figure 4.1). We know from hydrogen atomic wavefunctions in Chapter 3 that orbitals with more nodes have higher energies. The molecular orbital ψ_σ therefore corresponds to a bonding orbital with a lower energy than the ψ_{σ^*} orbital. The two electrons, one from ψ_{1s} and the other from p_x, enter the ψ_σ orbital with spins paired, thereby forming a bond between H and F.

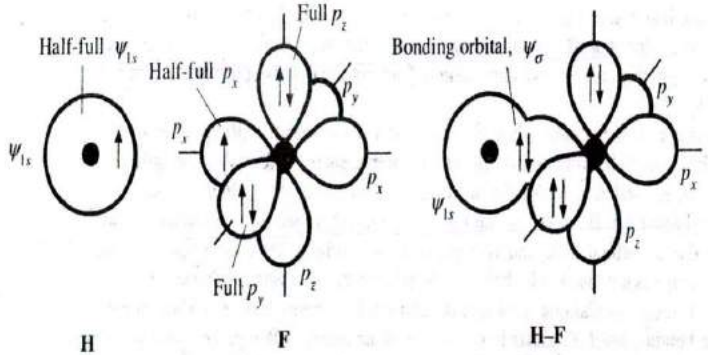


Figure 4.6 H has one half-empty ψ_{1s} orbital.
F has one half-empty p_x orbital but full p_y and p_z orbitals. The overlap between ψ_{1s} and p_x produces a bonding orbital and an antibonding orbital. The two electrons fill the bonding orbital and thereby form a covalent bond between H and F.

4.2 BAND THEORY OF SOLIDS

4.2.1 ENERGY BAND FORMATION

When we bring three hydrogen atoms (labeled A, B, and C) together, we generate three separate molecular orbital states, ψ_a, ψ_b, and ψ_c, from three ψ_{1s} atomic states. Again, this occurs in three different ways, as illustrated in Figure 4.7a. As in the case of the H_2 molecule, each molecular orbital must be either *symmetric* or *antisymmetric* with respect to center atom B.[1] The orbitals that satisfy even and odd requirements are

$$\psi_a = \psi_{1s}(A) + \psi_{1s}(B) + \psi_{1s}(C) \quad [4.3a]$$

$$\psi_b = \psi_{1s}(A) - \psi_{1s}(C) \quad [4.3b]$$

$$\psi_c = \psi_{1s}(A) - \psi_{1s}(B) + \psi_{1s}(C) \quad [4.3c]$$

where $\psi_{1s}(A)$, $\psi_{1s}(B)$, and $\psi_{1s}(C)$ are the $1s$ atomic wavefunctions centered around the atoms A, B, and C, respectively, as shown in Figure 4.7a. For example, the wavefunction $\psi_{1s}(A)$ represents $\psi_{1s}(r_A)$, which is centered around A and has the form $\exp(-r_A/a_o)$, where r_A is the distance from the nucleus of A, and a_o is the Bohr radius. Notice that $\psi_{1s}(B)$ is missing in Equation 4.3b, so ψ_b is antisymmetric.

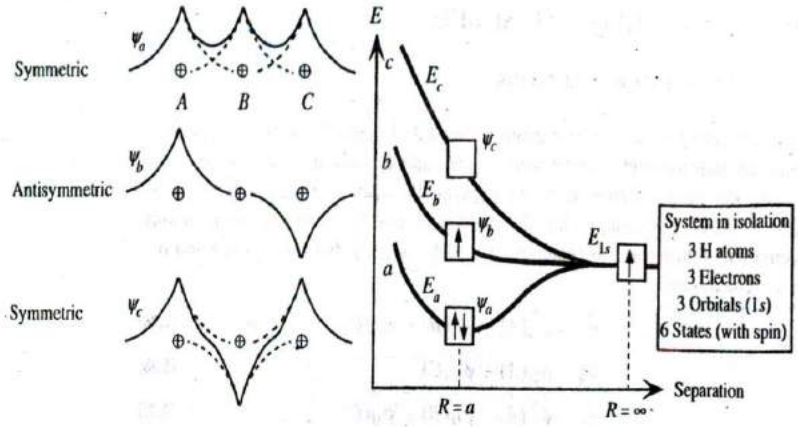
The energies E_a, E_b, and E_c of ψ_a, ψ_b, and ψ_c can be calculated from the Schrödinger equation by using the PE function of this system (the PE also includes proton–proton repulsions). It is clear that since ψ_a, ψ_b, and ψ_c are different, their energies E_a, E_b, and E_c are also different. Consequently, the $1s$ energy level splits into three separate levels, corresponding to the energies of ψ_a, ψ_b, and ψ_c, as depicted by Figure 4.7b. By analogy with the electron wavefunctions in the hydrogen atom, we can argue that if the molecular wavefunction has more nodes, its energy is higher. Thus, ψ_a has the lowest energy E_a, ψ_b has the next higher energy E_b, and ψ_c has the highest energy E_c, as shown in Figure 4.7b. There are three electrons in the three-hydrogen system. The first two pair their spins and enter orbital ψ_a at energy E_a, and the third enters orbital ψ_b at energy E_b. Comparing Figures 4.7 and 4.3, we notice that although H_2 and H_3 both have two electrons in the lowest energy level, H_3 also has an extra electron at the higher energy level (E_b), which tends to increase the net energy of the atom. Thus, the H_3 molecule is much less stable than the H_2 molecule.[2]

Now consider the formation of a solid. Take N Li (lithium) atoms from infinity and bring them together to form the Li metal. Lithium has the electronic configuration $1s^2 2s^1$, which is somewhat like the hydrogen atom, since the K shell is closed and the third electron is alone in the $2s$ orbital.

Based on our previous discussions, we assume that the atomic energy levels will split into N separate energy levels. Since the $1s$ subshell is full and is close to the nucleus, it will not be affected much by the interatomic interactions; consequently, the energy of

[1] The reason is that the molecule A-B-C, when A, B, and C are identical atoms, is symmetric with respect to B. Thus each wavefunction must have odd or even parity (Chapter 3).

[2] See G. Pimentel and R. Spratley, *Understanding Chemistry*, San Francisco: Holden-Day, Inc., 1972, pp. 682–687 for an excellent discussion.



(a) Three molecular orbitals from three ψ_{1s} atomic orbitals overlapping in three different ways.

(b) The energies of the three molecular orbitals, labeled *a*, *b*, and *c*, in a system with three H atoms.

Figure 4.7

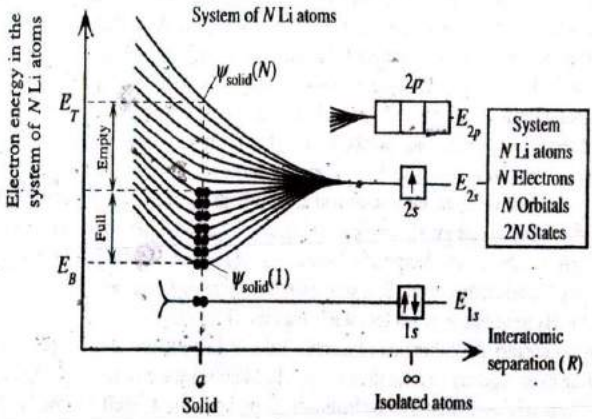


Figure 4.8 The formation of a 2*s* energy band from the 2*s* orbitals when *N* Li atoms come together to form the Li solid.

There are *N* 2*s* electrons, but 2*N* states in the band. The 2*s* band is therefore only half full. The atomic 1*s* orbital is close to the Li nucleus and remains undisturbed in the solid. Thus, each Li atom has a closed *K* shell (full 1*s* orbital).

this state will experience only negligible splitting, if any. Since the $1s$ electrons will stay close to their parent nuclei, we will not consider them during formation of the solid.

In the system of N isolated Li atoms, we have N electrons in N ψ_{2s} orbitals at the energy E_{2s}, as illustrated in Figure 4.8 (at infinite interatomic separation). Let us assume that N is large (typically, $\sim 10^{23}$). As N atoms are brought together to form the solid, the energy level at E_{2s} splits into N finely separated energy levels. The maximum width of the energy splitting depends on the closest interatomic distance a in the solid, as apparent in Figure 4.3a. The atoms separated by a distance greater than $R = a$ give rise to a lesser amount of energy splitting. The interatomic interactions between $N\psi_{2s}$ orbitals thus spread the N energy levels between the bottom and top levels, E_B and E_T, respectively, which are determined by the closest interatomic distance a. Put differently, E_B and E_T are determined by the distance between nearest neighbors. It is obvious that with N very large, the energy separation between two consecutive energy levels is very small; indeed, it is almost infinitesimal and not as exaggerated as in Figure 4.8.

Remember that each energy level E_i in the Li metal of Figure 4.8 is the energy of an electron wavefunction $\psi_{\text{solid}}(i)$ in the solid, where $\psi_{\text{solid}}(i)$ is one particular combination of the N atomic wavefunctions ψ_{2s}. There are N different ways to combine N atomic wavefunctions ψ_{2s}, since each can be added in phase or out of phase, as is apparent in Equations 4.3a to c (see also Figure 4.7a and b). For example, when all N ψ_{2s} are summed in phase, the resulting wavefunction $\psi_{\text{solid}}(1)$ is like ψ_a in Equation 4.3a, and it has the lowest energy. On the other hand, when N ψ_{2s} are summed with alternating phases, $+ - + \cdots$, the resulting wavefunction $\psi_{\text{solid}}(N)$ is like ψ_c, and it has the highest energy. Other combinations of ψ_{2s} give rise to different energy values between E_B and E_T.

The single $2s$ energy level E_{2s} therefore splits into N ($\sim 10^{23}$) finely separated energy levels, forming an **energy band,** as illustrated in Figure 4.8. Consequently, there are N separate energy levels, each of which can take two electrons with opposite spins. The N electrons fill all the levels up to and including the level at $N/2$. Therefore, the band is half full. We do not mean literally that the band is full to the half-energy point. The levels are not spread equally over the band from E_B to E_T, which means that the band cannot be full to the half-energy point. Half filled simply means half the states in the band are filled from the bottom up.

We have generated a half-filled band from a half-filled isolated $2s$ energy level. The energy band resulting from the splitting of the atomic $2s$ energy level is loosely termed the **$2s$ band.** By the same token, the atomic $1s$ levels are full, so any $1s$ band that forms from these $1s$ states will also be full. We can get an idea of the separation of energy levels in the $2s$ band by noting that the maximum separation, $E_T - E_B$, between the top and bottom of the band is on the order of 10 eV, but there are some 10^{23} atoms, giving rise to 10^{23} energy levels between E_B and E_T. Thus, the energy levels are finely separated, forming, for all practical purposes, a continuum of energy levels.

The $2p$ energy level, as well as the higher levels at $3s$ and so on, also split into finely separated energy levels, as shown in Figure 4.9. In fact, some of these energy levels overlap the $2s$ band; hence, they provide further energy levels and "extend" the $2s$ band into higher energy levels, as indicated in Figure 4.10, which shows how energy bands in metals are often represented. The vertical axis is the electron energy. The top of the $2s$ band, which is half full, overlaps the bottom of the $2p$ band, which itself

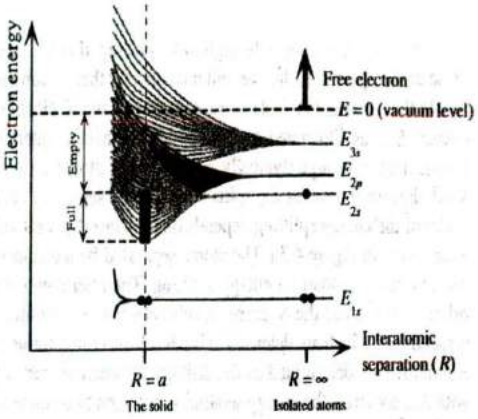


Figure 4.9 As Li atoms are brought together from infinity, the atomic orbitals overlap and give rise to bands.

Outer orbitals overlap first. The 3*s* orbitals give rise to the 3*s* band, 2*p* orbitals to the 2*p* band, and so on. The various bands overlap to produce a single band in which the energy is nearly continuous.

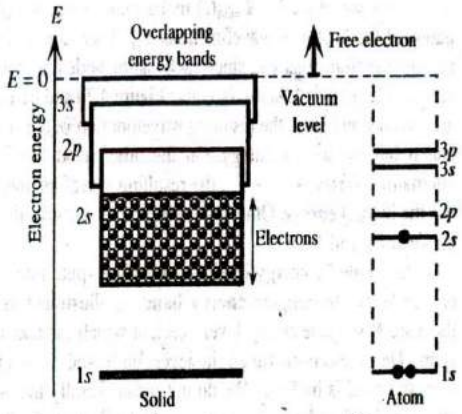


Figure 4.10 In a metal, the various energy bands overlap to give a single energy band that is only partially full of electrons.

There are states with energies up to the vacuum level, where the electron is free.

is overlapped near the top by the 3*s* band. We therefore have a band of energies that stretches from the bottom of the 2*s* band all the way to the vacuum level, as depicted in Figure 4.11. The reader may wonder what happened to the 3*d*, 4*s*, etc., bands. In the solid, the energies of these bands (including the top portion of the 3*s* band) are above the vacuum level, and the electron is free and far from the solid before it can acquire those energies.

At a temperature of absolute zero, or nearly so, the thermal energy is insufficient to excite the electrons to higher energy levels, so all the electrons pair their spins and fill each energy level from E_B up to an energy level E_{F0} that we call the Fermi level at 0 K, as shown in Figure 4.11. The energy value for the Fermi level depends on where we take the reference energy. For example, if we take the vacuum level as the zero reference, then for the Li metal, E_{F0} is at −2.5 eV. The Fermi level is normally measured with respect to the bottom of the band, in which case, it is simply termed the Fermi energy and denoted E_{F0}. For the Li metal, E_{F0} is 4.7 eV, which is with respect to the bottom of the band. The Fermi level has considerable significance, as we will discover later in this chapter.

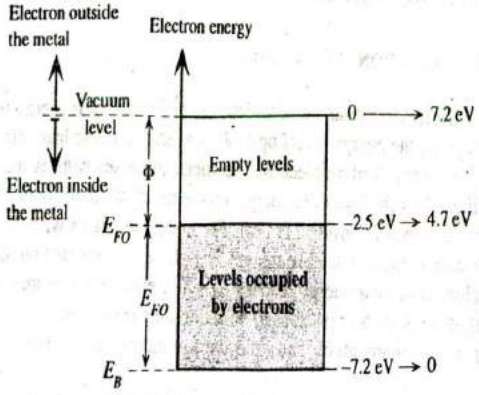


Figure 4.11 Typical electron energy band diagram for a metal.
All the valence electrons are in an energy band, which they only partially fill. The top of the band is the vacuum level, where the electron is free from the solid ($PE = 0$).

At absolute zero, all the energy levels up to the Fermi level are full. The energy required to excite an electron from the Fermi level to the vacuum level, that is, to liberate the electron from the metal, is called the **work function** Φ of the metal. As the temperature increases, some of the electrons get excited to higher energy levels. To determine the probability of finding an electron at an energy level E, we must consider what is called "particle statistics," a topic that is key to understanding the behavior of electronic devices. Clearly, the probability of finding an electron at 0 K at some energy $E < E_{FO}$ is unity, and at $E > E_{FO}$, the probability is zero. Table 4.1 summarizes the Fermi energy and work function of a few selected metals.

The electrons in the energy band of a metal are loosely bound valence electrons which become free in the crystal and thereby form a kind of **electron gas.** It is this electron gas that holds the metal ions together in the crystal structure and constitutes the metallic bond. This intuitive interpretation is shown in Figure 4.9. When solid Li is formed from N atoms, the N electrons fill all the lower energy levels up to $N/2$. The energy of the system of N Li atoms, according to Figure 4.9, is therefore much less than that of N isolated Li atoms by virtue of the N electrons taking up lower energy levels. It must be emphasized that the electrons within a band do not belong to any specific atom but to the whole solid. We cannot identify a given electron in the band with a certain Li atom. All the $2s$ electrons essentially form an electron gas and have energies that fall within the energy band. These electrons are constantly moving around in the metal which in terms of quantum mechanics means that their wavefunctions must be of the traveling wave type and not the type that localizes the electron around a given atom (*e.g.*, ψ_{n,l,m_l} in the hydrogen atom). We can represent each electron with a wavevector k so that its momentum p is $\hbar k$.

Table 4.1 Fermi energy and work function of selected metals

	Metal							
	Ag	Al	Au	Cs	Cu	Li	Mg	Na
Φ (eV)	4.5	4.28	5.0	2.14	4.65	2.3	3.7	2.75
E_{FO} (eV)	5.5	11.7	5.5	1.58	7.0	4.7	7.1	3.2

4.2.2 PROPERTIES OF ELECTRONS IN A BAND

Since the electrons inside the metal crystal are considered to be "free," their energy is *KE*. These electrons occupy all the energy levels up to E_{FO} as shown in the band diagram of Figure 4.12a. The energy E of an electron in a metal increases with its momentum p as $p^2/2m_e$. Figure 4.12b shows the energy versus momentum behavior of the electrons in a hypothetical one-dimensional crystal. The energy increases with momentum whether the electron is moving toward the left or right. Electrons take on all available momentum values until their energy reaches E_{FO}. For every electron that is moving right (such as *a*), there is another (such as *b*) with the same energy but moving left with the same magnitude of momentum. Thus, the average momentum is zero and there is no net current.

Consider what happens when an electric field $\mathcal{E}_x$ is applied in the $-x$ direction. The electron *a* at the Fermi level and moving along in the $+x$ direction experiences a force $e\mathcal{E}_x$ along the same direction. It therefore accelerates and gains momentum and hence has the energy as shown in Figure 4.12c. (The actual energy gained from the field is very small compared with E_{FO}, so Figure 4.12c is highly exaggerated.) The electron *a* at E_{FO} can move to higher energy levels because these adjacent higher levels are empty. The momentum state vacated by *a* is filled by the electron immediately below which now gains energy and moves up, and so on. An electron that is moving in the $-x$ direction, however, is decelerated (its momentum decreases) and hence loses energy as indicated by *b* moving to *b'* in Figure 4.12c. The electrons that are moving in the $+x$ direction gain energy, and those that are moving in the $-x$ direction, lose energy. The whole electron momentum distribution therefore shifts in the $+x$ direction as in Figure 4.12c. Eventually the electron *a*, now at *a'*, is scattered by a lattice vibration.

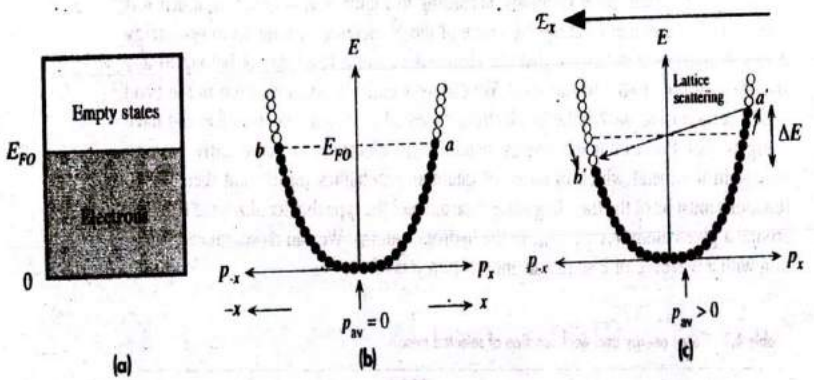


Figure 4.12

(a) Energy band diagram of a metal.

(b) In the absence of a field, there are as many electrons moving right as there are moving left. The motions of two electrons at each energy cancel each other as for *a* and *b*.

(c) In the presence of a field in the $-x$ direction, the electron *a* accelerates and gains energy to *a'* where it is scattered to an empty state near E_{FO} but moving in the $-x$ direction. The average of all momenta values is along the $+x$ direction and results in a net electric current.

Typically lattice vibrations have small energies but substantial momentum. The scattered electron must find an *unoccupied* momentum state with roughly the same energy, and it must change its momentum substantially. The electron at a' is therefore scattered to an empty state around E_{FO} but with a momentum in the opposite direction. Its momentum is *flipped* as shown in Figure 4.12c. The average momentum of the electrons is no longer zero but finite in the $+x$ direction. Consequently there is a current flow in the $-x$ direction, along the field, as determined by this average momentum p_{av}. Notice that a moves up to a' and b falls down to b'. Under steady-state conduction, lattice scattering simply replenishes the electrons at b' from a'. Notice that for energies below b', for every electron moving right there is another moving left with the same momentum magnitude that cancels it. Thus, electrons below the b' energy level do *not* contribute to conduction and are excluded from further consideration. Notice that electrons above the b' level are only moving right and their momenta are not canceled. Thus, the conductivity is determined by the electrons in the energy range ΔE from b' to a' about the Fermi level as shown in Figure 4.12c. Further, as the energy change from a to a' is orders of magnitude smaller than E_{FO}, we can summarize that conduction occurs by the drift of electrons at the Fermi level.[3] (If we were to calculate ΔE for a typical metal for typical currents, it would be $\sim 10^{-6}$ eV whereas E_{FO} is 1–10 eV. The shift in the distribution in Figure 4.12c is very small indeed; a' and b', for all practical purposes, are at the Fermi level.)

Conduction can be explained very simply and intuitively in terms of a band diagram as shown in Figure 4.13. Notice that the application of the electric field bends the energy band, because the electrostatic PE of the electron is $-eV(x)$ where $V(x)$ is the voltage at position x. However, $V(x)$ changes linearly from 0 to V, by virtue of $dV/dx = -\mathcal{E}_x$. Since $E = -eV(x)$ adds to the energy of the electron, the energy band must bend to account for the additional electrostatic energy. Since only the electrons near E_{FO} contribute to electrical conduction, we can represent this by drifting the electrons at E_{FO} down the potential hill. Although these electrons possess a very high mean velocity ($\sim 10^6$ m s^{-1}), as determined by the Fermi energy, they drift very slowly (10^{-2}–10^{-1} m s^{-1}) with a velocity that is drift mobility x field.

When a metal is illuminated, provided the wavelength of the radiation is correct, it will cause emission of electrons from the metal as in the photoelectric effect. Since Φ is the "minimum energy" required to excite an electron into the vacuum level (out from the metal), the longest wavelength radiation required is $hc/\lambda = \Phi$.

Addition of heat to a metal can excite some of the electrons in the band to higher energy levels. Thus heat can also be absorbed by the conduction electrons of a metal. We also know that the addition of heat increases the amplitude of atomic vibrations. We can therefore guess that the heat capacity of a metal has two terms which are due to energy absorption by the lattice vibrations and energy absorption by conduction electrons. It turns out that at room temperature the energy absorption by lattice vibrations dominates the heat capacity whereas at the lowest temperatures the electronic contribution is important.

[3] In some books (including the first edition of this textbook) it is stated that the electrons at E_{FO} can gain energy from the field and contribute to conduction but not those deep in the band (below b'). This is a simplified statement of the fact that at a level below E_{FO} there is one electron moving along in the $+x$ direction and gaining energy and another one at the same energy but moving along in the $-x$ direction and losing energy so that on average electron at this level does not gain energy.

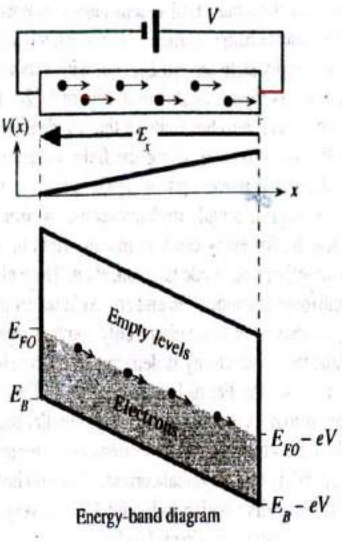


Figure 4.13 Conduction in a metal is due to the drift of electrons around the Fermi level.

When a voltage is applied, the energy band is bent to be lower at the positive terminal so that the electron's potential energy decreases as it moves toward the positive terminal.

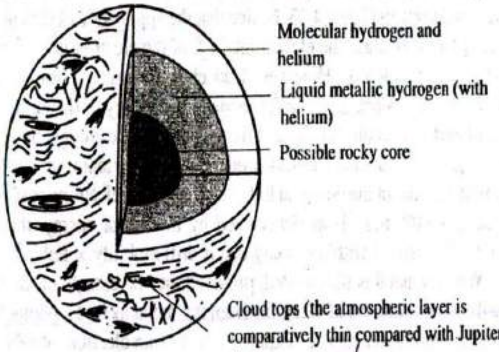


Figure 4.14 The interior of Jupiter is believed to contain liquid hydrogen, which is metallic.

SOURCE: Drawing adapted from T. Hey and P. Walters, *The Quantum Universe*, Cambridge, MA: Cambridge University Press, 1988, p. 96, figure 7.1.

EXAMPLE 4.2 **METALLIC LIQUID HYDROGEN IN JUPITER AND ITS MAGNETIC FIELD** The surface of Jupiter, as visualized schematically in Figure 4.14, mainly consists of a mixture of molecular hydrogen and He gases. Deep in the planet, however, the pressure is so tremendous that the hydrogen molecular bond breaks, leaving a dense ocean of hydrogen atoms. Hydrogen has only one electron in the 1*s* energy level. When atoms are densely packed, the 1*s* energy level forms an energy band, which is then only half filled. This is just like the Li metal, which means we can treat liquid hydrogen as a liquid metal, with electrical properties reminiscent of liquid mercury. Liquid hydrogen can sustain electric currents, which in turn can give rise to the magnetic fields on Jupiter. The origin of the electric currents are not known with certainty. We do know, however, that the core of the planet is hot and emanates heat, which causes convection currents. Temperature differences can readily give rise to electric currents, by virtue of thermoelectric effects, as discussed in Section 4.8.2.

EXAMPLE 4.3

WHAT MAKES A METAL? The Be atom has an electronic structure of $1s^2 2s^2$. Although the Be atom has a full $2s$ energy level, solid Be is a metal. Why?

SOLUTION

We will neglect the K shell ($1s$ state), which is full and very close to the nucleus, and consider only the higher energy states. In the solid, the $2s$ energy level splits into N levels, forming a $2s$ band. With $2N$ electrons, each level is occupied by spin-paired electrons. The $2s$ band is therefore full. However, the empty $2p$ band, from the empty $2p$ energy levels, overlaps the $2s$ band, thereby providing empty energy levels to these $2N$ electrons. Thus, the conduction electrons are in an energy band that is only partially filled; they can gain energy from the field to contribute to electrical conduction. Solid Be is therefore a metal.

EXAMPLE 4.4

FERMI SPEED OF CONDUCTION ELECTRONS IN A METAL In copper, the Fermi energy of conduction electrons is 7.0 eV. What is the speed of the conduction electrons around this energy?

SOLUTION

Since the conduction electrons are not bound to any one atom, their PE must be zero within the solid (but large outside), so all their energy is kinetic. For conduction electrons around the Fermi energy E_{FO} with a speed v_F, we have

$$\frac{1}{2}mv_F^2 = E_{FO}$$

so that

$$v_F = \sqrt{\frac{2E_{FO}}{m_e}} = \sqrt{\frac{2(1.6 \times 10^{-19}\ \text{J/eV})(7.0\ \text{eV})}{(9.1 \times 10^{-31}\ \text{kg})}} = 1.6 \times 10^6\ \text{m s}^{-1}$$

Although the Fermi energy depends on the properties of the energy band, to a good approximation it is only weakly temperature dependent, so v_F will be relatively temperature insensitive, as we will show later in Section 4.7.

4.3 SEMICONDUCTORS

The Si atom has 14 electrons, which distribute themselves in the various atomic energy levels as shown in Figure 4.15. The inner shells ($n = 1$ and $n = 2$) are full and therefore "closed." Since these shells are near the nucleus, when Si atoms come together to form the solid, they are not much affected and they stay around the parent Si atoms. They can therefore be excluded from further discussion. The $3s$ and $3p$ subshells are farther away from the nucleus. When two Si atoms approach, these electrons strongly interact with each other. Therefore, in studying the formation of bands in the Si solid, we will only consider the $3s$ and $3p$ levels.

The first task is to examine why Si actually bonds with four neighbors, since the $3s$ orbital is full and there are only two electrons in the $3p$ orbitals. The full $3s$ orbital should not overlap a neighbor and become involved in bonding. Since only two $3p$ orbitals are half full, bonds should be formed with two neighboring Si atoms. In reality,

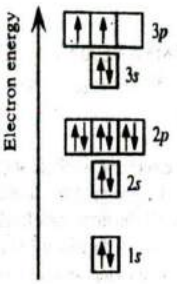
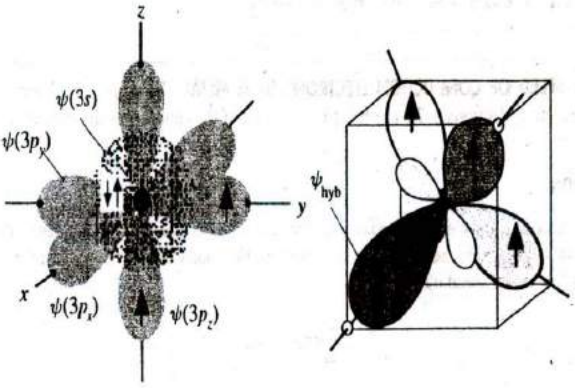


Figure 4.15 The electronic structure of Si.



(a) Isolated Si (b) Si just before bonding

Figure 4.16

(a) Si is in Group IV in the Periodic Table. An isolated Si atom has two electrons in the 3s and two electrons in the 3p orbitals.

(b) When Si is about to bond, the one 3s orbital and the three 3p orbitals become perturbed and mixed to form four hybridized orbitals, ψ_{hyb}, called sp^3 orbitals, which are directed toward the corners of a tetrahedron. The ψ_{hyb} orbital has a large major lobe and a small back lobe. Each ψ_{hyb} orbital takes one of the four valence electrons.

the $3s$ and $3p$ energy levels are quite close, and when five Si atoms approach each other, the interaction results in the four orbitals $\psi(3s)$, $\psi(3p_x)$, $\psi(3p_y)$, and $\psi(3p_z)$ mixing together to form four new **hybrid orbitals,** which are directed in tetrahedral directions; that is, each one is aimed as far away from the others as possible, as illustrated in Figure 4.16. We call this process sp^3 **hybridization,** since one s orbital and three p orbitals are mixed. (The superscript 3 on p has nothing to do with the number of electrons; it refers to the number of p orbitals used in the hybridization.)

The four sp^3 hybrid orbitals, ψ_{hyb}, each have one electron, so they are half occupied. This means that four Si atoms can have their orbitals ψ_{hyb} overlap to form bonds with one Si atom, which is what actually happens; thus, one Si atom bonds with four other Si atoms in tetrahedral directions.

In the same way, one Si atom bonds with four H atoms to form the important gas SiH_4, known as silane, which is widely used in the semiconductor technology to fabricate Si devices. In SiH_4, four hybridized orbitals of the Si atom overlap with the $1s$ orbitals of four H atoms. In exactly the same way, one carbon atom bonds with four hydrogen atoms to form methane, CH_4.

There are two ways in which the hybrid orbital ψ_{hyb} can overlap with that of the neighboring Si atom to form two molecular orbitals. They can add in phase (both positive or both negative) or out of phase (one positive and the other negative) to produce a bonding or an antibonding molecular orbital ψ_B and ψ_A, respectively, with energies E_B and E_A. Each Si–Si bond thus corresponds to two paired electrons in a bonding molecular orbital ψ_B. In the solid, there are $N(\sim 5 \times 10^{22}\,cm^{-3})$ Si atoms, and there are nearly as many such ψ_B bonds. The interactions between the ψ_B orbitals (*i.e.*, the Si–Si bonds) lead to the splitting of the E_B energy level to N levels, thereby forming an energy band labeled the **valence band** (VB) by virtue of the valence electrons it contains. Since the energy level E_B is full, so is the valence band. Figure 4.17 illustrates the formation of the VB from E_B.

In the solid, the interactions between the N number of ψ_A orbitals result in the splitting of the energy level E_A to N levels and the formation of an energy band that is

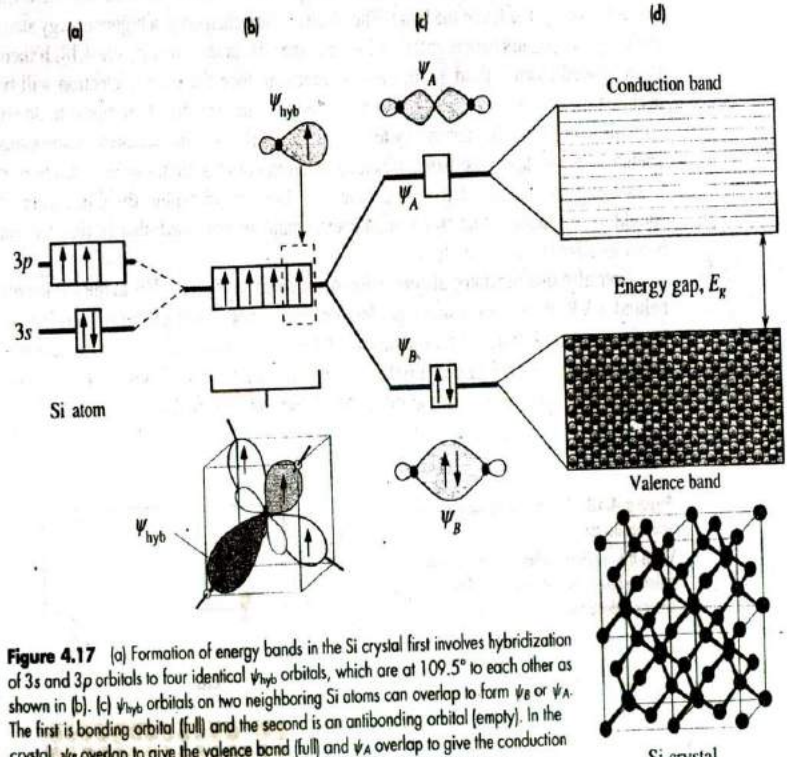


Figure 4.17 (a) Formation of energy bands in the Si crystal first involves hybridization of 3s and 3p orbitals to four identical ψ_{hyb} orbitals, which are at 109.5° to each other as shown in (b). (c) ψ_{hyb} orbitals on two neighboring Si atoms can overlap to form ψ_B or ψ_A. The first is bonding orbital (full) and the second is an antibonding orbital (empty). In the crystal, ψ_B overlap to give the valence band (full) and ψ_A overlap to give the conduction band (empty) (d).

completely empty and separated from the full valence band by a definite energy gap E_g. In this energy region, there are no states; therefore, the electron cannot have energy with a value within E_g. The energy band formed from $N\psi_A$ orbitals is a **conduction band** (CB), as also indicated in Figure 4.17.

The electronic states in the VB (and also in the CB) extend throughout the whole solid, because they result from $N\psi_B$ orbitals interfering and overlapping each other. As before $N\psi_B$, orbitals can overlap in N different ways to produce N distinct wavefunctions ψ_{vb} that extend throughout the solid. We cannot relate a particular electron to a particular bond or site because the wavefunctions ψ_{vb} corresponding to the VB energies are not concentrated at a single location. The electrical properties of solids are based on the fact that in solids, such as semiconductors and insulators, there are certain bands of allowed energies for the electrons, and these bands are separated by energy gaps, that is, bandgaps. The valence and conduction bands for the ideal Si crystal shown in Figure 4.17 are separated by an **energy gap,** or a **bandgap,** E_g, in which there are no allowed electron energy levels.

At temperatures above absolute zero, the atoms in a solid vibrate due to their thermal energy. Some of the atoms can acquire a sufficiently high energy from thermal fluctuations to strain and rupture their bonds. Physically, there is a possibility that the atomic vibration will impart sufficient energy to the electron for it to surmount the bonding energy and leave the bond. The electron must then enter a higher energy state. In the case of Si, this means entering a state in the CB, as shown in Figure 4.18. If there is an applied electric field $\mathcal{E}_x$ in the $+x$ direction, then the excited electron will be acted on by a force $-e\mathcal{E}_x$ and it will try to move in the $-x$ direction. For it to do so, there must be empty higher energy levels, so that as the electron accelerates and gains energy, it moves up in the band. When an electron collides with a lattice vibration, it loses the energy acquired from the field and drops down within the CB. Again, it should be emphasized that states in an energy band are extended; that is, the electron is not localized to any one atom.

Note also that the thermal generation of an electron from the VB to the CB leaves behind a VB state with a missing electron. This unoccupied electron state has an apparent positive charge, because this crystal region was neutral prior to the removal of the electron. The VB state with the missing electron is called a **hole** and is denoted h^+. The hole can "move" in the direction of the field by exchanging places with a

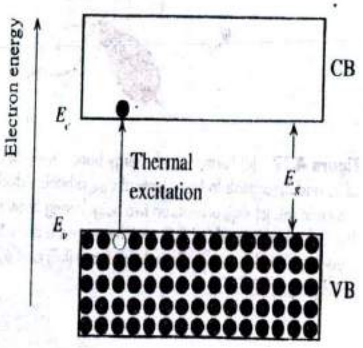


Figure 4.18 Energy band diagram of a semiconductor.
CB is the conduction band and VB is the valence band. At 0 K, the VB is full with all the valence electrons.

neighboring valence electron hence it contributes to conduction, as will be discussed in Chapter 5.

EXAMPLE 4.5

CUTOFF WAVELENGTH OF A Si PHOTODETECTOR What wavelengths of light can be absorbed by a Si photodetector given $E_g = 1.1$ eV? Can such a photodetector be used in fiber-optic communications at light wavelengths of 1.31 μm and 1.55 μm?

SOLUTION

The energy bandgap E_g of Si is 1.1 eV. A photon must have at least this much energy to excite an electron from the VB to the CB, where the electron can drift. Excitation corresponds to the breaking of a Si–Si bond. A photon of less energy does not get absorbed, because its energy will put the electron in the bandgap where there are no states. Thus, $hc/\lambda > E_g$ gives

$$\lambda < \frac{hc}{E_g} = \frac{(6.6 \times 10^{-34}\ \text{J s})(3 \times 10^{8}\ \text{m s}^{-1})}{(1.1\ \text{eV})(1.6 \times 10^{-19}\ \text{J/eV})}$$

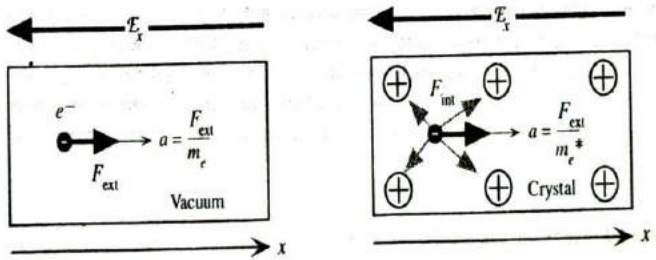
$$= 1.13 \times 10^{-6}\ \text{m} \quad \text{or} \quad 1.1\ \mu\text{m}$$

Since optical communications networks use wavelengths of 1.3 and 1.55 μm, these light waves will not be absorbed by Si and thus cannot be detected by a Si photodetector.

4.4 ELECTRON EFFECTIVE MASS

When an electric field $\mathcal{E}_x$ is applied to a metal, an electron near the Fermi level can gain energy from the field and move to higher energy levels, as shown in Figure 4.12. The external force $F_{ext} = e\mathcal{E}_x$ is in the x direction, and it drives the electron along x. The acceleration of the electron is still given by $a = F_{ext}/m_e$, where m_e is the mass of the electron in vacuum.

The law $F_{ext} = m_e a$ cannot strictly be valid for the electron inside a solid, because the electron interacts with the host ions and experiences internal forces F_{int} as it moves around, as depicted in Figure 4.19. The electron therefore has a PE that varies with distance. Recall that we interpret mass as inertial resistance against acceleration per unit



(a) An external force F_{ext} applied to an electron in a vacuum results in an acceleration $a_{vac} = F_{ext}/m_e$.

(b) An external force F_{ext} applied to an electron in a crystal results in an acceleration $a_{cryst} = F_{ext}/m_e^*$.

Figure 4.19

applied force. When an external force F_{ext} is applied to an electron in the vacuum level, as in Figure 4.19a, the electron will accelerate by an amount

$$a_{vac} = \frac{F_{ext}}{m_e} \tag{4.4}$$

as determined by its mass m_e in vacuum.

When the same force F_{ext} is applied to the electron inside a crystal, the acceleration of the electron will be different, because it will also experience internal forces, as shown in Figure 4.19b. Its acceleration in the crystal will be

$$a_{cryst} = \frac{F_{ext} + F_{int}}{m_e} \tag{4.5}$$

where F_{int} is the sum of all the internal forces acting on the electron, which is quite different than Equation 4.4. To the outside agent applying the force F_{ext}, the electron will appear to be exhibiting a different inertial mass, since its acceleration will be different. It would be most useful for the external agent if the effect of the internal forces in F_{int} could be accounted for in a simple way, and if the acceleration could be calculated from the external force F_{ext} alone, through something like Equation 4.4. This is indeed possible.

In a crystalline solid, the atoms are arranged periodically, and the variation of F_{int}, and hence the PE, or $V(x)$, of the electron with distance along x, is also periodic. In principle, then, the effect on the electron motion can be predicted and accounted for. When we solve the Schrödinger equation with the periodic PE, or $V(x)$, we essentially obtain the effect of these internal forces on the electron motion. It has been found that when the electron is in a band that is not full, we can still use Equation 4.4, but instead of the mass in vacuum m_e, we must use the effective mass m_e^* of the electron in that particular crystal. The effective mass is a quantum mechanical quantity that behaves in the same way as the inertial mass in classical mechanics. The acceleration of the electron in the crystal is then simply

$$a_{cryst} = \frac{F_{ext}}{m_e^*} \tag{4.6}$$

The effects of all internal forces are incorporated into m_e^*. It should be emphasized that m_e^*, is obtained theoretically from the solution of the Schrödinger equation for the electron in a particular crystal, a task that is by no means trivial. However, the effective mass can be readily measured. For some of the familiar metals, m_e^* is very close to m_e. For example, in copper, $m_e^* = m_e$ for all practical purposes, whereas in lithium $m_e^* = 1.28m_e$, as shown in Table 4.2. On the other hand, m_e^* for many metals and

Table 4.2 Effective mass m_e^* of electrons in some metals

Metal	Ag	Au	Bi	Cu	K	Li	Na	Ni	Pt	Zn
$\frac{m_e^*}{m_e}$	0.99	1.10	0.047	1.01	1.12	1.28	1.2	28	13	0.85

semiconductors is appreciably different than the electron mass in vacuum and can even be negative. (m_e^* depends on the properties of the band that contains the electron. This is further discussed in Section 5.11.)

4.5 DENSITY OF STATES IN AN ENERGY BAND

Although we know there are many energy levels (perhaps $\sim 10^{23}$) in a given band, we have not yet considered how many states (or electron wavefunctions) there are per unit energy per unit volume in that band. Consider the following *intuitive* argument. The crystal will have N atoms and there will be N electron wavefunctions $\psi_1, \psi_2, \ldots, \psi_N$ that represent the electron within the whole crystal. These wavefunctions are constructed from N different combinations of atomic wavefunctions, $\psi_A, \psi_B, \psi_C, \ldots$ as schematically illustrated in Figure 4.20a,[4] starting with

$$\psi_1 = \psi_A + \psi_B + \psi_C + \psi_D + \cdots$$

all the way to alternating signs

$$\psi_N = \psi_A - \psi_B + \psi_C - \psi_D + \cdots$$

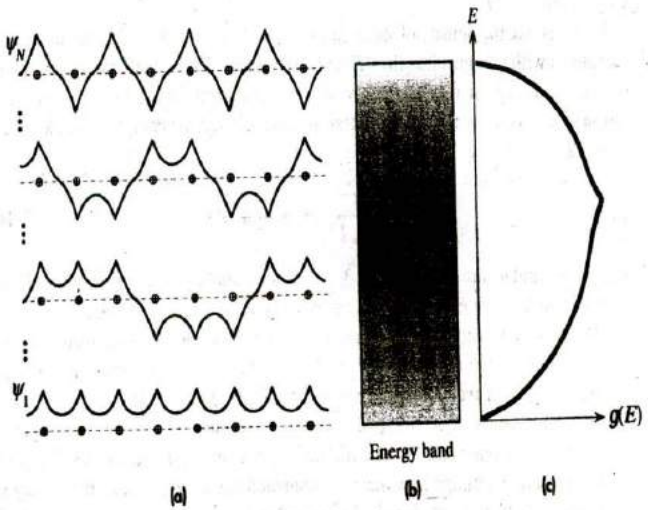


Figure 4.20

(a) In the solid there are N atoms and N extended electron wavefunctions from ψ_1 all the way to ψ_N. There are many wavefunctions, states, that have energies that fall in the central regions of the energy band.

(b) The distribution of states in the energy band; darker regions have a higher number of states.

(c) Schematic representation of the density of states $g(E)$ versus energy E.

[4] This intuitive argument, as schematically depicted in Figure 4.20a, is obviously highly simplified because the solid is three-dimensional (3-D) and we should combine the atomic wavefunctions not on a linear chain but on a 3-D lattice. In the 3-D case there are large numbers of wavefunctions with energies that fall in the central regions of the band.

and there are $N(\sim 10^{23})$ combinations. The lowest-energy wavefunction will be ψ_1 constructed by adding all atomic wavefunctions (all in phase), and the highest-energy wavefunction will be ψ_N from alternating the signs of the atomic wavefunctions, which will have the highest number of nodes. Between these two extremes, especially around $N/2$, there will be many combinations that will have comparable energies and fall near the middle of the band. (By analogy, if we arrange $N = 10$ coins by heads and tails, there will be many combinations of coins in which there are 5 heads and 5 tails, and only one combination in which there are 10 heads or 10 tails.) We therefore expect the number of energy levels, each corresponding to an electron wavefunction in the crystal, in the central regions of the band to be very large as depicted in Figure 4.20b and c.

Figure 4.20c illustrates schematically how the energy and volume density of electronic states change across an energy band. We define the **density of states** $g(E)$ such that $g(E)\,dE$ is the number of states (*i.e.*, wavefunctions) in the energy interval E to $(E + dE)$ per unit volume of the sample. Thus, the number of states per unit volume up to some energy E' is

$$S_v(E') = \int_0^{E'} g(E)\,dE \qquad [4.7]$$

which is called the total number of states per unit volume with energies less than E'. This is denoted $S_v(E')$.

To determine the density of states function $g(E)$, we must first determine the number of states with energies less than E' in a given band. This is tantamount to calculating $S_v(E')$ in Equation 4.7. Instead, we will improvise and use the energy levels for an electron in a three-dimensional potential well. Recall that the energy of an electron in a cubic *PE* well of size L is given by

$$E = \frac{h^2}{8m_e L^2}\left(n_1^2 + n_2^2 + n_3^2\right) \qquad [4.8]$$

where n_1, n_2, and n_3 are integers $1, 2, 3, \ldots$. The spatial dimension L of the well now refers to the size of the entire solid, as the electron is confined to be somewhere inside that solid. Thus, L is very large compared to atomic dimensions, which means that the separation between the energy levels is very small. We will use Equation 4.8 to describe the energies of **free electrons** inside the solid (as in a metal).

Each combination of n_1, n_2, and n_3 is one electron orbital state. For example, $\psi_{n_1,n_2,n_3} = \psi_{1,1,2}$ is one possible orbital state. Suppose that in Equation 4.8 E is given as E'. We need to determine how many combinations of n_1, n_2, n_3 (*i.e.*, how many ψ) have energies less than E', as given by Equation 4.8. Assume that $(n_1^2 + n_2^2 + n_3^2) = n'^2$. The object is to enumerate all possible choices of integers for n_1, n_2, and n_3 that satisfy $n_1^2 + n_2^2 + n_3^2 \leq n'^2$.

The two-dimensional case is easy to solve. Consider $n_1^2 + n_2^2 \leq n'^2$ and the two-dimensional ***n*-space** where the axes are n_1 and n_2, as shown in Figure 4.21. The two-dimensional space is divided by lines drawn at $n_1 = 1, 2, 3, \ldots$ and $n_2 = 1, 2, 3, \ldots$ into infinitely many boxes (squares), each of which has a unit area and represents a possible state ψ_{n_1,n_2}. For example, the state $n_1 = 1, n_2 = 3$ is shaded, as is that for $n_1 = 2, n_2 = 2$.

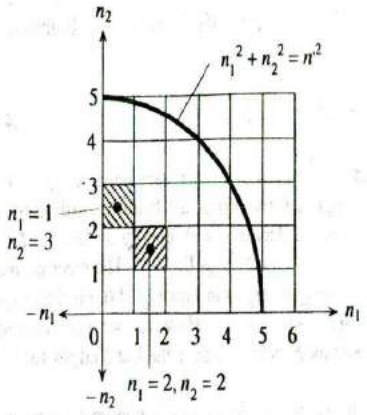


Figure 4.21 Each state, or electron wavefunction in the crystal, can be represented by a box at n_1, n_2.

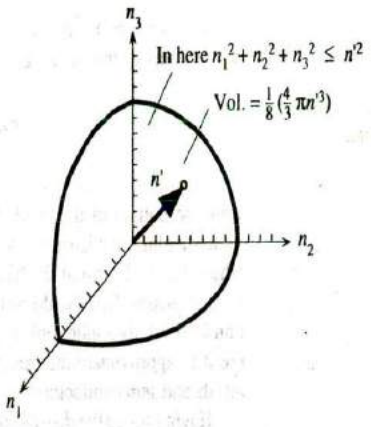


Figure 4.22 In three dimensions, the volume defined by a sphere of radius n' and the positive axes n_1, n_2, and n_3, contains all the possible combinations of positive n_1, n_2, and n_3 values that satisfy $n_1^2 + n_2^2 + n_3^2 \leq n'^2$.

Clearly, the area contained by n_1, n_2 and the circle defined by $n'^2 = n_1^2 + n_2^2$ (just like $r^2 = x^2 + y^2$) is the number of states that satisfy $n_1^2 + n_2^2 \leq n'^2$. This area is $\frac{1}{4}(\pi n'^2)$.

In the three-dimensional case, $n_1^2 + n_2^2 + n_3^2 \leq n'^2$ is required, as indicated in Figure 4.22. This is the volume contained by the positive n_1, n_2, and n_3 axes and the surface of a sphere of radius n'. Each state has a unit volume, and within the sphere, $n_1^2 + n_2^2 + n_3^2 \leq n'^2$ is satisfied. Therefore, the number of orbital states $S_{orb}(n')$ within this volume is given by

$$S_{orb}(n') = \frac{1}{8}\left(\frac{4}{3}\pi n'^3\right) = \frac{1}{6}\pi n'^3$$

Each orbital state can take two electrons with opposite spins, which means that the number of states, including spin, is given by

$$S(n') = 2S_{orb}(n') = \frac{1}{3}\pi n'^3$$

We need this expression in terms of energy. Substituting $n'^2 = 8m_e L^2 E'/h^2$ from Equation 4.8 in $S(n')$, we get

$$S(E') = \frac{\pi L^3 (8m_e E')^{3/2}}{3h^3}$$

Since L^3 is the physical volume of the solid, the number of states per unit volume $S_v(E')$ with energies $E \leq E'$ is

$$S_v(E') = \frac{\pi(8m_e E')^{3/2}}{3h^3} \qquad [4.9]$$

Furthermore, from Equation 4.7, $dS_v/dE = g(E)$. By differentiating Equation 4.9 with respect to energy, we get

Density of states

$$g(E) = (8\pi 2^{1/2})\left(\frac{m_e}{h^2}\right)^{3/2} E^{1/2} \qquad [4.10]$$

Equation 4.10 shows that the density of states $g(E)$ increases with energy as $E^{1/2}$ from the bottom of the band. As we approach the top of the band, according to our understanding in Figure 4.20d, $g(E)$ should decrease with energy as $(E_{top} - E)^{1/2}$, where E_{top} is the top of the band, so that as $E \to E_{top}$, $g(E) \to 0$. The electron mass m_e in Equation 4.10 should be the *effective mass* m_e^* as in Equation 4.6. Further, Equation 4.10 strictly applies only to *free electrons* in a crystal. However, we will frequently use it to approximate the true $g(E)$ versus E behavior near the band edges for both metals and semiconductors.

Having found the distribution of the electron energy states, Equation 4.10, we now wish to determine the number of states that actually contain electrons; that is, the probability of finding an electron at an energy level E. This is given by the Fermi–Dirac statistics.

As an example, one convenient way of calculating the population of a city is to find the density of houses in that city (*i.e.*, the number of houses per unit area), multiply that by the probability of finding a human in a house, and finally, integrate the result over the area of the city. The problem is working out the chances of actually finding someone at home, using a mathematical formula. For those who like analogies, if $g(A)$ is the density of houses and $f(A)$ is the probability that a house is occupied, then the population of the city is

$$n = \int_{City} f(A)g(A)\, dA$$

where the integration is done over the entire area of the city. This equation can be used to find the number of electrons per unit volume within a band. If E is the electron energy and $f(E)$ is the probability that a state with energy E is occupied, then

$$n = \int_{Band} f(E)g(E)\, dE$$

where the integration is done over all the energies of the band.

EXAMPLE 4.6 **X-RAY EMISSION AND THE DENSITY OF STATES IN A METAL** Consider what happens when a metal such as Al is bombarded with high-energy electrons. The inner atomic energy levels are not disturbed in the solid, so these inner levels remain as distinct single levels, each one localized to the parent atom. When an energetic electron hits an electron in one of the inner atomic energy levels, it knocks out this electron from the metal leaving behind a vacancy in the inner core as depicted in Figure 4.23a. An electron in the energy band of the solid can then fall down to occupy this empty state and emit a photon in the process. The energy difference between the energies in the band and the inner atomic level is in the X-ray range, so the emitted photon is an X-ray photon. Since electrons occupy the band from the bottom E_B to the Fermi level E_F, the

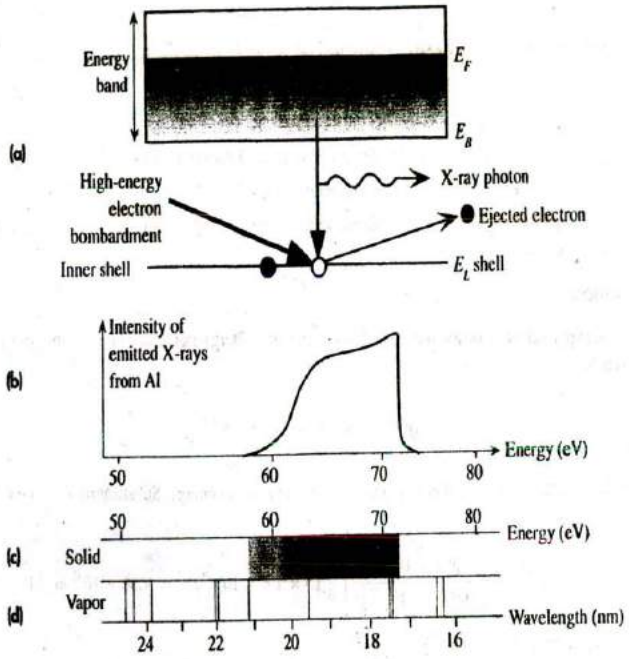


Figure 4.23

(a) High-energy electron bombardment knocks out an electron from the closed inner L shell leaving an empty state. An electron from the energy band of the metal drops into the L shell to fill the vacancy and emits a soft X-ray photon in the process.

(b) The spectrum (intensity versus photon energy) of soft X-ray emission from a metal involves a range of energies corresponding to transitions from the bottom of the band and from the Fermi level to the L shell. The intensity increases with energy until around E_F where it drops sharply.

(c) and (d) contrast the emission spectra from a solid and vapor (isolated gas atoms).

emitted X-ray photons have a range of energies corresponding to transitions from E_B and E_F to the inner atomic level as shown in Figure 4.23b. These energies are in the soft X-ray spectrum. We assumed that the levels above E_F are almost empty, though, undoubtedly, there is no sharp transition from full to empty levels at E_F. Further, since the density of states increases from E_B toward E_F, there are more and more electrons that can fall down to the atomic level as we move from E_B toward E_F. Therefore the intensity of the emitted X-ray radiation increases with energy until the energy reaches the Fermi level beyond which there are only a small number of electrons available for the transit. Figure 4.23c and d contrasts the emission spectra from an aluminum crystal (solid) and its vapor. The line spectra from a vapor become an emission band in the spectrum of the solid.

The X-ray intensity emitted from Al in Figure 4.23 starts to rise at around 60 eV and then sharply falls around 72 eV. Thus the energy range is 12 eV, which represents approximately the Fermi energy with respect to the bottom of the band, that is, $E_F \approx 72 - 60 = 12$ eV with respect to E_B.

EXAMPLE 4.7 **DENSITY OF STATES IN A BAND** Given that the width of an energy band is typically ~10 eV, calculate the following, in per cm^3 and per eV units:

a. The density of states at the center of the band.

b. The number of states per unit volume within a small energy range kT about the center.

c. The density of states at kT above the bottom of the band.

d. The number of states per unit volume within a small energy range of kT to $2kT$ from the bottom of the band.

SOLUTION

The density of states, or the number of states per unit energy range per unit volume $g(E)$, is given by

$$g(E) = (8\pi 2^{1/2})\left(\frac{m_e}{h^2}\right)^{3/2} E^{1/2}$$

which gives the number of states per cubic meter per Joule of energy. Substituting $E = 5$ eV, we have

$$g_{center} = (8\pi 2^{1/2})\left[\frac{9.1 \times 10^{-31}}{(6.626 \times 10^{-34})^2}\right]^{3/2} (5 \times 1.6 \times 10^{-19})^{1/2} = 9.50 \times 10^{46}\,m^{-3}J^{-1}$$

Converting to cm^{-3} and eV^{-1}, we get

$$g_{center} = (9.50 \times 10^{46}\,m^{-3}J^{-1})(10^{-6}\,m^3\,cm^{-3})(1.6 \times 10^{-19}\,J\,eV^{-1})$$

$$= 1.52 \times 10^{22}\,cm^{-3}\,eV^{-1}$$

If δE is a small energy range (such as kT), then, by definition, $g(E)\,\delta E$ is the number of states per unit volume in δE. To find the number of states per unit volume within kT at the center of the band, we multiply g_{center} by kT or $(1.52 \times 10^{22}\,cm^{-3}\,eV^{-1})(0.026\,eV)$ to get $3.9 \times 10^{20}\,cm^{-3}$. This is not a small number!

At kT above the bottom of the band, at 300 K ($kT = 0.026$ eV), we have

$$g_{0.026} = (8\pi 2^{1/2})\left[\frac{9.1 \times 10^{-31}}{(6.626 \times 10^{-34})^2}\right]^{3/2} (0.026 \times 1.6 \times 10^{-19})^{1/2}$$

$$= 6.84 \times 10^{45}\,m^{-3}J^{-1}$$

Converting to cm^{-3} and eV^{-1} we get

$$g_{0.026} = (6.84 \times 10^{45}\,m^{-3}J^{-1})(10^{-6}\,m^3\,cm^{-3})(1.6 \times 10^{-19}\,J\,eV^{-1})$$

$$= 1.10 \times 10^{21}\,cm^{-3}\,eV^{-1}$$

Within kT, the volume density of states is

$$(1.10 \times 10^{21}\,cm^{-3}\,eV^{-1})(0.026\,eV) = 2.8 \times 10^{19}\,cm^{-3}$$

This is very close to the bottom of the band and is still very large.

TOTAL NUMBER OF STATES IN A BAND

EXAMPLE 4.8

a. Based on the overlap of atomic orbitals to form the electron wavefunction in the crystal, how many states should there be in a band?

b. Consider the density of states function

$$g(E) = (8\pi 2^{1/2})\left(\frac{m_e}{h^2}\right)^{3/2} E^{1/2}$$

By integrating $g(E)$, estimate the total number of states in a band per unit volume, and compare this with the atomic concentration for silver. For silver, we have $E_{FO} = 5.5$ eV and $\Phi = 4.5$ eV. (Note that "state" means a distinct wavefunction, including spin.)

SOLUTION

a. We know that when N atoms come together to form a solid, N atomic orbitals can overlap N different ways to produce N orbitals or $2N$ states in the crystal, since each orbital has two states, spin up and spin down. These states form the band.

b. For silver, $E_{FO} = 5.5$ eV and $\Phi = 4.5$ eV, so the width of the energy band is 10 eV. To estimate the total volume density of states, we assume that the density of states $g(E)$ reaches its maximum at the center of the band $E = E_{center} = 5$ eV. Integrating $g(E)$ from the bottom of the band, $E = 0$, to the center, $E = E_{center}$, yields the number of states per unit volume up to the center of the band. This is half the total number of states in the whole band, that is, $\frac{1}{2}S_{band}$, where S_{band} is the number of states per unit volume in the band and is determined by

$$\frac{1}{2}S_{band} = \int_0^{E_{center}} g(E)\,dE = \frac{16\pi 2^{1/2}}{3}\left(\frac{m_e}{h^2}\right)^{3/2} E_{center}^{3/2}$$

or

$$\frac{1}{2}S_{band} = \frac{16\pi 2^{1/2}}{3}\left[\frac{9.1 \times 10^{-31}\ \text{kg}}{(6.626 \times 10^{-34}\ \text{J s})^2}\right]^{3/2} (5\ \text{eV} \times 1.6 \times 10^{-19}\ \text{J/eV})^{3/2}$$

$$= 5.08 \times 10^{28}\ \text{m}^{-3} = 5.08 \times 10^{22}\ \text{cm}^{-3}$$

Thus

$$S_{band} = 10.16 \times 10^{22}\ \text{states cm}^{-3}$$

We must now calculate the number of atoms per unit volume in silver. Given the density $d = 10.5$ g cm^{-3} and the atomic mass $M_{at} = 107.9$ g mol^{-1} of silver, the atomic concentration is

$$n_{Ag} = \frac{dN_A}{M_{at}} = 5.85 \times 10^{22}\ \text{atoms cm}^{-3}$$

As expected, the density of states is almost twice the atomic concentration, even though we used a crude approximation to estimate the density of states.

4.6 STATISTICS: COLLECTIONS OF PARTICLES

4.6.1 BOLTZMANN CLASSICAL STATISTICS

Given a collection of particles in random motion and colliding with each other,[5] we need to determine the concentration of particles in the energy range E to $(E + dE)$. Consider the process shown in Figure 4.24, in which two electrons with energies E_1 and E_2 interact and then move off in different directions, with energies E_3 and E_4. Let the probability of an electron having an energy E be $P(E)$, where $P(E)$ is the fraction of electrons with an energy E. Assume there are no restrictions to the electron energies, that is, we can ignore the Pauli exclusion principle. The probability of this event is then $P(E_1)P(E_2)$. The probability of the reverse process, in which electrons with energies E_3 and E_4 interact, is $P(E_3)P(E_4)$. Since we have thermal equilibrium, that is, the system is in equilibrium, the forward process must be just as likely as the reverse process, so

$$P(E_1)P(E_2) = P(E_3)P(E_4) \qquad [4.11]$$

Furthermore, the energy in this collision must be conserved, so we also need

$$E_1 + E_2 = E_3 + E_4 \qquad [4.12]$$

We therefore need to find the $P(E)$ that satisfies both Equations 4.11 and 4.12. Based on our experience with the distribution of energies among gas molecules, we can guess that the solution for Equations 4.11 and 4.12 would be

Boltzmann probability function

$$P(E) = A \exp\left(-\frac{E}{kT}\right) \qquad [4.13]$$

where k is the Boltzmann constant, T is the temperature, and A is a constant. We can show that Equation 4.13 is a solution to Equations 4.11 and 4.12 by a simple substitution. Equation 4.13 is the **Boltzmann probability function** and is shown in Figure 4.25. The probability of finding a particle at an energy E therefore decreases exponentially with energy. We assume, of course, that any number of particles may have a given energy E. In other words, there is no restriction such as permitting only one particle per state at an energy E, as in the Pauli exclusion principle. The term kT appears in Equation 4.13 because the average energy as calculated by using $P(E)$ then agrees with experiments. (There is no kT in Equations 4.11 and 4.12.)

Suppose that we have N_1 particles at energy level E_1 and N_2 particles at a higher energy E_2. Then, by Equation 4.13, we have

Boltzmann statistics

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{kT}\right) \qquad [4.14]$$

[5] From Chapter 1, we can associate this with the kinetic theory of gases. The energies of the gas molecules, which are moving around randomly, are distributed according to the Maxwell–Boltzmann statistics.

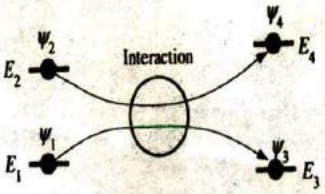


Figure 4.24 Two electrons with initial wavefunctions ψ_1 and ψ_2 at E_1 and E_2 interact and end up at different energies E_3 and E_4. Their corresponding wavefunctions are ψ_3 and ψ_4.

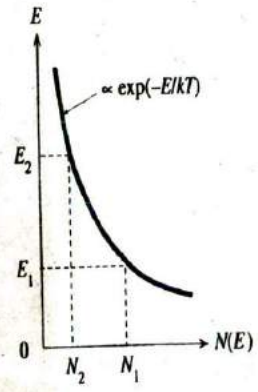


Figure 4.25 The Boltzmann energy distribution describes the statistics of particles, such as electrons, when there are many more available states than the number of particles.

If $E_2 - E_1 \gg kT$, then N_2 can be orders of magnitude smaller than N_1. As the temperature increases, N_2/N_1 also increases. Therefore, increasing the temperature populates the higher energy levels.

Classical particles obey the Boltzmann statistics. Whenever there are many more states (by orders of magnitude) than the number of particles, the likelihood of two particles having the same set of quantum numbers is negligible and we do not have to worry about the Pauli exclusion principle. In these cases, we can use the Boltzmann statistics. An important example is the statistics of electrons in the conduction band of a semiconductor where, in general, there are many more states than electrons.

4.6.2 FERMI–DIRAC STATISTICS

Now consider the interaction for which no two electrons can be in the same quantum state, which is essentially obedience to the Pauli exclusion principle, as shown in Figure 4.24. We assume that we can have only one electron in a particular quantum state ψ (including spin) associated with the energy value E. We therefore need those states that have energies E_3 and E_4 to be not occupied. Let $f(E)$ be the probability that an electron is in such a state, with energy E in this new interaction environment. The probability of the forward event in Figure 4.24 is

$$f(E_1)f(E_2)[1 - f(E_3)][1 - f(E_4)]$$

The square brackets represent the probability that the states with energies E_3 and E_4 are empty. In thermal equilibrium, the reverse process, the electrons with E_3 and E_4 interacting to transfer to E_1 and E_2, has just as equal a likelihood as the forward process.

Paul Adrien Maurice Dirac (1902–1984) received the 1933 Nobel prize for physics with Erwin Schrödinger. His first degree was in electrical engineering from Bristol University. He obtained his PhD in 1926 from Cambridge University under Ralph Fowler.

SOURCE: Courtesy of AIP Emilio Segrè Visual Archives.

Thus, $f(E)$ must satisfy the equation

$$f(E_1)f(E_2)[1 - f(E_3)][1 - f(E_4)] = f(E_3)f(E_4)[1 - f(E_1)][1 - f(E_2)] \quad [4.15]$$

In addition, for energy conservation, we must have

$$E_1 + E_2 = E_3 + E_4 \quad [4.16]$$

By an "intelligent guess," the solution to Equations 4.15 and 4.16 is

$$f(E) = \frac{1}{1 + A\exp\left(\dfrac{E}{kT}\right)} \quad [4.17]$$

where A is a constant. You can check that this is a solution by substituting Equation 4.17 into 4.15 and using Equation 4.16. The reason for the term kT in Equation 4.17 is not obvious from Equations 4.15 and 4.16. It appears in Equation 4.17 so that the mean properties of this system calculated by using $f(E)$ agree with experiments. Letting $A = \exp(-E_F/kT)$, we can write Equation 4.17 as

Fermi–Dirac statistics

$$f(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{kT}\right)} \quad [4.18]$$

where E_F is a constant called the **Fermi energy.** The probability of finding an electron in a state with energy E is given by Equation 4.18, which is called the **Fermi–Dirac function.**

The behavior of the Fermi–Dirac function is shown in Figure 4.26. Note the effect of temperature. As T increases, $f(E)$ extends to higher energies. At energies of a few kT (0.026 eV) above E_F, $f(E)$ behaves almost like the Boltzmann function

$$f(E) = \exp\left[-\frac{(E - E_F)}{kT}\right] \qquad (E - E_F) \gg kT \quad [4.19]$$

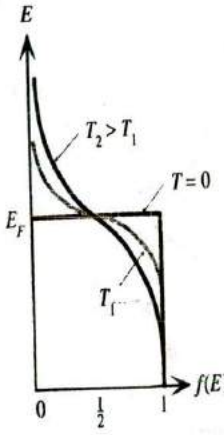


Figure 4.26
The Fermi-Dirac function f(E) describes the statistics of electrons in a solid. The electrons interact with each other and the environment, obeying the Pauli exclusion principle.

Above absolute zero, at $E = E_F$, $f(E_F) = \frac{1}{2}$. We define the Fermi energy as that energy for which the probability of occupancy $f(E_F)$ equals $\frac{1}{2}$. The approximation to $f(E)$ in Equation 4.19 at high energies is often referred to as the **Boltzmann tail** to the Fermi–Dirac function.

4.7 QUANTUM THEORY OF METALS

4.7.1 FREE ELECTRON MODEL[6]

We know that the number of states $g(E)$ for an electron, per unit energy per unit volume, increases with energy as $g(E) \propto E^{1/2}$. We have also calculated that the probability of an electron being in a state with an energy E is the Fermi–Dirac function $f(E)$. Consider the energy band diagram for a metal and the density of states $g(E)$ for that band, as shown in Figure 4.27a and b, respectively.

At absolute zero, all the energy levels up to E_F are full. At 0 K, $f(E)$ has the step form at E_F (Figure 4.26). This clarifies why E_F in $f(E)$ is termed the Fermi energy. At 0 K, $f(E) = 1$ for $E < E_F$, and $f(E) = 0$ for $E > E_F$, so at 0 K, E_F separates the empty and full energy levels. This explains why we restricted ourselves to 0 K or thereabouts when we introduced E_F in the band theory of metals.

At some finite temperature, $f(E)$ is *not* zero beyond E_F, as indicated in Figure 4.27c. This means that some of the electrons are excited to, and thereby occupy, energy levels above E_F. If we multiply $g(E)$, by $f(E)$, we obtain the number of electrons per unit energy per unit volume, denoted n_E. The distribution of electrons in the energy levels is described by $n_E = g(E)\, f(E)$.

Since $f(E) = 1$ for $E \ll E_F$, the states near the bottom of the band are all occupied; thus, $n_E \propto E^{1/2}$ initially. As E passes through E_F, $f(E)$ starts decreasing

[6] The free electron model of metals is also known as the Sommerfeld model.

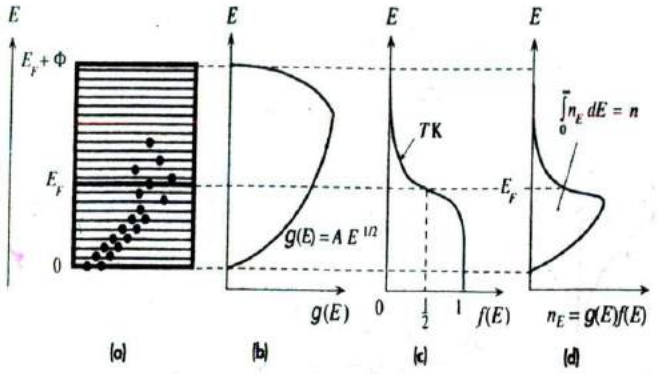


Figure 4.27

(a) Above 0 K, due to thermal excitation, some of the electrons are at energies above E_F.

(b) The density of states, $g(E)$ versus E in the band.

(c) The probability of occupancy of a state at an energy E is $f(E)$.

(d) The product $g(E)f(E)$ is the number of electrons per unit energy per unit volume, or the electron concentration per unit energy. The area under the curve on the energy axis is the concentration of electrons in the band.

sharply. As a result, n_E takes a turn and begins to decrease sharply as well, as depicted in Figure 4.27d.

In the small energy range E to $(E + dE)$, there are $n_E\, dE$ electrons per unit volume. When we sum all $n_E\, dE$ from the bottom to the top of the band ($E = 0$ to $E = E_F + \Phi$), we get the total number of valence electrons per unit volume, n, in the metal, as follows:

$$n = \int_0^{\text{Top of band}} n_E\, dE = \int_0^{\text{Top of band}} g(E) f(E)\, dE \qquad [4.20]$$

Since $f(E)$ falls very sharply when $E > E_F$, we can carry the integration to $E = \infty$, rather than to $(E_F + \Phi)$, because $f \to 0$ when $E \gg E_F$. Putting in the functional forms of $g(E)$ and $f(E)$ (*e.g.*, from Equations 4.10 and 4.18), we obtain

$$n = \frac{8\pi 2^{1/2} m_e^{3/2}}{h^3} \int_0^{\infty} \frac{E^{1/2} dE}{1 + \exp\left(\dfrac{E - E_F}{kT}\right)} \qquad [4.21]$$

If we could integrate this, we would obtain an expression relating n and E_F. At 0 K, however, $E_F = E_{F0}$ and the integrand exists only for $E < E_{F0}$. If we integrate at 0 K, Equation 4.21 yields

Fermi energy at $T = 0$ K

$$E_{F0} = \left(\frac{h^2}{8m_e}\right)\left(\frac{3n}{\pi}\right)^{2/3} \qquad [4.22]$$

It may be thought that E_F is temperature independent, since it was sketched that way in Figure 4.26. However, in our derivation of the Fermi–Dirac statistics, there was no restriction that demanded this. Indeed, since the number of electrons in a band is fixed, E_F at a temperature T is implicitly determined by Equation 4.21, which can be solved to express E_F in terms of n and T. It turns out that at 0 K, E_F is given by Equation 4.22, and it changes very little with temperature. In fact, by utilizing various mathematical approximations, it is not too difficult to integrate Equation 4.21 to obtain the **Fermi energy** at a temperature T, as follows:

$$E_F(T) = E_{FO}\left[1 - \frac{\pi^2}{12}\left(\frac{kT}{E_{FO}}\right)^2\right] \qquad [4.23]$$

Fermi energy at T (K)

which shows that $E_F(T)$ is only weakly temperature dependent, since $E_{FO} \gg kT$.

The Fermi energy has an important significance in terms of the average energy E_{av} of the conduction electrons in a metal. In the energy range E to $(E + dE)$, there are $n_E\,dE$ electrons with energy E. The average energy of an electron will therefore be

$$E_{av} = \frac{\int E n_E\,dE}{\int n_E\,dE} \qquad [4.24]$$

If we substitute $g(E)f(E)$ for n_E and integrate, the result at 0 K is

$$E_{av}(0) = \frac{3}{5}E_{FO} \qquad [4.25]$$

Average energy per electron at 0 K

Above absolute zero, the **average energy** is approximately

$$E_{av}(T) = \frac{3}{5}E_{FO}\left[1 + \frac{5\pi^2}{12}\left(\frac{kT}{E_{FO}}\right)^2\right] \qquad [4.26]$$

Average energy per electron at T (K)

Since $E_{FO} \gg kT$, the second term in the square brackets is much smaller than unity, and $E_{av}(T)$ shows only a very weak temperature dependence. Furthermore, in our model of the metal, the electrons are free to move around within the metal, where their potential energy PE is zero, whereas outside the metal, it is $E_F + \Phi$ (Figure 4.11). Therefore, their energy is purely kinetic. Thus, Equation 4.26 gives the average KE of the electrons in a metal

$$\frac{1}{2}m_e v_e^2 = E_{av} \approx \frac{3}{5}E_{FO}$$

where v_e is the root mean square (rms) speed of the electrons, which is simply called the **effective speed.** The effective speed v_e depends on the Fermi energy E_{FO} and is relatively insensitive to temperature. Compare this with the behavior of molecules in an ideal gas. In that case, the average $KE = \frac{3}{2}kT$, so $\frac{1}{2}mv^2 = \frac{3}{2}kT$. Clearly, the average speed of molecules in a gas increases with temperature.

The relationship $\frac{1}{2}mv_e^2 \approx \frac{3}{5}E_{FO}$ is an important conclusion that comes from the application of quantum mechanical concepts, ideas that lead to $g(E)$ and $f(E)$ and so on. It cannot be proved without invoking quantum mechanics. The fact that the average electronic speed is nearly constant is the only way to explain the observation that the resistivity of a metal is proportional to T (and not $T^{3/2}$), as we saw in Chapter 2.

4.7.2 CONDUCTION IN METALS

We know from our energy band discussions that in metals only those electrons in a small range ΔE around the Fermi energy E_F contribute to electrical conduction as shown in Figure 4.12c. The concentration n_F of these electrons is approximately $g(E_F)\,\Delta E$ inasmuch as ΔE is very small. The electron a moves to a', as shown in Figure 4.12b and c, and then it is scattered to an empty state above b'. In steady conduction, all the electrons in the energy range ΔE that are moving to the right are not canceled by any moving to the left and hence contribute to the current. An electron at the bottom of the ΔE range gains energy ΔE to move a' in a time interval Δt that corresponds to the scattering time τ. It gains a momentum Δp_x. Since $\Delta p_x/\Delta t =$ external force $= e\mathcal{E}_x$, we have $\Delta p_x = \tau e\mathcal{E}_x$. The electron a has an energy $E = p_x^2/(2m_e^*)$ which we can differentiate to obtain ΔE when the momentum changes by Δp_x,

$$\Delta E = \frac{p_x}{m_e^*}\Delta p_x = \frac{(m_e^* v_F)}{m_e^*}(\tau e\mathcal{E}_x) = e v_F \tau \mathcal{E}_x$$

The current J_x is due to all the electrons in the range ΔE which are moving toward the right in Figure 4.12c,

$$J_x = e n_F v_F = e[g(E_F)\,\Delta E]v_F = e[g(E_F) e v_F \tau \mathcal{E}_x]v_F = e^2 v_F^2 \tau\, g(E_F)\mathcal{E}_x$$

The conductivity is therefore

$$\sigma = e^2 v_F^2 \tau\, g(E_F)$$

However, the numerical factor is wrong because Figure 4.12c considers only a hypothetical one-dimensional crystal. In a three-dimensional crystal, the conductivity is one-third of the conductivity value just determined:

Conductivity of Fermi-level electrons

$$\sigma = \frac{1}{3} e^2 v_F^2 \tau\, g(E_F) \qquad \textbf{[4.27]}$$

This conductivity expression is in sharp contrast with the classical expression in which all the electrons contribute to conduction. According to Equation 4.27, what is important is the density of states at the Fermi energy $g(E_F)$. For example, Cu and Mg are metals with valencies I and II. Classically, Cu and Mg atoms each contribute one and two conduction electrons, respectively, into the crystal. Thus, we would expect Mg to have higher conductivity. However, the Fermi level in Mg is where the top tail of the $3s$ band overlaps the bottom tail of the $3p$ band where the density of states is small. In Cu, on the other hand, E_F is nearly in the middle of the $4s$ band where the density of states is high. Thus, Mg has a lower conductivity than Cu.

The scattering time τ in Equation 4.27 assumes that the scattered electrons at E_F remain in the same energy band. In certain metals, there are two different energy bands that overlap at E_F. For example, in Ni (see Figure 4.61), $3d$ and $4s$ bands overlap at E_F. An electron can be scattered from the $4s$ to the $3d$ band, and vice versa. Electrons in the $3d$ band have very low drift mobilities and effectively do not contribute to conduction, so only $g(E_F)$ of the $4s$ band operates in Equation 4.27.

Since 4*s* to 3*d* band scattering is an additional scattering mechanism, by virtue of Matthiessen's rule, the scattering time τ for the 4*s* band electrons is shortened. Thus, Ni has poorer conductivity than Cu.

In deriving Equation 4.27 we did not assume a particular density of states model. If we now apply the *free electron model* for $g(E_F)$ as in Equation 4.10, and also relate E_F to the total number of conduction electrons per unit volume n as in Equation 4.22, we would find that the conductivity is the same as the **Drude model**, that is,

$$\sigma = \frac{e^2 n \tau}{m_e} \qquad [4.28]$$

Drude model and free electrons

EXAMPLE 4.9

MEAN SPEED OF CONDUCTION ELECTRONS IN A METAL Calculate the Fermi energy E_{F0} at 0 K for copper and estimate the average speed of the conduction electrons in Cu. The density of Cu is 8.96 g cm^{-3} and the relative atomic mass (atomic weight) is 63.5.

SOLUTION

Assuming each Cu atom donates one free electron, we can find the concentration of electrons from the density d, atomic mass M_m, and Avogadro's number N_A, as follows:

$$n = \frac{dN_A}{M_{at}} = \frac{8.96 \times 6.02 \times 10^{23}}{63.5}$$

$$= 8.5 \times 10^{22}\ \text{cm}^{-3} \quad \text{or} \quad 8.5 \times 10^{28}\ \text{m}^{-3}$$

The Fermi energy at 0 K is given by Equation 4.22:

$$E_{F0} = \left(\frac{h^2}{8m_e}\right)\left(\frac{3n}{\pi}\right)^{2/3}$$

Substituting $n = 8.5 \times 10^{28}$ m^{-3} and the values for h and m_e, we obtain

$$E_{F0} = 1.1 \times 10^{-18}\ \text{J} \quad \text{or} \quad 7\ \text{eV}$$

To estimate the mean speed of the electrons, we calculate the rms speed v_e from $\frac{1}{2}m_e v_e^2 = \frac{3}{5}E_{F0}$. The mean speed will be close to the rms speed. Thus, $v_e = (6E_{F0}/5m_e)^{1/2}$. Substituting for E_{F0} and m_e, we find $v_e = 1.2 \times 10^6$ m s^{-1}.

EXAMPLE 4.10

CONDUCTION IN SILVER Consider silver whose density of states $g(E)$ was calculated in Example 4.8, assuming a *free electron model* for $g(E)$ as in Equation 4.10. For silver, $E_F = 5.5$ eV, so from Equation 4.10, the density of states at E_F is $g(E_F) = 1.60 \times 10^{28}$ m^{-3} eV^{-1}. The velocity of Fermi electrons, $v_F = (2E_F/m_e)^{1/2} = 1.39 \times 10^6$ m s^{-1}. The conductivity σ of Ag at room temperature is $62.5 \times 10^6\ \Omega^{-1}\ \text{m}^{-1}$. Substituting for σ, $g(E_F)$, and v_F in Equation 4.27,

$$\sigma = 62.5 \times 10^6 = \frac{1}{3}e^2 v_F^2 \tau g(E_F) = \frac{1}{3}(1.6 \times 10^{-19})^2 (1.39 \times 10^6)^2 \tau \left(\frac{1.60 \times 10^{28}}{1.6 \times 10^{-19}}\right)$$

we find $\tau = 3.79 \times 10^{-14}$ s. The *mean free path* $\ell = v_F \tau = 53$ nm. The *drift mobility* of E_F electrons is $\mu = e\tau/m_e = 67\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$.

From Example 4.8, since Ag has a valency of 1, the concentration of conduction electrons is $n = n_{Ag} = 5.85 \times 10^{28}$ m^{-3}. Substituting for n and σ in Equation 4.28 gives

$$\sigma = 62.5 \times 10^6 = \frac{e^2 n \tau}{m_e} = \frac{(1.6 \times 10^{-19})^2 (5.85 \times 10^{28}) \tau}{(9.1 \times 10^{-31})}$$

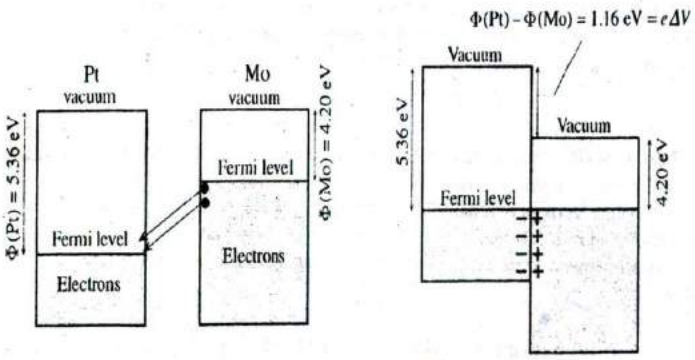
we find $\tau = 3.79 \times 10^{-14}$ s as expected because we have used the free electron model.

4.8 FERMI ENERGY SIGNIFICANCE

4.8.1 METAL–METAL CONTACTS: CONTACT POTENTIAL

Suppose that two metals, platinum (Pt) with a work function 5.36 eV and molybdenum (Mo) with a work function 4.20 eV, are brought together, as shown in Figure 4.28a. We know that in metals, all the energy levels up to the Fermi level are full. Since the Fermi level is higher in Mo (due to a smaller Φ), the electrons in Mo are more energetic. They therefore immediately go over to the Pt surface (by tunneling), where there are empty states at lower energies, which they can occupy. This electron transfer from Mo to the Pt surface reduces the total energy of the electrons in the Pt–Mo system, but at the same time, the Pt surface becomes negatively charged with respect to the Mo surface. Consequently, a contact voltage (or a potential difference) develops at the junction between Pt and Mo, with the Mo side being positive.

The electron transfer from Mo to Pt continues until the contact potential is large enough to prevent further electron transfer: the system reaches equilibrium. It should be apparent that the transfer of energetic electrons from Mo to Pt continues until the two Fermi levels are lined up, that is, until the Fermi level is uniform and the same in both metals, so that no part of the system has more (or less) energetic electrons, as



(a) Electrons are more energetic in Mo, so they tunnel to the surface of Pt.

(b) Equilibrium is reached when the Fermi levels are lined up.

Figure 4.28 When two metals are brought together, there is a contact potential ΔV.

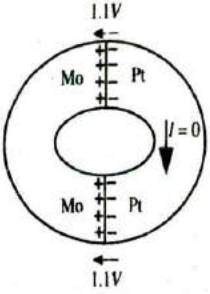


Figure 4.29 There is no current when a closed circuit is formed by two different metals, even though there is a contact potential at each contact.
The contact potentials oppose each other.

illustrated in Figure 4.28b. Otherwise, the energetic electrons in one part of the system will flow toward a region with lower energy states. Under these conditions, the Pt–Mo system is in equilibrium. The contact voltage ΔV is determined by the difference in the work functions, that is,

$$e\,\Delta V = \Phi(\text{Pt}) - \Phi(\text{Mo}) = 5.36\,\text{eV} - 4.20\,\text{eV} = 1.16\,\text{eV}$$

We should note that away from the junction on the Mo side, we must still provide an energy of $\Phi = 4.20\,\text{eV}$ to free an electron, whereas away from the junction on the Pt side, we must provide $\Phi = 5.36\,\text{eV}$ to free an electron. This means that the vacuum energy level going from Mo to Pt has a step $\Delta\Phi$ at the junction. Since we must do work equivalent to $\Delta\Phi$ to get a free electron (*e.g.*, on the metal surface) from the Mo surface to the Pt surface, this represents a voltage of $\Delta\Phi/e$ or 1.16 V.

From the second law of thermodynamics,[7] this contact voltage cannot do work; that is, it cannot drive current in an external circuit. To see this, we can close the Pt metal–Mo metal circuit to form a ring, as depicted in Figure 4.29. As soon as we close the circuit, we create another junction with a contact voltage that is equal and opposite to that of the first junction. Consequently, going around the circuit, the net voltage is zero and the current is therefore zero.

There is a deep significance to the Fermi energy E_F, which should at least be mentioned. For a given metal the Fermi energy represents the free energy per electron called the **electrochemical potential** μ. In other words, the Fermi energy is a measure of the potential of an electron to do electrical work ($e \times V$) or nonmechanical work, through chemical or physical processes.[8] In general, when two metals are brought into contact, the Fermi level (with respect to a vacuum) in each will be different. This difference means a difference in the chemical potential $\Delta\mu$, which in turn means that the system will do external work, which is obviously not possible. Instead, electrons are immediately transferred from one metal to the other, until the free energy per electron μ for the whole system is minimized and is uniform across the two metals, so that

[7] By the way, the second law of thermodynamics simply says that you cannot extract heat from a system in thermal equilibrium and do work (*i.e.*, charge × voltage).

[8] A change in any type of PE can, in principle, be used to do work, that is, $\Delta(PE)$ = work done. Chemical PE is the potential to do nonmechanical work (*e.g.*, electrical work) by virtue of physical or chemical processes. The chemical PE per electron is E_f and ΔE_f = electrical work per electron.

$\Delta\mu = 0$. We can guess that if the Fermi level in one metal could be maintained at a higher level than the other, by using an external energy source (*e.g.*, light or heat), for example, then the difference could be used to do electrical work.

4.8.2 THE SEEBECK EFFECT AND THE THERMOCOUPLE

Consider a conductor such as an aluminum rod that is heated at one end and cooled at the other end as depicted in Figure 4.30. The electrons in the hot region are more energetic and therefore have greater velocities than those in the cold region.[9]

Consequently there is a net diffusion of electrons from the hot end toward the cold end which leaves behind exposed positive metal ions in the hot region and accumulates electrons in the cold region. This situation prevails until the electric field developed between the positive ions in the hot region and the excess electrons in the cold region prevents further electron motion from the hot to the cold end. A voltage therefore develops between the hot and cold ends, with the hot end at positive potential. The potential difference ΔV across a piece of metal due to a temperature difference ΔT is called the **Seebeck effect.**[10] To gauge the magnitude of this effect we introduce a special coefficient which is defined as the potential difference developed per unit temperature difference, or

Thermoelectric power or Seebeck coefficient

$$S = \frac{dV}{dT} \qquad [4.29]$$

By convention, the sign of S represents the potential of the cold side with respect to the hot side. If electrons diffuse from the hot end to the cold end as in Figure 4.30, then the cold side is negative with respect to the hot side and the Seebeck coefficient is *negative* (as for aluminum).

In some metals, such as copper, this intuitive explanation fails to explain why electrons actually diffuse from the cold to the hot region, giving rise to *positive* Seebeck coefficients; the polarity of the voltage in Figure 4.30 is actually reversed for copper. The reason is that the net diffusion process depends on how the mean free path ℓ and the mean free time (due to scattering from lattice vibrations) change with the electron energy, which can be quite complicated. Typical Seebeck coefficients for various selected metals are listed in Table 4.3.

Consider two neighboring regions H (hot) and C (cold) with widths corresponding to the mean free paths ℓ and ℓ' in H and C as depicted in Figure 4.31a. Half the electrons in H would be moving in the $+x$ direction and the other half in the $-x$ direction. Half of the electrons in H therefore cross into C, and half in C cross into H. Suppose that, very roughly, the electron concentration n in H and C is about the same. The number of electrons crossing from H to C is $\frac{1}{2}n\ell$, and the number crossing from C to H is $\frac{1}{2}n\ell'$. Then,

$$\text{Net diffusion from H to C} \propto \tfrac{1}{2}n(\ell - \ell') \qquad [4.30]$$

[9] The conduction electrons around the Fermi energy have a mean speed that has only a small temperature dependence. This small change in the mean speed with temperature is, nonetheless, intuitively significant in appreciating the thermoelectric effect. The actual effect, however, depends on the mean free path as discussed later.

[10] Thomas Seebeck observed this thermoelectric effect in 1821 using two different metals as in the thermocouple, which is the only way to observe the phenomenon. It was William Thomson (Lord Kelvin) who explained the observed effect.

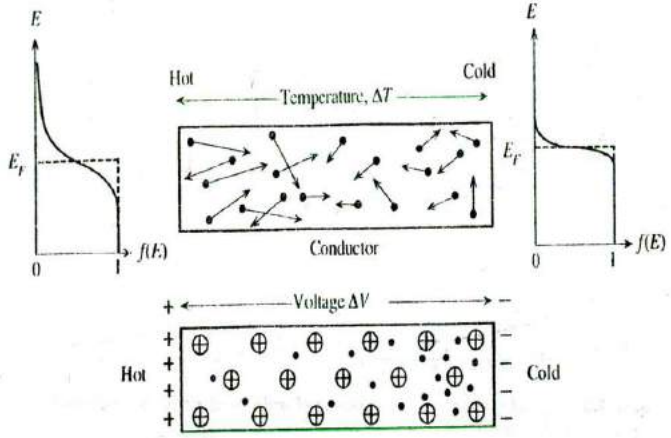


Figure 4.30 The Seebeck effect.
A temperature gradient along a conductor gives rise to a potential difference.

Suppose that the scattering of electrons is such that ℓ increases strongly with the electron energy. Then electrons in H, which are more energetic, have a longer mean free path, that is, $\ell > \ell'$ as shown in Figure 4.31a. This means that the net migration is from H to C and S is negative, as in aluminum. In those metals such as copper in which ℓ decreases strongly with the energy, electrons in the cold region have a longer mean free path, $\ell' > \ell$ as shown in Figure 4.31b. The net electron migration is then from C to H and S is positive. Even this qualitative explanation is not quite correct because n is not the same in H and C (diffusion changes n) and, further, we neglected the change in the mean scattering time with the electron energy.

The coefficient S is widely referred to as the **thermoelectric power** even though this term is misleading, as it refers to a voltage difference rather than power. A more appropriate recent term is the **Seebeck coefficient.** S is a material property that depends on temperature, $S = S(T)$, and is tabulated for many materials as a function of

Table 4.3 Seebeck coefficients of selected metals (from various sources)

Metal	S at 0 °C ($\mu V K^{-1}$)	S at 27 °C ($\mu V K^{-1}$)	E_F (eV)	x
Al	−1.6	−1.8	11.6	2.78
Au	+1.79	+1.94	5.5	−1.48
Cu	+1.70	+1.84	7.0	−1.79
K		−12.5	2.0	3.8
Li	+14		4.7	−9.7
Mg	−1.3		7.1	1.38
Na		−5	3.1	2.2
Pd	−9.00	−9.99		
Pt	−4.45	−5.28		

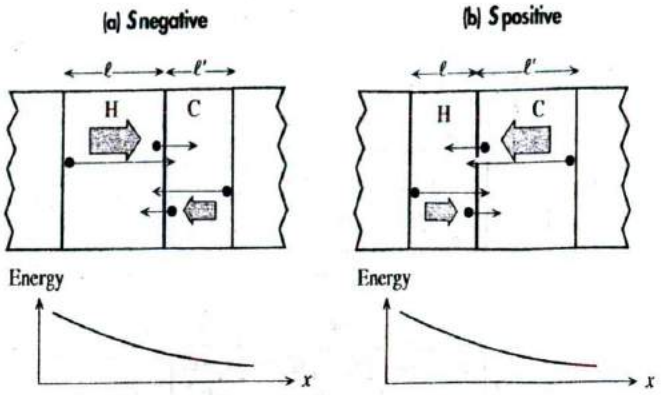


Figure 4.31 Consider two neighboring regions H (hot) and C (cold) with widths corresponding to the mean free paths ℓ and ℓ' in H and C.
Half the electrons in H would be moving in the $+x$ direction and the other half in the $-x$ direction. Half of the electrons in H therefore cross into C, and half in C cross into H.

temperature. Given the Seebeck coefficient $S(T)$ for a material, Equation 4.29 yields the voltage difference between two points where temperatures are T_o and T as follows:

$$\Delta V = \int_{T_o}^{T} S\,dT \qquad [4.31]$$

A proper explanation of the Seebeck effect has to consider how electrons around the Fermi energy E_F, which contribute to electrical conduction, are scattered by lattice vibrations, impurities, and crystal defects. This scattering process controls the mean free path and hence the Seebeck coefficient (Figure 4.31). The scattered electrons need empty states, which in turn requires that we consider how the density of states changes with the energy as well. Moreover, in certain metals such as Ni, there are overlapping partially filled bands and the Fermi electron can be scattered from one electronic band to another, for example from the 4*s* band to the 3*d* band, which must also be considered (see Question 4.25). The Seebeck coefficient for many metals is given by the **Mott and Jones equation,**

Mott and Jones thermoelectric power

$$S \approx -\frac{\pi^2 k^2 T}{3eE_{FO}} x \qquad [4.32]$$

where x is a numerical constant that takes into account how various charge transport parameters (such as ℓ) depend on the electron energy. A few examples for x are given in Table 4.3. The reason for the kT/E_{FO} factor in Equation 4.32 is that only those electrons about a kT around the Fermi level E_{FO} are involved in the transport and scattering processes. Equation 4.32 does not apply directly to transition metals (Ni, Pd, Pt) that have overlapping bands. These metals have a negative Seebeck coefficient that is proportional to temperature as in Equation 4.32, but the exact expression depends on the band structure.

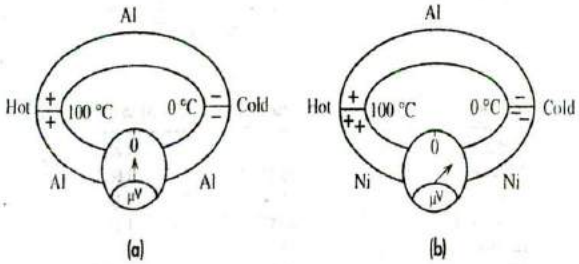


Figure 4.32

(a) If Al wires are used to measure the Seebeck voltage across the Al rod, then the net emf is zero.

(b) The Al and Ni have different Seebeck coefficients. There is therefore a net emf in the Al–Ni circuit between the hot and cold ends that can be measured.

Suppose that we try to measure the voltage difference ΔV across the aluminum rod by using aluminum connecting wires to a voltmeter as indicated in Figure 4.32a. The same temperature difference now also exists across the aluminum connecting wires; therefore an identical voltage also develops across the connecting wires, opposing that across the aluminum rod. Consequently no net voltage will be registered by the voltmeter. It is, however, possible to read a net voltage difference, if the connecting wires are of different material, that is, have a different Seebeck coefficient from that of aluminum. Then the thermoelectric voltage across this material is different than that across the aluminum rod, as in Figure 4.32b.

The Seebeck effect is fruitfully utilized in the thermocouple (TC), shown in Figure 4.32b, which uses two different metals with one junction maintained at a reference temperature T_a and the other used to sense the temperature T. The voltage across each metal element depends on its Seebeck coefficient. The potential difference between the two wires will depend on $S_A - S_B$. By virtue of Equation 4.31, the electromotive force (emf) between the two wires, $V_{AB} = \Delta V_A - \Delta V_B$, is then given by

Thermocouple emf between metals A and B

$$V_{AB} = \int_{T_a}^{T} (S_A - S_B)\, dT = \int_{T_a}^{T} S_{AB}\, dT \qquad [4.33]$$

where $S_{AB} = S_A - S_B$ is defined as the thermoelectric power for the thermocouple pair A-B. For the chromel-alumel (K-type) TC, for example, $S_{AB} \approx 40\ \mu\text{V K}^{-1}$ at 300 K.

The output voltage from a TC pair obviously depends on the two metals used. Instead of tabulating the emf from all possible pairs of materials in the world, which would be a challenging task, engineers have tabulated the emfs available when a given material is used with a reference metal which is chosen to be platinum. The reference junction is kept at 0 °C (273.16 K) which corresponds to a mixture of ice and water. Some typical materials and their emfs are listed in Table 4.4.

Thermocouple equation

Using the expression for the Seebeck coefficient, Equation 4.32, in Equation 4.33, and then integrating, leads to the familiar thermocouple equation,

$$V_{AB} = a\,\Delta T + b(\Delta T)^2 \qquad [4.34]$$

Table 4.4 Thermoelectric emf for metals at 100 and 200 °C with respect to Pt and the reference junction at 0 °C

Material	emf (mV)	
	At 100 °C	At 200 °C
Copper, Cu	0.76	1.83
Aluminum, Al	0.42	1.06
Nickel, Ni	−1.48	−3.10
Palladium, Pd	−0.57	−1.23
Platinum, Pt	0	0
Silver, Ag	0.74	1.77
Alumel	−1.29	−2.17
Chromel	2.81	5.96
Constantan	−3.51	−7.45
Iron, Fe	1.89	3.54
90% Pt–10% Rh (platinum-rhodium)	0.643	1.44

where a and b are the thermocouple coefficients and $\Delta T = T - T_o$ is the temperature with respect to the reference temperature T_o (273.16 K). The inference from Equation 4.34 is that the emf output from the thermocouple wires does not depend linearly on the temperature difference ΔT. Figure 4.33 shows the emf output versus temperature for various thermocouples. It should be immediately obvious that the voltages are small, typically a few tens of a microvolt per degree temperature difference. At

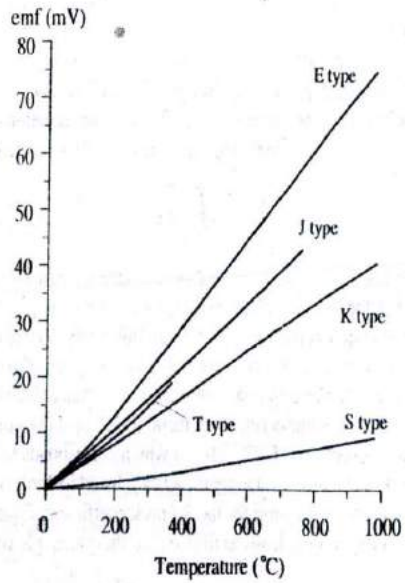


Figure 4.33 Output emf versus temperature (°C) for various thermocouples between 0 to 1000 °C.

0 °C, by definition, the TC emf is zero. The K-type thermocouple, the chromel-alumel pair, is a widely employed general-purpose thermocouple sensor up to about 1200 °C.

EXAMPLE 4.11

THE THERMOCOUPLE EMF Consider a thermocouple pair from Al and Cu which have Fermi energies and x as in Table 4.3. Estimate the emf available from this thermocouple if one junction is held at 0 °C and the other at 100 °C.

SOLUTION

We essentially have the arrangement shown in Figure 4.32b but with Cu replacing Ni and Cu having the cold end positive (S is positive). For each metal there will be a voltage across it, given by integrating the Seebeck coefficient from T_o (at the low temperature end) to T. From the Mott and Jones equation,

$$\Delta V = \int_{T_o}^{T} S\,dT = \int_{T_o}^{T} -\frac{x\pi^2 k^2 T}{3eE_{FO}}\,dT = -\frac{x\pi^2 k^2}{6eE_{FO}}\left(T^2 - T_o^2\right)$$

The available emf (V_{AB}) is the difference in ΔV for the two metals (A and B), so

$$V_{AB} = \Delta V_A - \Delta V_B = -\frac{\pi^2 k^2}{6e}\left[\frac{x_A}{E_{FAO}} - \frac{x_B}{E_{FBO}}\right]\left(T^2 - T_o^2\right)$$

where in this example $T = 373$ K and $T_o = 273$ K.

For Al (A), $E_{FAO} = 11.6$ eV, $x_A = 2.78$, and for copper (B), $E_{FBO} = 7.0$ eV, $x_B = -1.79$. Thus,

$$V_{AB} = -189\ \mu\text{V} - (+201\ \mu\text{V}) = -390\ \mu\text{V}$$

Thermocouple emf calculations that closely represent experimental observations require thermocouple voltages for various metals listed against some reference metal. The reference is usually Pt with the reference junction at 0 °C. From Table 4.4 we can read Al–Pt and Cu–Pt emfs as $V_{\text{Al-Pt}} = 0.42$ mV and $V_{\text{Cu-Pt}} = 0.76$ mV at 100 °C with the experimental error being around ±0.01 mV, so that for the Al–Cu pair,

$$V_{\text{Al-Cu}} = V_{\text{Al-Pt}} - V_{\text{Cu-Pt}} = 0.42\ \text{mV} - 0.76\ \text{mV} = -0.34\ \text{mV}$$

There is a reasonable agreement with the calculation using the Mott and Jones equation.

EXAMPLE 4.12

THE THERMOCOUPLE EQUATION We know that we can only measure differences between thermoelectric powers of materials. When two different metals A and B are connected to make a thermocouple, as in Figure 4.32b, then the net emf is the voltage difference between the two elements. From Example 4.11,

$$\Delta V_{AB} = \Delta V_A - \Delta V_B = \int_{T_o}^{T} (S_A - S_B)\,dT = \int_{T_o}^{T} S_{AB}\,dT$$

$$= -\frac{\pi^2 k^2}{6e}\left[\frac{x_A}{E_{FAO}} - \frac{x_B}{E_{FBO}}\right]\left(T^2 - T_o^2\right)$$

$$= C\left(T^2 - T_o^2\right)$$

where C is a constant that is independent of T but dependent on the material properties (x, E_{FO} for the metals).

We can now expand V_{AB} about T_o by using Taylor's expansion

$$F(T) \approx F(T_o) + \Delta T\,(dF/dT)_o + \tfrac{1}{2}(\Delta T)^2(d^2F/dT^2)_o$$

where the function $F = V_{AB}$ and $\Delta T = T - T_o$ and the derivatives are evaluated at T_o. The result is the thermocouple equation:

$$V_{AB}(T) = a(\Delta T) + b(\Delta T)^2$$

where the coefficients a and b are $2CT_o$ and C, respectively.

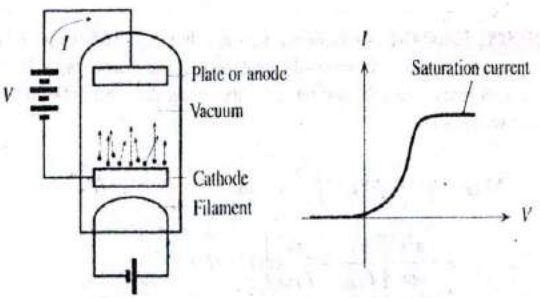
It is clear that the magnitude of the emf produced depends on C or $S_A - S_B$, which we can label as S_{AB}. The greater the thermoelectric power difference S_{AB} for the TC, the larger the emf produced. For the copper constantan TC, S_{AB} is about 43 μV K^{-1}.

4.9 THERMIONIC EMISSION AND VACUUM TUBE DEVICES

4.9.1 THERMIONIC EMISSION: RICHARDSON–DUSHMAN EQUATION

Even though most of us view vacuum tubes as electrical antiques, their basic principle of operation (electrons emitted from a heated cathode) still finds application in cathode ray and X-ray tubes and various RF microwave vacuum tubes, such as triodes, tetrodes, klystrons, magnetrons, and traveling wave tubes and amplifiers. Therefore, it is useful to examine how electrons are emitted when a metal is heated.

When a metal is heated, the electrons become more energetic as the Fermi–Dirac function extends to higher temperatures. Some of the electrons have sufficiently large energies to leave the metal and become free. This situation is self-limiting because as the electrons accumulate outside the metal, they prevent more electrons from leaving the metal. (Put differently, emitted electrons leave a net positive charge behind, which pulls the electrons in.) Consequently, we need to replenish the "lost" electrons and collect the emitted ones, which is done most conveniently using the vacuum tube arrangement in a closed circuit, as shown in Figure 4.34a. The cathode, heated by a filament, emits electrons. A battery connected between the cathode and the anode replenishes



(a) Thermionic electron emission in a vacuum tube.

(b) Current-voltage characteristics of a vacuum diode.

Figure 4.34

the cathode electrons and provides a positive bias to the anode to collect the thermally emitted electrons from the cathode. The vacuum inside the tube ensures that the electrons do not collide with the air molecules and become dispersed, with some even being returned to the cathode by collisions. Therefore, the vacuum is essential. The current due to the flow of emitted electrons from the cathode to the anode depends on the anode voltage as indicated in Figure 4.34b. The current increases with the anode voltage until, at sufficiently high voltages, all the emitted electrons are collected by the anode and the current *saturates*. The **saturation current** of the vacuum diode depends on the rate of thermionic emission of electrons which we will derive below. The vacuum tube in Figure 4.34a acts as a **rectifier** because there is no current flow when the anode voltage becomes negative; the anode then repels the electrons.

We know that only those electrons with energies greater than $E_F + \Phi$ (Fermi energy + work function) which are moving toward the surface can leave the metal. Their number depends on the temperature, by virtue of the Fermi–Dirac statistics. Figure 4.35 shows how the concentration of conduction electrons with energies above $E_F + \Phi$ increases with temperature. We know that conduction electrons behave as if they are free within the metal. We can therefore take the PE to be zero within the metal, but $E_F + \Phi$ outside the metal. The energy E of the electron within the metal is then purely kinetic, or

$$E = \frac{1}{2}m_e v_x^2 + \frac{1}{2}m_e v_y^2 + \frac{1}{2}m_e v_z^2 \qquad [4.35]$$

Suppose that the surface of the metal is perpendicular to the direction of emission, say along x. For an electron to be emitted from the surface, its $KE = \frac{1}{2}mv_x^2$ along x must be greater than the potential energy barrier $E_F + \Phi$, that is,

$$\frac{1}{2}mv_x^2 > E_F + \Phi \qquad [4.36]$$

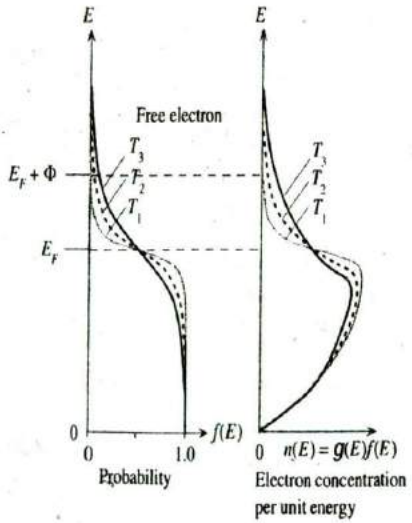


Figure 4.35 Fermi-Dirac function $f(E)$ and the energy density of electrons $n(E)$ (electrons per unit energy and per unit volume) at three different temperatures.
The electron concentration extends more and more to higher energies as the temperature increases. Electrons with energies in excess of $E_F + \Phi$ can leave the metal (thermionic emission).

Left to right: Owen Williams Richardson, Robert Andrews Millikan, and Arthur Holly Compton at an international conference on nuclear physics, Rome, 1931. Richardson won the physics Nobel prize in 1928 for thermionic emission.

SOURCE: Amaldi Archives, Dipartimento di Fisica, Università La Sapienza, Rome; courtesy of AIP Emilio Segrè Visual Archives.

Let $dn(v_x)$ be the number of electrons moving along x with velocities in the range v_x to $(v_x + dv_x)$, with v_x satisfying emission in Equation 4.36. These electrons will be emitted when they reach the surface. Their number $dn(v_x)$ can be determined from the density of states and the Fermi–Dirac statistics, since energy and velocity are related through Equation 4.35. Close to $E_F + \Phi$, the Fermi–Dirac function will approximate the Boltzmann distribution, $f(E) = \exp[-(E - E_F)/kT]$. The number $dn(v_x)$ is therefore at least proportional to this exponential energy factor.

The emission of $dn(v_x)$ electrons will give a thermionic current density $dJ_x = ev_x\, dn(v_x)$. This must be integrated (summed) for all velocities satisfying Equation 4.36 to obtain the total current density J_x, or simply J. Since $dn(v_x)$ includes an exponential energy function, the integration also leads to an exponential. The final result is

Richardson–Dushman thermionic emission equation

$$J = B_0 T^2 \exp\left(-\frac{\Phi}{kT}\right) \qquad [4.37]$$

where $B_0 = 4\pi e m_e k^2/h^3$. Equation 4.37 is called the **Richardson–Dushman equation,** and B_0 is the Richardson–Dushman constant, whose value is $1.20 \times 10^6\ \mathrm{A\,m^{-2}\,K^{-2}}$. We see from Equation 4.37 that the emitted current from a heated cathode varies exponentially with temperature and is sensitive to the work function Φ of the cathode material. Both factors are apparent in Equation 4.37.

The wave nature of electrons means that when an electron approaches the surface, there is a probability that it may be reflected back into the metal, instead of being emitted over the potential barrier. As the potential energy barrier becomes very large, $\Phi \to \infty$, the electrons are totally reflected and there is no emission. Taking into account that waves can be reflected, the thermionic emission equation is appropriately modified to

Thermionic emission

$$J = B_e T^2 \exp\left(-\frac{\Phi}{kT}\right) \qquad [4.38]$$

where $B_e = (1 - R)B_o$ is the **emission constant** and R is the reflection coefficient. The value of R will depend on the material and the surface conditions. For most metals, B_e is about half of B_o, whereas for some oxide coatings on Ni cathodes used in thermionic tubes, B_e can be as low as 1×10^2 A m^{-2} K^{-2}.

Equation 4.37 was derived by neglecting the effect of the applied field on the emission process. Since the anode is positively biased with respect to the cathode, the field will not only collect the emitted electrons (by drifting them to the anode), but will also enhance the process of thermal emission by lowering the potential energy barrier Φ.

There are many thermionic emission–based vacuum tubes that find applications in which it is not possible or practical to use semiconductor devices, especially at high-power and high-frequency operation at the same time, such as in radio and TV broadcasting, radars, microwave communications; for example, a tetrode vacuum tube in radio broadcasting equipment has to handle hundreds of kilowatts of power. X-ray tubes operate on the thermionic emission principle in which electrons are thermally emitted, and then accelerated and impacted on a metal target to generate X-ray photons.

VACUUM TUBES It is clear from the Richardson–Dushman equation that to obtain an efficient thermionic cathode, we need high temperatures and low work functions. Metals such as tungsten (W) and tantalum (Ta) have high melting temperatures but high work functions. For example, for W, the melting temperature T_m is 3680 °C and its work function is about 4.5 eV. Some metals have low work functions, but also low melting temperatures, a typical example being Cs with $\Phi = 1.8$ eV and $T_m = 28.5$ °C. If we use a thin film coating of a low Φ material, such as ThO or BaO, on a high-melting-temperature base metal such as W, we can maintain the high melting properties and obtain a lower Φ. For example, Th on W has a $\Phi = 2.6$ eV and $T_m = 1845$ °C. Most vacuum tubes use indirectly heated cathodes that consist of the oxides of B, Sr, and Ca on a base metal of Ni. The operating temperatures for these cathodes are typically 800 °C.

EXAMPLE 4.13

A certain transmitter-type vacuum tube has a cylindrical Th-coated W (thoriated tungsten) cathode, which is 4 cm long and 2 mm in diameter. Estimate the saturation current if the tube is operated at a temperature of 1600 °C, given that the emission constant is $B_e = 3.0 \times 10^4$ A m^{-2} K^{-2} for Th on W.

SOLUTION

We apply the Richardson–Dushman equation with $\Phi = 2.6$ eV, $T = (1600 + 273)$ K $= 1873$ K, and $B_e = 3.0 \times 10^4$ A m^{-2} K^{-2}, to find the maximum current density that can be obtained from the cathode at 1873 K, as follows:

$$J = (3.0 \times 10^4 \text{ A m}^{-2} \text{ K}^{-2})(1873 \text{ K})^2 \exp\left[-\frac{(2.6 \times 1.6 \times 10^{-19})}{(1.38 \times 10^{-23} \times 1873)}\right]$$

$$= 1.08 \times 10^4 \text{ A m}^{-2}$$

The emission surface area is

$$A = \pi(\text{diameter})(\text{length}) = \pi(2 \times 10^{-3})(4 \times 10^{-2}) = 2.5 \times 10^{-4} \text{ m}^2$$

so the saturation current, which is the maximum current obtainable (*i.e.*, the thermionic current), is

$$I = JA = (1.08 \times 10^4 \text{ A m}^{-2})(2.5 \times 10^{-4} \text{ m}^2) = 2.7 \text{ A}$$

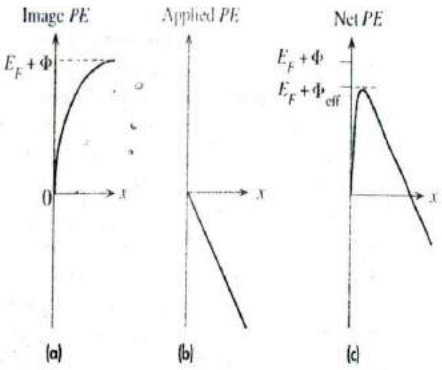


Figure 4.36
(a) PE of the electron near the surface of a conductor.
(b) Electron PE due to an applied field, that is, between cathode and anode.
(c) The overall PE is the sum.

4.9.2 SCHOTTKY EFFECT AND FIELD EMISSION

When a positive voltage is applied to the anode with respect to the cathode, the electric field at the cathode helps the thermionic emission process by lowering the *PE* barrier Φ. This is called the **Schottky effect.** Consider the *PE* of the electron just outside the surface of the metal. The electron is pulled in by the effective positive charge left in the metal. To represent this attractive *PE* we use the **theorem of image charges** in electrostatics,[11] which says that an electron at a distance x from the surface of a conductor possesses a potential energy that is

$$PE_{\text{image}}(x) = -\frac{e^2}{16\pi\varepsilon_o x} \qquad [4.39]$$

where ε_o is the absolute permittivity.

This equation is valid for x much greater than the atomic separation a; otherwise, we must consider the interaction of the electron with the individual ions. Further, Equation 4.39 has a reference level of zero *PE* at infinity ($x = \infty$), but we defined $PE = 0$ to be inside the metal. We must therefore modify Equation 4.39 to conform to our definition of zero *PE* as a reference. Figure 4.36a shows how this "image *PE*" varies with x in this system. In the region $x < x_o$, we artificially bring $PE_{\text{image}}(x)$ to zero at $x = 0$, so our definition $PE = 0$ within the metal is maintained. Far away from the surface, the *PE* is expected to be $(E_F + \Phi)$ (and not zero, as in Equation 4.39), so we modify Equation 4.39 to read

$$PE_{\text{image}}(x) = (E_F + \Phi) - \frac{e^2}{16\pi\varepsilon_o x} \qquad [4.40]$$

The present model, which takes $PE_{\text{image}}(x)$ from 0 to $(E_F + \Phi)$ along Equation 4.40, is in agreement with the thermionic emission analysis, since the electron must still overcome a *PE* barrier of $E_F + \Phi$ to escape.

[11] An electron at a distance x from the surface of a conductor experiences a force as if there were a positive charge of $+e$ at a distance $2x$ from it. The force is $e^2/[4\pi\varepsilon_o(2x)^2]$ or $e^2/[16\pi\varepsilon_o x^2]$. The result is called the image charge theorem. Integrating the force gives the potential energy in Equation 4.39.

From the definition of potential, which is potential energy per unit charge, when a voltage difference is applied between the anode and cathode, there is a PE gradient just outside the surface of the metal, given by $eV(x)$, or

$$PE_{\text{applied}}(x) = -ex\mathcal{E} \qquad [4.41]$$

where $\mathcal{E}$ is the applied field and is assumed, for all practical purposes, to be uniform. The variation of $PE_{\text{applied}}(x)$ with x is depicted in Figure 4.36b. The total $PE(x)$ of the electron outside the metal is the sum of Equations 4.40 and 4.41, as sketched in Figure 4.36c,

$$PE(x) = (E_F + \Phi) - \frac{e^2}{16\pi\varepsilon_o x} - ex\mathcal{E} \qquad [4.42]$$

Note that the $PE(x)$ outside the metal no longer goes up to $(E_F + \Phi)$, and the PE barrier against thermal emission is effectively reduced to $(E_F + \Phi_{\text{eff}})$, where Φ_{eff} is a new effective work function that takes into account the effect of the applied field. The new barrier $(E_F + \Phi_{\text{eff}})$ can be found by locating the maximum of $PE(x)$, that is, by differentiating Equation 4.42 and setting it to zero. The **effective work function** in the presence of an applied field is therefore

$$\Phi_{\text{eff}} = \Phi - \left(\frac{e^3\mathcal{E}}{4\pi\varepsilon_o}\right)^{1/2} \qquad [4.43]$$

This lowering of the work function by the applied field, as predicted by Equation 4.43, is the **Schottky effect.** The current density is given by the Richardson–Dushman equation, but with Φ_{eff} instead of Φ,

Field-assisted thermionic emission

$$J = B_c T^2 \exp\left[-\frac{(\Phi - \beta_S \mathcal{E}^{1/2})}{kT}\right] \qquad [4.44]$$

where $\beta_S = [e^3/4\pi\varepsilon_o]^{1/2}$ is the **Schottky coefficient,** whose value is 3.79×10^{-5} $(\text{eV}/\sqrt{\text{V m}^{-1}})$.

When the field becomes very large, for example, $\mathcal{E} > 10^7$ V cm^{-1}, the $PE(x)$ outside the metal surface may bend sufficiently steeply to give rise to a narrow PE barrier. In this case, there is a distinct probability that an electron at an energy E_F will tunnel through the barrier and escape into vacuum, as depicted in Figure 4.37. The likelihood of tunneling depends on the effective height Φ_{eff} of the PE barrier above E_F, as well as the width x_F of the barrier at energy level E_F. Since tunneling is temperature independent, the emission process is termed **field emission.** The tunneling probability P was calculated in Chapter 3, and depends on Φ_{eff} and x_F through the equation[12]

$$P \approx \exp\left[\frac{-2(2m_e\Phi_{\text{eff}})^{1/2}x_F}{\hbar}\right]$$

We can easily find x_F by noting that when $x = x_F$, $PE(x_F)$ is level with E_F, as shown in Figure 4.37. From Equation 4.42, when the field is very strong, then around

[12] In Chapter 3 we showed that the transmission probability $T = T_o \exp(-2\alpha a)$ where $\alpha^2 = 2m(V_o - E)/\hbar^2$ and a is the barrier width. The pre-exponential constant T_o can be taken to be ~ 1. Clearly $V_o - E = \Phi_{\text{eff}}$ since electrons with $E = E_F$ are tunneling and $a = x_F$.

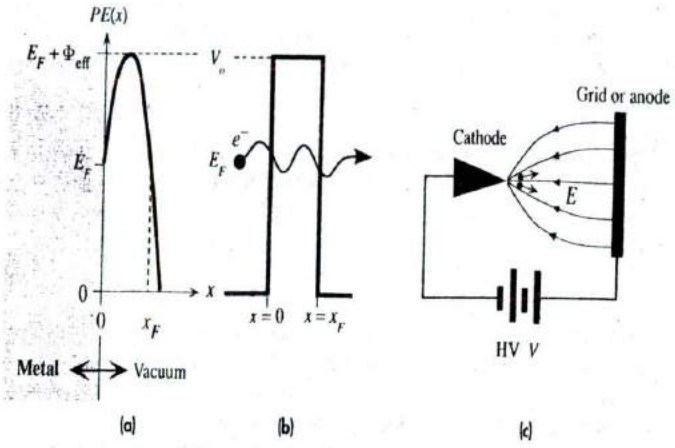


Figure 4.37

(a) Field emission is the tunneling of an electron at an energy E_F through the narrow PE barrier induced by a large applied field.

(b) For simplicity, we take the barrier to be rectangular.

(c) A sharp point cathode has the maximum field at the tip where the field emission of electrons occurs.

$x \approx x_F$ the second term is negligible compared to the third, so putting $x = x_F$ and $PE(x_F) = E_F$ in Equation 4.42 yields $\Phi = e\mathcal{E}x_F$. Substituting $x_F = \Phi/e\mathcal{E}$ in Equation 4.45, we can obtain the tunneling probability P

Field-assisted tunneling probability

$$P \approx \exp\left[-\frac{2(2m_e\Phi_{eff})^{1/2}\Phi}{e\hbar\mathcal{E}}\right] \qquad [4.45]$$

Equation 4.45 represents the probability P that an electron in the metal at E_F will tunnel out from the metal, as in Figure 4.37a and b, and become field-emitted. In a more rigorous analysis we have to consider that electrons not just at E_F but at energies below E_F can also tunnel out (though with lower probability) and we have to abandon the rough rectangular $PE(x)$ approximation in Figure 4.37b.

To calculate the current density J we have to consider how many electrons are moving toward the surface per second and per unit area, the electron flux, and then multiply this flow by the probability that they will tunnel out. The final result of the calculations is the **Fowler–Nordheim equation,** which still has the exponential field dependence in Equation 4.45,

Field-assisted tunneling: Fowler–Nordheim equation

$$J_{\text{field-emission}} \approx CE^2 \exp\left(-\frac{\mathcal{E}_c}{\mathcal{E}}\right) \qquad [4.46a]$$

in which C and $\mathcal{E}_c$ are temperature-independent constants

$$C = \frac{e^3}{8\pi h\Phi} \quad \text{and} \quad \mathcal{E}_c = \frac{8\pi(2m_e\Phi^3)^{1/2}}{3eh} \qquad [4.46b]$$

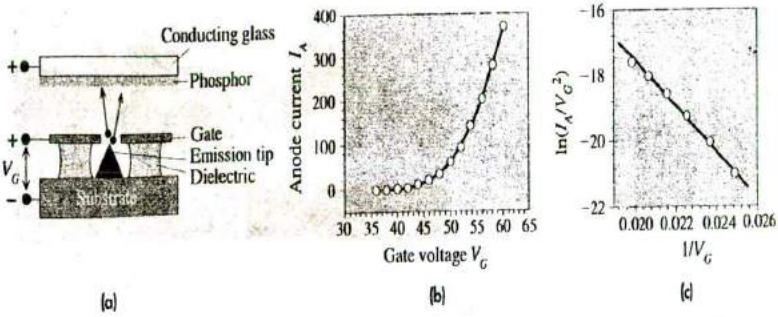


Figure 4.38

(a) Spindt-type cathode and the basic structure of one of the pixels in the FED.

(b) Emission (anode) current versus gate voltage.

(c) Fowler–Nordheim plot that confirms field emission.

that depend on the work function Φ of the metal. Equation 4.46a can also be used for field emission of electrons from a metal into an insulating material by using the electron *PE* barrier Φ_B from metal's E_F into the insulator's conduction band (where the electron is free) instead of Φ.

Notice that the field $\mathcal{E}$ in Equation 4.46a has taken over the role of temperature in thermionic emission in Equation 4.38. Since field-assisted emission depends exponentially on the field via Equation 4.46a, it can be enhanced by shaping the cathode into a cone with a sharp point where the field is maximum and the electron emission occurs from the tip as depicted in Figure 4.37c. The field $\mathcal{E}$ in Equation 4.46a is the *effective field* at the tip of the cathode that emits the electrons.

A popular field-emission tip design is based on the **Spindt tip cathode,** named after its originator. As shown in Figure 4.38a, the emission cathode is an iceberg-type sharp cone and there is a positively biased **gate** above it with a hole to extract the emitted electrons. A positively biased **anode** draws and accelerates the electrons passing through the gate toward it, which impinge on a phosphor screen to generate light by **cathodoluminescence,** a process in which light is emitted from a material when it is bombarded with electrons. Arrays of such electron field-emitters are used in field emission displays (FEDs) to generate bright images with vivid colors. Color is obtained by using red, green, and blue phosphors. The field at the tip is controlled by the potential difference between the gate and the cathode, the gate voltage V_G, which therefore controls field emission. Since $\mathcal{E} \propto V_G$, Equation 4.46a can be written to obtain the emission current or the anode current I_A as

Fowler–Nordheim anode current in a field emission device

$$I_A = aV_G^2 \exp\left(-\frac{b}{V_G}\right) \qquad [4.47]$$

where a and b are constants that depend on the particular field-emitting structure and cathode material. Figure 4.38b shows the dependence of I_A on V_G. There is a very sharp increase with the voltage once the threshold voltages (around ~45 V in Figure 4.38b) are reached to start the electron emission. Once the emission is fully operating,

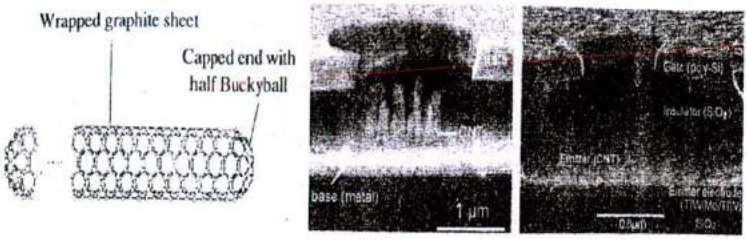


Figure 4.39

(a) A carbon nanotube (CNT) is a whisker-like, very thin and long carbon molecule with rounded ends, almost the perfect shape to be an electron field-emitter.

(b) Multiple CNTs as electron emitters.

(c) A single CNT as an emitter.

SOURCE: Courtesy of Professor W. I. Milne, University of Cambridge; G. Pirio *et al.*, *Nanotechnology*, **13**, 1, 2002.

I_A versus V_G follows the Fowler–Nordheim emission. A plot of $\ln(I_A/V_G^2)$ versus $1/V_G$ is a straight line as shown in Figure 4.38c.

Field emission has a number of distinct advantages. It is much more power efficient than thermionic emission which requires heating the cathode to high temperatures. In principle, field emission can be operated at high frequencies (fast switching times) by reducing various capacitances in the emission device or controlling the electron flow with a grid. Field emission has a number of important realized and potential applications: field emission microscopy, microwave amplifiers (high power and wide bandwidth), parallel electron beam microscopy, nanolithography, portable X-ray generators, and FEDs. For example, FEDs are thin flat displays (~2 mm thick), that have a low power consumption, quick start, and most significantly, a wide viewing angle of about 170°. Monochrome FEDs are already on the market, and color FEDs are expected to be commercialized soon, probably before the fourth edition of this text.

Typically molybdenum, tungsten, and hafnium have been used as the field-emission tip materials. Micromachining (microfabrication) has lead to the use of Si emission tips as well. Good electron emission characteristics have been also reported for diamond-like carbon films. Recently there has been a particular interest in using carbon nanotubes as emitters. A **carbon nanotube** (CNT) is a very thin filament-like carbon molecule whose diameter is in the nanometer range but whose length can be quite long, *e.g.*, 10–100 microns, depending on how it is grown or prepared. A CNT is made by rolling a graphite sheet into a tube and then capping the ends with hemispherical buckminsterfullerene molecules (a half Buckyball) as shown in Figure 4.39a. Depending on how the graphite sheet is rolled up, the CNT may be a metal or a semiconductor[13]. The high aspect ratio (length/diameter) of the CNT makes it an efficient

[13] Carbon nanotubes can be single-walled or multiwalled (when the graphite sheets are wrapped more than once) and can have quite complicated structures. There is no doubt that they possess some remarkable properties, so it is likely that CNTs will eventually be used in various engineering applications. See, for example, M. Baxendale, *J. Mater. Sci.: Mater Electron*, **14**, 657, 2003.

electron emitter. If one were to wonder what is the best shape for an efficient field emission tip, one might guess that it should be a sharp cone with some suitable apex angle. However, it turns out that the best emitter is actually a whisker-type thin filament with a rounded tip, much like a CNT. It is as if the CNT has been designed by nature to be the best field emitter. Figure 4.39b and c shows SEM photographs of two CNT Spindt-type emitters. Figure 4.39b has several CNTs, and Figure 4.39c just one CNT for electron emission. (Which is more efficient?)

EXAMPLE 4.14

FIELD EMISSION Field emission displays operate on the principle that electrons can be readily emitted from a microscopic sharp point source (*cathode*) that is biased negatively with respect to a neighboring electrode (*gate* or *grid*) as depicted in Figure 4.38a. Emitted electrons impinge on colored phosphors on a screen and cause light emission by cathodoluminescence. There are millions of these microscopic field emitters to constitute the image. A particular field emission cathode in a field-emission-type flat panel display gives a current of 61.0 μA when the voltage between the cathode and the grid is 50 V. The current is 279 μA when the voltage is 58.2 V. What is the current when the voltage is 56.2 V?

SOLUTION

Equation 4.47 related I_A to V_G,

$$I_A = aV_G^2 \exp\left(-\frac{b}{V_G}\right)$$

where a and b are constants that can be determined from the two sets of data given. Thus,

$$61.0\ \mu\text{A} = a50^2 \exp\left(-\frac{b}{50}\right) \quad \text{and} \quad 279\ \mu\text{A} = a58.2^2 \exp\left(-\frac{b}{58.2}\right)$$

Dividing the first by the second gives

$$\frac{61.0}{279} = \frac{50^2}{58.2^2} \exp\left[-b\left(\frac{1}{50} - \frac{1}{58.2}\right)\right]$$

which can be solved to obtain $b = 431.75$ V and hence $a = 137.25\ \mu\text{A/V}^2$. At $V = 58.2$ V,

$$I = (137.25)(56.2)^2 \exp\left(-\frac{431.75}{56.2}\right) = 200\ \mu\text{A}$$

The experimental value for this device was 202 μA, which happens to be the device in Figure 4.37b (close).

4.10 PHONONS

4.10.1 HARMONIC OSCILLATOR AND LATTICE WAVES

Quantum Harmonic Oscillator In the classical picture of a solid, the constituent atoms are held together by bonds which can be represented by springs. According to the kinetic molecular theory, the atoms in a solid are constantly vibrating about their equilibrium positions by stretching and compressing their springs. The oscillations are

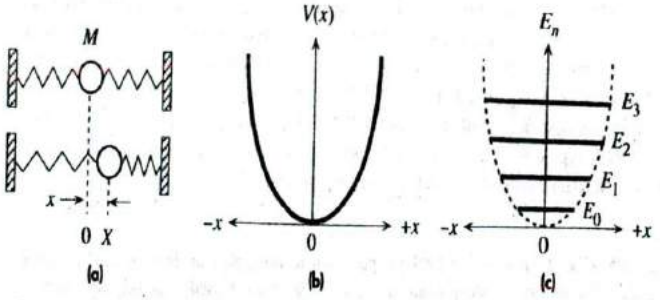


Figure 4.40

(a) Harmonic vibrations of an atom about its equilibrium position assuming its neighbors are fixed.
(b) The *PE* curve $V(x)$ versus displacement from equilibrium, x.
(c) The energy is quantized.

assumed to be simple harmonic so that the average kinetic and potential energies are the same. Figure 4.40a shows a one-dimensional independent simple harmonic oscillator that represents an atom of mass M attached by springs to fixed neighbors. The potential energy $V(x)$ is a function of displacement x from equilibrium. For small displacements, $V(x)$ is parabolic in x, as indicated in Figure 4.40b, that is,

Harmonic potential energy

$$V(x) = \frac{1}{2}\beta x^2 \qquad [4.48]$$

where β is a spring constant. The instantaneous energy, in principle, can be of any value. Equation 4.48 neglects the cubic term and is therefore symmetric about the equilibrium position at $x = 0$. It is called a **harmonic** approximation to the *PE* curve.

In modern physics, the energy of such a harmonic oscillator must be calculated using the *PE* in Equation 4.48 in the Schrödinger equation so that

Schrödinger equation: harmonic oscillator

$$\frac{d^2\psi}{dx^2} + \frac{2M}{\hbar^2}\left(E - \frac{1}{2}\beta x^2\right)\psi = 0 \qquad [4.49]$$

The solution of Equation 4.49 shows that the energy E_n of such a harmonic oscillator is quantized,

Energy of a harmonic oscillator

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega \qquad [4.50]$$

where ω is the angular frequency of the vibrations[14] and n is a quantum number $0, 1, 2, 3, \ldots$. The oscillation frequency is determined by the spring constant β and the mass M through $\omega = (\beta/M)^{1/2}$. Figure 4.40c shows the allowed energies of the quantum mechanical harmonic oscillator.

[14] Henceforth frequency will imply ω.

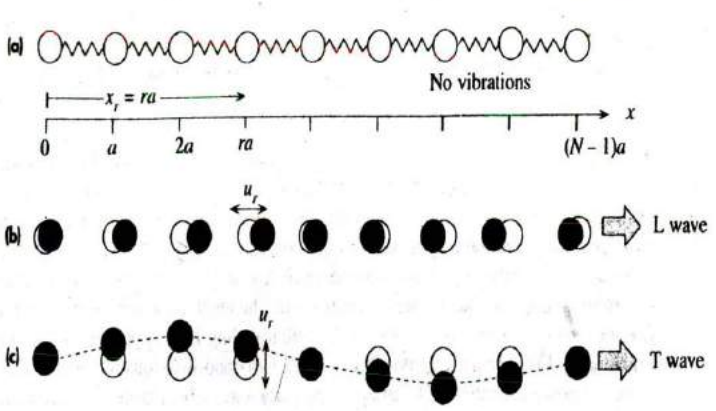


Figure 4.41

(a) A chain of N atoms through a crystal in the absence of vibrations.

(b) Coupled atomic vibrations generate a traveling longitudinal (L) wave along x. Atomic displacements (u_r) are parallel to x.

(c) A transverse (T) wave traveling along x. Atomic displacements (u_r) are perpendicular to the x axis. (b) and (c) are snapshots at one instant.

It is apparent that the minimum energy of the oscillator can never be zero but must be a finite value that is $E_0 = \frac{1}{2}\hbar\omega$. This energy is called the **zero-point energy.** As the temperature approaches 0 K, the harmonic oscillator would have an energy of E_0 and not zero. The energy levels are equally spaced by an amount $\hbar\omega$, which represents the amount of energy absorbed or emitted by the oscillator when it is excited and de-excited to a neighboring energy level. The vibrational energies of a molecule due to its atoms vibrating relative to each other, *e.g.*, the vibrations of the Cl_2 molecule in which the Cl–Cl bond is stretched and compressed, can also be described by Equation 4.50.

Phonons Atoms in a solid are coupled to each other by bonds. Atomic vibrations are therefore also coupled. These coupled vibrations lead to waves that involve cooperative vibrations of many atoms and cannot be represented by independent vibrations of individual atoms. Figure 4.41a shows a chain of atoms in a crystal. As an atom vibrates it transfers its energy to neighboring vibrating atoms and the coupled vibrations produce traveling wave-trains in the crystal.[15] (Consider grabbing and strongly vibrating the first atom in the atomic chain in Figure 4.41a. Your vibrations will be coupled and transferred by the springs to neighboring atoms in the chain along x.) Two examples are shown in Figure 4.41b and c. In the first, the atomic vibrations are parallel to the direction of propagation x and the wave is **a longitudinal wave.** In the second, the vibrations are transverse to the direction of propagation and the corresponding wave is a **transverse wave.** Suppose that x_r is the position of the rth atom in the absence of vibrations, that is, $x_r = ra$, where r is an integer from 0 to N, the number of atoms in the chain, as indicated in Figure 4.41a. By writing the mechanical equations (Newton's

[15] In the presence of coupling, the individual atoms do not execute simple harmonic motion.

second law) for the coupled atoms in Figure 4.41a, we can show that the displacement u_r from equilibrium at a location x_r is given by a **traveling-wave**-like behavior,[16]

Traveling-wave-type lattice vibrations

$$u_r = A\exp[j(Kx_r - \omega t)] \quad [4.51]$$

where A is the amplitude, K is a wavevector, and ω is the angular frequency. Notice that the Kx_r term is very much like the usual kx phase term of a traveling wave propagating in a continuous medium; the only difference is that Kx_r exists at discrete x_r locations. The wave-train described by Equation 4.51 in the crystal is called a **lattice wave.** Along the x direction it has a **wavelength** $\Lambda = 2\pi/K$ over which the longitudinal (or transverse) displacement u_r repeats itself. The displacement u_r repeats itself at one location over a time period $2\pi/\omega$. A wave traveling in the opposite direction to Equation 4.51 is of course also possible. Indeed, two oppositely traveling waves of the same frequency can interfere to set up a stationary wave which is also a lattice wave.

The lattice wave described by Equation 4.51 is a *harmonic oscillation* with a frequency ω that itself has no coupling to another lattice wave. The energy possessed by this lattice vibration is *quantized* in much the same way as the energy of the quantized harmonic oscillator in Equation 4.50. The energy of a lattice vibration therefore can only be multiples of $\hbar\omega$ above the zero-point energy, $\frac{1}{2}\hbar\omega$. The quantum of energy $\hbar\omega$ is therefore the smallest unit of lattice vibrational energy that can be added or subtracted from a lattice wave. The quantum of lattice vibration $\hbar\omega$ is called a **phonon** in analogy with the quantum of electromagnetic radiation, the photon. Whenever a lattice vibration interacts with another lattice vibration, an electron or a photon, in the crystal, it does so as if it had possessed a momentum of $\hbar K$. Thus,

Phonon energy

$$E_{\text{phonon}} = \hbar\omega = h\nu \quad [4.52]$$

Phonon momentum

$$p_{\text{phonon}} = \hbar K \quad [4.53]$$

The frequency of vibrations ω and the wavevector K of a lattice wave are related. If we were to use Equation 4.51 in the mechanical equations that describe the coupled atomic vibrations, we would find that

Dispersion relation

$$\omega = 2\left(\frac{\beta}{M}\right)^{1/2}\left|\sin\left(\frac{1}{2}Ka\right)\right| \quad [4.54]$$

which relates ω and K and is called the **dispersion relation.** Figure 4.42 shows how the frequency ω of the lattice waves increases with increasing wavevector K, or decreasing wavelength Λ. From Equation 4.54, there can be no frequencies higher than $\omega_{\max} = 2(\beta/M)^{1/2}$, which is the **lattice cut-off frequency.** Both longitudinal and transverse waves exhibit this type of dispersion relationship shown in Figure 4.42a though their exact ω–K curves would be different depending on the nature of interatomic bonding and the crystal structure. The dispersion relation in Equation 4.54 is periodic in K with a period $2\pi/a$. Only values of K in the range $-\pi/a < K < \pi/a$ are physically meaningful. A point A with K_A is the same as a point B with K_B because we can shift K by the period, $2\pi/a$ as shown in Figure 4.42a.

[16] The exponential notation for a wave is convenient, but we have to consider only the real part to actually represent the wave in the physical world.

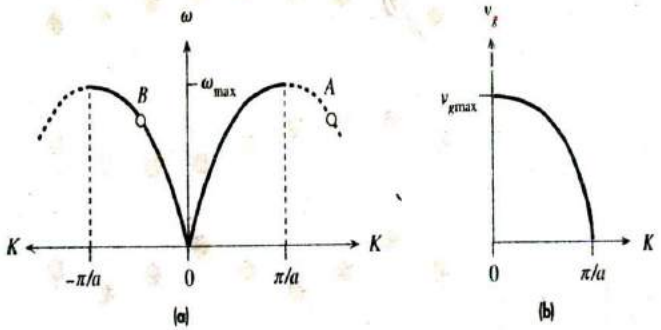


Figure 4.42
(a) Frequency ω versus wavevector K relationship for lattice waves.
(b) Group velocity v_g versus wavevector K.

The velocity at which traveling waves carry energy is called the **group velocity** v_g of the wave.[17] It depends on the slope $d\omega/dK$ of the ω–K dispersion curve, so for lattice waves,

$$v_g = \frac{d\omega}{dK} = \left(\frac{\beta}{M}\right)^{1/2} a\cos\left(\frac{1}{2}Ka\right) \qquad [4.55]$$

Group velocity

which is shown in Figure 4.42b. Points A and B in Figure 4.42a have the same group velocity and are equivalent.

The number of distinct or independent lattice waves, with different wavevectors, in a crystal is not infinite but depends on the number of atoms N. Consider a linear crystal as in Figure 4.43 with many atoms. We will take N to be large and ignore the difference between N and $N-2$. The lattice waves in this crystal would be standing waves represented by two oppositely traveling waves. The crystal length $L = Na$ can support multiples of the half-wavelength $\frac{1}{2}\Lambda$ as indicated in Figure 4.43,

$$q\frac{\Lambda}{2} = L = Na \qquad q = 1, 2, 3, \ldots \qquad [4.56a]$$

Vibrational modes

or

$$K = \frac{q\pi}{L} = \frac{q\pi}{Na} \qquad q = 1, 2, 3, \ldots \qquad [4.56b]$$

Vibrational modes

where q is an integer. Each particular K value K_q represents one distinct lattice wave with a particular frequency as determined by the dispersion relation. Four examples are shown in Figure 4.43. Each of these K_q values defines a **mode** or **state of lattice vibration.** Each mode is an independent lattice vibration. Its energy can be increased or decreased only by a quantum amount of $\hbar\omega$. Since K_q values outside the range $-\pi/a < K < \pi/a$ are the same as those in that range (A and B are the same

[17] For those readers who are not familiar with the group velocity concept, this is discussed in Chapter 9 without prerequisite material.

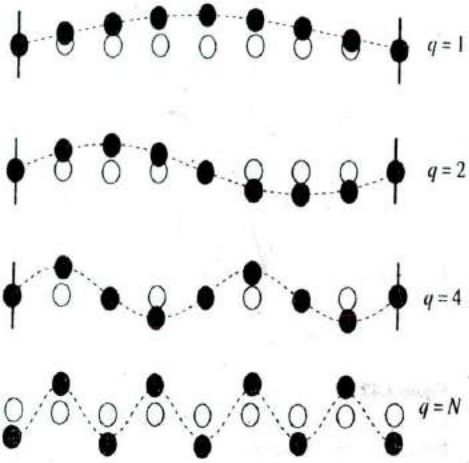


Figure 4.43 Four examples of standing waves in a linear crystal corresponding to $q = 1, 2, 4$, and N. q is maximum when alternating atoms are vibrating in opposite directions. A portion from a very long crystal is shown.

in Figure 4.42a), it is apparent that the maximum value of q is N and thus the **number of modes** is also N. Notice that as q increases, Λ decreases. The smallest Λ occurs when alternating atoms in the crystal are moving in opposite directions which corresponds to $\frac{1}{2}\Lambda = a$, that is, $q = N$, as shown in Figure 4.43. In terms of the wavevector, $K = 2\pi/\Lambda = \pi/a$. Smaller wavelengths or longer wavevectors are meaningless and correspond to shifting K by a multiple of $2\pi/a$. Since N is large, the ω versus K curve in Figure 4.42a consists of very finely separated distinct points, each corresponding to a particular q, analogous to the energy levels in an energy band.

The above ideas for the linear chain of atoms can be readily extended to a three-dimensional crystal. If L_x, L_y, and L_z are the sides of the solid along the x, y, and z axes, with N_x, N_y, and N_z number of atoms, respectively, then the wavevector components along x, y, and z are

Lattice vibrational modes in 3-D

$$K_x = \frac{q_x\pi}{L_x} \qquad K_y = \frac{q_y\pi}{L_y} \qquad K_z = \frac{q_z\pi}{L_z} \tag{4.57}$$

where the integers q_x, q_y, and q_z run from 1 to N_x, N_y, and N_z, respectively. The total number of permitted modes is $N_xN_yN_z$ or N, the total number of atoms in the solid. Vibrations however can be set up independently along the x, y, and z directions so that the actual *number of independent modes* is $3N$.

4.10.2 DEBYE HEAT CAPACITY

The heat capacity of a solid represents the increase in the internal energy of the crystal per unit increase in the temperature. The increase in the internal energy is due to an increase in the energy of lattice vibrations. This is generally true for all the solids except metals at very low temperatures where the heat capacity is due to the electrons

near the Fermi level becoming excited to higher energies. For most practical temperature ranges of interest, the heat capacity of solids is determined by the excitation of lattice vibrations. The **molar heat capacity** C_m is the increase in the internal energy U_m of a crystal of N_A atoms per unit increase in the temperature at constant volume,[18] that is, $C_m = dU_m/dT$.

The simplest approach to calculating the average energy is first to assume that all the lattice vibrational modes have the same frequency ω. (We will account for different modes having different frequencies later.) If E_n is the energy of a harmonic oscillator such as a lattice vibration, then the average energy, by definition, is given by

$$\overline{E} = \frac{\sum_{n=0}^{\infty} E_n P(E_n)}{\sum_{n=0}^{\infty} P(E_n)} \qquad [4.58]$$

Average energy of oscillators

where $P(E_n)$ is the probability that the vibration has the energy E_n which is proportional to the Boltzmann factor. Thus we can use $P(E_n) \propto \exp(-E_n/kT)$ and $E_n = (n + \frac{1}{2})\hbar\omega$ in Equation 4.58. We can drop the zero-point energy as this does not affect the heat capacity (which deals with energy *changes*). The substitution and calculation of Equation 4.58 yields the vibrational mean energy at a frequency ω,

$$\overline{E}(\omega) = \frac{\hbar\omega}{\exp\left(\frac{\hbar\omega}{kT}\right) - 1} \qquad [4.59]$$

Average energy of oscillators at ω

This energy increases with temperature. Each phonon has an energy of $\hbar\omega$. Thus, the *phonon concentration in the crystal increases with temperature*; increasing the temperature creates more phonons.

To find the internal energy due to *all* the lattice vibrations we must also consider how many modes there are at various frequencies, that is, the distribution of the modes over the possible frequencies, the spectrum of the vibrations. Suppose that $g(\omega)$ is the number of modes per unit frequency, that is, $g(\omega)$ is the **density of vibrational states** or modes. Then $g(\omega)\,d\omega$ is the number of states in the range $d\omega$. The internal energy U_m of all lattice vibrations for 1 mole of solid is

$$U_m = \int_0^{\omega_{max}} \overline{E}(\omega) g(\omega)\, d\omega \qquad [4.60]$$

Internal energy of all lattice vibrations

The integration is up to a certain allowed maximum frequency ω_{max} (Figure 4.42a). The density of states $g(\omega)$ for the lattice vibrations can be found in a similar fashion to the density of states for electrons in an energy band, and we will simply quote the result,

$$g(\omega) \approx \frac{3V}{2\pi^2} \frac{\omega^2}{v^3} \qquad [4.61]$$

Density of states for lattice vibrations

[18] Constant volume in the definition means that the heat added to the system increases the internal energy without doing mechanical work by changing the volume.

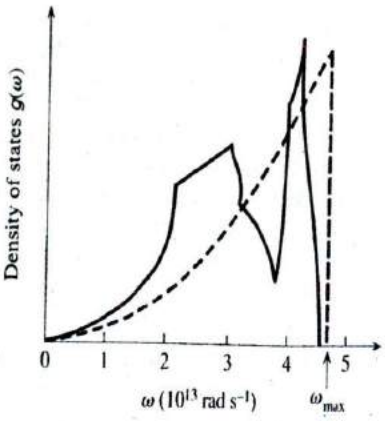


Figure 4.44 Density of states for phonons in copper. The solid curve is deduced from experiments on neutron scattering. The broken curve is the three-dimensional Debye approximation, scaled so that the areas under the two curves are the same.

This requires that $\omega_{max} \approx 4.5 \times 10^{13}$ rad s^{-1}, or a Debye characteristic temperature $T_D = 344$ K.

where v is the mean velocity of longitudinal and transverse waves in the solid and V is the volume of the crystal. Figure 4.44 shows the spectrum $g(\omega)$ for a real crystal such as Cu and the expression in Equation 4.61. The maximum frequency is ω_{max} and is determined by the fact that the total number of modes up to ω_{max} must be $3N_A$. It is called the **Debye frequency.** Thus, integrating $g(\omega)$ up to ω_{max} we find,

Debye frequency

$$\omega_{max} \approx v(6\pi^2 N_A/V)^{1/3} \qquad [4.62]$$

This maximum frequency ω_{max} corresponds to an energy $\hbar\omega_{max}$ and to a temperature T_D defined by,

Debye temperature

$$T_D = \frac{\hbar\omega_{max}}{k} \qquad [4.63]$$

and is called the **Debye temperature.** Qualitatively, it represents the temperature above which all vibrational frequencies are executed by the lattice waves.

Thus, by using Equations 4.59 to 4.63 in Equation 4.60 we can evaluate U_m and hence differentiate U_m with respect to temperature to obtain the molar heat capacity at constant volume,

Heat capacity: lattice vibrations

$$C_m = 9R\left(\frac{T}{T_D}\right)^3 \int_0^{T_D/T} \frac{x^4 e^x\, dx}{(e^x - 1)^2} \qquad [4.64]$$

which is the Debye heat capacity expression.

Figure 4.45 represents the constant-volume molar heat capacity C_m of nearly all crystals, Equation 4.64, as a function of temperature, normalized with respect to the Debye temperature. The **Dulong–Petit rule** of $C_m = 3R$ is only obeyed when $T > T_D$. Notice that C_m at $T = 0.5T_D$ is $0.825(3R)$ whereas at $T = T_D$ it is $0.952(3R)$. For most practical purposes, C_m is to within 6 percent of $3R$ when the temperature is at $0.9T_D$. For example, for copper $T_D = 315$ K and above

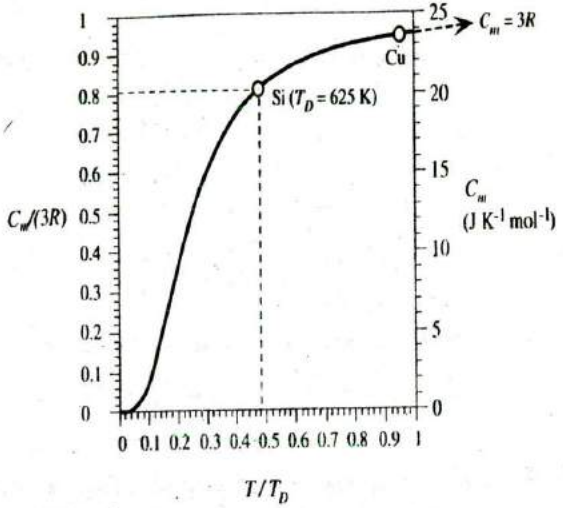


Figure 4.45 Debye constant-volume molar heat capacity curve.
The dependence of the molar heat capacity C_m on temperature with respect to the Debye temperature: C_m versus T/T_D. For Si, $T_D = 625$ K, so at room temperature (300 K), $T/T_D = 0.48$ and C_m is only 0.81 (3R).

about $0.9T_D$, that is, above 283 K (or 10 °C), $C_m \approx 3R$, as borne out by experiments.[19] Table 4.5 provides typical values for T_D, and heat capacities for a few selected elements. It is left as an exercise to check the accuracy of Equation 4.64 for predicting the heat capacity given the T_D values. At the lowest temperatures when $T \ll T_D$, Equation 4.64 predicts that $C_m \propto T^3$, and this is indeed observed in low-temperature heat capacity experiments on a variety of crystals.[20]

It is useful to provide a physical picture of the Debye model inherent in Equation 4.64. As the temperature increases from near zero, the increase in the crystal's vibrational energy is due to *more* phonons being created and *higher* frequencies being excited. The phonon concentration increases as T^3, and the mean phonon energy increases as T. Thus, the internal energy increases as T^4. At temperatures above T_D, increasing the temperature creates *more* phonons but does not increase the mean phonon energy and does not excite higher frequencies. All frequencies up to ω_{max} have now been excited. The internal energy increases only due to more phonons being created. The phonon concentration and hence the internal energy increase as T; the heat capacity is constant as expected from Equation 4.64.

[19] Sometimes it is stated that the Debye temperature is a characteristics temperature for each material at which all the atoms are able to possess vibrational kinetic energies in accordance with the Maxwell equipartition of energy principle; that is, the average vibrational kinetic energy will be $\frac{3}{2}kT$ per atom and average potential energy will also be $\frac{3}{2}kT$. This means that the average energy per atom is $3kT$, and hence the heat capacity is $3kN_A$ or $3R$ per mole which is the *Dulong–Petit rule*.

[20] Well-known exceptions are glasses, noncrystalline solids, whose heat capacity is proportional to $a_1T + a_2T^3$, where a_1 and a_2 are constants.

Table 4.5 Debye temperatures T_D, heat capacities, and thermal conductivities of selected elements

	Crystal							
	Ag	Be	Cu	Diamond	Ge	Hg	Si	W
T_D(K)*	215	1000	315	1860	360	100	625	310
C_m(J K^{-1} mol^{-1})†	25.6	16.46	24.5	6.48	23.38	27.68	19.74	24.45
c_s(J K^{-1} g^{-1})†	0.237	1.825	0.385	0.540	0.322	0.138	0.703	0.133
κ (W m^{-1} K^{-1})†	429	183	385	1000	60	8.65	148	173

* T_D is obtained by fitting the Debye curve to the experimental molar heat capacity data at the point $C_m = \frac{1}{2}$ [38].
† C_m, c_s, and κ are at 25 °C.
SOURCE: T_D data from J. De Launay, *Solid State Physics*, vol. 2, F. Seitz and D. Turnbull, eds., Academic Press, New York, 1956.

It is apparent that, above the Debye temperature, the increase in temperature leads to the creation of more phonons. In Chapters 1 and 2, using classical concepts only, we had mentioned that increasing the temperature increases the magnitude of atomic vibrations. This simple and intuitive classical concept in terms of modern physics corresponds to creating more phonons with temperature. We can use the photon analogy from Chapter 3. When we increase the intensity of light of a given frequency, classically we simply increase the electric field (magnitude of the vibrations), but in modern physics we have to increase the number of photons flowing per unit area.

EXAMPLE 4.15 **SPECIFIC HEAT CAPACITY OF Si** Find the specific heat capacity c_s of a silicon crystal at room temperature given $T_D = 625$ K for Si.

SOLUTION

At room temperature, $T = 300$ K, $(T/T_D) = 0.48$, and, from Figure 4.45, the molar heat capacity is

$$C_{at} = 0.81(3R) = 20.2 \text{ J K}^{-1} \text{ mol}^{-1}$$

The specific heat capacity c_s from the Debye curve is

$$c_s = \frac{C_{at}}{M_{at}} \approx \frac{(0.81 \times 25 \text{ J K}^{-1} \text{ mol}^{-1})}{(28.09 \text{ g mol}^{-1})} = 0.72 \text{ J K}^{-1} \text{ g}^{-1}$$

The experimental value of 0.70 J K^{-1} g^{-1} is very close to the Debye value.

EXAMPLE 4.16 **SPECIFIC HEAT CAPACITY OF GaAs** Example 4.15 applied Equation 4.64, the Debye molar heat capacity C_m, to the silicon crystal in which all atoms are of the same type. It was relatively simple to calculate the specific heat capacity c_s (what is really used in engineering) from the molar heat capacity C_{at} by using $c_s = C_m/M_{at}$ where M_{at} is the atomic mass of the type of atom (only one) in the crystal. When the crystal has two types of atoms, we must modify the specific heat capacity derivation. We can still keep the symbol C_{at} to represent the Debye molar heat capacity given in Equation 4.64. Consider a GaAs crystal that has N_A units of GaAs, that is,

1 mole of GaAs. There will be 1 mole (N_A atoms) of Ga and 1 mole of As atoms. To a reasonable approximation we can assume that each mole of Ga and As contributes a C_m amount of heat capacity so that the total heat capacity of 1 mole GaAs will be $C_m + C_m$ or $2C_m$, a maximum of 50 J K^{-1} mol^{-1}. The total mass of this 1 mole of GaAs is $M_{Ga} + M_{As}$. Thus, the specific heat capacity of GaAs is

Specific heat capacity of GaAs

$$c_s = \frac{C_{total}}{M_{total}} = \frac{C_m + C_m}{M_{Ga} + M_{As}} = \frac{2C_m}{M_{Ga} + M_{As}}$$

which can alternatively be written as

Specific heat capacity of a polyatomic crystal

$$c_s = \frac{C_m}{\frac{1}{2}(M_{Ga} + M_{As})} = \frac{C_m}{\overline{M}}$$

where $\overline{M} = (M_{Ga} + M_{As})/2$ is the average atomic mass of the constituent atoms. Although we derived c_s for GaAs, it can also be applied to other compounds by suitably calculating an average atomic mass $\overline{M}$. GaAs has a Debye temperature $T_D = 344$ K, so that at a room temperature of 300 K, $T/T_D = 0.87$, and from Figure 4.45, $C_m/(3R) = 0.94$. Therefore,

$$c_s = \frac{C_m}{\overline{M}} = \frac{(0.94)(25 \text{ J K}^{-1} \text{ mol}^{-1})}{\frac{1}{2}(69.72 \text{ g mol}^{-1} + 74.92 \text{ g mol}^{-1})} = 0.325 \text{ J K}^{-1} \text{ g}^{-1}$$

At −40 °C, $T/T_D = 0.68$, and $C_m/(3R) = 0.90$, so the new $c_s = (0.90/0.94)(0.325) = 0.311$ J K^{-1} g^{-1}, which is not a large change in c_s.

The heat capacity per unit volume C_v can be found from $C_v = c_s\rho$, where ρ is the density. Thus, at 300 K, $C_v = (0.325 \text{ J K}^{-1} \text{ g}^{-1})(5.32 \text{ g cm}^{-3}) = 1.73 \text{ J K}^{-1} \text{ cm}^{-3}$. The calculated c_s match the reported experimental values very closely.

EXAMPLE 4.17

LATTICE WAVES AND SOUND VELOCITY Consider *longitudinal* waves in a linear crystal and three atoms at $r - 1$, r, and $r + 1$ as in Figure 4.46. The displacement of each atom from equilibrium in the $+x$ direction is u_{r-1}, u_r, and u_{r+1}, respectively. Consider the rth atom. Its bond with the left neighbor stretches by $(u_r - u_{r-1})$. Its bond with the right neighbor stretches by $(u_{r+1} - u_r)$. The left spring exerts a force $\beta(u_r - u_{r-1})$, and the right spring exerts a force $\beta(u_{r+1} - u_r)$. The net force on the rth atom is mass × acceleration,

$$\text{Net force} = \beta(u_{r+1} - u_r) - \beta(u_r - u_{r-1}) = M\frac{d^2u_r}{dt^2}$$

so

$$M\frac{d^2u_r}{dt^2} = \beta(u_{r+1} - 2u_r + u_{r-1}) \qquad [4.65]$$

Wave equation

This is the **wave equation** that describes the coupled longitudinal vibrations of the atoms in the crystal. A similar expression can also be derived for transverse vibrations. We can substitute Equation 4.51 in Equation 4.65 to show that Equation 4.51 is indeed a solution of the wave

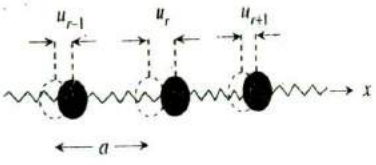


Figure 4.46 Atoms executing longitudinal vibrations parallel to x.

equation. It is assumed that the crystal response is **linear**, that is, the net force is proportional to net displacement.

The **group velocity** of lattice waves is given by Equation 4.55. For sufficiently small K, or long wavelengths, such that $\frac{1}{2}Ka \ll 1$,

Long-wavelength group velocity

$$v_g = \left(\frac{\beta}{M}\right)^{1/2} a\cos\left(\frac{1}{2}Ka\right) \approx \left(\frac{\beta}{M}\right)^{1/2} a \qquad [4.66]$$

which is a constant. It is the slope of the straight-line region of ω versus K curve for small K values in Figure 4.42. Furthermore, the elastic modulus Y depends on the slope of the net force versus displacement curve as derived in Example 1.5. From Equation 4.48 $F_N = dV/dx = \beta x$ and hence $Y = \beta/a$. Moreover, each atom occupies a volume of a^3, so the density ρ is M/a^3. Substituting both of these results in Equation 4.66 yields

Longitudinal elastic wave velocity

$$v_g \approx \left(\frac{Y}{\rho}\right)^{1/2} \qquad [4.67]$$

The relationship has to be modified for an actual crystal incorporating a small numerical factor multiplying Y. Aluminum has a density of 2.7 g cm^{-3} and $Y = 70$ GPa, so the long-wavelength longitudinal velocity from Equation 4.67 is 5092 m s^{-1}. The sound velocity in Al is 5100 m s^{-1}, which is very close.

4.10.3 THERMAL CONDUCTIVITY OF NONMETALS

In nonmetals the heat transfer involves lattice vibrations, that is, phonons. The heat absorbed in the hot region increases the amplitudes of the lattice vibrations, which is the same as generating more phonons. These new phonons travel toward the cold regions and thereby transport the lattice energy from the hot to cold end. The **thermal conductivity** κ measures the rate at which heat can be transported through a medium per unit area per unit temperature gradient. It is proportional to the rate at which a medium can absorb energy; that is, κ is proportional to the heat capacity. κ is also proportional to the rate at which phonons are transported which is determined by their mean velocity v_{ph}. In addition, of course, κ is proportional to the *mean free path* ℓ_{ph} that a phonon has to travel before losing its momentum just as the electrical conductivity is proportional to the electron's mean free path. A rigorous classical treatment gives κ as

Thermal conductivity due to phonons

$$\kappa = \tfrac{1}{3}C_v v_{ph}\ell_{ph} \qquad [4.68]$$

where C_v is the heat capacity per unit volume. The mean free path ℓ_{ph} depends on various processes that can scatter the phonons and *hinder* their propagation along the direction of heat flow. Phonons collide with other phonons, crystal defects, impurities, and crystal surfaces.

The mean phonon velocity v_{ph} is constant and approximately independent of temperature. At temperatures above the Debye temperature, C_v is constant and, thus, $\kappa \propto \ell_{ph}$. The mean free path of phonons at these temperatures is determined by phonon–phonon collisions, that is, phonons interacting with other phonons as depicted in Figure 4.47. Since the phonon concentration n_{ph} increases with temperature, $n_{ph} \propto T$, the mean free path decreases as $\ell_{ph} \propto 1/T$. Thus, κ decreases with increasing temperature as observed for most crystals at sufficiently high temperatures.

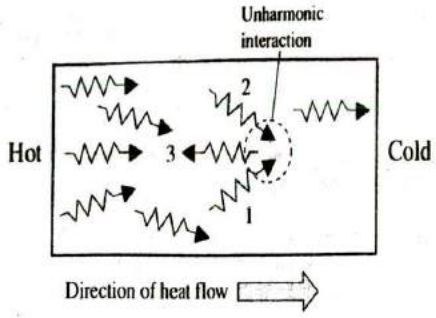


Figure 4.47 Phonons generated in the hot region travel toward the cold region and thereby transport heat energy. Phonon–phonon unharmonic interaction generates a new phonon whose momentum is toward the hot region.

The phonon–phonon collisions that are responsible for limiting the thermal conductivity, that is, scattering the phonon momentum in the opposite direction to the heat flow, are due to the **unharmonicity (asymmetry)** of the interatomic potential energy curve. Stated differently, the net force F acting on an atom is not simply βx but also has an x^2 term; it is **nonlinear.** The greater the asymmetry or nonlinearity, the larger is the effect of such momentum flipping collisions. The same asymmetry that is responsible for thermal expansion of solids is also responsible for determining the thermal conductivity. When two phonons 1 and 2 interact in a crystal region as in Figure 4.47, the *nonlinear* behavior and the *periodicity* of the lattice cause a new phonon 3 to be generated. This new phonon 3 has the same energy as the sum of 1 and 2, but it is traveling in the wrong direction! (The frequency of 3 is the sum of the frequencies of 1 and 2.)

At low temperatures there are two factors. The phonon concentration is too low for phonon–phonon collisions to be significant. Instead, the mean free path ℓ_{ph} is determined by phonon collisions with crystal imperfections, most significantly, crystal surfaces and grain boundaries. Thus, ℓ_{ph} depends on the sample geometry and crystallinity. Further, as we expect from the Debye model, C_v depends on T^3, so κ has the same temperature dependence as C_v, that is, $\kappa \propto T^3$. Between the two temperature regimes κ exhibits a peak as shown in Figure 4.48 for sapphire (crystalline Al_2O_3) and

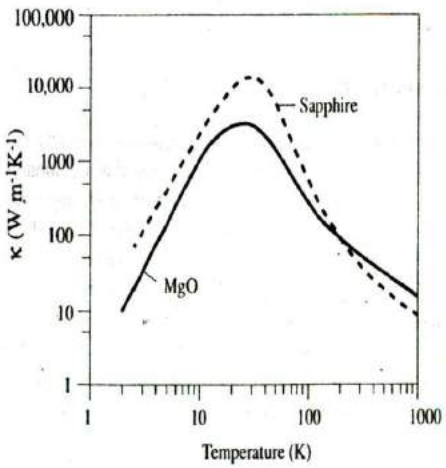


Figure 4.48 Thermal conductivity of sapphire and MgO as a function of temperature.

MgO crystals. Even though there are no conduction electrons in these two example crystals, they nonetheless exhibit substantial thermal conductivity.

EXAMPLE 4.18 **PHONONS IN GaAs** Estimate the phonon mean free path in GaAs at room temperature 300 K and at 20 K from its κ, C_v, and v_{ph}, using Equation 4.68. At room temperature, semiconductor data handbooks list the following for GaAs: $\kappa = 45\ \text{W m}^{-1}\ \text{K}^{-1}$, elastic modulus $Y = 85$ GPa, density $\rho = 5.32\ \text{g cm}^{-3}$, and specific heat capacity $c_v = 0.325\ \text{J K}^{-1}\ \text{g}^{-1}$. At 20 K, $\kappa = 4000\ \text{W m}^{-1}\ \text{K}^{-1}$ and $c_v = 0.0052\ \text{J K}^{-1}\ \text{g}^{-1}$. Y and ρ and hence v_{ph} do not change significantly with temperature compared with the changes in κ and C_v with temperature.

SOLUTION

The phonon velocity v_{ph} from Equation 4.67 is approximately

$$v_{ph} \approx \sqrt{\frac{Y}{\rho}} = \sqrt{\frac{85 \times 10^9\ \text{N m}^{-2}}{5.32 \times 10^3\ \text{kg m}^{-3}}} = 4000\ \text{m s}^{-1}$$

Heat capacity per unit volume $C_v = c_v\rho = (325\ \text{J K}^{-1}\ \text{kg}^{-1})(5320\ \text{kg m}^{-3}) = 1.73 \times 10^6\ \text{J K}^{-1}\ \text{m}^{-3}$. From Equation 4.68, $\kappa = \frac{1}{3}C_v v_{ph}\ell_{ph}$.

$$\ell_{ph} = \frac{3\kappa}{C_v v_{ph}} = \frac{(3)(45\ \text{W m}^{-1}\ \text{K}^{-1})}{(1.73 \times 10^6\ \text{J K}^{-1}\ \text{m}^{-3})(4000\ \text{m s}^{-1})} = 2.0 \times 10^{-8}\text{m} \quad \text{or} \quad 20\ \text{nm}$$

We can easily repeat the calculation at 20 K, given $\kappa \approx 4000\ \text{W m}^{-1}\ \text{K}^{-1}$ and $c_v = 5.2\ \text{J K}^{-1}\ \text{kg}^{-1}$, so $C_v = c_v\rho \approx (5.2\ \text{J K}^{-1}\ \text{kg}^{-1})(5320\ \text{kg m}^{-3}) = 2.77 \times 10^4\ \text{J K}^{-1}\ \text{m}^{-3}$. Y and ρ and hence $v_{ph}(\approx 4000\ \text{m s}^{-1})$, do not change significantly with temperature compared with κ and C_v. Thus,

$$\ell_{ph} = \frac{3\kappa}{C_v v_{ph}} \approx \frac{(3)(4 \times 10^3\ \text{W m}^{-1}\ \text{K}^{-1})}{(2.77 \times 10^4\ \text{J K}^{-1}\ \text{m}^{-3})(4000\ \text{m s}^{-1})} = 1.1 \times 10^{-4}\text{m} \quad \text{or} \quad 0.011\ \text{cm}$$

For small specimens, the above phonon mean free path will be comparable to the sample size, which means that ℓ_{ph} will actually be limited by the sample size. Consequently κ will depend on the sample dimensions, being smaller for smaller samples, similar to the dependence of the electrical conductivity of thin films on the film thickness.

4.10.4 ELECTRICAL CONDUCTIVITY

Except at low temperatures, the electrical conductivity of metals is primarily controlled by scattering of electrons around E_F by lattice vibrations, that is, phonons. These electrons have a speed $v_F = (2E_F/m_e)^{1/2}$ and a momentum of magnitude $m_e v_F$. We know that the electrical conductivity σ is proportional to the mean collision time τ of the electrons, that is, $\sigma \propto \tau$. This scattering time assumes that each scattering process is 100 percent efficient in randomizing the electron's momentum, that is, destroying the momentum gained from the field, which may not be the case. If it takes on average N collisions to randomize the electron's momentum, and τ is the mean time between the scattering events, then the *effective* scattering time is simply $N\tau$ and $\sigma \propto N\tau$. ($1/N$ indicates the efficiency of each scattering process in randomizing the velocity.)

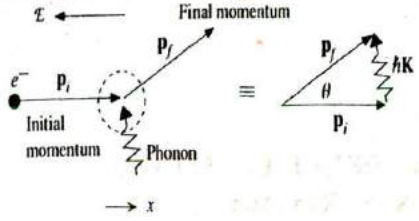


Figure 4.49 Low-angle scattering of a conduction electron by a phonon.

Figure 4.49 shows an example in which an electron with an initial momentum $\mathbf{p}_i$ collides with a lattice vibration of momentum $\hbar\mathbf{K}$. The result of the interaction is that the electron's momentum is deflected through a small angle θ to $\mathbf{p}_f$ which still has a component along the original direction x. This is called a low-angle scattering process. It will take many such collisions to reverse the electron's momentum which corresponds to flipping the momentum along the $+x$ direction to the $-x$ direction. Recall that the momentum gained from the field is actually very small compared with the momentum of the electron which is $m_e v_F$. A scattered electron must have an energy close to E_F because lower energy states are filled. Thus, $\mathbf{p}_i$ and $\mathbf{p}_f$ have approximately the same magnitude $p_i = p_f = m_e v_F$ as shown in Figure 4.49.

At temperatures above the Debye temperature, we can assume that most of the phonons are vibrating with the Debye frequency ω_{max} and the phonon concentration n_{ph} increases as T. These phonons have sufficient energies and momenta to fully scatter the electron on impact. Thus,

$$\sigma \propto \tau \propto \frac{1}{n_{ph}} \propto \frac{1}{T} \qquad [4.69a]$$

Electrical conductivity $T > T_D$

When $T < T_D$, the phonon concentration follows $n_{ph} \propto T^3$, and the mean phonon energy $\overline{E}_{ph} \propto T$, because, as the temperature is raised, higher frequencies are excited. However, these phonons have low energy and small momenta, thus they only cause small-angle scattering processes as in Figure 4.49. The average phonon momentum $\hbar K$ is also proportional to the temperature (recall that at low frequencies Figure 4.42a shows that $\hbar\omega \propto \hbar K$). It will take many such collisions, say N, to flip the electron's momentum by $2m_e v_F$ from $+m_e v_F$ to $-m_e v_F$. During each collision, a phonon of momentum $\hbar K$ is absorbed as shown in Figure 4.49. Thus, if all phonons deflected the electron in the same angular direction, the collisions would sequentially add to θ in Figure 4.49, and we will need $(2m_e v_F)/(\hbar K)$ number of steps to flip the electron's momentum. The actual collisions add θ's randomly and the process is similar to particle diffusion, random walk, in Example 1.12 ($L^2 = Na^2$, where L = displaced distance after N jumps and a = jump step). Thus,

$$N = \frac{(2m_e v_F)^2}{(\hbar K)^2} \propto \frac{1}{T^2}$$

The conductivity is therefore given by

$$\sigma \propto N\tau \propto \frac{N}{n_{ph}} \propto \frac{1}{T^5} \qquad [4.69b]$$

Electrical conductivity $T < T_D$

which is indeed observed for Cu in Figure 2.8 when $T < T_D$ over the range where impurity scattering is negligible.

ADDITIONAL TOPICS

4.11 BAND THEORY OF METALS: ELECTRON DIFFRACTION IN CRYSTALS

A rigorous treatment of the band theory of solids involves extensive quantum mechanical analysis and is beyond the scope of this book. However, we can attain a satisfactory understanding through a semiquantitative treatment.

We know that the wavefunction of the electron moving freely along x in space is a traveling wave of the spatial form $\psi_k(x) = \exp(jkx)$, where k is the wavevector $k = 2\pi/\lambda$ of the electron and $\hbar k$ is its momentum. Here, $\psi_k(x)$ represents a traveling wave because it must be multiplied by $\exp(-j\omega t)$, where $\omega = E/\hbar$, to get the total wavefunction $\Psi(x, t) = \exp[j(kx - \omega t)]$.

We will assume that an electron moving freely within the crystal and within a given energy band should also have a traveling wave type of wavefunction,

$$\psi_k(x) = A \exp(jkx) \qquad [4.70]$$

where k is the electron wavevector in the crystal and A is the amplitude. This is a reasonable expectation, since, to a first order, we can take the PE of the electron inside a solid as zero, $V = 0$. Yet, the PE must be large outside, so the electron is contained within the crystal. When the PE is zero, Equation 4.70 is a solution to the Schrödinger equation. The momentum of the electron described by the traveling wave Equation 4.70 is then $\hbar k$ and its energy is

$$E_k = \frac{(\hbar k)^2}{2m_e} \qquad [4.71]$$

The electron, as a traveling wave, will freely propagate through the crystal. However, not all traveling waves, can propagate in the lattice. The electron cannot have any k value in Equation 4.70 and still move through the crystal. Waves can be reflected and diffracted, whether they are electron waves, X-rays, or visible light. Diffraction occurs when reflected waves interfere constructively. Certain k values will cause the electron wave to be diffracted, preventing the wave from propagating.

The simplest illustration that certain k values will result in the electron wave being diffracted is shown in Figure 4.50 for a hypothetical linear lattice in which diffraction is simply a reflection (what we call diffraction becomes Bragg reflection). The electron is assumed to be propagating in the forward direction along x with a traveling wave function of the type in Equation 4.70. At each atom, some of this wave will be reflected. At A, the reflected wave is A' and has a magnitude A'. If the reflected waves A', B', and C' will reinforce each other, a full reflected wave will be created, traveling in the backward direction. The reflected waves A', B', C', ... will reinforce each other if the path difference between A', B', C', ... is $n\lambda$, where λ is the wavelength and $n = 1, 2, 3, \ldots$ is an integer. When wave B' reaches A', it has traveled an additional

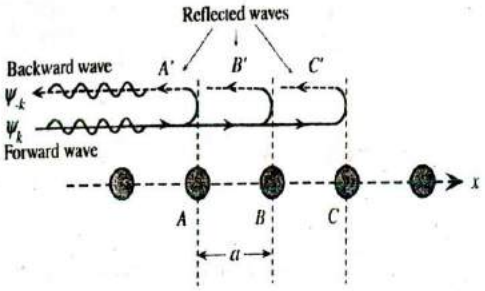


Figure 4.50 An electron wave propagation through a linear lattice.
For certain k values, the reflected waves at successive atomic planes reinforce each other, giving rise to a reflected wave traveling in the backward direction. The electron cannot then propagate through the crystal.

distance of $2a$. The path difference between A' and B' is therefore $2a$. For A' and B' to reinforce each other, we need

$$2a = n\lambda \qquad n = 1, 2, 3, \ldots$$

Substituting $\lambda = 2\pi/k$, we obtain the condition in terms of k

$$k = \frac{n\pi}{a} \qquad n = 1, 2, 3, \ldots \qquad [4.72]$$

Thus, whenever k is such that it satisfies the condition in Equation 4.72, all the reflected waves reinforce each other and produce a backward-traveling, reflected wave of the following form (with a negative k value):

$$\psi_{-k}(x) = A\exp(-jkx) \qquad [4.73]$$

This wave will also probably suffer a reflection, since its k satisfies Equation 4.72, and the reflections will continue. The crystal will then contain waves traveling in the forward and backward directions. These waves will interfere to give **standing waves** inside the crystal. Hence, whenever the k value satisfies Equation 4.72, traveling waves cannot propagate through the lattice. Instead, there can only be standing waves. For k satisfying Equation 4.72, the electron wavefunction consists of waves ψ_k and ψ_{-k} interfering in two possible ways to give two possible standing waves:

$$\psi_c(x) = A\exp(jkx) + A\exp(-jkx) = A_c\cos\left(\frac{n\pi x}{a}\right) \qquad [4.74]$$

$$\psi_s(x) = A\exp(jkx) - A\exp(-jkx) = A_s\sin\left(\frac{n\pi x}{a}\right) \qquad [4.75]$$

The probability density distributions $|\psi_c(x)|^2$ and $|\psi_s(x)|^2$ for the two standing waves are shown in Figure 4.51. The first standing wave $\psi_c(x)$ is at a maximum on the ion cores, and the other $\psi_s(x)$ is at a maximum between the ion cores. Note also that both the standing waves $\psi_c(x)$ and $\psi_s(x)$ are solutions to the Schrödinger equation.

The closer the electron is to a positive nucleus, the lower is its electrostatic PE, by virtue of $-e^2/4\pi\varepsilon_o r$. The PE of the electron distribution in $\psi_c(x)$ is lower than that in $\psi_s(x)$, because the maxima for $\psi_c(x)$ are nearer the positive ions. Therefore, the energy of the electron in $\psi_c(x)$ is lower than that of the electron in $\psi_s(x)$, or $E_c < E_s$.

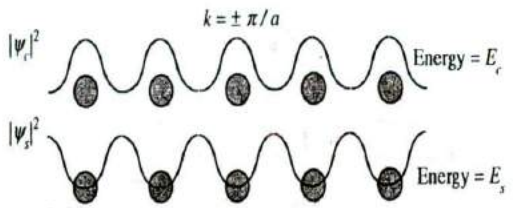


Figure 4.51 Forward and backward waves in the crystal with $k = \pm \pi/a$ give rise to two possible standing waves ψ_c and ψ_s. Their probability density distributions $|\psi_c|^2$ and $|\psi_s|^2$ have maxima either at the ions or between the ions, respectively.

It is not difficult to evaluate the energies E_c and E_s. The kinetic energy of the electron is the same in both $\psi_c(x)$ and $\psi_s(x)$, because these wavefunctions have the same k value and KE is given by $(\hbar k)^2/2m_e$. However, there is an electrostatic PE arising from the interaction of the electron with the ion cores, and this PE is different for the two wavefunctions. Suppose that $V(x)$ is the electrostatic PE of the electron at position x. We then must find the average, using the probability density distribution. Given that $|\psi_c(x)|^2\,dx$ is the probability of finding the electron at x in dx, the potential energy V_c of the electron is simply $V(x)$ averaged over the entire linear length L of the crystal. Thus, the potential energy V_c for $\psi_c(x)$ is

$$V_c = \frac{1}{L}\int_0^L V(x)|\psi_c(x)|^2\,dx = -V_n \qquad [4.76]$$

where V_n is the numerical result of the integration, which depends on $k = n\pi/a$ or n, by virtue of Equation 4.74. The integration in Equation 4.76 is a negative number that depends on n. We do not need to evaluate the integral, as we only need its final numerical result.

Using $|\psi_s(x)|^2$, we can also find V_s, the PE associated with $\psi_s(x)$. The result is that V_s is a positive quantity given by $+V_n$, where V_n is again the numerical result of the integration in Equation 4.76, which depends on n. The energies of the wavefunctions ψ_c and ψ_s whenever $k = n\pi/a$ are

$$E_c = \frac{(\hbar k)^2}{2m_e} - V_n \qquad k = \frac{n\pi}{a} \qquad [4.77]$$

$$E_s = \frac{(\hbar k)^2}{2m_e} + V_n \qquad k = \frac{n\pi}{a} \qquad [4.78]$$

Clearly, whenever k has the critical values $n\pi/a$, there are only two possible values for the energies E_c and E_s as determined by Equations 4.77 and 4.78; no other energies are allowed in between. These two energies are separated by $2V_n$.

Away from the critical k values determined by $k = n\pi/a$, the electron simply propagates as a traveling wave; the wave does not get reflected. The energy is then given by the free-running wave solution to the Schrödinger equation, that is, Equation 4.71,

$$E_k = \frac{(\hbar k)^2}{2m_e} \qquad \text{Away from } k = \frac{n\pi}{a} \qquad [4.79]$$

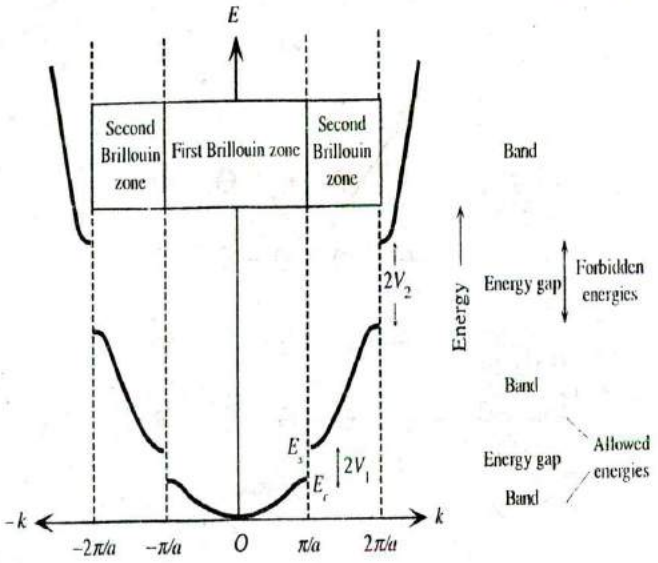


Figure 4.52 The energy of the electron as a function of its wavevector k inside a one-dimensional crystal.

There are discontinuities in the energy at $k = \pm n\pi/a$, where the waves suffer Bragg reflections in the crystal. For example, there can be no energy value for the electron between E_c and E_s. Therefore, $E_s - E_c$ is an energy gap at $k = \pm\pi/a$. Away from the critical k values, the E–k behavior is like that of a free electron, with E increasing with k as $E = (\hbar k)^2/2m_e$. In a solid, these energies fall within an energy band.

It seems that the energy of the electron increases parabolically with k along Equation 4.79 and then suddenly, at $k = n\pi/a$, it suffers a sharp discontinuity and increases parabolically again. Although the discontinuities at the critical points $k = n\pi/a$ are expected, by virtue of the Bragg reflection of waves, reflection effects will still be present to a certain extent, even within a small region around $k = n\pi/a$. The individual reflections shown in Figure 4.50 do not occur exactly at the origins of the atoms at $x = a, 2a, 3a, \ldots$. Rather, they occur over some distance, since the wave must interact with the electrons in the ion cores to be reflected. We therefore expect E–k behavior to deviate from Equation 4.79 in the neighborhood of the critical points, even if k is not exactly $n\pi/a$. Figure 4.52 shows the E–k behavior we expect, based on these arguments.

In Figure 4.52, we notice that there are certain energy ranges occurring at $k = \pm(n\pi/a)$ in which there are no allowed energies for the electron. As we saw previously, the electron cannot possess an energy between E_c and E_s at $k = \pi/a$. These energy ranges form **energy gaps** at the critical points $k = \pm(n\pi/a)$.

The range of k values from zero to the first energy gap at $k = \pm(\pi/a)$ defines a zone of k values called the **first Brillouin zone.** The zone between the first and second energy gap defines the **second Brillouin zone,** and so on. The Brillouin zone boundaries therefore identify where the energy discontinuities, or gaps, occur along the k axis.

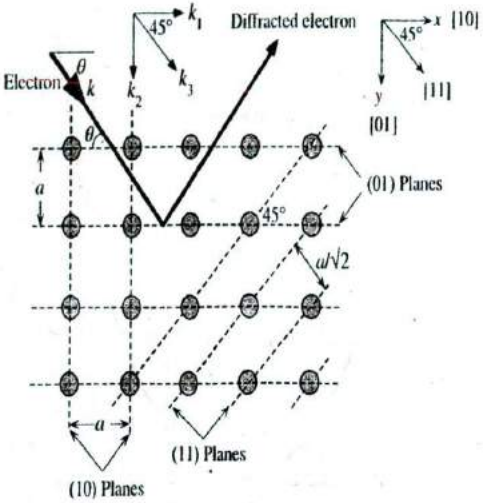


Figure 4.53 Diffraction of the electron in a two-dimensional crystal.

Diffraction occurs whenever k has a component satisfying $k_1 = \pm n\pi/a$, $k_2 = \pm n\pi/a$, or $k_3 = \pm n\pi\sqrt{2}/a$. In general terms, diffraction occurs when $k \sin\theta = n\pi/a$.

Electron motion in the three-dimensional crystal can be readily understood based on the concepts described here. For simplicity, we consider an electron propagating in a two-dimensional crystal, which is analogous, for example, to propagation in the xy plane of a crystal, as depicted in Figure 4.53. For certain k values and in certain directions, the electron will suffer diffraction and will be unable to propagate in the crystal.

Suppose that the electron's k vector along x is k_1. Whenever $k_1 = \pm n\pi/a$, the electron will be diffracted by the planes perpendicular to x, that is, the (10) planes.[21] Similarly, it will be diffracted by the (01) planes whenever its k vector along y is $k_2 = \pm n\pi/a$. The electron can also be diffracted by the (11) planes, whose separation is $a/\sqrt{2}$. If the component of k perpendicular to the (11) plane is k_3, then whenever $k_3 = \pm n\pi(\sqrt{2}/a)$, the electron will experience diffraction. These diffraction conditions can all be expressed through the **Bragg diffraction condition** $2d \sin\theta = n\lambda$, or

Bragg diffraction condition

$$k \sin\theta = \frac{n\pi}{d} \qquad [4.80]$$

where d is the interplanar separation and n is an integer; $d = a$ for (10) planes, and $d = a/\sqrt{2}$ for (11) planes.

When we plot the energy of the electron as a function of k, we must consider the direction of k, since the diffraction behavior in Equation 4.80 depends on $\sin\theta$. Along x, at $\theta = 0$, the energy gap occurs at $k = \pm(n\pi/a)$. Along $\theta = 45°$, it is at $k = \pm n\pi(\sqrt{2}/a)$, which is farther away. The E–k behavior for the electron in the two-dimensional lattice is shown in Figure 4.54 for the [10] and [11] directions. The figure shows that the first energy gap along x, in the [10] direction, is at $k = \pi/a$. Along the [11] direction, which is at 45° to the x axis, the first gap is at $k = \pi\sqrt{2}/a$.

[21] We use Miller indices in two dimensions by dropping the third digit but keeping the same interpretation. The direction along x is [10] and the plane perpendicular to x is (10).

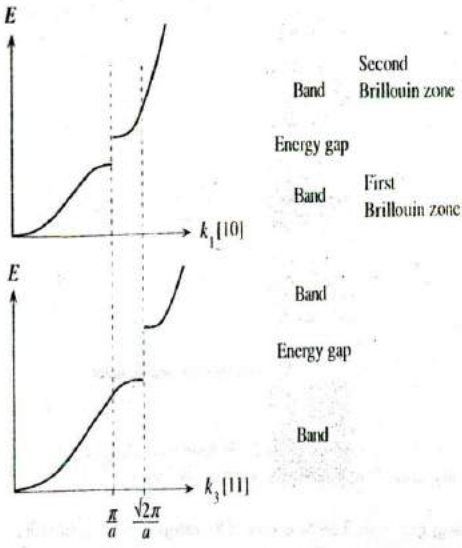


Figure 4.54 The E–k behavior for the electron along different directions in the two-dimensional crystal.
The energy gap along [10] is at π/a whereas it is at $\sqrt{2}\pi/a$ along [11].

When we consider the overlap of the energy bands along [10] and [11], in the case of a metal, there is no apparent energy gap. The electron can always find any energy simply by changing its direction.

The effects of overlap between energy bands and of energy gaps in different directions are illustrated in Figure 4.55. In the case of a semiconductor, the energy gap along [10] overlaps that along [11], so there is an overall energy gap. The electron in the semiconductor cannot have an energy that falls into this energy gap.

The first and second Brillouin zones for the two-dimensional lattice of Figure 4.53 are shown in Figure 4.56. The zone boundaries mark the occurrences of energy gaps in k space (space defined by k axes along the x and y directions). When we look at the E–k behavior, we must consider the crystal directions. This is most conveniently done by plotting energy contours in k space, as in Figure 4.57. Each contour connects all those values of k that possess the same energy. A point such as P on an energy contour gives the value of k for that energy along the direction OP. Initially, the energy contours are circles, as the energy follows $(\hbar k)^2/2m_e$ behavior, whatever the direction of k. However, near the critical values, that is, near the Brillouin zone boundaries, E increases more slowly than the parabolic relationship, as is apparent in Figure 4.52. Therefore, the circles begin to bulge as critical k values are approached. In Figure 4.57, the high-energy contours are concentrated in the corners of the zone, simply because the critical value is reached last along [11]. The energy contours do not continue smoothly across the zone boundary, because of the energy discontinuity in the E–k relationship at the boundary. Indeed, Figure 4.54 shows that the lowest energy in the second Brillouin zone may be lower than the highest energy in the first Brillouin zone.

There are two cases of interest. In the first, there is no apparent energy gap, as in Figure 4.57a, which corresponds to Figure 4.55a. The electron can have any energy

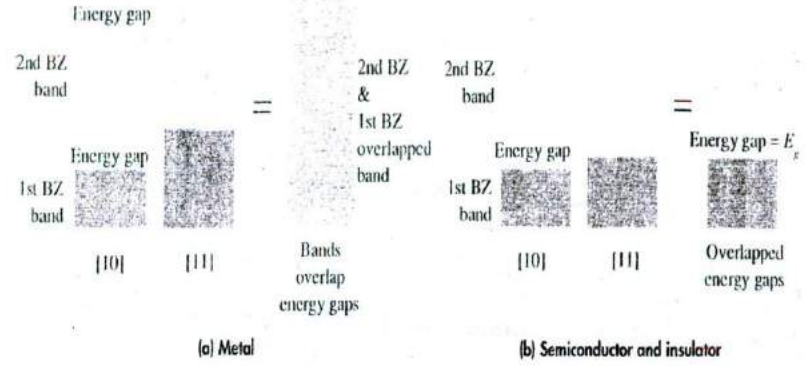


Figure 4.55

(a) For the electron in a metal, there is no apparent energy gap because the second BZ (Brillouin zone) along [10] overlaps the first BZ along [11]. Bands overlap the energy gaps. Thus, the electron can always find any energy by changing its direction.

(b) For the electron in a semiconductor, there is an energy gap arising from the overlap of the energy gaps along the [10] and [11] directions. The electron can never have an energy within this energy gap E_g.

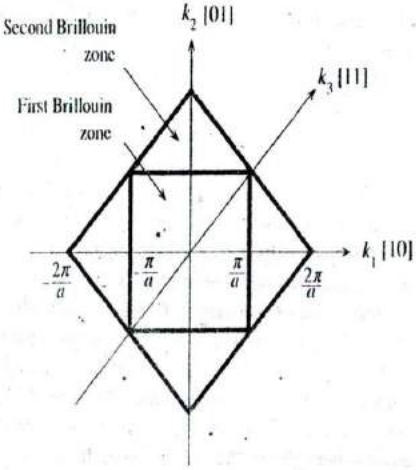


Figure 4.56 The Brillouin zones in two dimensions for the cubic lattice.

The Brillouin zones identify the boundaries where there are discontinuities in the energy (energy gaps).

value. In the second case, there is a range of energies that are not allowed, as shown in Figure 4.57b, which corresponds to Figure 4.55b.

In three dimensions, the E–k energy contour in Figure 4.57 becomes a surface in three-dimensional k space. To understand the use of such E–k contours or surfaces, consider that an E–k contour (or a surface) is made of many finely separated individual points, each representing a possible electron wavefunction ψ_k with a possible

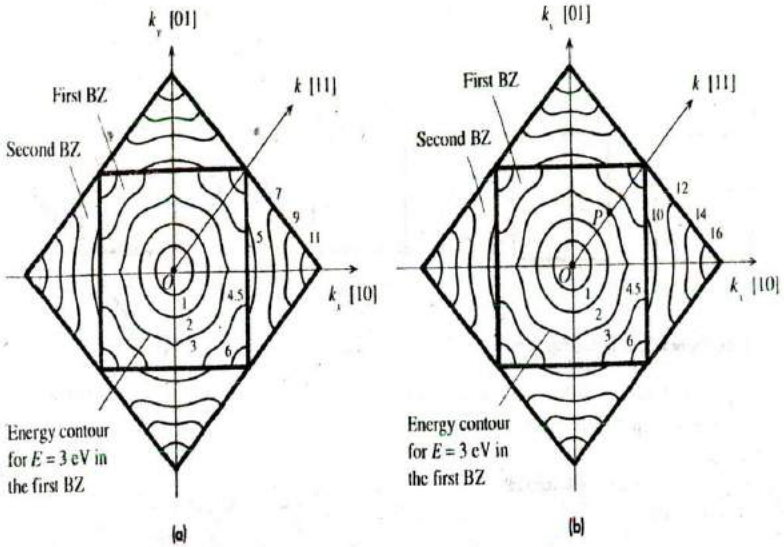


Figure 4.57 Energy contours in k space (space defined by k_x, k_y).

Each contour represents the same energy value. Any point P on the contour gives the values of k_x and k_y for that energy in that direction from O. For point P, $E = 3$ eV and OP along [11] is k.

(a) In a metal, the lowest energy in the second zone (5 eV) is lower than the highest energy (6 eV) in the first zone. There is an overlap of energies between the Brillouin zones.

(b) In a semiconductor or an insulator, there is an energy gap between the highest energy contour (6 eV) in the first zone and the lowest energy contour (10 eV) in the second zone.

energy E. At absolute zero, all the energies up to the Fermi energy are taken by the valence electrons. In k space, the energy surface, corresponding to the Fermi energy is termed the **Fermi surface.** The shape of this Fermi surface provides a means of interpreting the electrical and magnetic properties of solids.

For example, Na has one $3s$ electron per atom. In the solid, the $3s$ band is half full. The electrons take energies up to E_F, which corresponds to a spherical Fermi surface within the first Brillouin zone, as indicated in Figure 4.58a. We can then say that all the valence electrons (or nearly all) in this alkali solid exhibit an $E = (\hbar k)^2/2m_e$ type of behavior, as if they were free. When an external force is applied, such as an electric or magnetic field, we can treat the electron behavior as if it were free inside the metal with a constant mass. This is a desirable simplification for studying such metals. We can illustrate this desirability with an example. The Hall coefficient R_H derived in Chapter 2 was based on treating the electron as if it were a free particle inside the metal, or

$$R_H = -\frac{1}{en} \qquad [4.81]$$

For Na, the experimental value of R_H is -2.50×10^{-10} m^3 C^{-1}. Using the density (0.97 g cm^{-3}) and atomic mass (23) of Na and one valence electron per atom, we can

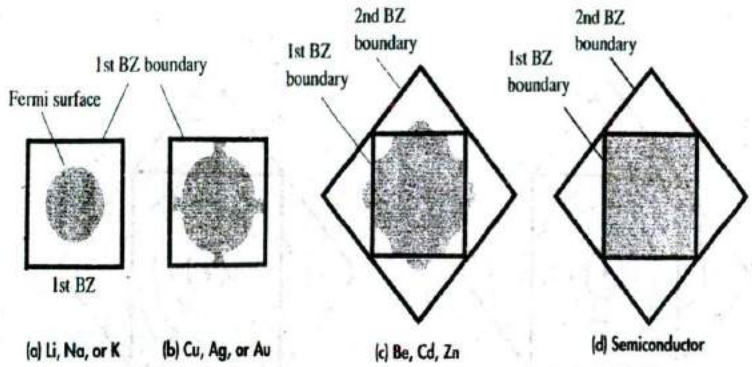


Figure 4.58 Schematic sketches of Fermi surfaces in two dimensions, representing various materials qualitatively.
(a) Monovalent group IA metals.
(b) Group IB metals.
(c) Be (Group IIA), Zn, and Cd (Group IIB).
(d) A semiconductor.

calculate $n = 2.54 \times 10^{28}$ m^{-3} and $R_H = -2.46 \times 10^{-10}$ m^3 C^{-1}, which is very close to the experimental value.

In the case of Cu, Ag, and Au (the IB metals in the Periodic Table), the Fermi surface is inside the first Brillouin zone, but it is not spherical as depicted in Figure 4.58b. Also, it touches the centers of the zone boundaries. Some of those electrons near the zone boundary behave quite differently than $E = (\hbar k)^2/2m_e$, although the majority of the electrons in the sphere do exhibit this type of behavior. To an extent, we can expect the free electron derivations to hold. The experimental value of R_H for Cu is -0.55×10^{-10} m^3 C^{-1}, whereas the expected value, based on Equation 4.81 with one electron per atom, is -0.73×10^{-10} m^3 C^{-1}, which is noticeably greater than the experimental value.

The divalent metals Be, Mg, and Ca have closed outer s subshells and should have a full s band in the solid. Recall that electrons in a full band cannot respond to an applied field and drift. We also know that there should be an overlap between the s and p bands, forming one partially filled continuous energy band, so these metals are indeed conductors. In terms of Brillouin zones, their structure is based on Figure 4.55a, which has the second zone overlapping the first Brillouin zone. The Fermi surface extends into the second zone and the corners of the first zone are empty, as depicted in Figure 4.58c. Since there are empty energy levels next to the Fermi surface, the electrons can gain energy and drift in response to an applied field. But the surface is not spherical; indeed, near the corners of the first zone, it even has the wrong curvature. Therefore, it is no longer possible to describe these electrons on the Fermi surface as obeying $E = (\hbar k)^2/2m_e$. When a magnetic field is applied to a drifting electron to bend its trajectory, its total behavior is different than that expected when it is acting as a free particle. The external force changes the momentum $\hbar k$ and the corresponding

change in the energy depends on the Fermi surface and can be quite complicated. To finish the example on the Hall coefficient, we note that based on two valence electrons per atom (Group IIA), the Hall coefficient for Be should be -0.25×10^{-10} m^3 C^{-1}, but the measured value is a positive coefficient of $+2.44 \times 10^{-10}$ m^3 C^{-1}. Equation 4.81 is therefore useless. It seems that the electrons moving at the Fermi surface of Be are equivalent to the motion of positive charges (like holes), so the Hall effect registers a positive coefficient.

The Fermi surface of a semiconductor is simply the boundary of the first Brillouin zone, because there is an energy gap between the first and the second Brillouin zones, as depicted in Figure 4.55b. In a semiconductor, all the energy levels up to the energy gap are taken up by the valence electrons. The first Brillouin zone forms the valence band and the second forms the conduction band.

4.12 GRÜNEISEN'S MODEL OF THERMAL EXPANSION

We considered thermal expansion in Section 1.4.2 where the principle is illustrated in Figure 1.18, which shows the potential energy curve $U(r)$ for two atoms separated by a distance r in a crystal. At temperature T_1 we know that the atoms will be vibrating about their equilibrium positions between positions B and C, compressing (B) and stretching (C) the bond between them. The line BC corresponds to the total energy E of the pair of atoms. The average separation at T_1 is at A, halfway between B and C. We also know that the PE curve $U(r)$ is *asymmetric*, and it is this asymmetry that leads to the phenomenon of thermal expansion. When the temperature increases from T_1 to T_2, the atoms vibrate between B' and C' and the average separation between the atoms also increases, from A to A', which we identified as *thermal expansion*. If the PE curve were symmetric, then there would be no thermal expansion.

Since the linear expansion coefficient λ is related to the shape of the PE curve, $U(r)$, it is also related to the elastic bulk modulus K that measures how difficult it is to stretch or compress the bonds. K depends on $U(r)$ in the same way that the elastic modulus Y depends on $U(r)$ as explained in Example 1.5.[22] Further, λ also depends on the amount of increase from BC to $B'C'$ per degree of increase in the temperature. λ must therefore also depend on the heat capacity. When the temperature increases by a small amount δT, the energy per atom increases by $(C_v \delta T)/N$ where C_v is the heat capacity per unit volume and N is the number of atoms per unit volume. If $C_v \delta T$ is large, then the line $B'C'$ in Figure 1.18 will be higher up on the energy curve and the average separation A' will therefore be larger. Thus, the larger is the heat capacity, the greater is the interatomic separation, which means $\lambda \propto C_v$. Further, the average separation, point A, depends on how much the bonds are stretched and compressed. For large

[22] K is a measure of the elastic change in the volume of a body in response to an applied pressure; large K means a small change in volume for a given pressure. Y is a measure of the elastic change in the length of the body in response to an applied stress, large Y means a small change in length. Both involve stretching or compressing bonds.

amounts of displacement from equilibrium, the average A will be greater as more asymmetry of the PE curve is used. Thus, the smaller is the elastic modulus K, the greater is λ; we see that $\lambda \propto C_v/K$.

Asymmetric potential energy curve

If we were to expand $U(r)$ about its minimum value U_{min} at $r = r_o$, we would obtain the Taylor expansion,

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

where a_2 and a_3 are coefficients related to the second and third derivatives of U at r_o. The term $(r - r_o)$ is missing because we are expanding a series about U_{min} where $dU/dr = 0$. The U_{min} and the $a_2(r - r_o)^2$ term give a parabola about U_{min} which is a symmetric curve around r_o and therefore does not lead to thermal expansion. It is the a_3 term that gives the expansion because it leads to asymmetry. Thus the amount of expansion λ also depends on the amount of asymmetry with respect to symmetry, that is a_3/a_2. Thus,

Linear expansion coefficient

$$\lambda \propto \frac{a_3}{a_2}\frac{C_v}{K}$$

The ratio of a_3 and a_2 depends on the nature of the bond. A simplified analytical treatment (beyond the scope of this book) gives λ as

Grüneisen's law

$$\lambda \approx 3\gamma \frac{C_v}{K} \qquad [4.82]$$

where γ is a "constant" called the *Grüneisen parameter*. The Grüneisen constant γ is approximately $-(r_o a_3)/(2a_2)$ where r_o is the equilibrium atomic separation, and thus γ represents the asymmetry of the energy curve. The approximate equality simply emphasizes the number of assumptions that are typically made in deriving Equation 4.82. The Grüneisen parameter γ is of the order of unity for many materials; experimentally, $\gamma = 0.1 - 1$. We can also write the Grüneisen law in terms of the molar heat capacity C_m (heat capacity per mole) or the specific heat capacity c_s (heat capacity per unit mass). If ρ is the density, and M_{at} is the atomic mass of the constituent atoms of the crystal, then

Grüneisen's law

$$\lambda = 3\gamma \frac{\rho C_m}{M_{at} K} = 3\gamma \frac{\rho c_s}{K} \qquad [4.83]$$

We can calculate the Grüneisen parameter γ for materials that possess different types of interatomic bonding and thereby obtain typical values for γ. This would also expose the extent of unharmonicity in the bonding. Given the experimental values for λ, K, ρ and c_s, the Grüneisen parameters have been calculated from Equation 4.83 and are listed in Table 4.6. An interesting feature of the results is that the experimental γ values, within a factor of 2–3, are about the same, at least to an order of magnitude. Equation 4.83 also indicates that the λ versus T behavior should resemble the C_v versus T dependence, which is approximately the case if one compares Figure 1.20 with Figure 4.45. (K does not change much with temperature.) There is one notable difference. At very low temperatures λ can change sign and become negative for certain crystals, whereas C_v cannot.

Table 4.6 The Grüneisen parameter for some selected materials with different types of interatomic bonding

Material	ρ (g cm^{-3})	λ ($\times 10^{-6}$ K^{-1})	K (GPa)	c_v (J kg^{-1} K^{-1})	γ
Iron (metallic, BCC)	7.9	12.1	170	444	0.20
Copper (metallic, FCC)	8.96	17	140	380	0.23
Germanium (covalent)	5.32	6	77	322	0.09
Glass (covalent-ionic)	2.45	8	70	800	0.10
NaCl (ionic)	2.16	39.5	28	880	0.19
Tellurium (mixed)	6.24	18.2	40	202	0.19
Polystyrene (van der Waals)	1.05	100	3	1200	0.08

CD Selected Topics and Solved Problems

Selected Topics

Hall Effect
Thermal Conductivity
Thermoelectric Effects in Metals:
Thermocouples
Thermal Expansion (Grüneisen's Law)

Solved Problems

The Water Molecule

DEFINING TERMS

Average energy E_{av} of an electron in a metal is determined by the Fermi–Dirac statistics and the density of states. It increases with the Fermi energy and also with the temperature.

Boltzmann statistics describes the behavior of a collection of particles (*e.g.*, gas atoms) in terms of their energy distribution. It specifies the number of particles $N(E)$ with given energy, through $N(E) \propto \exp(-E/kT)$, where k is the Boltzmann constant. The description is nonquantum mechanical in that there is no restriction on the number of particles that can have the same state (the same wavefunction) with an energy E. Also, it applies when there are only a few particles compared to the number of possible states, so the likelihood of two particles having the same state becomes negligible. This is generally the case for thermally excited electrons in the conduction band of a semiconductor, where there are many more states than electrons. The kinetic energy distribution of gas molecules in a tank obeys the Boltzmann statistics.

Cathode is a negative electrode. It emits electrons or attracts positive charges, that is, cations.

Debye frequency is the maximum frequency of lattice vibrations that can exist in a particular crystal. It is the cut-off frequency for lattice vibrations.

Debye temperature is a characteristic temperature of a particular crystal above which nearly all the atoms are vibrating in accordance with the kinetic molecular theory, that is, each atom has an average energy (potential + kinetic) of $3kT$ due to atomic vibrations, and the heat capacity is determined by the Dulong–Petit rule.

Density of states $g(E)$ is the number of electron states [*e.g.*, wavefunctions, $\psi(n, \ell, m_\ell, m_s)$] per unit energy per unit volume. Thus, $g(E)\,dE$ is the number of states in the energy range E to $(E + dE)$ per unit volume.

Density of vibrational states is the number of lattice vibrational modes per unit angular frequency range.

Dispersion relation relates the angular frequency ω and the wavevector K of a wave. In a crystal lattice, the coupling of atomic oscillations leads to a particular relationship between ω and K which determines the allowed lattice waves and their group velocities. The dispersion relation is specific to the crystal structure, that is, it depends on the lattice, basis, and bonding.

Effective electron mass m_e^* represents the inertial resistance of an electron inside a crystal against an acceleration imposed by an external force, such as the applied electric field. If $F_{ext} = eE_x$ is the external applied force due to the applied field $\mathcal{E}_x$, then the effective mass m_e^* determines the acceleration a of the electron by $eE_x = m_e^* a$. This takes into account the effect of the internal fields on the motion of the electron. In vacuum where there are no internal fields, m_e^* is the mass in vacuum m_e.

Fermi–Dirac statistics determines the probability of an electron occupying a state at an energy level E. This takes into account that a collection of electrons must obey the Pauli exclusion principle. The Fermi–Dirac function quantifies this probability via $f(E) = 1/[1 + \exp[(E - E_F)/kT]]$, where E_F is the Fermi energy.

Fermi energy is the maximum energy of the electrons in a metal at 0 K.

Field emission is the tunneling of an electron from the surface of a metal into vacuum, due to the application of a strong electric field (typically $\mathcal{E} > 10^9$ V m^{-1}).

Group velocity is the velocity at which traveling waves carry energy. If ω is the angular frequency and K is the wavevector of a wave, then the group velocity $v_g = d\omega/dK$.

Harmonic oscillator is an oscillating system, for example, two masses joined by a spring, that can be described by *simple harmonic motion*. In quantum mechanics, the energy of a harmonic oscillator is quantized and can only increase or decrease by a discrete amount $\hbar\omega$. The minimum energy of a harmonic oscillator is not zero but $\frac{1}{2}\hbar\omega$ (see **zero-point energy**).

Lattice wave is a wave in a crystal due to coupled oscillations of the atoms. Lattice waves may be traveling or stationary waves.

Linear combination of atomic orbitals (LCAO) is a method for obtaining the electron wavefunction in the molecule from a linear combination of individual atomic wavefunctions. For example, when two H atoms A and B come together, the electron wavefunctions, based on LCAO, are

$$\psi_a = \psi_{1s}(A) + \psi_{1s}(B)$$

$$\psi_b = \psi_{1s}(A) - \psi_{1s}(B)$$

where $\psi_{1s}(A)$ and $\psi_{1s}(B)$ are atomic wavefunctions centered around the H atoms A and B, respectively. The ψ_a and ψ_b represent molecular orbital wavefunctions for the electron; they reflect the behavior of the electron, or its probability distribution, in the molecule.

Mode or **state of lattice vibration** is a distinct, independent way in which a crystal lattice can vibrate with its own particular frequency ω and wavevector K. There are only a finite number of vibrational modes in a crystal.

Molecular orbital wavefunction, or simply molecular orbital, is a wavefunction for an electron within a system of two or more nuclei (*e.g.*, molecule). A molecular orbital determines the probability distribution of the electron within the molecule, just as the atomic orbital determines the electron's probability distribution within the atom. A molecular orbital can take two electrons with opposite spins.

Orbital is a region of space in an atom or molecule where an electron with a given energy may be found. An orbit, which is a well-defined path for an electron, cannot be used to describe the whereabouts of the electron in an atom or molecule because the electron has a probability distribution. Orbitals are generally represented by a surface within which the total probability is high, for example, 90 percent.

Orbital wavefunction, or simply orbital, describes the spatial dependence of the electron. The orbital is $\psi(r, \theta, \phi)$, which depends on n, ℓ, and m_ℓ, and the spin dependence m_s is excluded.

Phonon is a quantum of lattice vibrational energy of magnitude $\hbar\omega$, where ω is the vibrational angular frequency. A phonon has a momentum $\hbar K$ where K is the wavevector of the lattice wave.

Seebeck effect is the development of a built-in potential difference across a material as a result of a temperature gradient. If dV is the built-in potential across a

temperature difference dT, then the Seebeck coefficient S is defined as $S = dV/dT$. The coefficient gauges the magnitude of the Seebeck effect. Only the net Seebeck voltage difference between different metals can be measured. The principle of the thermocouple is based on the Seebeck effect.

State is a possible wavefunction for the electron that defines its spatial (orbital) and spin properties, for example, $\psi(n, \ell, m_\ell, m_s)$ is a state of the electron. From the Schrödinger equation, each state corresponds to a certain electron energy E. We thus speak of a state with energy E, state of energy E, or even an energy state. Generally there may be more than one state ψ with the same energy E.

Thermionic emission is the emission of electrons from the surface of a heated metal.

Work function is the minimum energy needed to free an electron from the metal at a temperature of absolute zero. It is the energy separation of the Fermi level from the vacuum level.

Zero-point energy is the minimum energy of a harmonic oscillator $\frac{1}{2}\hbar\omega$. Even at 0 K, an oscillator in quantum mechanics will have a finite amount of energy which is its zero-point energy. Heisenberg's uncertainty principle does not allow a harmonic oscillator to have zero energy because that would mean no uncertainty in the momentum and consequently an infinite uncertainty in space ($\Delta p, \Delta x > \hbar$).

QUESTIONS AND PROBLEMS

4.1 Phase of an atomic orbital

a. What is the functional form of a $1s$ wavefunction $\psi(r)$? Sketch schematically the atomic wavefunction $\psi_{1s}(r)$ as a function of distance from the nucleus.

b. What is the total wavefunction $\Psi_{1s}(r, t)$?

c. What is meant by two wavefunctions $\Psi_{1s}(A)$ and $\Psi_{1s}(B)$ that are out of phase?

d. Sketch schematically the two wavefunctions $\Psi_{1s}(A)$ and $\Psi_{1s}(B)$ at one instant.

4.2 Molecular orbitals and atomic orbitals Consider a linear chain of four identical atoms representing a hypothetical molecule. Suppose that each atomic wavefunction is a $1s$ wavefunction. This system of identical atoms has a center of symmetry C with respect to the center of the molecule (midway between the second and the third atom), and all molecular wavefunctions must be either symmetric or antisymmetric about C.

a. Using the LCAO principle, sketch the possible molecular orbitals.

b. Sketch the probability distributions $|\psi|^2$.

c. If more nodes in the wavefunction lead to greater energies, order the energies of the molecular orbitals.

Note: The electron wavefunctions, and the related probability distributions, in a simple potential energy well that are shown in Figure 3.15 can be used as a rough *guide* toward finding the appropriate molecular wavefunctions in the four-atom symmetric molecule. For example, if we were to smooth the electron potential energy in the four-atom molecule into a constant potential energy, that is, generate a potential energy well, we should be able to modify or distort, without flipping, the molecular orbitals to somewhat resemble ψ_1 to ψ_4 sketched in Figure 3.15. Consider also that the number of nodes increases from none for ψ_1 to three for ψ_4 in Figure 3.15.

4.3 Diamond and tin Germanium, silicon, and diamond have the same crystal structure, that of diamond. Bonding in each case involves sp^3 hybridization. The bonding energy decreases as we go from C to Si to Ge, as noted in Table 4.7.

a. What would you expect for the bandgap of diamond? How does it compare with the experimental value of 5.5 eV?

b. Tin has a tetragonal crystal structure, which makes it different than its group members, diamond, silicon, and germanium.

1. Is it a metal or a semiconductor?
2. What experiments do you think would expose its semiconductor properties?

Table 4.7

Property	Diamond	Silicon	Germanium	Tin
Melting temperature, °C	3800	1417	937	232
Covalent radius, nm	0.077	0.117	0.122	0.146
Bond energy, eV	3.60	1.84	1.7	1.2
First ionization energy, eV	11.26	8.15	7.88	7.33
Bandgap, eV	?	1.12	0.67	?

4.4 **Compound III–V Semiconductors** Indium as an element is a metal. It has a valency of III. Sb as an element is a metal and has a valency of V. InSb is a semiconductor, with each atom bonding to four neighbors, just like in silicon. Explain how this is possible and why InSb is a semiconductor and not a metal alloy. (Consider the electronic structure and sp^3 hybridization for each atom.)

4.5 **Compound II–VI semiconductors** CdTe is a semiconductor, with each atom bonding to four neighbors, just like in silicon. In terms of covalent bonding and the positions of Cd and Te in the Periodic Table, explain how this is possible. Would you expect the bonding in CdTe to have more ionic character than that in III–V semiconductors?

***4.6** **Density of states for a two-dimensional electron gas** Consider a two-dimensional electron gas in which the electrons are restricted to move freely within a square area a^2 in the x y plane. Following the procedure in Section 4.5, show that the density of states $g(E)$ is constant (independent of energy).

4.7 **Fermi energy of Cu** The Fermi energy of electrons in copper at room temperature is 7.0 eV. The electron drift mobility in copper, from Hall effect measurements, is 33 $cm^2\ V^{-1}\ s^{-1}$.

a. What is the speed v_F of conduction electrons with energies around E_F in copper? By how many times is this larger than the average thermal speed $v_{thermal}$ of electrons, if they behaved like an ideal gas (Maxwell–Boltzmann statistics)? Why is v_F much larger than $v_{thermal}$?

b. What is the De Broglie wavelength of these electrons? Will the electrons get diffracted by the lattice planes in copper, given that interplanar separation in Cu = 2.09 Å? (Solution guide: Diffraction of waves occurs when $2d \sin\theta = \lambda$, which is the Bragg condition. Find the relationship between λ and d that results in $\sin\theta > 1$ and hence no diffraction.)

c. Calculate the mean free path of electrons at E_F and comment.

4.8 **Free electron model, Fermi energy, and density of states** Na and Au both are valency I metals; that is, each atom donates one electron to the sea of conduction electrons. Calculate the Fermi energy (in eV) of each at 300 K and 0 K. Calculate the mean speed of all the conduction electrons and also the speed of electrons at E_F for each metal. Calculate the density of states as states per eV cm^{-3} at the Fermi energy and also at the center of the band, to be taken at $(E_F + \Phi)/2$. (See Table 4.1 for Φ.)

4.9 **Fermi energy and electron concentration** Consider the metals in Table 4.8 from Groups I, II, and III in the Periodic Table. Calculate the Fermi energies at absolute zero, and compare the values with the experimental values. What is your conclusion?

Table 4.8

Metal	Group	M_{at}	Density (g cm^{-3})	E_F (eV) [Calculated]	E_F (eV) [Experiment]
Cu	I	63.55	8.96	—	6.5
Zn	II	65.38	7.14	—	11.0
Al	III	27	2.70	—	11.8

4.10 Temperature dependence of the Fermi energy

a. Given that the Fermi energy for Cu is 7.0 eV at absolute zero, calculate the E_F at 300 K. What is the percentage change in E_F and what is your conclusion?

b. Given the Fermi energy for Cu at absolute zero, calculate the average energy and mean speed per conduction electron at absolute zero and 300 K, and comment.

4.11 X-ray emission spectrum from sodium Structure of the Na atom is $[Ne]3s^1$. Figure 4.59a shows the formation of the $3s$ and $3p$ energy bands in Na as a function of internuclear separation. Figure 4.59b shows the X-ray emission spectrum (called the L-band) from crystalline sodium in the soft X-ray range as explained in Example 4.6.

a. From Figure 4.59a, estimate the nearest neighbor equilibrium separation between Na atoms in the crystal if some electrons in the $3s$ band spill over into the states in the $3p$ band.

b. Explain the origin of the X-ray emission band in Figure 4.59b and the reason for calling it the L-band.

c. What is the Fermi energy of the electrons in Na from Figure 4.59b?

d. Taking the valency of Na to be 1, what is the expected Fermi energy and how does it compare with that in part (c)?

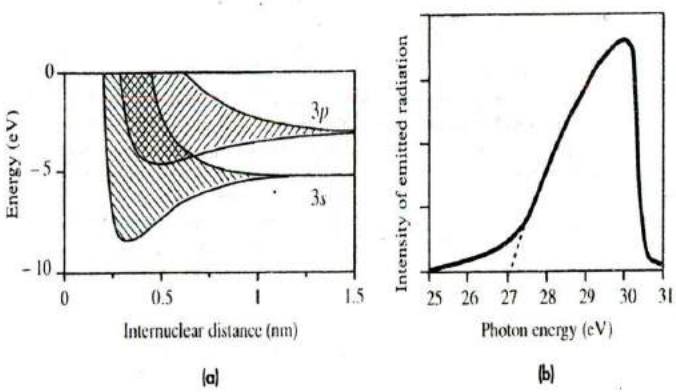


Figure 4.59

(a) Energy band formation in sodium.

(b) L-emission band of X-rays from sodium.

SOURCE: (b) Data extracted from W. M. Cadt and D. H. Tomboulian, *Phys. Rev.*, **59**, 1941, p. 381.

4.12 Conductivity of metals in the free electron model Consider the general expression for the conductivity of metals in terms of the density of states $g(E_F)$ at E_F given by

$$\sigma = \tfrac{1}{3}e^2 v_F^2 \tau g(E_F)$$

Show that within the free electron theory, this reduces to $\sigma = e^2 n\tau/m_e$, the Drude expression.

4.13 Mean free path of conduction electrons in a metal Show that within the free electron theory, the mean free path ℓ and conductivity σ are related by

Mean free path and conductivity in the free electron model

$$\sigma = \frac{e^2}{3^{1/3}\pi^{2/3}h}\ell n^{2/3} = 7.87 \times 10^{-5}\ell n^{2/3}$$

Calculate ℓ for Cu and Au, given each metal's resistivity of 17 nΩ m and 22 nΩ m, respectively, and that each has a valency of 1. We are used to seeing $\sigma \propto n$. Can you explain why $\sigma \propto n^{2/3}$?

*4.14 **Low-temperature heat capacity of metals** The heat capacity of conduction electrons in a metal is proportional to the temperature. The overall heat capacity of a metal is determined by the lattice heat capacity, except at the lowest temperatures. If δE_t is the increase in the total energy of the conduction electrons (per unit volume) and δT is the increase in the temperature of the metal as a result of heat addition, E_t has been calculated as follows:

$$E_t = \int_0^\infty E g(E) f(E)\, dE = E_t(0) + \left(\frac{\pi^2}{4}\right)\frac{n(kT)^2}{E_{FO}}$$

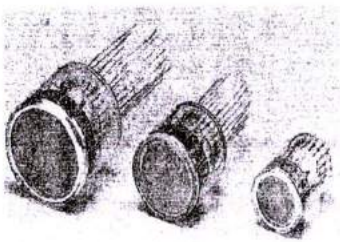
where $E_t(0)$ is the total energy per unit volume at 0 K, n is the concentration of conduction electrons, and E_{FO} is the Fermi energy at 0 K. Show that the heat capacity per unit volume due to conduction electrons in the free electron model of metals is

Heat capacity of conduction electrons

$$C_e = \frac{\pi^2}{2}\left(\frac{nk^2}{E_{FO}}\right)T = \gamma T \qquad [4.84]$$

where $\gamma = (\pi^2/2)(nk^2/E_{FO})$. Calculate C_e for Cu, and then using the Debye equation for the lattice heat capacity, find C_v for Cu at 10 K. Compare the two values and comment. What is the comparison at room temperature? (Note: $C_{volume} = C_{molar}(\rho/M_{at})$, where ρ is the density in g cm^{-3}, C_{volume} is in J K^{-1} cm^{-3}, and M_{at} is the atomic mass in g mol^{-1}.)

*4.15 **Secondary emission and photomultiplier tubes** When an energetic (high velocity) projectile electron collides with a material with a low work function, it can cause electron emission from the surface. This phenomenon is called **secondary emission.** It is fruitfully utilized in photomultiplier tubes as illustrated in Figure 4.60. The tube is evacuated and has a photocathode for receiving photons as a signal. An incoming photon causes photoemission of an electron from the photocathode material. The electron is then accelerated by a positive voltage applied to an electrode called a dynode which has a work function that easily allows secondary emission. When the accelerated electron strikes dynode D_1, it can release several electrons. All these electrons, the original and the secondary electrons, are then accelerated by the more positive voltage applied to dynode D_2. On impact with D_2, further electrons are released by secondary emission. The secondary emission process continues at each dynode stage until the final electrode, called the anode, is reached whereupon all the electrons are collected which results in a signal. Typical applications for photomultiplier tubes are in X-ray and nuclear medical instruments



Photomultiplier tubes.
| SOURCE: Courtesy of Hamamatsu.

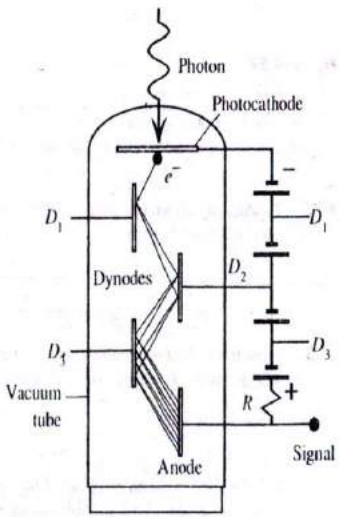


Figure 4.60 The photomultiplier tube.

(X-ray CT scanner, positron CT scanner, gamma camera, etc.), radiation measuring instruments (*e.g.*, radon counter), X-ray diffractometers, and radiation measurement in high-energy physics research.

A particular photomultiplier tube has the following properties. The photocathode is made of a semiconductor-type material with $E_x \approx 1$ eV, an electron affinity χ of 0.4 eV, and a quantum efficiency of 20 percent at 400 nm. *Quantum efficiency* is defined as the number of photoemitted electrons per absorbed photon. The diameter of the photocathode is 18 mm. There are 10 dynode electrodes and an applied voltage of 1250 V between the photocathode and anode. Assume that this voltage is equally distributed among all the electrodes.

a. What is the longest threshold wavelength for the phototube?

b. What is the maximum kinetic energy of the emitted electron if the photocathode is illuminated with a 400 nm radiation?

c. What is the emission current from the photocathode at 400 nm illumination per unit intensity of radiation?

d. What is the *KE* of the electron as it strikes the first dynode electrode?

e. It has been found that the tube has a gain of 10^6 electrons per incident photon. What is the average number of secondary electrons released at each dynode?

4.16 **Thermoelectric effects and E_F** Consider a thermocouple pair that consists of gold and aluminum. One junction is at 100 °C and the other is at 0 °C. A voltmeter (with a very large input resistance) is inserted into the aluminum wire. Use the properties of Au and Al in Table 4.3 to estimate the emf registered by the voltmeter and identify the positive end.

4.17 **The thermocouple equation** Although inputting the measured emf for V in the thermocouple equation $V = a\Delta T + b(\Delta T)^2$ leads to a quadratic equation, which in principle can be solved for ΔT, in general ΔT is related to the measured emf via

$$\Delta T = a_1 V + a_2 V^2 + a_3 V^3 + \cdots$$

with the coefficients a_1, a_2, etc., determined for each pair of TCs. By carrying out a Taylor's expansion of the TC equation, find the first two coefficients a_1 and a_2. Using an emf table for the K-type thermocouple or Figure 4.33, evaluate a_1 and a_2.

4.18 **Thermionic emission** A vacuum tube is required to have a cathode operating at 800 °C and providing an emission (saturation) current of 10 A. What should be the surface area of the cathode for the two materials in Table 4.9? What should be the operating temperature for the Th on W cathode, if it is to have the same surface area as the oxide-coated cathode?

Table 4.9

	B_e (A m^{-2} K^{-2})	Φ (eV)
Th on W	3×10^4	2.6
Oxide coating	100	1

4.19 **Field-assisted emission in MOS devices** Metal-oxide-semiconductor (MOS) transistors in microelectronics have a metal gate on an SiO_2 insulating layer on the surface of a doped Si crystal. Consider this as a parallel plate capacitor. Suppose the gate is an Al electrode of area 50 μm × 50 μm and has a voltage of 10 V with respect to the Si crystal. Consider two thicknesses for the SiO_2, (a) 100 Å and (b) 40 Å, where (1 Å = 10^{-10} m). The work function of Al is 4.2 eV, but this refers to electron emission into vacuum, whereas in this case, the electron is emitted into the oxide. The potential energy barrier Φ_B between Al and SiO_2 is about 3.1 eV, and the field-emission current density is given by Equation 4.46a and b. Calculate the field-emission current for the two cases. For simplicity, take m_e to be the electron mass in free space. What is your conclusion?

4.20 **CNTs and field emission** The electric field at the tip of a sharp emitter is much greater than the "applied field," $\mathcal{E}_a$. The applied field is simply defined as V_G/d where d is the distance from the cathode tip to the gate or the grid; it represents the average nearly uniform field that would exist if the tip were replaced by a flat surface so that the cathode and the gate would almost constitute a parallel plate capacitor. The tip experiences an effective field $\mathcal{E}$ that is much greater than $\mathcal{E}_a$, which is expressed by a **field enhancement** factor β that depends on the geometry of the cathode–gate emitter, and the shape of the emitter: $\mathcal{E} = \beta\mathcal{E}_a$. Further, we can take $\Phi_{\text{eff}}^{1/2}\Phi \approx \Phi^{3/2}$ in Equation 4.46. The final expression for the field-emission current density then becomes

Fowler–Nordheim field emission current

$$J = \frac{1.5 \times 10^6}{\Phi}\beta^2\mathcal{E}_a^2 \exp\left(\frac{10.4}{\Phi^{1/2}}\right)\exp\left(-\frac{6.44 \times 10^7\Phi^{3/2}}{\beta\mathcal{E}_a}\right) \qquad [4.85]$$

where Φ is in eV. For a particular CNT emitter, $\Phi = 4.9$ eV. Estimate the applied field required to achieve a field-emission current density of 100 mA cm^{-2} in the absence of field enhancement ($\beta = 1$) and with a field enhancement of $\beta = 800$ (typical value for a CNT emitter).

4.21 **Nordheim–Fowler field emission in an FED** Table 4.10 shows the results of I–V measurements on a Motorola FED microemitter. By a suitable plot show that the I–V follows the Nordheim–Fowler emission characteristics.

Table 4.10 Tests on a Motorola FED micro field emitter

V_G (V)	40.0	42	44	46	48	50	52	53.8	56.2	58.2	60.4
I_{emission} (μA)	0.40	2.14	9.40	20.4	34.1	61	93.8	142.5	202	279	367

4.22 **Lattice waves and heat capacity**

a. Consider an aluminum sample. The nearest separation $2R$ (2 × atomic radius) between the Al–Al atoms in the crystal is 0.286 nm. Taking a to be $2R$, and given the sound velocity in Al as 5100 m s^{-1}, calculate the force constant β in Equation 4.66. Use the group velocity v_g from the actual dispersion relation, Equation 4.55, to calculate the "sound velocity" at wavelengths of $\lambda = 1$ mm, 1 μm, and 1 nm. What is your conclusion?

b. Aluminum has a Debye temperature of 394 K. Calculate its specific heat at 30 °C (Darwin, Australia) and at −30 °C (January, Resolute Nunavut, Canada).

c. Calculate the specific heat capacity of a germanium crystal at 25 °C and compare it with the experimental value in Table 2.5.

4.23 **Specific heat capacity of GaAs and InSb**

a. The Debye temperature T_D of GaAs is 344 K. Calculate its specific heat capacity at 300 K and at 30 °C.

b. For InSb, $T_D = 203$ K. Calculate the room temperature specific heat capacity of InSb and compare it with the value expected from the Dulong–Petit rule ($T > T_D$).

4.24 **Thermal conductivity**

a. Given that silicon has a Young's modulus of about 110 GPa and a density of 2.3 g cm^{-3}, calculate the mean free path of phonons in Si at room temperature.

b. Diamond has the same crystal structure as Si but has a very large thermal conductivity, about 1000 W m^{-1} K^{-1} at room temperature. Given that diamond has a specific heat capacity c_v of 0.50 J K^{-1} g^{-1}, Young's modulus Y of 830 GPa, and density ρ of 0.35 g cm^{-3}, calculate the mean free path of phonons in diamond.

c. GaAs has a thermal conductivity of 200 W m^{-1} K^{-1} at 100 K and 80 W m^{-1} K^{-1} at 200 K. Calculate its thermal conductivity at 25 °C and compare with the experimental value of 44 W m^{-1} K^{-1}. (Hint: Take $\kappa \propto T^{-n}$ in the temperature region of interest; see Figure 4.48.)

*4.25 **Overlapping bands** Consider Cu and Ni with their density of states as schematically sketched in Figure 4.61. Both have overlapping $3d$ and $4s$ bands, but the $3d$ band is very narrow compared to the $4s$ band. In the case of Cu the band is full, whereas in Ni, it is only partially filled.

a. In Cu, do the electrons in the $3d$ band contribute to electrical conduction? Explain.

b. In Ni, do electrons in both bands contribute to conduction? Explain.

c. Do electrons have the same effective mass in the two bands? Explain.

d. Can an electron in the $4s$ band with energy around E_F become scattered into the $3d$ band as a result of a scattering process? Consider both metals.

e. Scattering of electrons from the $4s$ band to the $3d$ band and vice versa can be viewed as an additional scattering process. How would you expect the resistivity of Ni to compare with that of Cu, even though Ni has two valence electrons and nearly the same density as Cu? In which case would you expect a stronger temperature dependence for the resistivity?

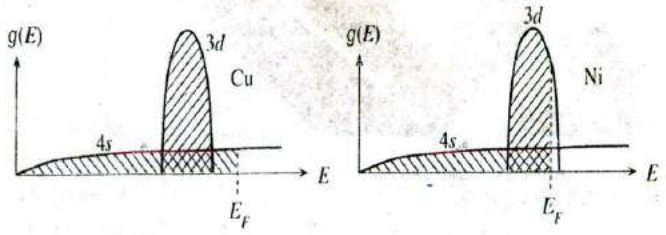
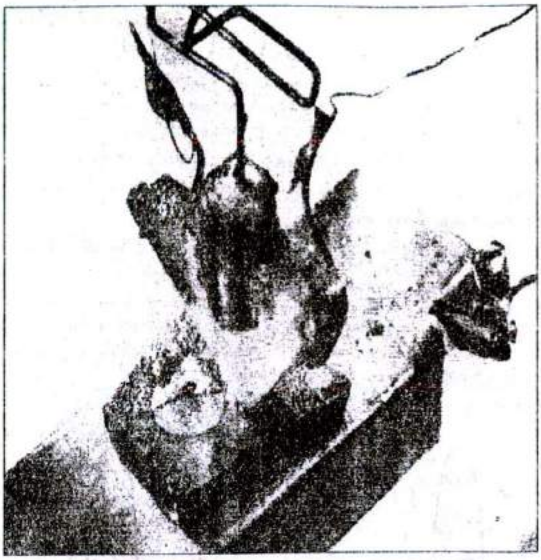


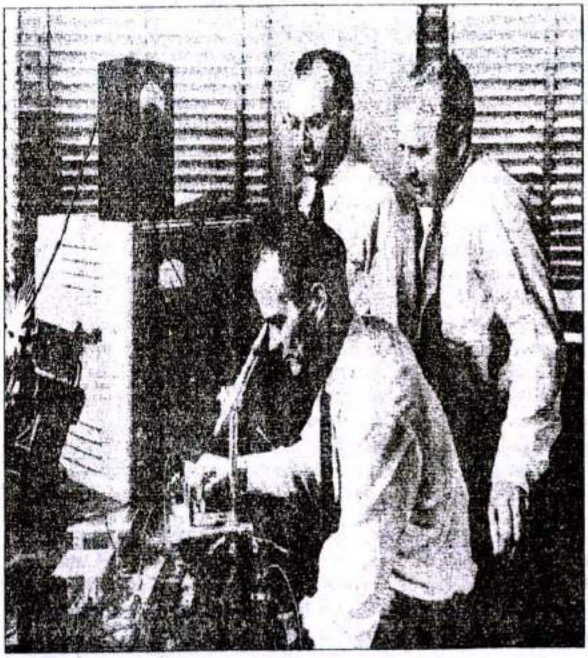
Figure 4.61 Density of states and electron filling in Cu and Ni.

*4.26 **Overlapping bands at E_F and higher resistivity** Figure 4.61 shows the density of states for Cu (or Ag) and Ni (or Pd). The d band in Cu is filled, and only electrons at E_F in the s band make a contribution to the conductivity. In Ni, on the other hand, there are electrons at E_F both in the s and d bands. The d band is narrow compared with the s band, and the electron's effective mass in this d band is large; for simplicity, we will assume m_e^* is "infinite" in this band. Consequently, the d-band electrons cannot be accelerated by the field (infinite m_e^*), have a negligible drift mobility, and make no contribution to the conductivity. Electrons in the s band can become scattered by phonons into the d band, and hence become relatively immobile until they are scattered back into the s band when they can drift again. Consider Ni and one particular conduction electron at E_F starting in the s band. Sketch schematically the magnitude of the velocity gained $|v_i - u_x|$ from the field E_x as a function of time for 10 scattering events; v_x and u_x are the instantaneous and initial velocities, and $|v_i - u_x|$ increases linearly with time, as the electron accelerates in the s band and then drops to zero upon scattering. If τ_{ss} is the mean time for s to s-band scattering, τ_{sd} is for s-band to d-band scattering, τ_{ds} is for d-band to s-band scattering, assume the following sequence of 10 events in your sketch: $\tau_{ss}, \tau_{ss}, \tau_{sd}, \tau_{ds}, \tau_{ss}, \tau_{sd}, \tau_{ds}, \tau_{ss}, \tau_{sd}, \tau_{ds}$. What would a similar sketch look like for Cu? Suppose that we wish to apply Equation 4.27. What does $g(E_F)$ and τ represent? What is the most important factor that makes Ni more resistive than Cu? Consider Matthiessen's rule. (Note: There are also electron spin related effects on the resistivity of Ni, but for simplicity these have been neglected.)

4.27 **Grüneisen's law** Al and Cu both have metallic bonding and the same crystal structure. Assuming that the Grüneisen's parameter γ for Al is the same as that for Cu, $\gamma = 0.23$, estimate the linear expansion coefficient λ of Al, given that its bulk modulus $K = 75$ GPa, $c_v = 900\ \text{J K}^{-1}\ \text{kg}^{-1}$, and $\rho = 2.7\ \text{g cm}^{-3}$. Compare your estimate with the experimental value of $23.5 \times 10^{-6}\ \text{K}^{-1}$.



First point-contact transistor invented at Bell Labs.
| SOURCE: Courtesy of Bell Labs



The three inventors of the transistor: William Shockley (seated), John Bardeen (left), and Walter Brattain (right) in 1948; the three inventors shared the Nobel prize in 1956.
| SOURCE: Courtesy of Bell Labs.

CHAPTER

5

Semiconductors

In this chapter we develop a basic understanding of the properties of intrinsic and extrinsic semiconductors. Although most of our discussions and examples will be based on Si, the ideas are applicable to Ge and to the compound semiconductors such as GaAs, InP, and others. By intrinsic Si we mean an ideal perfect crystal of Si that has no impurities or crystal defects such as dislocations and grain boundaries. The crystal thus consists of Si atoms perfectly bonded to each other in the diamond structure. At temperatures above absolute zero, we know that the Si atoms in the crystal lattice will be vibrating with a distribution of energies. Even though the average energy of the vibrations is at most $3kT$ and incapable of breaking the Si–Si bond, a few of the lattice vibrations in certain crystal regions may nonetheless be sufficiently energetic to "rupture" a Si–Si bond. When a Si–Si bond is broken, a "free" electron is created that can wander around the crystal and also contribute to electrical conduction in the presence of an applied field. The broken bond has a missing electron that causes this region to be positively charged. The vacancy left behind by the missing electron in the bonding orbital is called a **hole.** An electron in a neighboring bond can readily tunnel into this broken bond and fill it, thereby effectively causing the hole to be displaced to the original position of the tunneling electron. By electron tunneling from a neighboring bond, holes are therefore also free to wander around the crystal and also contribute to electrical conduction in the presence of an applied field. In an intrinsic semiconductor, the number of thermally generated electrons is equal to the number of holes (broken bonds). In an extrinsic semiconductor, impurities are added to the semiconductor that can contribute either excess electrons or excess holes. For example, when an impurity such as arsenic is added to Si, each As atom acts as a donor and contributes a free electron to the crystal. Since these electrons do not come from broken bonds, the numbers of electrons and holes are not equal in an extrinsic semiconductor, and the As-doped Si in this example will have excess electrons. It will be an n-type Si since electrical conduction will be mainly due to the motion of electrons. It is also possible to obtain a p-type Si crystal in which hole concentration is in excess of the electron concentration due to, for example, boron doping.

5.1 INTRINSIC SEMICONDUCTORS

5.1.1 SILICON CRYSTAL AND ENERGY BAND DIAGRAM

The electronic configuration of an isolated Si atom is $[Ne]3s^2p^2$. However, in the vicinity of other atoms, the $3s$ and $3p$ energy levels are so close that the interactions result in the *four* orbitals $\psi(3s)$, $\psi(3p_x)$, $\psi(3p_y)$, and $\psi(3p_z)$ mixing together to form *four* new hybrid orbitals (called ψ_{hyb}) that are symmetrically directed as far away from each other as possible (toward the corners of a tetrahedron). In two dimensions, we can simply view the orbitals pictorially as in Figure 5.1a. The four hybrid orbitals, ψ_{hyb}, each have one electron so that they are half-occupied. Therefore, a ψ_{hyb} orbital of one Si atom can overlap a ψ_{hyb} orbital of a neighboring Si atom to form a covalent bond with two spin-paired electrons. In this manner one Si atom bonds with four other Si atoms by overlapping the half-occupied ψ_{hyb} orbitals, as illustrated in Figure 5.1b.

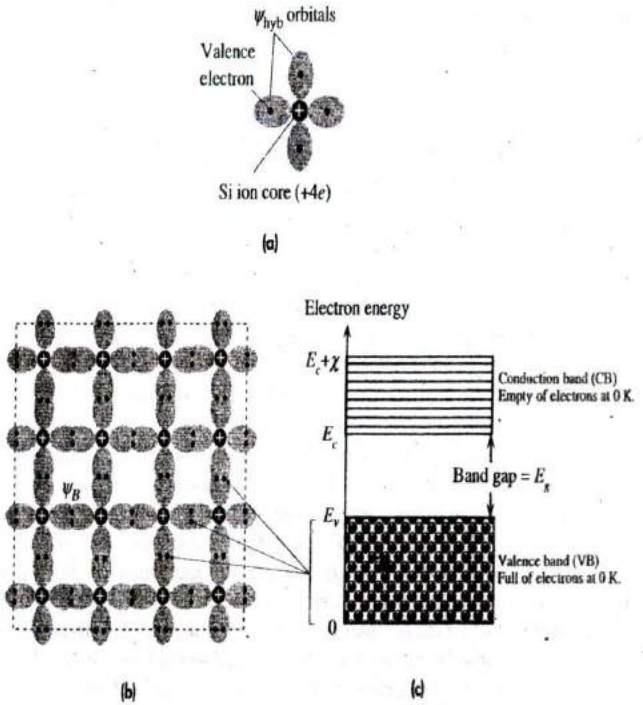


Figure 5.1

(a) A simplified two-dimensional illustration of a Si atom with four hybrid orbitals ψ_{hyb}. Each orbital has one electron.

(b) A simplified two-dimensional view of a region of the Si crystal showing covalent bonds.

(c) The energy band diagram at absolute zero of temperature.

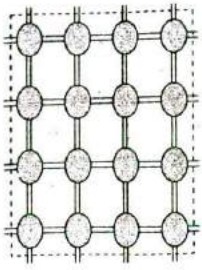


Figure 5.2 A two-dimensional pictorial view of the Si crystal showing covalent bonds as two lines where each line is a valence electron.

Each Si–Si bond corresponds to a bonding orbital, ψ_B, obtained by overlapping two neighboring ψ_{hyb} orbitals. Each bonding orbital (ψ_B) has two spin-paired electrons and is therefore *full*. Neighboring Si atoms can also form covalent bonds with other Si atoms, thus forming a three-dimensional network of Si atoms. The resulting structure is the Si crystal in which each Si atom bonds with four Si atoms in a tetrahedral arrangement. The crystal structure is that of a *diamond*, which was described in Chapter 1. We can imagine the Si crystal in two dimensions as depicted in Figure 5.1b. The electrons in the covalent bonds are the valence electrons.

The energy band diagram of the silicon crystal is shown in Figure 5.1c.[1] The vertical axis is the electron energy in the crystal. The valence band (VB) contains those electronic states that correspond to the overlap of bonding orbitals (ψ_B). Since all the bonding orbitals (ψ_B) are full with valence electrons in the crystal, the VB is also full with these valence electrons at a temperature of absolute zero. The conduction band (CB) contains electronic states that are at higher energies, those corresponding to the overlap of antibonding orbitals. The CB is separated from the VB by an energy gap E_g, called the **bandgap.** The energy level E_v marks the top of the VB and E_c marks the bottom of the CB. The energy distance from E_c to the vacuum level, the width of the CB, is called the **electron affinity** χ. The general energy band diagram in Figure 5.1c applies to all crystalline semiconductors with appropriate changes in the energies.

The electrons shown in the VB in Figure 5.1c are those in the covalent bonds between the Si atoms in Figure 5.1b. An electron in the VB, however, is not localized to an atomic site but extends throughout the whole solid. Although the electrons appear localized in Figure 5.1b, at the bonding orbitals between the Si atoms this is not, in fact, true. In the crystal, the electrons can tunnel from one bond to another and exchange places. If we were to work out the wavefunction of a valence electron in the Si crystal, we would find that it extends throughout the whole solid. This means that the electrons in the covalent bonds are indistinguishable. We cannot label an electron from the start and say that the electron is in the covalent bond between these two atoms.

We can crudely represent the silicon crystal in two dimensions as shown in Figure 5.2. Each covalent bond between Si atoms is represented by two lines corresponding to two spin-paired electrons. Each line represents a valence electron.

[1] The formation of energy bands in the silicon crystal was described in detail in Chapter 4.

5.1.2 ELECTRONS AND HOLES

The only empty electronic states in the silicon crystal are in the CB (Figure 5.1c). An electron placed in the CB is free to move around the crystal and also respond to an applied electric field because there are plenty of neighboring empty energy levels. An electron in the CB can easily gain energy from the field and move to higher energy levels because these states are empty. Generally we can treat an electron in the CB as if it were free within the crystal with certain modifications to its mass, as explained later in Section 5.1.3.

Since the only empty states are in the CB, the excitation of an electron from the VB requires a minimum energy of E_g. Figure 5.3a shows what happens when a photon of energy $h\nu > E_g$ is incident on an electron in the VB. This electron absorbs the incident photon and gains sufficient energy to surmount the energy gap E_g and reach the CB. Consequently, a free electron and a "hole," corresponding to a missing electron in the VB, are created. In some semiconductors such as Si and Ge, the photon absorption process also involves lattice vibrations (vibrations of the Si atoms), which we have not shown in Figure 5.3b.

Although in this specific example a photon of energy $h\nu > E_g$ creates an electron-hole pair, this is not necessary. In fact, in the absence of radiation, there is an electron-hole generation process going on in the sample as a result of **thermal generation.** Due to thermal energy, the atoms in the crystal are constantly vibrating, which corresponds to the bonds between the Si atoms being periodically deformed. In a certain region, the atoms, at some instant, may be moving in such a way that a bond becomes overstretched, as pictorially depicted in Figure 5.4. This will result in the overstretched bond rupturing and hence releasing an electron into the CB (the electron effectively

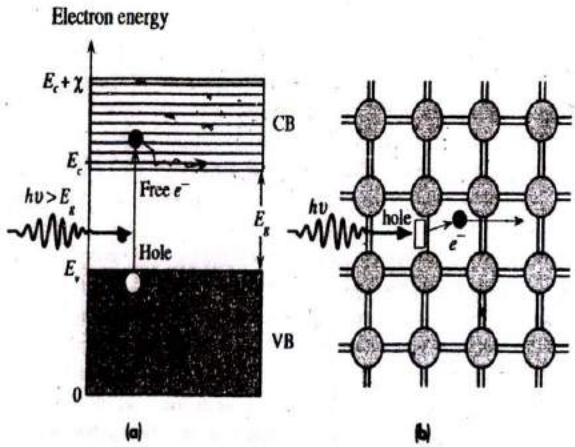


Figure 5.3

(a) A photon with an energy greater than E_g can excite an electron from the VB to the CB. (b) When a photon breaks a Si–Si bond, a free electron and a hole in the Si–Si bond are created.

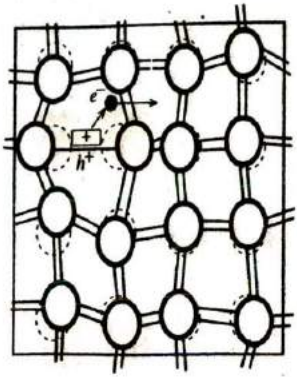


Figure 5.4 Thermal vibrations of atoms can break bonds and thereby create electron–hole pairs.

becomes "free"). The empty electronic state of the missing electron in the bond is what we call a **hole** in the valence band. The free electron, which is in the CB, can wander around the crystal and contribute to the electrical conduction when an electric field is applied. The region remaining around the hole in the VB is positively charged because a charge of $-e$ has been removed from an otherwise neutral region of the crystal. This hole, denoted as h^+, can also wander around the crystal as if it were free. This is because an electron in a neighboring bond can "jump," that is, tunnel, into the hole to fill the vacant electronic state at this site and thereby create a hole at its original position. This is effectively equivalent to the hole being displaced in the opposite direction, as illustrated in Figure 5.5a. This single step can reoccur, causing the hole to be further displaced. As a result, the hole moves around the crystal as if it were a free positively charged entity, as pictured in Figure 5.5a to d. Its motion is quite independent from that of the original electron. When an electric field is applied, the hole will drift in the direction of the field and hence contribute to electrical conduction. It is now apparent that there are essentially two types of charge carriers in semiconductors: *electrons* and *holes*. A hole is effectively an empty electronic state in the VB that behaves as if it were a positively charged "particle" free to respond to an applied electric field.

When a wandering electron in the CB meets a hole in the VB, the electron has found an empty state of lower energy and therefore occupies the hole. The electron falls from the CB to the VB to fill the hole, as depicted in Figure 5.5e and f. This is called **recombination** and results in the annihilation of an electron in the CB and a hole in the VB. The excess energy of the electron falling from CB to VB in certain semiconductors such as GaAs and InP is emitted as a photon. In Si and Ge the excess energy is lost as lattice vibrations (heat).

It must be emphasized that the illustrations in Figure 5.5 are pedagogical pictorial visualizations of hole motion based on classical notions and cannot be taken too seriously, as discussed in more advanced texts (see also Section 5.11). We should remember that the electron has a wavefunction in the crystal that is extended and not localized, as the pictures in Figure 5.5 imply. Further, the hole is a concept that corresponds to an empty valence band wavefunction that normally has an electron. Again, we cannot localize the hole to a particular site, as the pictures in Figure 5.5 imply.

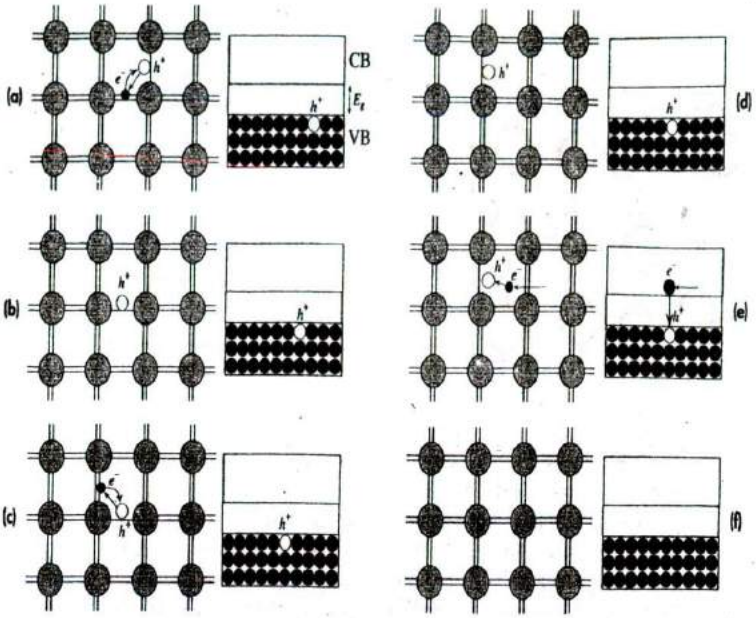


Figure 5.5 A pictorial illustration of a hole in the valence band wandering around the crystal due to the tunneling of electrons from neighboring bonds.

5.1.3 CONDUCTION IN SEMICONDUCTORS

When an electric field is applied across a semiconductor as shown in Figure 5.6, the energy bands bend. The total electron energy E is $KE + PE$, but now there is an additional electrostatic PE contribution that is not constant in an applied electric field. A uniform electric field $\mathcal{E}_x$ implies a linearly decreasing potential $V(x)$, by virtue of $(dV/dx) = -\mathcal{E}_x$, that is, $V = -Ax + B$. This means that the PE, $-eV(x)$, of the electron is now $eAx - eB$, which increases linearly across the sample. All the energy levels and hence the energy bands must therefore tilt up in the x direction, as shown in Figure 5.6, in the presence of an applied field.

Under the action of $\mathcal{E}_x$, the electron in the CB moves to the left and immediately starts gaining energy from the field. When the electron collides with a thermal vibration of a Si atom, it loses some of this energy and thus "falls" down in energy in the CB. After the collision, the electron starts to accelerate again, until the next collision, and so on. We recognize this process as the drift of the electron in an applied field, as illustrated in Figure 5.6. The drift velocity v_{de} of the electron is $\mu_e \mathcal{E}_x$ where μ_e is the drift mobility of the electron. In a similar fashion, the holes in the VB also drift in an applied field, but here the drift is along the field. Notice that when a hole gains energy, it moves "down" in the VB because the potential energy of the hole is of opposite sign to that of the electron.

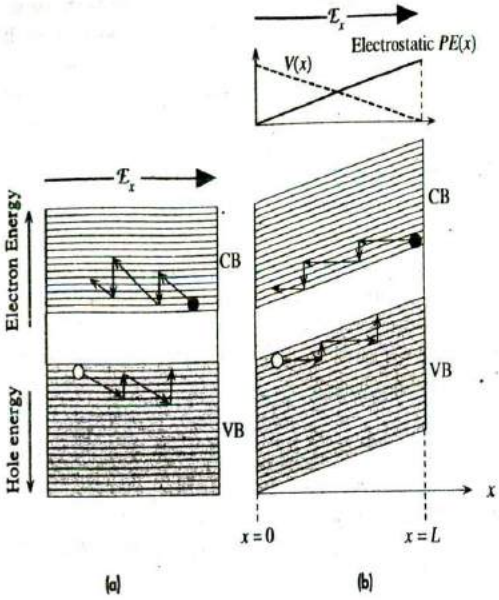


Figure 5.6 When an electric field is applied, electrons in the CB and holes in the VB can drift and contribute to the conductivity.
(a) A simplified illustration of drift in $\mathcal{E}_x$.
(b) Applied field bends the energy bands since the electrostatic PE of the electron is $-eV(x)$ and $V(x)$ decreases in the direction of $\mathcal{E}_x$, whereas PE increases.

Since both electrons and holes contribute to electrical conduction, we may write the current density J, from its definition, as

$$J = env_{de} + epv_{dh} \qquad [5.1]$$

where n is the electron concentration in the CB, p is the hole concentration in the VB, and v_{de} and v_{dh} are the drift velocities of electrons and holes in response to an applied electric field $\mathcal{E}_x$. Thus,

Electron and hole drift velocities

$$v_{de} = \mu_e \mathcal{E}_x \quad \text{and} \quad v_{dh} = \mu_h \mathcal{E}_x \qquad [5.2]$$

where μ_e and μ_h are the electron and hole drift mobilities. In Chapter 2 we derived the drift mobility μ_e of the electrons in a conductor as

$$\mu_e = \frac{e\tau_e}{m_e}. \qquad [5.3]$$

where τ_e is the mean free time between scattering events and m_e is the electronic mass. The ideas on electron motion in metals can also be applied to the electron motion in the CB of a semiconductor to rederive Equation 5.3. We must, however, use an effective mass m_e^* for the electron in the crystal rather than the mass m_e in free space. A "free" electron in a crystal is not entirely free because as it moves it interacts with the potential energy (PE) of the ions in the solid and therefore experiences various internal forces. The effective mass m_e^* accounts for these internal forces in such a way that we can relate the acceleration a of the electron in the CB to an external force F_{ext} (*e.g.*, $-e\mathcal{E}_x$) by $F_{ext} = m_e^* a$ just as we do for the electron in vacuum by $F_{ext} = m_e a$. In applying the

$F_{ext} = m_e^* a$ type of description to the motion of the electron, we are assuming, of course, that the effective mass of the electron can be calculated or measured experimentally. It is important to remark that the true behavior is governed by the solution of the Schrödinger equation in a periodic lattice (crystal) from which it can be shown that we can indeed describe the inertial resistance of the electron to acceleration in terms of an effective mass m_e^*. The effective mass depends on the interaction of the electron with its environment within the crystal.

We can now speculate on whether the hole can also have a mass. As long as we view mass as resistance to acceleration, that is, inertia, there is no reason why the hole should not have a mass. Accelerating the hole means accelerating electrons tunneling from bond to bond in the opposite direction. Therefore it is apparent that the hole will have a nonzero finite inertial mass because otherwise the smallest external force will impart an infinite acceleration to it. If we represent the effective mass of the hole in the VB by m_h^*, then the hole drift mobility will be

$$\mu_h = \frac{e\tau_h}{m_h^*} \qquad [5.4]$$

where τ_h is the mean free time between scattering events for holes.

Taking Equation 5.1 for the current density further, we can write the **conductivity of a semiconductor** as

Conductivity of a semiconductor

$$\sigma = en\mu_e + ep\mu_h \qquad [5.5]$$

where n and p are the electron and hole concentrations in the CB and VB, respectively. This is a general equation valid for all semiconductors.

5.1.4 Electron and Hole Concentrations

The general equation for the conductivity of a semiconductor, Equation 5.5, depends on n, the electron concentration, and p, the hole concentration. How do we determine these quantities? We follow the procedure schematically shown in Figure 5.7a to d in which the density of states is multiplied by the probability of a state being occupied and integrated over the entire CB for n and over the entire VB for p.

We define $g_{cb}(E)$ as the **density of states** in the CB, that is, the number of states per unit energy per unit volume. The probability of finding an electron in a state with energy E is given by the Fermi–Dirac function $f(E)$, which is discussed in Chapter 4. Then $g_{cb}(E)f(E)$ is the actual number of electrons per unit energy per unit volume $n_E(E)$ in the CB. Thus,

$$n_E\, dE = g_{cb}(E) f(E)\, dE$$

is the number of electrons in the energy range E to $E + dE$. Integrating this from the bottom (E_c) to the top ($E_c + \chi$) of the CB gives the electron concentration n, number of electrons per unit volume, in the CB. In other words,

$$n = \int_{E_c}^{E_c+\chi} n_E(E)\, dE = \int_{E_c}^{E_c+\chi} g_{cb}(E) f(E)\, dE$$

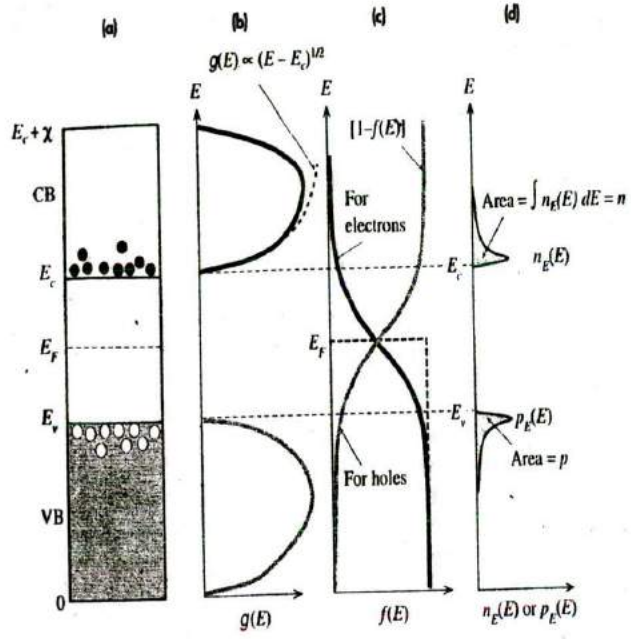


Figure 5.7

(a) Energy band diagram.

(b) Density of states (number of states per unit energy per unit volume).

(c) Fermi–Dirac probability function (probability of occupancy of a state).

(d) The product of $g(E)$ and $f(E)$ is the energy density of electrons in the CB (number of electrons per unit energy per unit volume). The area under $n_E(E)$ versus E is the electron concentration.

We will assume that $(E_c - E_F) \gg kT$ (*i.e.*, E_F is at least a few kT below E_c) so that

$$f(E) \approx \exp[-(E - E_F)/kT]$$

We are thus replacing Fermi–Dirac statistics by Boltzmann statistics and thereby inherently assuming that the number of electrons in the CB is far less than the number of states in this band.

Further, we will take the upper limit to be $E = \infty$ rather than $E_c + \chi$ since $f(E)$ decays rapidly with energy so that $g_{cb}(E)f(E) \to 0$ near the top of the band. Furthermore, since $g_{cb}(E)f(E)$ is significant only close to E_c, we can use

Density of states in conduction band

$$g_{cb}(E) = \frac{(\pi 8\sqrt{2})m_e^{*3/2}}{h^3}(E - E_c)^{1/2}$$

for an electron in a three-dimensional *PE* well without having to consider the exact form of $g_{cb}(E)$ across the whole band. Thus

$$n \approx \frac{(\pi 8\sqrt{2})m_e^{*3/2}}{h^3} \int_{E_c}^{\infty} (E - E_c)^{1/2} \exp\left[-\frac{(E - E_F)}{kT}\right] dE$$

which leads to

Electron concentration in CB

$$n = N_c \exp\left[-\frac{(E_c - E_F)}{kT}\right] \tag{5.6}$$

where

Effective density of states at CB edge

$$N_c = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{3/2} \tag{5.7}$$

The result of the integration in Equation 5.6 seems to be simple, but it is an approximation as it assumes that $(E_c - E_F) \gg kT$. N_c is a temperature-dependent constant, called the **effective density of states at the CB edge.** Equation 5.6 can be interpreted as follows. If we take all the states in the conduction band and replace them with an effective concentration N_c (number of states per unit volume) at E_c and then multiply this simply by the Boltzmann probability function, $f(E_c) = \exp[-(E_c - E_F)/kT]$, we obtain the concentration of electrons at E_c, that is, in the conduction band. N_c is thus an effective density of states at the CB band edge.

We can carry out a similar analysis for the concentration of holes in the VB. Multiplying the density of states $g_{vb}(E)$ in the VB with the probability of occupancy by a hole $[1 - f(E)]$, that is, the probability that an electron is absent, gives p_E, the hole concentration per unit energy. Integrating this over the VB gives the hole concentration

$$p = \int_0^{E_v} p_E \, dE = \int_0^{E_v} g_{vb}(E)[(1 - f(E)]\, dE$$

With the assumption that E_F is a few kT above E_v, the integration simplifies to

Hole concentration in VB

$$p = N_v \exp\left[-\frac{(E_F - E_v)}{kT}\right] \tag{5.8}$$

where N_v is the effective density of states at the VB edge and is given by

Effective density of states at VB edge

$$N_v = 2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{3/2} \tag{5.9}$$

We can now see the virtues of studying the density of states $g(E)$ as a function of energy E and the Fermi–Dirac function $f(E)$. Both were central factors in deriving the expressions for n and p. There are no specific assumptions in our derivations, except for E_F being a few kT away from the band edges, which means that Equations 5.6 and 5.8 are generally valid.

The general equations that determine the free electron and hole concentrations are thus given by Equations 5.6 and 5.8. It is interesting to consider the product np,

$$np = N_c \exp\left[-\frac{(E_c - E_F)}{kT}\right] N_v \exp\left[-\frac{(E_F - E_v)}{kT}\right] = N_c N_v \exp\left[-\frac{(E_c - E_v)}{kT}\right]$$

or

$$np = N_c N_v \exp\left(-\frac{E_g}{kT}\right) \tag{5.10}$$

where $E_g = E_c - E_v$ is the bandgap energy. First, we note that this is a general expression in which the right-hand side, $N_c N_v \exp(-E_g/kT)$, is a constant that depends on the temperature and the material properties, for example, E_g, and not on the position of the Fermi level. In the special case of an intrinsic semiconductor, $n = p$, which we can denote as n_i, the **intrinsic concentration,** so that $N_c N_v \exp(-E_g/kT)$ must be n_i^2. From Equation 5.10 we therefore have

$$np = n_i^2 = N_c N_v \exp\left(-\frac{E_g}{kT}\right) \qquad [5.11]$$

Mass action law

This is a general equation that is valid as long as we have thermal equilibrium. External excitation, such as photogeneration, is excluded. It states that the product np is a temperature-dependent constant. If we somehow increase the electron concentration, then we inevitably reduce the hole concentration. The constant n_i has a special significance because it represents the free electron and hole concentrations in the intrinsic material.

An **intrinsic semiconductor** is a pure semiconductor crystal in which the electron and hole concentrations are equal. By pure we mean virtually no impurities in the crystal. We should also exclude crystal defects that may capture carriers of one sign and thus result in unequal electron and hole concentrations. Clearly in a pure semiconductor, electrons and holes are generated in pairs by thermal excitation across the bandgap. It must be emphasized that Equation 5.11 is generally valid and therefore applies to both intrinsic and nonintrinsic ($n \neq p$) semiconductors.

When an electron and hole meet in the crystal, they "recombine." The electron falls in energy and occupies the empty electronic state that the hole represents. Consequently, the broken bond is "repaired," but we lose two free charge carriers. **Recombination** of an electron and hole results in their annihilation. In a semiconductor we therefore have thermal generation of electron–hole pairs by thermal excitation from the VB to the CB, and we also have recombination of electron–hole pairs that removes them from their conduction and valence bands, respectively. The rate of recombination R will be proportional to the number of electrons and also to the number of holes. Thus

$$R \propto np$$

The rate of generation G will depend on how many electrons are available for excitation at E_v, that is, N_v; how many empty states are available at E_c, that is, N_c; and the probability that the electron will make the transition, that is, $\exp(-E_g/kT)$, so that

$$G \propto N_c N_v \exp\left(-\frac{E_g}{kT}\right)$$

Since in thermal equilibrium we have no continuous increase in n or p, we must have the rate of generation equal to the rate of recombination, that is, $G = R$. This is equivalent to Equation 5.11.

In sketching the diagrams in Figure 5.7a to d to illustrate the derivation of the expressions for n and p (in Equations 5.6 and 5.8), we assumed that the Fermi level E_F is somewhere around the middle of the energy bandgap. This was not an assumption in the mathematical derivations but only in the sketches. From Equations 5.6 and 5.8 we

also note that the position of Fermi level is important in determining the electron and hole concentrations. It serves as a "mathematical crank" to determine n and p.

We first consider an intrinsic semiconductor, $n = p = n_i$. Setting $p = n_i$ in Equation 5.8, we can solve for the Fermi energy in the intrinsic semiconductor, E_{Fi}, that is,

$$N_v \exp\left[-\frac{(E_{Fi} - E_v)}{kT}\right] = (N_c N_v)^{1/2} \exp\left(-\frac{E_g}{2kT}\right)$$

which leads to

Fermi energy in intrinsic semiconductor

$$E_{Fi} = E_v + \frac{1}{2}E_g - \frac{1}{2}kT \ln\left(\frac{N_c}{N_v}\right) \qquad [5.12]$$

Furthermore, substituting the proper expressions for N_c and N_v we get

Fermi energy in intrinsic semiconductor

$$E_{Fi} = E_v + \frac{1}{2}E_g - \frac{3}{4}kT \ln\left(\frac{m_e^*}{m_h^*}\right) \qquad [5.13]$$

It is apparent from these equations that if $N_c = N_v$ or $m_e^* = m_h^*$, then

$$E_{Fi} = E_v + \frac{1}{2}E_g$$

that is, E_{Fi} is right in the middle of the energy gap. Normally, however, the effective masses will not be equal and the Fermi level will be slightly shifted down from midgap by an amount $\frac{3}{4}kT \ln(m_e^*/m_h^*)$, which is quite small compared with $\frac{1}{2}E_g$. For Si and Ge, the hole effective mass (for density of states) is slightly smaller than the electron effective mass, so E_{Fi} is slightly below the midgap.

The condition $np = n_i^2$ means that if we can somehow increase the electron concentration in the CB over the intrinsic value—for example, by adding impurities into the Si crystal that donate additional electrons to the CB—we will then have $n > p$. The semiconductor is then called ***n*-type.** The Fermi level must be closer to E_c than E_v, so that

$$E_c - E_F < E_F - E_v$$

and Equations 5.6 and 5.8 yield $n > p$. The np product always yields n_i^2 in thermal equilibrium in the absence of external excitation, for example, illumination.

It is also possible to have an excess of holes in the VB over electrons in the CB, for example, by adding impurities that remove electrons from the VB and thereby generate holes. In that case E_F is closer to E_v than to E_c. A semiconductor in which $p > n$ is called a ***p*-type semiconductor.** The general band diagrams with the appropriate Fermi levels for intrinsic, n-type, and p-type semiconductors (*e.g.*, *i*-Si, n-Si, and p-Si, respectively) are illustrated in Figure 5.8a to c.

It is apparent that if we know where E_F is, then we have effectively determined n and p by virtue of Equations 5.6 and 5.8. We can view E_F as a *material property* that is related to the concentration of charge carriers that contribute to electrical conduction. Its significance, however, goes beyond n and p. It also determines the energy needed to remove an electron from the semiconductor. The energy difference between the vacuum level (where the electron is free) and E_F is the **work function** Φ of the semiconductor, the energy required to remove an electron even though there are no electrons at E_F in a semiconductor.

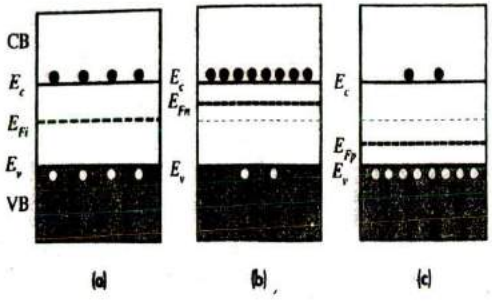


Figure 5.8 Energy band diagrams for (a) intrinsic, (b) n-type, and (c) p-type semiconductors. In all cases, $np = n_i^2$.

The Fermi level can also be interpreted in terms of the potential energy per electron for electrical work similar to the interpretation of electrostatic PE. Just as $e\,\Delta V$ is the electrical work involved in taking a charge e across a potential difference ΔV, any difference in E_F in going from one end of a material (or system) to another is available to do an amount ΔE_F of external work. A corollary to this is that if electrical work is done on the material, for example, by passing a current through it, then the Fermi level is not uniform in the material. ΔE_F then represents the work done per electron. For a material in thermal equilibrium and not subject to any external excitation such as illumination or connections to a voltage supply, the Fermi level in the material must therefore be uniform, $\Delta E_F = 0$.

What is the average energy of an electron in the conduction band of a semiconductor? Also, what is the mean speed of an electron in the conduction band? We note that the concentration of electrons with energies E to $E + dE$ is $n_E(E)\,dE$ or $g_{cb}(E)\,f(E)\,dE$. Thus the average energy of electrons in the CB, by definition of the mean, is

$$\bar{E}_{CB} = \frac{1}{n}\int_{CB} E g_{cb}(E) f(E)\,dE$$

where the integration must be over the CB. Substituting the proper expressions for $g_{cb}(E)$ and $f(E)$ in the integrand and carrying out the integration from E_c to the top of the band, we find the very simple result that

Average electron energy in CB

$$\bar{E}_{CB} = E_c + \frac{3}{2}kT \qquad [5.14]$$

Thus, an electron in the conduction band has an average energy of $\frac{3}{2}kT$ *above* E_c. Since we know that an electron at E_c is "free" in the crystal, $\frac{3}{2}kT$ must be its average kinetic energy.

This is just like the average kinetic energy of gas atoms (such as He atoms) in a tank assuming that the atoms (or the "spheres") do not interact, that is, they are independent. We know from the kinetic theory that the statistics of a collection of independent gas atoms obeys the classical Maxwell–Boltzmann description with an average energy given by $\frac{3}{2}kT$. We should also recall that the description of electron statistics in a metal involves the Fermi–Dirac function, which is based on the Pauli exclusion principle. In a metal the average energy of the conduction electron is $\frac{3}{5}E_F$ and, for all practical purposes, temperature independent. We see that the collective electron behavior is completely different in the two solids. We can explain the difference by noting that the conduction band in a

Table 5.1 Selected typical properties of Ge, Si, and GaAs at 300 K

	E_g (eV)	χ (eV)	N_c (cm^{-3})	N_v (cm^{-3})	n_i (cm^{-3})	μ_e ($cm^2\,V^{-1}\,s^{-1}$)	μ_h ($cm^2\,V^{-1}\,s^{-1}$)	m_e^*/m_e	m_h^*/m_e	ϵ_r
Ge	0.66	4.13	1.04×10^{19}	6.0×10^{18}	2.5×10^{13}	3900	1900	0.12*a* 0.56*b*	0.23*a* 0.40*b*	16
Si	1.10	4.01	2.8×10^{19}	1.2×10^{19}	1.0×10^{10}	1350	450	0.26*a* 1.08*b*	0.38*a* 0.60*b*	11.9
GaAs	1.42	4.07	4.7×10^{17}	7×10^{18}	2.1×10^{6}	8500	400	0.067*a,b*	0.40*a* 0.50*b*	13.1

NOTE: Effective mass related to conductivity (labeled *a*) is different than that for density of states (labeled *b*). In numerous textbooks, n_i is taken as $1.45 \times 10^{10}\ cm^{-3}$ and is therefore the most widely used value of n_i for Si, though the correct value is actually $1.0 \times 10^{10}\ cm^{-3}$. (M. A. Green, *J. Appl. Phys.*, **67**, 2944, 1990.)

semiconductor is only scarcely populated by electrons, which means that there are many more electronic states than electrons and thus the likelihood of two electrons trying to occupy the same electronic state is practically nil. We can then neglect the Pauli exclusion principle and use the Boltzmann statistics. This is not the case for metals where the number of conduction electrons and the number of states are comparable in magnitude.

Table 5.1 is a comparative table of some of the properties of the important semiconductors, Ge, Si, and GaAs.

EXAMPLE 5.1 **INTRINSIC CONCENTRATION AND CONDUCTIVITY OF Si** Given that the density of states related effective masses of electrons and holes in Si are approximately $1.08m_e$ and $0.60m_e$, respectively, and the electron and hole drift mobilities at room temperature are 1350 and $450\ cm^2\,V^{-1}\,s^{-1}$, respectively, calculate the intrinsic concentration and intrinsic resistivity of Si.

SOLUTION

We simply calculate the effective density of states N_c and N_v by

$$N_c = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{3/2} \qquad \text{and} \qquad N_v = 2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{3/2}$$

Thus

$$N_c = 2\left[\frac{2\pi(1.08 \times 9.1 \times 10^{-31}\ \text{kg})(1.38 \times 10^{-23}\ \text{J K}^{-1})\,(300\ \text{K})}{(6.63 \times 10^{-34}\ \text{J s})^2}\right]^{3/2}$$

$$= 2.81 \times 10^{25}\ \text{m}^{-3} \qquad \text{or} \qquad 2.81 \times 10^{19}\ \text{cm}^{-3}$$

and

$$N_v = 2\left[\frac{2\pi(0.60 \times 9.1 \times 10^{-31}\ \text{kg})(1.38 \times 10^{-23}\ \text{J K}^{-1})(300\ \text{K})}{(6.63 \times 10^{-34}\ \text{J s})^2}\right]^{3/2}$$

$$= 1.16 \times 10^{25}\ \text{m}^{-3} \qquad \text{or} \qquad 1.16 \times 10^{19}\ \text{cm}^{-3}$$

The intrinsic concentration is

$$n_i = (N_c N_v)^{1/2} \exp\left(-\frac{E_g}{2kT}\right)$$

so that

$$n_i = [(2.81 \times 10^{19}\ \text{cm}^{-3})(1.16 \times 10^{19}\ \text{cm}^{-3})]^{1/2} \exp\left[-\frac{(1.10\ \text{eV})}{2(300\ \text{K})(8.62 \times 10^{-5}\ \text{eV K}^{-1})}\right]$$
$$= 1.0 \times 10^{10}\text{cm}^{-3}$$

The conductivity is

$$\sigma = en\mu_e + ep\mu_h = en_i(\mu_e + \mu_h)$$

that is,

$$\sigma = (1.6 \times 10^{-19}\ \text{C})(1.0 \times 10^{10}\ \text{cm}^{-3})(1350 + 450\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1})$$
$$= 2.9 \times 10^{-6}\ \Omega^{-1}\ \text{cm}^{-1}$$

The resistivity is

$$\rho = \frac{1}{\sigma} = 3.5 \times 10^5\ \Omega\ \text{cm}$$

Although we calculated $n_i = 1.0 \times 10^{10}\ \text{cm}^{-3}$, the most widely used n_i value in the literature has been $1.45 \times 10^{10}\ \text{cm}^{-3}$. The difference arises from a number of factors but, most importantly, from what exact value of the effective hole mass should be used in calculating N_v. Henceforth we will simply use[2] $n_i = 1.0 \times 10^{10}\ \text{cm}^{-3}$, which seems to be the "true" value.

EXAMPLE 5.2

MEAN SPEED OF ELECTRONS IN THE CB Estimate the mean speed of electrons in the conduction band of Si at 300 K. If a is the magnitude of lattice vibrations, then the kinetic theory predicts $a^2 \propto T$; or stated differently, the mean energy associated with lattice vibrations (proportional to a^2) increases with kT. Given the temperature dependence of the mean speed of electrons in the CB, what should be the temperature dependence of the drift mobility? The effective mass of an electron in the conduction band is $0.26m_e$.

SOLUTION

The fact that the average KE, $\frac{1}{2}m_e^* v_e^2$, of an electron in the CB of a semiconductor is $\frac{3}{2}kT$ means that the effective mean speed v_e must be

$$v_e = \left(\frac{3kT}{m_e^*}\right)^{1/2} = \left[\frac{(3 \times 1.38 \times 10^{-23} \times 300)}{(0.26 \times 9.1 \times 10^{-31})}\right]^{1/2} = 2.3 \times 10^5\ \text{m s}^{-1}$$

The effective mean speed v_e is called the **thermal velocity** v_{th} of the electron.

The mean free time τ of the electron between scattering events due to thermal vibrations of the atoms is inversely proportional to both the mean speed v_e of the electron and the scattering cross section of the thermal vibrations, that is,

$$\tau \propto \frac{1}{v_e(\pi a^2)}$$

where a is the amplitude of the atomic thermal vibrations. But, $v_e \propto T^{1/2}$ and $(\pi a^2) \propto kT$, so that $\tau \propto T^{-3/2}$ and consequently $\mu_e \propto T^{-3/2}$.

Experimentally μ_e is not exactly proportional to $T^{-3/2}$ but to $T^{-2.4}$, a higher power index. The effective mass used in the density of states calculations is actually different than that used in transport calculations such as the mean speed, drift mobility, and so on.

---

[2] The correct value appears to be $1.0 \times 10^{10}\ \text{cm}^{-3}$ as discussed by M. A. Green [*J. Appl. Phys.*, **67**, 2944, 1990] and A. B. Sproul and M. A. Green [*J. Appl. Phys.*, **70**, 846, 1991].

5.2 EXTRINSIC SEMICONDUCTORS

By introducing small amounts of impurities into an otherwise pure Si crystal, it is possible to obtain a semiconductor in which the concentration of carriers of one polarity is much in excess of the other type. Such semiconductors are referred to as **extrinsic semiconductors** vis-à-vis the intrinsic case of a pure and perfect crystal. For example, by adding pentavalent impurities, such as arsenic, which have a valency of more than four, we can obtain a semiconductor in which the electron concentration is much larger than the hole concentration. In this case we will have an *n*-type semiconductor. If we add trivalent impurities, such as boron, which have a valency of less than four, then we find that we have an excess of holes over electrons. We now have a *p*-type semiconductor. How do impurities change the concentrations of holes and electrons in a semiconductor?

5.2.1 *n*-TYPE DOPING

Consider what happens when small amounts of a pentavalent (valency of 5) element from Group V in the Periodic Table, such as As, P, Sb, are introduced into a pure Si crystal. We only add small amounts (*e.g.*, one impurity atom for every million host atoms) because we wish to surround each impurity atom by millions of Si atoms, thereby forcing the impurity atoms to bond with Si atoms in the same diamond crystal structure. Arsenic has five valence electrons, whereas Si has four. Thus when an As atom bonds with four Si atoms, it has one electron left unbonded. It cannot find a bond to go into, so it is left orbiting around the As atom, as illustrated in Figure 5.9. The As^+ ionic center with an electron e^- orbiting it is just like a hydrogen atom in a silicon environment. We can easily calculate how much energy is required to free this electron away from the As site, thereby ionizing the As impurity. Had this been a hydrogen atom in free space, the energy required to remove the electron from its ground state (at $n = 1$) to far away from the positive center would have been given by $-E_n$ with $n = 1$. The binding energy of the electron in the H atom is thus

$$E_b = -E_1 = \frac{m_e e^4}{8\varepsilon_o^2 h^2} = 13.6 \text{ eV}$$

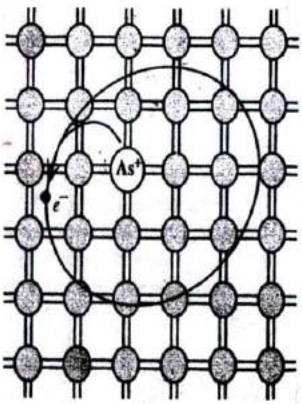


Figure 5.9 Arsenic-doped Si crystal.
The four valence electrons of As allow it to bond just like Si, but the fifth electron is left orbiting the As site. The energy required to release the free fifth electron into the CB is very small.

If we wish to apply this to the electron around an As^+ core in the Si crystal environment, we must use $\varepsilon_r\varepsilon_o$ instead of ε_o, where ε_r is the relative permittivity of silicon, and also the effective mass of the electron m_e^* in the silicon crystal. Thus, the binding energy of the electron to the As^+ site in the Si crystal is

Electron binding energy at a donor

$$E_b^{Si} = \frac{m_e^* e^4}{8\varepsilon_r^2\varepsilon_o^2 h^2} = (13.6\text{ eV})\left(\frac{m_e^*}{m_e}\right)\left(\frac{1}{\varepsilon_r^2}\right) \qquad [5.15]$$

With $\varepsilon_r = 11.9$ and $m_e^* \approx \frac{1}{3}m_e$ for silicon, we find $E_b^{Si} = 0.032$ eV, which is comparable with the average thermal energy of atomic vibrations at room temperature, $\sim 3kT$ (~ 0.07 eV). Thus, the fifth valence electron can be readily freed by thermal vibrations of the Si lattice. The electron will then be "free" in the semiconductor, or, in other words, it will be in the CB. The energy required to excite the electron to the CB is therefore 0.032 eV. The addition of As atoms introduces localized electronic states at the As sites because the fifth electron has a localized wavefunction, of the hydrogenic type, around As^+. The energy E_d of these states is 0.032 eV below E_c because this is how much energy is required to take the electron away into the CB. Thermal excitation by the lattice vibrations at room temperature is sufficient to ionize the As atom, that is, excite the electron from E_d into the CB. This process creates free electrons but immobile As^+ ions, as shown in the energy band diagram of an *n*-type semiconductor in Figure 5.10. Because the As atom donates an electron into the CB, it is called a **donor atom.** E_d is the electron energy around the donor atom. E_d is close to E_c, so the spare fifth electron from the dopant can be readily donated to the CB. If N_d is the donor atom concentration in the crystal, then provided that $N_d \gg n_i$, at room temperature the electron concentration in the CB will be nearly equal to N_d, that is $n \approx N_d$. The hole concentration will be $p = n_i^2/N_d$, which is less than the intrinsic concentration because a few of the large number of electrons in the CB recombine with holes in the VB so as to maintain $np = n_i^2$. The conductivity will then be

n-type conductivity

$$\sigma = eN_d\mu_e + e\left(\frac{n_i^2}{N_d}\right)\mu_h \approx eN_d\mu_e \qquad [5.16]$$

At low temperatures, however, not all the donors will be ionized and we need to know the probability, denoted as $f_d(E_d)$, of finding an electron in a state with energy

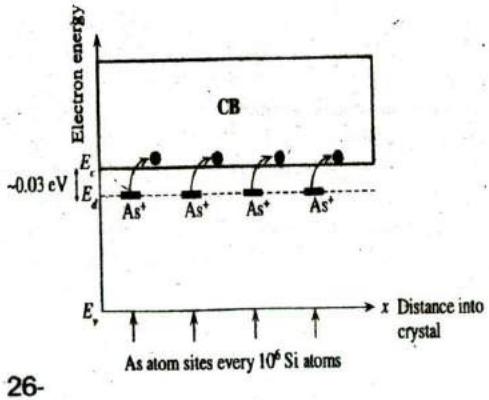


Figure 5.10 Energy band diagram for an *n*-type Si doped with 1 ppm As. There are donor energy levels just below E_c around As^+ sites.

E_d at a donor. This probability function is similar to the Fermi–Dirac function $f(E_d)$ except that it has a factor of $\frac{1}{2}$ multiplying the exponential term,

Occupation probability at a donor

$$f_d(E_d) = \frac{1}{1 + \frac{1}{2}\exp\left[\frac{(E_d - E_F)}{kT}\right]} \quad [5.17]$$

The factor $\frac{1}{2}$ is due to the fact that the electron state at the donor can take an electron with spin either up or down but not both[3] (once the donor has been occupied, a second electron cannot enter this site). Thus, the number of ionized donors at a temperature T is given by

$$\begin{aligned} N_d^+ &= N_d \times (\text{probability of not finding an electron at } E_d) \\ &= N_d[1 - f_d(E_d)] \\ &= \frac{N_d}{1 + 2\exp\left[\frac{(E_F - E_d)}{kT}\right]} \end{aligned} \quad [5.18]$$

5.2.2 *p*-TYPE DOPING

We saw that introducing a pentavalent atom into a Si crystal results in *n*-type doping because the fifth electron cannot go into a bond and escapes from the donor into the CB by thermal excitation. By similar arguments, we should anticipate that doping a Si crystal with a trivalent atom (valency of 3) such as B, Al, Ga, or In will result in a *p*-type Si crystal. We consider doping Si with small amounts of B as shown in Figure 5.11a. Because B has only three valence electrons, when it shares them with four neighboring Si atoms, one of the bonds has a missing electron, which of course is a hole. A nearby electron can tunnel into this hole and displace the hole further away from the boron atom. As the hole moves away, it gets attracted by the negative charge left behind on the boron atom and therefore takes an orbit around the B^- ion, as shown in Figure 5.11b. The binding energy of this hole to the B^- ion can be calculated using the hydrogenic atom analogy as in the *n*-type Si case. This binding energy turns out to be very small, ~0.05 eV, so at room temperature the thermal vibrations of the lattice can free the hole away from the B^- site. A free hole, we recall, exists in the VB. The escape of the hole from the B^- site involves the B atom *accepting* an electron from a neighboring Si–Si bond (from the VB), which effectively results in the hole being displaced away and its eventual escape to freedom in the VB. The B atom introduced into the Si crystal therefore acts as an electron acceptor and, because of this, it is called **an acceptor impurity.** The electron accepted by the B atom comes from a nearby bond. On the energy band diagram, an electron leaves the VB and gets accepted by a B atom, which becomes negatively charged. This process leaves a hole in the VB that is free to wander away, as illustrated in Figure 5.12.

It is apparent that doping a silicon crystal with a trivalent impurity results in a *p*-type material. We have many more holes than electrons for electrical conduction

[3] The proof can be found in advanced solid-state physics texts.

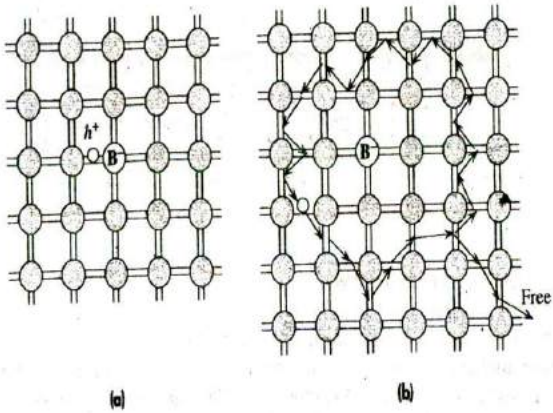


Figure 5.11 Boron-doped Si crystal.
B has only three valence electrons. When it substitutes for a Si atom, one of its bonds has an electron missing and therefore a hole, as shown in (a). The hole orbits around the B^- site by the tunneling of electrons from neighboring bonds, as shown in (b). Eventually, thermally vibrating Si atoms provide enough energy to free the hole from the B^- site into the VB, as shown.

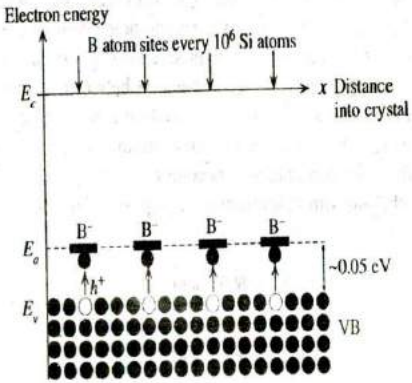


Figure 5.12 Energy band diagram for a p-type Si doped with 1 ppm B.
There are acceptor energy levels E_a just above E_v around B^- sites. These acceptor levels accept electrons from the VB and therefore create holes in the VB.

since the negatively charged B atoms are immobile and hence cannot contribute to the conductivity. If the concentration of acceptor impurities N_a in the crystal is much greater than the intrinsic concentration n_i, then at room temperature all the acceptors would have been ionized and thus $p \approx N_a$. The electron concentration is then determined by the mass action law, $n = n_i^2/N_a$, which is much smaller than p, and consequently the conductivity is simply given by $\sigma = eN_a\mu_h$.

Typical ionization energies for donor and acceptor atoms in the silicon crystal are summarized in Table 5.2.

Table 5.2 Examples of donor and acceptor ionization energies (eV) in Si

Donors			Acceptors		
P	As	Sb	B	Al	Ga
0.045	0.054	0.039	0.045	0.057	0.072

5.2.3 COMPENSATION DOPING

What happens when a semiconductor contains both donors and acceptors? **Compensation doping** is a term used to describe the doping of a semiconductor with both donors and acceptors to control the properties. For example, a p-type semiconductor doped with N_a acceptors can be converted to an n-type semiconductor by simply adding donors until the concentration N_d exceeds N_a. The effect of donors compensates for the effect of acceptors and vice versa. The electron concentration is then given by $N_d - N_a$ provided the latter is larger than n_i. When both acceptors and donors are present, what essentially happens is that electrons from donors recombine with the holes from the acceptors so that the mass action law $np = n_i^2$ is obeyed. Remember that we cannot simultaneously increase the electron and hole concentrations because that leads to an increase in the recombination rate that returns the electron and hole concentrations to satisfy $np = n_i^2$. When an acceptor atom accepts a valence band electron, a hole is created in the VB. This hole then recombines with an electron from the CB. Suppose that we have more donors than acceptors. If we take the initial electron concentration as $n = N_d$, then the recombination between the electrons from the donors and N_a holes generated by N_a acceptors results in the electron concentration reduced by N_a to $n = N_d - N_a$. By a similar argument, if we have more acceptors than donors, the hole concentration becomes $p = N_a - N_d$, with electrons from N_d donors recombining with holes from N_a acceptors. Thus there are two compensation effects:

Compensation doping

1. More donors: $N_d - N_a \gg n_i$ $\quad n = (N_d - N_a)$ and $p = \dfrac{n_i^2}{(N_d - N_a)}$

2. More acceptors: $N_a - N_d \gg n_i$ $\quad p = (N_a - N_d)$ and $n = \dfrac{n_i^2}{(N_a - N_d)}$

These arguments assume that the temperature is sufficiently high for donors and acceptors to have been ionized. This will be the case at room temperature. At low temperatures, we have to consider donor and acceptor statistics and the charge neutrality of the whole crystal, as in Example 5.8.

EXAMPLE 5.3 **RESISTIVITY OF INTRINSIC AND DOPED Si** Find the resistance of a 1 cm^3 pure silicon crystal. What is the resistance when the crystal is doped with arsenic if the doping is 1 in 10^9, that is, 1 part per billion (ppb) (note that this doping corresponds to one foreigner living in China)? Given data: Atomic concentration in silicon is 5×10^{22} cm^{-3}, $n_i = 1.0 \times 10^{10}$ cm^{-3}, $\mu_e = 1350$ cm^2 V^{-1} s^{-1}, and $\mu_h = 450$ cm^2 V^{-1} s^{-1}.

SOLUTION

For the intrinsic case, we apply

$$\sigma = en\mu_e + ep\mu_h = en(\mu_e + \mu_h)$$

so
$$\sigma = (1.6 \times 10^{-19}\,\text{C})(1.0 \times 10^{10}\,\text{cm}^{-3})(1350 + 450\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1})$$
$$= 2.88 \times 10^{-6}\,\Omega^{-1}\,\text{cm}^{-1}$$

Since $L = 1$ cm and $A = 1\ \text{cm}^2$, the resistance is

$$R = \frac{L}{\sigma A} = \frac{1}{\sigma} = 3.47 \times 10^5\,\Omega \quad \text{or} \quad 347\,\text{k}\Omega$$

When the crystal is doped with 1 in 10^9, then

$$N_d = \frac{N_{\text{Si}}}{10^9} = \frac{5 \times 10^{22}}{10^9} = 5 \times 10^{13}\,\text{cm}^{-3}$$

At room temperature all the donors are ionized, so

$$n = N_d = 5 \times 10^{13}\,\text{cm}^{-3}$$

The hole concentration is

$$p = \frac{n_i^2}{N_d} = \frac{(1.0 \times 10^{10})^2}{(5 \times 10^{13})} = 2.0 \times 10^6\,\text{cm}^{-3} \ll n_i$$

Therefore,

$$\sigma = en\mu_e = (1.6 \times 10^{-19}\,\text{C})(5 \times 10^{13}\,\text{cm}^{-3})(1350\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1})$$
$$= 1.08 \times 10^{-2}\,\Omega^{-1}\,\text{cm}^{-1}$$

Further,
$$R = \frac{L}{\sigma A} = \frac{1}{\sigma} = 92.6\,\Omega$$

Notice the drastic fall in the resistance when the crystal is doped with only 1 in 10^9 atoms.

Doping the silicon crystal with boron instead of arsenic, but still in amounts of 1 in 10^9, means that $N_a = 5 \times 10^{13}\,\text{cm}^{-3}$, which results in a conductivity of

$$\sigma = ep\mu_h = (1.6 \times 10^{-19}\,\text{C})(5 \times 10^{13}\,\text{cm}^{-3})(450\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1})$$
$$= 3.6 \times 10^{-3}\,\Omega^{-1}\,\text{cm}^{-1}$$

Therefore,
$$R = \frac{L}{\sigma A} = \frac{1}{\sigma} = 278\,\Omega$$

The reason for a higher resistance with p-type doping compared with the same amount of n-type doping is that $\mu_h < \mu_e$.

EXAMPLE 5.4

COMPENSATION DOPING An n-type Si semiconductor containing 10^{16} phosphorus (donor) atoms cm^{-3} has been doped with 10^{17} boron (acceptor) atoms cm^{-3}. Calculate the electron and hole concentrations in this semiconductor.

SOLUTION

This semiconductor has been compensation doped with excess acceptors over donors, so

$$N_a - N_d = 10^{17} - 10^{16} = 9 \times 10^{16}\,\text{cm}^{-3}$$

This is much larger than the intrinsic concentration $n_i = 1.0 \times 10^{10}\ \text{cm}^{-3}$ at room temperature, so

$$p = N_a - N_d = 9 \times 10^{16}\ \text{cm}^{-3}$$

The electron concentration

$$n = \frac{n_i^2}{p} = \frac{(1.0 \times 10^{10}\ \text{cm}^{-3})^2}{(9 \times 10^{16}\ \text{cm}^{-3})} = 1.1 \times 10^3\ \text{cm}^{-3}$$

Clearly, the electron concentration and hence its contribution to electrical conduction is completely negligible compared with the hole concentration. Thus, by excessive boron doping, the n-type semiconductor has been converted to a p-type semiconductor.

EXAMPLE 5.5 **THE FERMI LEVEL IN *n*- AND *p*-TYPE Si** An n-type Si wafer has been doped uniformly with 10^{16} antimony (Sb) atoms cm^{-3}. Calculate the position of the Fermi energy with respect to the Fermi energy E_{Fi} in intrinsic Si. The above n-type Si sample is further doped with 2×10^{17} boron atoms cm^{-3}. Calculate the position of the Fermi energy with respect to the Fermi energy E_{Fi} in intrinsic Si. (Assume that $T = 300$ K, and $kT = 0.0259$ eV.)

SOLUTION

Sb gives n-type doping with $N_d = 10^{16}\ \text{cm}^{-3}$, and since $N_d \gg n_i\ (= 1.0 \times 10^{10}\ \text{cm}^{-3})$, we have

$$n = N_d = 10^{16}\ \text{cm}^{-3}$$

For intrinsic Si,

$$n_i = N_c \exp\left[-\frac{(E_c - E_{Fi})}{kT}\right]$$

whereas for doped Si,

$$n = N_c \exp\left[-\frac{(E_c - E_{Fn})}{kT}\right] = N_d$$

where E_{Fi} and E_{Fn} are the Fermi energies in the intrinsic and n-type Si. Dividing the two expressions,

$$\frac{N_d}{n_i} = \exp\left[\frac{(E_{Fn} - E_{Fi})}{kT}\right]$$

so that

$$E_{Fn} - E_{Fi} = kT \ln\left(\frac{N_d}{n_i}\right) = (0.0259\ \text{eV}) \ln\left(\frac{10^{16}}{1.0 \times 10^{10}}\right) = 0.36\ \text{eV}$$

When the wafer is further doped with boron, the acceptor concentration is

$$N_a = 2 \times 10^{17}\ \text{cm}^{-3} > N_d = 10^{16}\ \text{cm}^{-3}$$

The semiconductor is compensation doped and compensation converts the semiconductor to p-type Si. Thus

$$p = N_a - N_d = (2 \times 10^{17} - 10^{16}) = 1.9 \times 10^{17}\ \text{cm}^{-3}$$

For intrinsic Si,

$$n_i = N_v \exp\left[-\frac{(E_{Fi} - E_v)}{kT}\right]$$

whereas for doped Si,

$$p = N_v \exp\left[-\frac{(E_{Fp} - E_v)}{kT}\right] = N_a - N_d$$

where E_{Fi} and E_{Fp} are the Fermi energies in the intrinsic and p-type Si, respectively. Dividing the two expressions,

$$\frac{p}{n_i} = \exp\left[-\frac{(E_{Fp} - E_{Fi})}{kT}\right]$$

so that

$$E_{Fp} - E_{Fi} = -kT \ln\left(\frac{p}{n_i}\right) = -(0.0259\,\text{eV}) \ln\left(\frac{1.9 \times 10^{17}}{1.0 \times 10^{10}}\right)$$
$$= -0.43\,\text{eV}$$

ENERGY BAND DIAGRAM OF AN n-TYPE SEMICONDUCTOR CONNECTED TO A VOLTAGE SUPPLY Consider the energy band diagram for an n-type semiconductor that is connected to a voltage supply of V and is carrying a current. The applied voltage drops uniformly along the semiconductor, so the electrons in the semiconductor now also have an imposed electrostatic potential energy that decreases toward the positive terminal, as depicted in Figure 5.13. The whole band structure, the CB and the VB, therefore tilts. When an electron drifts from A toward

EXAMPLE 5.6

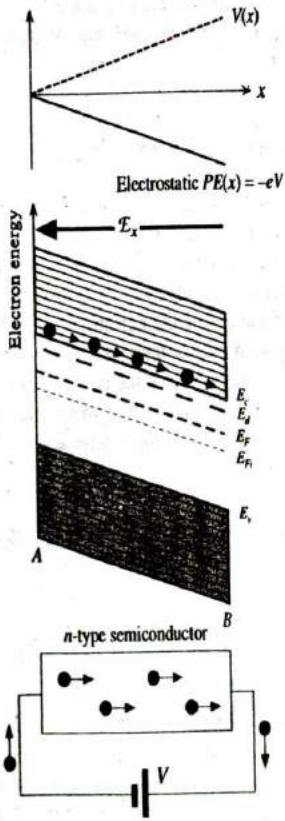


Figure 5.13 Energy band diagram of an n-type semiconductor connected to a voltage supply of V volts. The whole energy diagram tilts because the electron now also has an electrostatic potential energy.

B, its PE decreases because it is approaching the positive terminal. The Fermi level E_F is above that for the intrinsic case, E_{Fi}.

We should remember that an important property of the Fermi level is that a change in E_F within a system is available externally to do electrical work. As a corollary we note that when electrical work is done on the system, for example, when a battery is connected to a semiconductor, then E_F is not uniform throughout the whole system. A change in E_F within a system ΔE_F is equivalent to electrical work per electron or eV. E_F therefore follows the electrostatic PE behavior, and the change in E_F from one end to the other, $E_F(A) - E_F(B)$, is just eV, the energy expended in taking an electron through the semiconductor, as shown in Figure 5.13. Electron concentration in the semiconductor is uniform, so $E_c - E_F$ must be constant from one end to the other. Thus the CB, VB, and E_F all bend by the same amount.

5.3 TEMPERATURE DEPENDENCE OF CONDUCTIVITY

So far we have been calculating conductivities and resistivities of doped semiconductors at room temperature by simply assuming that $n \approx N_d$ for n-type and $p \approx N_a$ for p-type doping, with the proviso that the concentration of dopants is much greater than the intrinsic concentration n_i. To obtain the conductivity at other temperatures we have to consider two factors: the temperature dependence of the carrier concentration and the drift mobility.

5.3.1 CARRIER CONCENTRATION TEMPERATURE DEPENDENCE

Consider an n-type semiconductor doped with N_d donors per unit volume where $N_d \gg n_i$. We take the semiconductor down to very low temperatures until its conductivity is practically nil. At this temperature, the donors will *not* be ionized because the thermal vibrational energy is insufficiently small. As the temperature is increased, some of the donors become ionized and donate their electrons to the CB, as shown in Figure 5.14a. The Si–Si bond breaking, that is, thermal excitation from E_v to E_c, is unlikely because it takes too much energy. Since the donor ionization energy $\Delta E = E_c - E_d$ is very small ($\ll E_g$), thermal generation involves exciting electrons from E_d to E_c. The electron concentration at low temperatures is given by the expression

$$n = \left(\frac{1}{2}N_c N_d\right)^{1/2} \exp\left(-\frac{\Delta E}{2kT}\right) \qquad [5.19]$$

similar to the intrinsic case, that is,

$$n = (N_c N_v)^{1/2} \exp\left(-\frac{E_g}{2kT}\right) \qquad [5.20]$$

Equation 5.20 is valid when thermal generation occurs across the bandgap E_g from E_v to E_c. Equation 5.19 is the counterpart of Equation 5.20 taking into account that at low temperatures the excitation is from E_d to E_c (across ΔE) and that instead of N_v, we have N_d as the number of available electrons. The numerical factor $\frac{1}{2}$ in

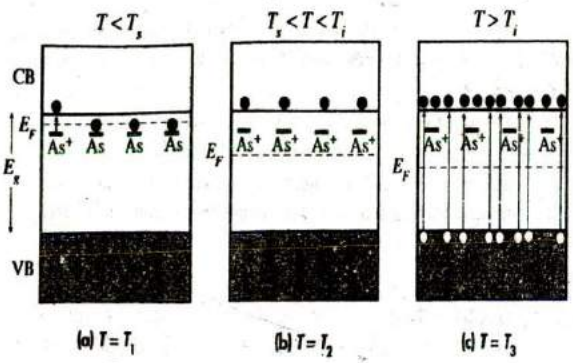


Figure 5.14

(a) Below T_s the electron concentration is controlled by the ionization of the donors.

(b) Between T_s and T_i the electron concentration is equal to the concentration of donors since they would all have ionized.

(c) At high temperatures, thermally generated electrons from the VB exceed the number of electrons from ionized donors and the semiconductor behaves as if intrinsic.

Equation 5.19 arises because donor occupation statistics is different by this factor from the usual Fermi–Dirac function, as mentioned earlier.

As the temperature is increased further, eventually all the donors become ionized and the electron concentration is equal to the donor concentration, that is, $n = N_d$, as depicted in Figure 5.14b. This state of affairs remains unchanged until very high temperatures are reached, when thermal generation across the bandgap begins to dominate. At very high temperatures, thermal vibrations of the atoms will be so strong that many Si–Si bonds will be broken and thermal generation across E_g will dominate. The electron concentration in the CB will then be mainly due to thermal excitation from the VB to the CB, as illustrated in Figure 5.14c. But this process also generates an equal concentration of holes in the VB. Accordingly, the semiconductor behaves as if it were intrinsic. The electron concentration at these temperatures will therefore be equal to the intrinsic concentration n_i, which is given by Equation 5.20.

The dependence of the electron concentration on temperature thus has three regions:

1. **Low-temperature range ($T < T_s$).** The increase in temperature at these low temperatures ionizes more and more donors. The donor ionization continues until we reach a temperature T_s, called the **saturation temperature,** when all donors have been ionized and we have saturation in the concentration of ionized donors. The electron concentration is given by Equation 5.19. This temperature range is often referred to as the **ionization range.**

2. **Medium-temperature range ($T_s < T < T_i$).** Since nearly all the donors have been ionized in this range, $n = N_d$. This condition remains unchanged until $T = T_i$, when n_i, which is temperature dependent, becomes equal to N_d. It is this

temperature range $T_s < T < T_i$ that utilizes the n-type doping properties of the semiconductor in pn junction device applications. This temperature range is often referred to as the **extrinsic range.**

3. **High-temperature range ($T > T_i$).** The concentration of electrons generated by thermal excitation across the bandgap n_i is now much larger than N_d, so the electron concentration $n = n_i(T)$. Furthermore, as excitation occurs from the VB to the CB, the hole concentration $p = n$. This temperature range is referred to as the **intrinsic range.**

Figure 5.15 shows the behavior of the electron concentration with temperature in an n-type semiconductor. By convention we plot $\ln(n)$ versus the reciprocal temperature T^{-1}. At low temperatures, $\ln(n)$ versus T^{-1} is almost a straight line with a slope $-(\Delta E/2k)$, since the temperature dependence of $N_c^{1/2}(\propto T^{3/4})$ is negligible compared with the $\exp(-\Delta E/2kT)$ part in Equation 5.19. In the high-temperature range, however, the slope is quite steep and almost $-E_g/2k$ since Equation 5.20 implies that

$$n \propto T^{3/2} \exp\left(-\frac{E_g}{2kT}\right)$$

and the exponential part again dominates over the $T^{3/2}$ part. In the intermediate range, n is equal to N_d and practically independent of the temperature.

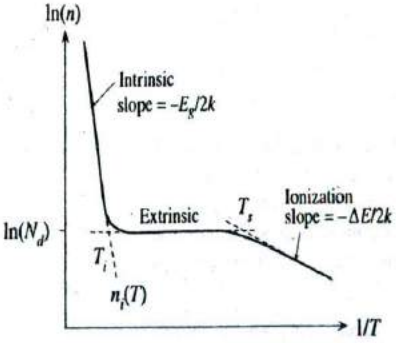


Figure 5.15 The temperature dependence of the electron concentration in an n-type semiconductor.

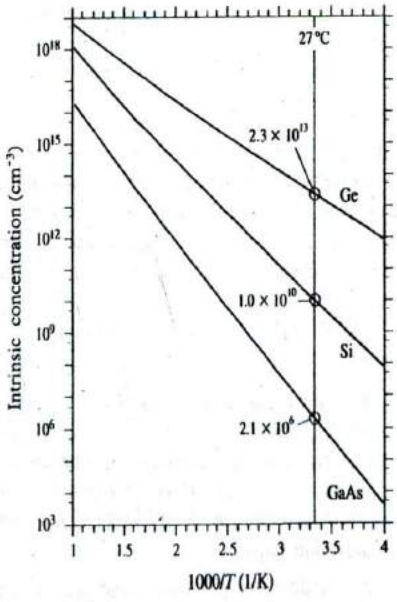


Figure 5.16 The temperature dependence of the intrinsic concentration.

Figure 5.16 displays the temperature dependence of the intrinsic concentration in Ge, Si, and GaAs as $\log(n_i)$ versus $1/T$ where the slope of the lines is, of course, a measure of the bandgap energy E_g. The $\log(n_i)$ versus $1/T$ graphs can be used to find, for example, whether the dopant concentration at a given temperature is more than the intrinsic concentration. As we will find out in Chapter 6, the reverse saturation current in a *pn* junction diode depends on n_i^2, so Figure 5.16 also indicates how this saturation current varies with temperature.

EXAMPLE 5.7

SATURATION AND INTRINSIC TEMPERATURES An *n*-type Si sample has been doped with 10^{15} phosphorus atoms cm^{-3}. The donor energy level for P in Si is 0.045 eV below the conduction band edge energy.

a. Estimate the temperature above which the sample behaves as if intrinsic.

b. Estimate the lowest temperature above which most of the donors are ionized.

SOLUTION

Remember that $n_i(T)$ is highly temperature dependent, as shown in Figure 5.16 so that as T increases, eventually at $T \approx T_i$, n_i becomes comparable to N_d. Beyond T_i, $n_i(T > T_i) \gg N_d$. Thus we need to solve

$$n_i(T_i) = N_d = 10^{15}\ \text{cm}^{-3}$$

From the $\log(n_i)$ versus $10^3/T$ graph for Si in Figure 5.16, when $n_i = 10^{15}\ \text{cm}^{-3}$, $(10^3/T_i) \approx 1.85$, giving $T_i \approx 541$ K or 268 °C.

We will assume that most of the donors are ionized, say at $T \approx T_s$, where the extrinsic and the extrapolated ionization lines intersect in Figure 5.15:

$$n = \left(\frac{1}{2}N_c N_d\right)^{1/2} \exp\left(-\frac{\Delta E}{2kT_s}\right) \approx N_d$$

This is the temperature at which the ionization behavior intersects the extrinsic region. In the above equation, $N_d = 10^{15}\ \text{cm}^{-3}$, $\Delta E = 0.045$ eV, and $N_c \propto T^{3/2}$, that is,

$$N_c(T_s) = N_c(300\ \text{K})\left(\frac{T_s}{300}\right)^{3/2}$$

Clearly, then, the equation can only be solved numerically. Similar equations occur in a wide range of physical problems where one term has the strongest temperature dependence. Here, $\exp(-\Delta E/kT_s)$ has the strongest temperature dependence. First assume N_c is that at 300 K, $N_c = 2.8 \times 10^{19}\ \text{cm}^{-3}$, and evaluate T_s,

$$T_s = \frac{\Delta E}{k \ln\left(\frac{N_c}{2N_d}\right)} = \frac{0.045\ \text{eV}}{(8.62 \times 10^{-5}\ \text{eV K}^{-1}) \ln\left[\frac{2.8 \times 10^{19}\ \text{cm}^{-3}}{2(1.0 \times 10^{15}\ \text{cm}^{-3})}\right]} = 54.7\ \text{K}$$

At $T = 54.7$ K,

$$N_c(54.7\ \text{K}) = N_c(300\ \text{K})\left(\frac{54.7}{300}\right)^{3/2} = 2.18 \times 10^{18}\ \text{cm}^{-3}$$

With this new N_c at a lower temperature, the improved T_s is 74.6 K. Since we only need an estimate of T_s, the extrinsic range of this semiconductor is therefore from about 75 to 541 K or −198 to about 268 °C.

EXAMPLE 5.8 **TEMPERATURE DEPENDENCE OF THE ELECTRON CONCENTRATION** By considering the mass action law, charge neutrality within the crystal, and occupation statistics of electronic states, we can show that at the lowest temperatures the electron concentration in an n-type semiconductor is given by

Electron concentration in the ionization region

$$n = \left(\frac{1}{2}N_c N_d\right)^{1/2} \exp\left(-\frac{\Delta E}{2kT}\right)$$

where $\Delta E = E_c - E_d$. Furthermore, at the lowest temperatures, the Fermi energy is midway between E_d and E_c.

There are only a few physical principles that must be considered to arrive at the effect of doping on the electron and hole concentrations. For an n-type semiconductor, these are

1. **Charge carrier statistics.**

$$n = N_c \exp\left[-\frac{(E_c - E_F)}{kT}\right] \tag{1}$$

2. **Mass action law.**

$$np = n_i^2 \tag{2}$$

3. **Electrical neutrality of the crystal.** We must have the same number of positive and negative charges:

$$p + N_d^+ = n \tag{3}$$

where N_d^+ is the concentration of *ionized* donors.

4. **Statistics of ionization of the dopants.**

$$N_d^+ = N_d \times (\text{probability of not finding an electron at } E_d) = N_d[1 - f_d(E_d)]$$

$$= \frac{N_d}{1 + 2\exp\left[\dfrac{(E_F - E_d)}{kT}\right]} \tag{4}$$

Solving Equations 1 to 4 for n will give the dependence of n on T and N_d. For example, from the mass action law, Equation 2, and the charge neutrality condition, Equation 3, we get

$$\frac{n_i^2}{n} + N_d^+ = n$$

This is a quadratic equation in n. Solving this equation gives

$$n = \frac{1}{2}(N_d^+) + \left[\frac{1}{4}(N_d^+)^2 + n_i^2\right]^{1/2}$$

Clearly, this equation should give the behavior of n as a function of T and N_d when we also consider the statistics in Equation 4. In the low-temperature region ($T < T_s$), n_i^2 is negligible in the expression for n and we have

$$n = N_d^+ = \frac{N_d}{1 + 2\exp\left[\dfrac{(E_F - E_d)}{kT}\right]} \approx \frac{1}{2}N_d \exp\left[-\frac{(E_F - E_d)}{kT}\right]$$

But the statistical description in Equation 1 is generally valid, so multiplying the low-temperature region equation by Equation 1 and taking the square root eliminates E_F from the expression, giving

$$n = \left(\frac{1}{2}N_c N_d\right)^{1/2} \exp\left[-\frac{(E_c - E_d)}{2kT}\right]$$

To find the location of the Fermi energy, consider the general expression

$$n = N_c \exp\left[-\frac{(E_c - E_F)}{kT}\right]$$

which must now correspond to n at low temperatures. Equating the two and rearranging to obtain E_F we find

$$E_F = \frac{E_c + E_d}{2} + \frac{1}{2}kT \ln\left(\frac{N_d}{2N_c}\right)$$

which puts the Fermi energy near the middle of $\Delta E = E_c - E_d$ at low temperatures.

5.3.2 DRIFT MOBILITY: TEMPERATURE AND IMPURITY DEPENDENCE

The temperature dependence of the drift mobility follows two distinctly different temperature variations. In the high-temperature region, it is observed that the drift mobility is limited by scattering from lattice vibrations. As the magnitude of atomic vibrations increases with temperature, the drift mobility decreases in the fashion $\mu \propto T^{-3/2}$. However, at low temperatures the lattice vibrations are not sufficiently strong to be the major limitation to the mobility of the electrons. It is observed that at low temperatures the scattering of electrons by ionized impurities is the major mobility limiting mechanism and $\mu \propto T^{3/2}$, as we will show below.

We recall from Chapter 2 that the electron drift mobility μ depends on the mean free time τ between scattering events via

$$\mu = \frac{e\tau}{m_c^*} \qquad [5.21]$$

in which

$$\tau = \frac{1}{S v_{th} N_s} \qquad [5.22]$$

where S is the cross-sectional area of the scatterer; v_{th} is the mean speed of the electrons, called the **thermal velocity**; and N_s is the number of scatterers per unit volume. If a is the amplitude of the atomic vibrations about the equilibrium, then $S = \pi a^2$. As the temperature increases, so does the amplitude a of the lattice vibrations following $a^2 \propto T$ behavior, as shown in Chapter 2. An electron in the CB is free to wander around and therefore has only *KE*. We also know that the mean kinetic energy per electron in the CB is $\frac{3}{2}kT$, just as if the kinetic molecular theory could be applied to all those electrons in the CB. Therefore,

$$\frac{1}{2}m_c^* v_{th}^2 = \frac{3}{2}kT$$

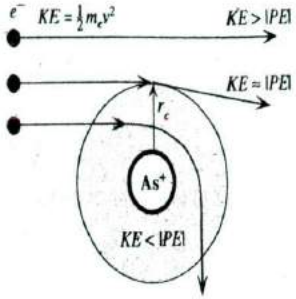


Figure 5.17 Scattering of electrons by an ionized impurity.

so that $v_{th} \propto T^{1/2}$. Thus the mean time τ_L between scattering events from lattice vibrations is

$$\tau_L = \frac{1}{(\pi a^2) v_{th} N_s} \propto \frac{1}{(T)(T^{1/2})} \propto T^{-3/2}$$

Lattice-scattering-limited mobility

which leads to a **lattice vibration scattering limited mobility**, denoted as μ_L, of the form

$$\mu_L \propto T^{-3/2} \qquad [5.23]$$

At low temperatures, scattering of electrons by thermal vibrations of the lattice will not be as strong as the electron scattering brought about by ionized donor impurities. As an electron passes by an ionized donor As^+, it is attracted and thus deflected from its straight path, as schematically shown in Figure 5.17. This type of scattering of an electron is what limits the drift mobility at low temperatures.

The *PE* of an electron at a distance r from an As^+ ion is due to the Coulombic attraction, and its magnitude is given by

$$|PE| = \frac{e^2}{4\pi\varepsilon_o\varepsilon_r r}$$

If the *KE* of the electron approaching an As^+ ion is larger than its *PE* at distance r from As^+, then the electron will essentially continue without feeling the *PE* and therefore without being deflected, and we can say that it has not been scattered. Effectively, due to its high *KE*, the electron does not feel the Coulombic pull of the donor. On the other hand, if the *KE* of the electron is less than its *PE* at r from As^+, then the *PE* of the Coulombic interaction will be so strong that the electron will be strongly deflected. This is illustrated in Figure 5.17. The critical radius r_c corresponds to the case when the electron is just scattered, which is when $KE \approx |PE(r_c)|$. But average $KE = \frac{3}{2}kT$, so at $r = r_c$

$$\frac{3}{2}kT = |PE(r_c)| = \frac{e^2}{4\pi\varepsilon_o\varepsilon_r r_c}$$

from which $r_c = e^2/(6\pi\varepsilon_o\varepsilon_r kT)$. As the temperature increases, the scattering radius decreases. The scattering cross section $S = \pi r_c^2$ is thus given by

$$S = \frac{\pi e^4}{(6\pi\varepsilon_o\varepsilon_r kT)^2} \propto T^{-2}$$

Incorporating $v_{th} \propto T^{1/2}$ as well, the temperature dependence of the mean scattering time τ_I between impurities, from Equation 5.22, must be

$$\tau_I = \frac{1}{S v_{th} N_I} \propto \frac{1}{(T^{-2})(T^{1/2}) N_I} \propto \frac{T^{3/2}}{N_I}$$

where N_I is the concentration of ionized impurities (all ionized impurities including donors and acceptors). Consequently, the **ionized impurity scattering limited mobility** from Equation 5.21 is

Ionized impurity scattering limited mobility

$$\mu_I \propto \frac{T^{3/2}}{N_I} \qquad [5.24]$$

Note also that μ_I decreases with increasing ionized dopant concentration N_I, which itself may be temperature dependent. Indeed, at the lowest temperatures, below the saturation temperature T_s, N_I will be strongly temperature dependent because not all the donors would have been fully ionized.

The overall temperature dependence of the drift mobility is then, simply, the reciprocal additions of the μ_I and μ_L by virtue of Matthiessen's rule, that is,

Effective mobility

$$\frac{1}{\mu_e} = \frac{1}{\mu_I} + \frac{1}{\mu_L} \qquad [5.25]$$

so the scattering process having the lowest mobility determines the overall (effective) drift mobility.

The experimental temperature dependence of the electron drift mobility in both Ge and Si is shown in Figure 5.18 as a log–log plot for various donor concentrations. The slope on this plot corresponds to the index n in $\mu_e \propto T^n$. The simple theoretical sketches in the insets show how μ_L and μ_I from Equations 5.23 and 5.24 depend on the temperature. For Ge, at low doping concentrations (*e.g.*, $N_d = 10^{13}\,\text{cm}^{-3}$), the experiments indicate a $\mu_e \propto T^{-1.5}$ type of behavior, which is in agreement with μ_e determined by μ_L in Equation 5.23. Curves for Si at low-level doping (μ_I negligible)

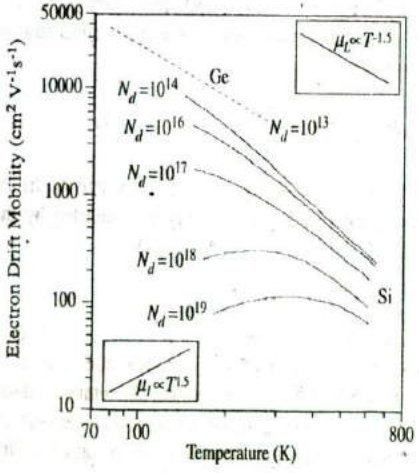


Figure 5.18 Log–log plot of drift mobility versus temperature for n-type Ge and n-type Si samples. Various donor concentrations for Si are shown. N_d are in cm^{-3}. The upper right inset is the simple theory for lattice limited mobility, whereas the lower left inset is the simple theory for impurity scattering limited mobility.

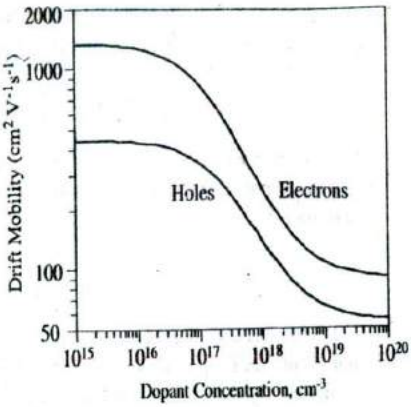


Figure 5.19 The variation of the drift mobility with dopant concentration in Si for electrons and holes at 300 K.

at high temperatures, however, exhibit a $\mu_e \propto T^{-2.5}$ type of behavior rather than $T^{-1.5}$, which can be accounted for in a more rigorous theory. As the donor concentration increases, the drift mobility decreases by virtue of μ_I getting smaller. At the highest doping concentrations and at low temperatures, the electron drift mobility in Si exhibits almost a $\mu_e \propto T^{3/2}$ type of behavior. Similar arguments can be extended to the temperature dependence of the hole drift mobility.

The dependences of the room temperature electron and hole drift mobilities on the dopant concentration for Si are shown in Figure 5.19 where, as expected, past a certain amount of impurity addition, the drift mobility is overwhelmingly controlled by μ_I in Equation 5.25.

5.3.3 Conductivity Temperature Dependence

The conductivity of an extrinsic semiconductor doped with donors depends on the electron concentration and the drift mobility, both of which have been determined above. At the lowest temperatures in the ionization range, the electron concentration depends exponentially on the temperature by virtue of

Electron concentration in ionization region

$$n = \left(\frac{1}{2}N_c N_d\right)^{1/2} \exp\left[-\frac{(E_c - E_d)}{2kT}\right]$$

which then also dominates the temperature dependence of the conductivity. In the intrinsic range at the highest temperatures, the conductivity is dominated by the temperature dependence of n_i since

$$\sigma = en_i(\mu_e + \mu_h)$$

and n_i is an exponential function of temperature in contrast to $\mu \propto T^{-3/2}$. In the extrinsic temperature range, $n = N_d$ and is constant, so the conductivity follows the temperature dependence of the drift mobility. Figure 5.20 shows schematically the semilogarithmic plot of the conductivity against the reciprocal temperature where through the extrinsic range σ exhibits a broad "S" due to the temperature dependence of the drift mobility.

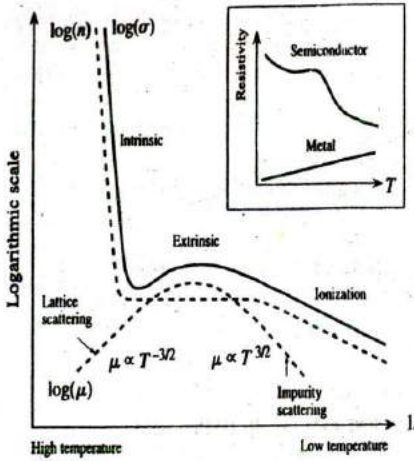


Figure 5.20 Schematic illustration of the temperature dependence of electrical conductivity for a doped (*n*-type) semiconductor.

COMPENSATION-DOPED Si EXAMPLE 5.9

a. A Si sample has been doped with 10^{17} arsenic atoms cm^{-3}. Calculate the conductivity of the sample at 27 °C (300 K) and at 127 °C (400 K).

b. The above *n*-type Si sample is further doped with 9×10^{16} boron atoms cm^{-3}. Calculate the conductivity of the sample at 27 °C and 127 °C.

SOLUTION

a. The arsenic dopant concentration, $N_d = 10^{17}\ cm^{-3}$, is much larger than the intrinsic concentration n_i, which means that $n = N_d$ and $p = (n_i^2/N_d) \ll n$ and can be neglected. Thus $n = 10^{17}\ cm^{-3}$ and the electron drift mobility at $N_d = 10^{17}\ cm^{-3}$ is 800 $cm^2\ V^{-1}\ s^{-1}$ from the drift mobility versus dopant concentration graph in Figure 5.19, so

$$\sigma = en\mu_e + ep\mu_h = eN_d\mu_e$$
$$= (1.6 \times 10^{-19}\ C)(10^{17}\ cm^{-3})(800\ cm^2\ V^{-1}s^{-1}) = 12.8\ \Omega^{-1}\ cm^{-1}$$

At $T = 127\ °C = 400\ K$,

$$\mu_e \approx 420\ cm^2\ V^{-1}\ s^{-1}$$

(from the μ_e versus T graph in Figure 5.18). Thus

$$\sigma = eN_d\mu_e = 6.72\ \Omega^{-1}\ cm^{-1}$$

b. With further doping we have $N_a = 9 \times 10^{16}\ cm^{-3}$, so from the compensation effect

$$N_d - N_a = 1 \times 10^{17} - 9 \times 10^{16} = 10^{16}\ cm^{-3}$$

Since $N_d - N_a \gg n_i$, we have an *n*-type material with $n = N_d - N_a = 10^{16}\ cm^{-3}$. But the drift mobility now is about $\sim 600\ cm^2\ V^{-1}\ s^{-1}$ because, even though $N_d - N_a$ is now $10^{16}\ cm^{-3}$ and not $10^{17}\ cm^{-3}$, all the donors and acceptors are still ionized and hence still scatter the charge carriers. The recombination of electrons from the donors and holes from the acceptors does not alter the fact that at room temperature all the dopants will be ionized.

Effectively, the compensation effect is as if all electrons from the donors were being accepted by the acceptors. Although with compensation doping the net electron concentration is $n = N_d - N_a$, the drift mobility scattering is determined by $(N_d + N_a)$, which in this case is $10^{17} + 9 \times 10^{16}\ \text{cm}^{-3} = 1.9 \times 10^{17}\ \text{cm}^{-3}$, which gives an electron drift mobility of ~600 $\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$ at 300 K and ~400 $\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$ at 400 K. Then, neglecting the hole concentration $p = n_i^2/(N_d - N_a)$, we have

At 300 K, $\sigma = e(N_d - N_a)\mu_e \approx (1.6 \times 10^{-19}\ \text{C})(10^{16}\ \text{cm}^{-3})(600\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1})$

$= 0.96\ \Omega^{-1}\ \text{cm}^{-1}$

At 400 K, $\sigma = e(N_d - N_a)\mu_e \approx (1.6 \times 10^{-19}\ \text{C})(10^{16}\ \text{cm}^{-3})(400\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1})$

$= 0.64\ \Omega^{-1}\ \text{cm}^{-1}$

5.3.4 DEGENERATE AND NONDEGENERATE SEMICONDUCTORS

The general exponential expression for the concentration of electron in the CB,

$$n \approx N_c \exp\left[-\frac{(E_c - E_F)}{kT}\right] \qquad [5.26]$$

is based on replacing Fermi–Dirac statistics with Boltzmann statistics, which is only valid when E_c is several kT above E_F. In other words, we assumed that the number of states in the CB far exceeds the number of electrons there, so the likelihood of two electrons trying to occupy the same state is almost nil. This means that the Pauli exclusion principle can be neglected and the electron statistics can be described by the Boltzmann statistics. N_c is a measure of the density of states in the CB. The Boltzmann expression for n is valid only when $n \ll N_c$. Those semiconductors for which $n \ll N_c$ and $p \ll N_v$ are termed **nondegenerate semiconductors.** They essentially follow all the discussions above and exhibit all the normal semiconductor properties outlined above.

When the semiconductor has been excessively doped with donors, then n may be so large, typically 10^{19}–$10^{20}\ \text{cm}^{-3}$, that it may be comparable to or greater than N_c. In that case the Pauli exclusion principle becomes important in the electron statistics and we have to use the Fermi–Dirac statistics. Equation 5.26 for n is then no longer valid. Such a semiconductor exhibits properties that are more metal-like than semiconductor-like; for example, the resistivity follows $\rho \propto T$. Semiconductors that have $n > N_c$ or $p > N_v$ are called **degenerate semiconductors.**

The large carrier concentration in a degenerate semiconductor is due to its heavy doping. For example, as the donor concentration in an n-type semiconductor is increased, at sufficiently high doping levels, the donor atoms become so close to each other that their orbitals overlap to form a narrow energy band that overlaps and becomes part of the conduction band. E_c is therefore slightly shifted down and E_g becomes slightly narrower. The valence electrons from the donors fill the band from E_c. This situation is reminiscent of the valence electrons filling overlapping energy bands in a metal. In a degenerate n-type semiconductor, the Fermi level is therefore within the CB, or above E_c just like E_F is within the band in a metal. The

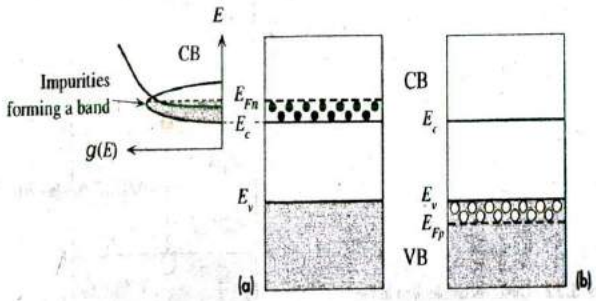


Figure 5.21

(a) Degenerate *n*-type semiconductor. Large number of donors form a band that overlaps the CB.

(b) Degenerate *p*-type semiconductor.

majority of the states between E_c and E_F are full of electrons as indicated in Figure 5.21. In the case of a *p*-type degenerate semiconductor, the Fermi level lies in the VB below E_v. It should be emphasized that one cannot simply assume that $n = N_d$ or $p = N_a$ in a degenerate semiconductor because the dopant concentration is so large that they interact with each other. Not all dopants are able to become ionized, and the carrier concentration eventually reaches a saturation typically around $\sim 10^{20}$ cm^{-3}. Furthermore, the mass action law $np = n_i^2$ is not valid for degenerate semiconductors.

Degenerate semiconductors have many important uses. For example, they are used in laser diodes, zener diodes, and ohmic contacts in ICs, and as metal gates in many microelectronic MOS devices.

5.4 RECOMBINATION AND MINORITY CARRIER INJECTION

5.4.1 DIRECT AND INDIRECT RECOMBINATION

Above absolute zero of temperature, the thermal excitation of electrons from the VB to the CB continuously generates free electron–hole pairs. It should be apparent that in equilibrium there should be some annihilation mechanism that returns the electron from the CB down to an empty state (a hole) in the VB. When a free electron, wandering around in the CB of a crystal, "meets" a hole, it falls into this low-energy empty electronic state and fills it. This process is called **recombination.** Intuitively, recombination corresponds to the free electron finding an incomplete bond with a missing electron. The electron then enters and completes this bond. The free electron in the CB and the free hole in the VB are consequently annihilated. On the energy band diagram, the recombination process is represented by returning the electron from the CB (where it is free) into a hole in the VB (where it is in a bond). Figure 5.22

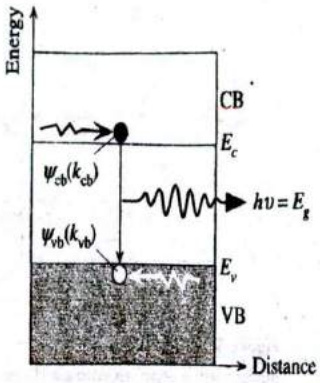


Figure 5.22 Direct recombination in GaAs. $k_{cb} = k_{vb}$ so that momentum conservation is satisfied.

shows a direct recombination mechanism, for example, as it occurs in GaAs, in which a free electron recombines with a free hole when they meet at one location in the crystal. The excess energy of the electron is lost as a photon of energy $h\nu = E_g$. In fact, it is this type of recombination that results in the emitted light from light emitting diodes (LEDs).

The recombination process between an electron and a hole, like every other process in nature, must obey the momentum conservation law. The wavefunction of an electron in the CB, $\psi_{cb}(k_{cb})$, will have a certain momentum $\hbar k_{cb}$ associated with the wavevector k_{cb} and, similarly, the electron wavefunction $\psi_{vb}(k_{vb})$ in the VB will have a momentum $\hbar k_{vb}$ associated with the wavevector k_{vb}. Conservation of linear momentum during recombination requires that when the electron drops from the CB to the VB, its wavevector should remain the same, $k_{vb} = k_{cb}$. For the elemental semiconductors, Si and Ge, the electronic states $\psi_{vb}(k_{vb})$ with $k_{vb} = k_{cb}$ are right in the middle of the VB and are therefore fully occupied. Consequently, there are no empty states in the VB that can satisfy $k_{vb} = k_{cb}$, and so direct recombination in Si and Ge is next to impossible. For some compound semiconductors, such as GaAs and InSb, for example, the states with $k_{vb} = k_{cb}$ are right at the top of the valence band, so they are essentially empty (contain holes). Consequently, an electron in the CB of GaAs can drop down to an empty electronic state at the top of the VB and maintain $k_{vb} = k_{cb}$. Thus **direct recombination** is highly probable in GaAs, and it is this very reason that makes GaAs an LED material.

In elemental semiconductor crystals, for example, in Si and Ge, electrons and holes usually recombine through recombination centers. A recombination center increases the probability of recombination because it can "take up" any momentum difference between a hole and electron. The process essentially involves a third body, which may be an impurity atom or a crystal defect. The electron is captured by the recombination center and thus becomes localized at this site. It is "held" at the center until some hole arrives and recombines with it. In the energy band diagram picture shown in Figure 5.23a, the recombination center provides a localized electronic state below E_c in the bandgap, which is at a certain location in the crystal. When an electron approaches the center, it is captured. The electron is then localized and bound to this

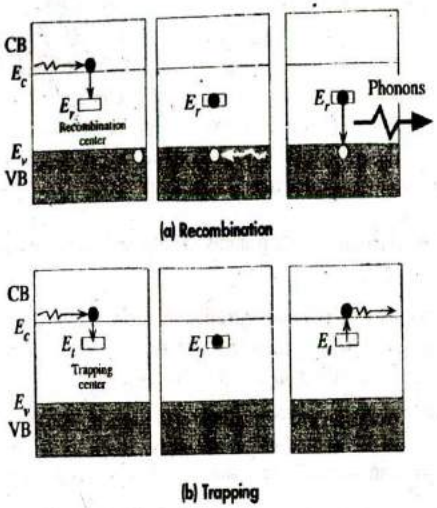


Figure 5.23 Recombination and trapping.
(a) Recombination in Si via a recombination center that has a localized energy level at E_r in the bandgap, usually near the middle.
(b) Trapping and detrapping of electrons by trapping centers. A trapping center has a localized energy level in the bandgap.

center and "waits" there for a hole with which it can recombine. In this recombination process, the energy of the electron is usually lost to lattice vibrations (as "sound") via the "recoiling" of the third body. Emitted lattice vibrations are called phonons. A **phonon** is a quantum of energy associated with atomic vibrations in the crystal analogous to the photon.

Typical recombination centers, besides the donor and acceptor impurities, might be metallic impurities and crystal defects such as dislocations, vacancies, or interstitials. Each has its own peculiar behavior in aiding recombination, which will not be described here.

It is instructive to mention briefly the phenomenon of charge carrier **trapping** since in many devices this can be the main limiting factor on the performance. An electron in the conduction band can be captured by a localized state, just like a recombination center, located in the bandgap, as shown in Figure 5.23b. The electron falls into the trapping center at E_t and becomes temporarily removed from the CB. At a later time, due to an incident energetic lattice vibration, it becomes excited back into the CB and is available for conduction again. Thus trapping involves the temporary removal of the electron from the CB, whereas in the case of recombination, the electron is permanently removed from the CB since the capture is followed by recombination with a hole. We can view a trap as essentially being a flaw in the crystal that results in the creation of a localized electronic state, around the flaw site, with an energy in the bandgap. A charge carrier passing by the flaw can be captured and lose its freedom. The flaw can be an impurity or a crystal imperfection in the same way as a recombination center. The only difference is that when a charge carrier is captured at a recombination site, it has no possibility of escaping again because the center aids recombination. Although Figure 5.23b illustrates an electron trap, similar arguments also apply to hole traps, which are normally closer to E_v. In general, flaws and defects that give localized states near the middle of the bandgap tend to act as recombination centers.

5.4.2 MINORITY CARRIER LIFETIME

Consider what happens when an n-type semiconductor, doped with 5×10^{16} cm^{-3} donors, is uniformly illuminated with appropriate wavelength light to photogenerate electron–hole pairs (EHPs), as shown in Figure 5.24. We will now define thermal equilibrium majority and minority carrier concentrations in an extrinsic semiconductor. In general, the subscript n or p is used to denote the type of semiconductor, and o to refer to thermal equilibrium in the dark.

In an n-type semiconductor, electrons are the majority carriers and holes are the minority carriers

n_{no} is defined as the **majority carrier concentration** (electron concentration in an n-type semiconductor) in thermal equilibrium in the dark. These electrons, constituting the majority carriers, are thermally ionized from the donors.

p_{no} is termed the **minority carrier concentration** (hole concentration in an n-type semiconductor) in thermal equilibrium in the dark. These holes that constitute the minority carriers are thermally generated across the bandgap.

In both cases the subscript *no* refers to an n-type semiconductor and thermal equilibrium conditions, respectively. Thermal equilibrium means that the mass action law is obeyed and $n_{no}p_{no} = n_i^2$.

When we illuminate the semiconductor, we create *excess* EHPs by photogeneration. Suppose that the electron and hole concentrations at any instant are denoted by n_n and p_n, which are defined as the *instantaneous* majority (electron) and minority (hole) concentrations, respectively. At any instant and at any location in the semiconductor, we define the departure from the equilibrium by **excess concentrations** as follows:

Δn_n is the *excess* electron (majority carrier) concentration: $\Delta n_n = n_n - n_{no}$

Δp_n is the *excess* hole (minority carrier) concentration: $\Delta p_n = p_n - p_{no}$

Under illumination, at any instant, therefore

$$n_n = n_{no} + \Delta n_n \qquad \text{and} \qquad p_n = p_{no} + \Delta p_n$$

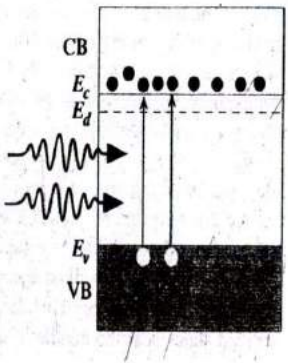


Figure 5.24 Low-level photoinjection into an n-type semiconductor in which $\Delta n_n < n_{no}$.

Photoexcitation creates EHPs or an equal number of electrons and holes, as shown in Figure 5.24, which means that

$$\Delta p_n = \Delta n_n$$

and obviously the mass action law is not obeyed: $n_n p_n \neq n_i^2$. It is worth remembering that

$$\frac{dn_n}{dt} = \frac{d\Delta n_n}{dt} \quad \text{and} \quad \frac{dp_n}{dt} = \frac{d\Delta p_n}{dt}$$

since n_{no} and p_{no} depend only on temperature.

Let us assume that we have "weak" illumination, which causes, say, only a 10 percent change in n_{no}, that is,

$$\Delta n_n = 0.1 n_{no} = 0.5 \times 10^{16}\,\text{cm}^{-3}$$

Then

$$\Delta p_n = \Delta n_n = 0.5 \times 10^{16}\,\text{cm}^{-3}$$

Figure 5.25 shows a single-axis plot of the majority (n_n) and minority (p_n) concentrations in the dark and in light. The scale is logarithmic to allow large orders of magnitude changes to be recorded. Under illumination, the minority carrier concentration is

$$p_n = p_{no} + \Delta p_n = 2.0 \times 10^3 + 0.5 \times 10^{16} \approx 0.5 \times 10^{16} = \Delta p_n$$

That is, $p_n \approx \Delta p_n$, which shows that although n_n changes by only 10 percent, p_n changes *drastically*, that is, by a factor of $\sim 10^{12}$.

Figure 5.26 shows a pictorial view of what is happening inside an n-type semiconductor when light is switched on at a certain time and then later switched off again. Obviously when the light is switched off, the condition $p_n = \Delta p_n$ (state B in Figure 5.26) must eventually revert back to the dark case (state A) where $p_n = p_{no}$. In other words, the excess minority carriers Δp_n and excess majority carriers Δn_n must

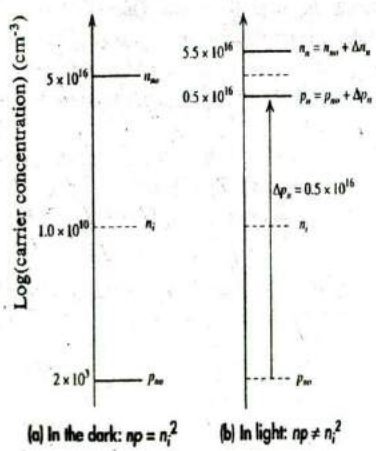


Figure 5.25 Low-level injection in an n-type semiconductor does not significantly affect n_n but drastically affects the minority carrier concentration p_n.

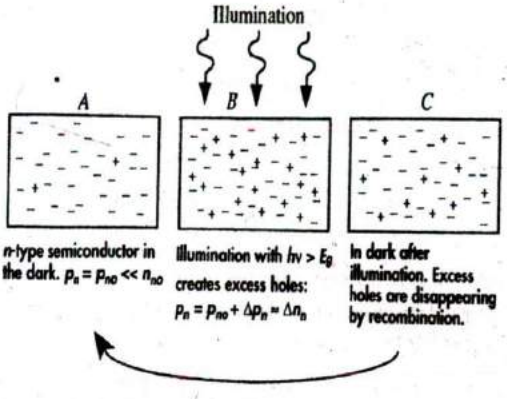


Figure 5.26 Illumination of an *n*-type semiconductor results in excess electron and hole concentrations.
After the illumination, the recombination process restores equilibrium; the excess electrons and holes simply recombine.

be removed. This removal occurs by recombination. Excess holes recombine with the electrons available and disappear. This, however, takes time because the electrons and holes have to find each other. In order to describe the rate of recombination, we introduce a temporal quantity, denoted by τ_h and called the **minority carrier lifetime (mean recombination time)**, which is defined as follows: τ_h is the average time a hole exists in the VB from its generation to its recombination, that is, the mean time the hole is free before recombining with an electron. An alternative and equivalent definition is that $1/\tau_h$ is the average probability per unit time that a hole will recombine with an electron. We must remember that the recombination process occurs through recombination centers, so the recombination time τ_h will depend on the concentration of these centers and their effectiveness in capturing the minority carriers. Once a minority carrier has been captured by a recombination center, there are many majority carriers available to recombine with it, so τ_h in an indirect process is independent of the majority carrier concentration. This is the reason for defining the recombination time as a minority carrier lifetime.

If the minority carrier recombination time is, say, 10 s, and if there are some 1000 excess holes, then it is clear that these excess holes will be disappearing at a rate of 1000/10 s = 100 per second. The rate of recombination of excess minority carriers is simply $\Delta p_n/\tau_h$. At any instant, therefore,

$$\text{Rate of increase in excess hole concentration} = \text{Rate of photogeneration} - \text{Rate of recombination of excess holes}$$

Excess minority carrier concentration

If G_{ph} is the rate of photogeneration, then clearly the net rate of change of Δp_n is

$$\frac{d\Delta p_n}{dt} = G_{ph} - \frac{\Delta p_n}{\tau_h} \qquad [5.27]$$

This is a general expression that describes the time evolution of the excess minority carrier concentration given the photogeneration rate G_{ph}, the minority carrier lifetime τ_h, and the initial condition at $t = 0$. The only assumption is weak injection ($\Delta p_n < n_{no}$).

We should note that the recombination time τ_h depends on the semiconductor material, impurities, crystal defects, temperature, and so forth, and there is no typical value to quote. It can be anywhere from nanoseconds to seconds. Later it will be shown that certain applications require a short τ_h, as in fast switching of pn junctions, whereas others require a long τ_h, for example, persistent luminescence.

EXAMPLE 5.10

PHOTORESPONSE TIME Sketch the hole concentration when a step illumination is applied to an n-type semiconductor at time $t = 0$ and switched off at time $t = t_{off} (\gg \tau_h)$.

SOLUTION

We use Equation 5.27 with G_{ph} = constant in $0 \leq t \leq t_{off}$. Since Equation 5.27 is a first-order differential equation, integrating it we simply find

$$\ln\left[G_{ph} - \left(\frac{\Delta p_n}{\tau_h}\right)\right] = -\frac{t}{\tau_h} + C_1$$

where C_1 is the integration constant. At $t = 0$, $\Delta p_n = 0$, so $C_1 = \ln G_{ph}$. Therefore the solution is

$$\Delta p_n(t) = \tau_h G_{ph}\left[1 - \exp\left(-\frac{t}{\tau_h}\right)\right] \qquad 0 \leq t < t_{off} \qquad [5.28]$$

We see that as soon as the illumination is turned on, the minority carrier concentration rises exponentially toward its steady-state value $\Delta p_n(\infty) = \tau_h G_{ph}$. This is reached after a time $t > \tau_h$.

At the instant the illumination is switched off, we assume that $t_{off} \gg \tau_h$ so that from Equation 5.28,

$$\Delta p_n(t_{off}) = \tau_h G_{ph}$$

We can define t' to be the time measured from $t = t_{off}$, that is, $t' = t - t_{off}$. Then

$$\Delta p_n(t' = 0) = \tau_h G_{ph}$$

Solving Equation 5.27 with $G_{ph} = 0$ in $t > t_{off}$ or $t' > 0$, we get

$$\Delta p_n(t') = \Delta p_n(0) \exp\left(-\frac{t'}{\tau_h}\right)$$

where $\Delta p_n(0)$ is actually an integration constant that is equivalent to the boundary condition on Δp_n at $t' = 0$. Putting $t' = 0$ and $\Delta p_n = \tau_h G_{ph}$ gives

$$\Delta p_n(t') = \tau_h G_{ph} \exp\left(-\frac{t'}{\tau_h}\right) \qquad [5.29]$$

We see that the excess minority carrier concentration decays exponentially from the instant the light is switched off with a time constant equal to the minority carrier recombination time. The time evolution of the minority carrier concentration is sketched in Figure 5.27.

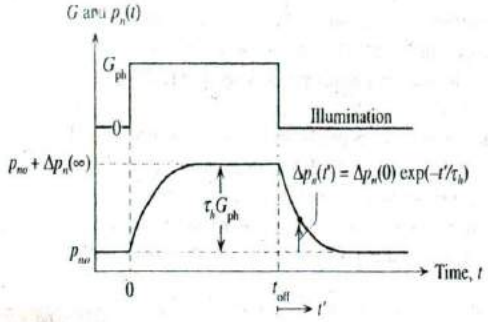


Figure 5.27 Illumination is switched on at time $t = 0$ and then off at $t = t_{off}$.
The excess minority carrier concentration $\Delta p_n(t)$ rises exponentially to its steady-state value with a time constant τ_h. From t_{off}, the excess minority carrier concentration decays exponentially to its equilibrium value.

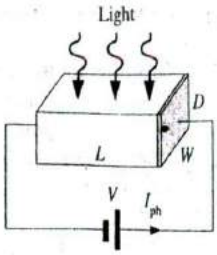


Figure 5.28 A semiconductor slab of length L, width W, and depth D is illuminated with light of wavelength λ. I_{ph} is the steady-state photocurrent.

EXAMPLE 5.11 **PHOTOCONDUCTIVITY** Suppose that a direct bandgap semiconductor with no traps is illuminated with light of intensity $I(\lambda)$ and wavelength λ that will cause photogeneration as shown in Figure 5.28. The area of illumination is $A = (L \times W)$, and the thickness (depth) of the semiconductor is D. If η is the quantum efficiency (number of free EHPs generated per absorbed photon) and τ is the recombination lifetime of the photogenerated carriers, show that the **steady-state photoconductivity**, defined as

$$\Delta\sigma = \sigma(\text{in light}) - \sigma(\text{in dark})$$

is given by

Steady-state photoconductivity

$$\Delta\sigma = \frac{e\eta I\lambda\tau(\mu_e + \mu_h)}{hcD} \qquad [5.30]$$

A photoconductive cell has a CdS crystal 1 mm long, 1 mm wide, and 0.1 mm thick with electrical contacts at the end, so the receiving area of radiation is 1 mm^2, whereas the area of each contact is 0.1 mm^2. The cell is illuminated with a blue radiation of wavelength 450 nm and intensity 1 mW/cm^2. For unity quantum efficiency and an electron recombination time of 1 ms, calculate

a. The number of EHPs generated per second

b. The photoconductivity of the sample

c. The photocurrent produced if 50 V is applied to the sample

Note that a CdS photoconductor is a direct bandgap semiconductor with an energy gap $E_g = 2.6$ eV, electron mobility $\mu_e = 0.034\ m^2\ V^{-1}\ s^{-1}$, and hole mobility $\mu_h = 0.0018\ m^2\ V^{-1}\ s^{-1}$.

SOLUTION

If Γ_{ph} is the number of photons arriving per unit area per unit second (the photon flux), then $\Gamma_{ph} = I/h\nu$ where I is the light intensity (energy flowing per unit area per second) and $h\nu$ is the energy per photon. The quantum efficiency η is defined as the number of free EHPs

generated per absorbed photon. Thus, the number of EHPs generated *per unit volume per second*, the photogeneration rate per unit volume G_{ph} is given by

$$G_{ph} = \frac{\eta A \Gamma_{ph}}{AD} = \frac{\eta\left(\frac{I}{h\nu}\right)}{D} = \frac{\eta I \lambda}{hcD}$$

In the steady state,

$$\frac{d\Delta n}{dt} = G_{ph} - \frac{\Delta n}{\tau} = 0$$

so

$$\Delta n = \tau G_{ph} = \frac{\tau \eta I \lambda}{hcD}$$

But, by definition,

$$\Delta\sigma = e\mu_e \Delta n + e\mu_h \Delta p = e\,\Delta n(\mu_e + \mu_h)$$

since electrons and holes are generated in pairs, $\Delta n = \Delta p$. Thus, substituting for Δn in the $\Delta\sigma$ expression. we get Equation 5.30:

$$\Delta\sigma = \frac{e\eta I \lambda \tau(\mu_e + \mu_h)}{hcD}$$

a. The photogeneration rate per unit time is not G_{ph}, which is per unit time per unit volume. We define EHP_{ph} as the total number of EHPs photogenerated per unit time in the whole volume (AD). Thus

$$\begin{aligned}
EHP_{ph} &= \text{Total photogeneration rate} \\
&= (AD)G_{ph} = (AD)\frac{\eta I \lambda}{hcD} = \frac{A\eta I \lambda}{hc} \\
&= [(10^{-3} \times 10^{-3}\,m^2)(1)(10^{-3} \times 10^{4}\,J\,s^{-1}\,m^{-2})(450 \times 10^{-9}\,m)] \\
&\quad \div [(6.63 \times 10^{-34}\,J\,s)(3 \times 10^{8}\,m\,s^{-1})] \\
&= 2.26 \times 10^{13}\,EHP\,s^{-1}
\end{aligned}$$

b. From Equation 5.30,

$$\Delta\sigma = \frac{e\eta I \lambda \tau(\mu_e + \mu_h)}{hcD}$$

That is

$$\Delta\sigma = \frac{(1.6 \times 10^{-19}\,C)(1)(10^{-3} \times 10^{4}\,J\,s^{-1}\,m^{-2})(450 \times 10^{-9}\,m)(1 \times 10^{-3}\,s)(0.0358\,m^2\,V^{-1}\,s^{-1})}{(6.63 \times 10^{-34}\,J\,s)(3 \times 10^{8}\,m\,s^{-1})(0.1 \times 10^{-3}\,m)}$$

$$= 1.30\,\Omega^{-1}m^{-1}$$

c. Photocurrent density will be

$$\Delta J = \mathcal{E}\,\Delta\sigma = (1.30\,\Omega^{-1}m^{-1})(50\,V/10^{-3}\,m) = 6.50 \times 10^{4}\,A\,m^{-2}$$

Thus the photocurrent

$$\Delta I = A\,\Delta J = (10^{-3} \times 0.1 \times 10^{-3}\,\text{m}^2)(6.50 \times 10^{4}\,\text{A m}^{-2})$$
$$= 6.5 \times 10^{-3}\,\text{A} \quad \text{or} \quad 6.5\,\text{mA}$$

We assumed that all the incident radiation is absorbed.

5.5 DIFFUSION AND CONDUCTION EQUATIONS, AND RANDOM MOTION

It is well known that, by virtue of their random motion, gas particles diffuse from high-concentration regions to low-concentration regions. When a perfume bottle is opened at one end of a room, the molecules diffuse out from the bottle and, after a while, can be smelled at the other end of the room. Whenever there is a concentration gradient of particles, there is a net diffusional motion of particles in the direction of decreasing concentration. The origin of diffusion lies in the random motion of particles. To quantify particle flow, we define the **particle flux** Γ just like current, as the number of particles (not charges) crossing unit area per unit time. Thus if ΔN particles cross an area A in time Δt, then, by definition, the particle flux is

Definition of particle flux

$$\Gamma = \frac{\Delta N}{A\,\Delta t} \qquad [5.31]$$

Clearly if the particles are charged with a charge Q ($-e$ for electrons and $+e$ for holes), then the electric current density J, which is basically a charge flux, is related to the particle flux Γ by

Definition of current density

$$J = Q\Gamma \qquad [5.32]$$

Suppose that the electron concentration at some time t in a semiconductor decreases in the x direction and has the profile $n(x, t)$ shown in Figure 5.29a. This may have been achieved, for example, by photogeneration at one end of a semiconductor. We will assume that the electron concentration changes only in the x direction so that the diffusion of electrons can be simplified to a one-dimensional problem as depicted in Figure 5.29a. We know that in the absence of an electric field, the electron motion is random and involves scattering from lattice vibrations and impurities. Suppose that ℓ is the mean free path in the x direction and τ is the mean free time between the scattering events. The electron moves a mean distance ℓ in the $+x$ or $-x$ direction and then it is scattered and changes direction. Its mean speed along x is $v_x = \ell/\tau$. Let us evaluate the flow of electrons in the $+x$ and $-x$ directions through the plane at x_o and hence find the net flow in the $+x$ direction.

We can divide the x axis into hypothetical segments of length ℓ so that each segment corresponds to a mean free path. Going across a segment, the electron experiences one scattering process. Consider what happens during one mean free time, the time it takes for the electrons to move across a segment toward the left or right. Half of the electrons in $(x_o - \ell)$ would be moving toward x_o and the other half away from x_o, and in time τ half of them will reach x_o and cross as shown in Figure 5.29b. If n_1 is the concentration of electrons at $x_o - \frac{1}{2}\ell$, then the number of electrons moving toward the right to

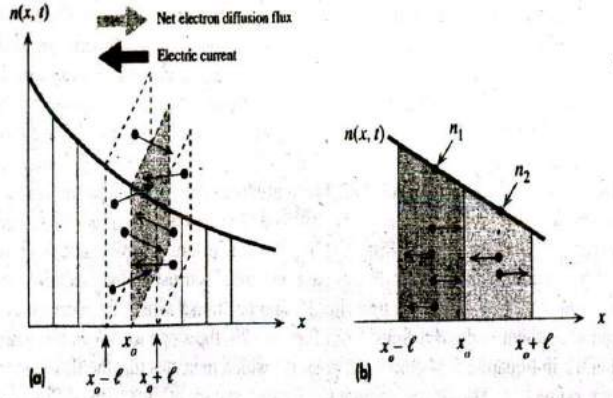


Figure 5.29

(a) Arbitrary electron concentration $n(x, t)$ profile in a semiconductor. There is a net diffusion (flux) of electrons from higher to lower concentrations.

(b) Expanded view of two adjacent sections at x_o. There are more electrons crossing x_o coming from the left $(x_o - \ell)$ than coming from the right $(x_o + \ell)$.

cross x_o is $\frac{1}{2}n_1 A\ell$ where A is the cross-sectional area and hence $A\ell$ is the volume of the segment. Similarly half of the electrons in $(x_o + \ell)$ would be moving toward the left and in time τ would reach x_o. Their number is $\frac{1}{2}n_2 A\ell$ where n_2 is the concentration at $x_o + \frac{1}{2}\ell$. The net number of electrons crossing x_o per unit time per unit area in the $+x$ direction is the electron flux Γ_e,

$$\Gamma_e = \frac{\frac{1}{2}n_1 A\ell - \frac{1}{2}n_2 A\ell}{A\tau}$$

that is,

$$\Gamma_e = -\frac{\ell}{2\tau}(n_2 - n_1) \qquad [5.33]$$

As far as calculus of variations is concerned, the mean free path ℓ is small, so we can calculate $n_2 - n_1$ from the concentration gradient using

$$n_2 - n_1 \approx \left(\frac{dn}{dx}\right)\Delta x = \left(\frac{dn}{dx}\right)\ell$$

We can now write the flux in Equation 5.33 in terms of the concentration gradient as

$$\Gamma_e = -\frac{\ell^2}{2\tau}\left(\frac{dn}{dx}\right)$$

or

$$\Gamma_e = -D_e\frac{dn}{dx} \qquad [5.34]$$

Fick's first law

where the quantity $(\ell^2/2\tau)$ has been defined as the diffusion coefficient of electrons and denoted by D_e. Thus, the net electron flux Γ_e at a position x is proportional to the concentration gradient and the diffusion coefficient. The steeper this gradient, the larger the flux Γ_e. In fact, we can view the concentration gradient dn/dx as the driving force for the diffusion flux, just like the electric field $-(dV/dx)$ is the driving force for the electric current: $J = \sigma\mathcal{E} = -\sigma(dV/dx)$.

Equation 5.34 is called **Fick's first law** and represents the relationship between the net particle flux and the driving force, which is the concentration gradient. It is the counterpart of Ohm's law for diffusion. D_e has the dimensions of $m^2\ s^{-1}$ and is a measure of how readily the particles (in this case, electrons) diffuse in the medium. Note that Equation 5.34 gives the electron flux Γ_e at a position x where the electron concentration gradient is dn/dx. Since from Figure 5.29, the slope dn/dx is a negative number, Γ_e in Equation 5.34 comes out positive, which indicates that the flux is in the positive x direction. The electric current (conventional current) due to the diffusion of electrons to the right will be in the negative direction by virtue of Equation 5.32. Representing this electric current density due to diffusion as $J_{D,e}$ we can write

Electron diffusion current density

$$J_{D,e} = -e\Gamma_e = eD_e\frac{dn}{dx} \qquad [5.35]$$

In the case of a hole concentration gradient, as shown in Figure 5.30, the hole flux $\Gamma_h(x)$ is given by

$$\Gamma_h = -D_h\frac{dp}{dx}$$

where D_h is the hole diffusion coefficient. Putting in a negative number for the slope dp/dx, as shown in Figure 5.30, results in a positive hole flux (in the positive x direction), which in turn implies a diffusion current density toward the right. The current density due to hole diffusion is given by

Hole diffusion current density

$$J_{D,h} = e\Gamma_h = -eD_h\frac{dp}{dx} \qquad [5.36]$$

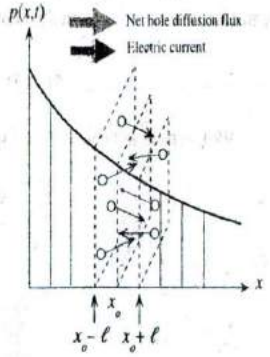


Figure 5.30 Arbitrary hole concentration $p(x, t)$ profile in a semiconductor.
There is a net diffusion (flux) of holes from higher to lower concentrations. There are more holes crossing x_o coming from the left $(x_o - \ell)$ than coming from the right $(x_o + \ell)$.

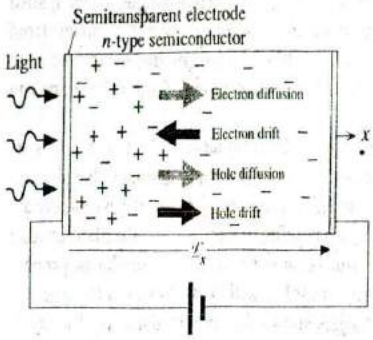


Figure 5.31 When there is an electric field and also a concentration gradient, charge carriers move both by diffusion and drift.

Suppose that there is also a positive electric field $\mathcal{E}_x$ acting along $+x$ in Figures 5.29 and 5.30. A practical example is shown in Figure 5.31 in which a semiconductor is sandwiched between two electrodes, the left one semitransparent. By connecting a battery to the electrodes, an applied field of $\mathcal{E}_x$ is set up in the semiconductor along $+x$. The left electrode is continuously illuminated, so excess EHPs are generated at this surface that give rise to concentration gradients in n and p. The applied field imposes an electrical force on the charges, which then try to drift. Holes drift toward the right and electrons toward the left. Charge motion then involves both drift and diffusion. The total current density due to the electrons drifting, driven by $\mathcal{E}_x$, and also diffusing, driven by dn/dx, is then given by adding Equation 5.35 to the usual electron drift current density,

Total electron current due to drift and diffusion

$$J_e = en\mu_e\mathcal{E}_x + eD_e\frac{dn}{dx} \qquad [5.37]$$

We note that as $\mathcal{E}_x$ is along x, so is the drift current (first term), but the diffusion current (second term) is actually in the opposite direction by virtue of a negative dn/dx.

Similarly, the hole current due to holes drifting and diffusing, Equation 5.36, is given by

Total hole current due to drift and diffusion

$$J_h = ep\mu_h\mathcal{E}_x - eD_h\frac{dp}{dx} \qquad [5.38]$$

In this case the drift and diffusion currents are in the same direction.

We mentioned that the diffusion coefficient is a measure of the ease with which the diffusing charge carriers move in the medium. But drift mobility is also a measure of the ease with which the charge carriers move in the medium. The two quantities are related through the **Einstein relation,**

Einstein relation

$$\frac{D_e}{\mu_e} = \frac{kT}{e} \quad \text{and} \quad \frac{D_h}{\mu_h} = \frac{kT}{e} \qquad [5.39]$$

In other words, the diffusion coefficient is proportional to the temperature and mobility. This is a reasonable expectation since increasing the temperature will

increase the mean speed and thus accelerate diffusion. The randomizing effect against diffusion in one particular direction is introduced by the scattering of the carriers from lattice vibrations, impurities, and so forth, so that the longer the mean free path between scattering events, the larger the diffusion coefficient. This is examined in Example 5.12.

We equated the diffusion coefficient D to $\ell^2/2\tau$ in Equation 5.34. Our analysis, as represented in Figure 5.29, is oversimplified because we simply assumed that all electrons move a distance ℓ before scattering and all are free for a time τ. We essentially assumed that all those at a distance ℓ from x_o and moving toward x_o cross the plane exactly in time τ. This assumption is not entirely true because scattering is a stochastic process and consequently not all electrons moving toward x_o will cross it even in the segment of thickness ℓ. A rigorous statistical analysis shows that the diffusion coefficient is given by

Diffusion coefficient

$$D = \frac{\ell^2}{\tau} \qquad [5.40]$$

EXAMPLE 5.12 **THE EINSTEIN RELATION** Using the relation between the drift mobility and the mean free time τ between scattering events and the expression for the diffusion coefficient $D = \ell^2/\tau$, derive the Einstein relation for electrons.

SOLUTION

In one dimension, for example, along x, the diffusion coefficient for electrons is given by $D_e = \ell^2/\tau$ where ℓ is the mean free path along x and τ is the mean free time between scattering events for electrons. The mean free path $\ell = v_x\tau$, where v_x is the mean (or effective) speed of the electrons along x. Thus,

$$D_e = v_x^2\tau$$

In the conduction band and in one dimension, the mean KE of electrons is $\frac{1}{2}kT$, so $\frac{1}{2}kT = \frac{1}{2}m_e^*v_x^2$ where m_e^* is the effective mass of the electron in the CB. This gives

$$v_x^2 = \frac{kT}{m_e^*}$$

Substituting for v_x in the D_e equation, we get,

$$D_e = \frac{kT\tau}{m_e^*} = \frac{kT}{e}\left(\frac{e\tau}{m_e^*}\right)$$

Further, we know from Chapter 2 that the electron drift mobility μ_e is related to the mean free time τ via $\mu_e = e\tau/m_e^*$, so we can substitute for τ to obtain

$$D_e = \frac{kT}{e}\mu_e$$

which is the Einstein relation. We assumed that Boltzmann statistics, that is, $v_x^2 = kT/m_e^*$ is applicable, which, of course, is true for the conduction band electrons in a semiconductor but not for the conduction electrons in a metal. Thus, the Einstein relation is only valid for electrons and holes in a nondegenerate semiconductor and certainly not valid for electrons in a metal.

DIFFUSION COEFFICIENT OF ELECTRONS IN Si Calculate the diffusion coefficient of electrons at 27 °C in *n*-type Si doped with 10^{15} As atoms cm^{-3}. **EXAMPLE 5.13**

SOLUTION

From the μ_e versus dopant concentration graph, the electron drift mobility μ_e with 10^{15} cm^{-3} of dopants is about 1300 $cm^2\ V^{-1}\ s^{-1}$, so

$$D_e = \frac{\mu_e kT}{e} = (1300\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1})(0.0259\ \text{V}) = 33.7\ \text{cm}^2\ \text{s}^{-1}$$

BUILT-IN POTENTIAL DUE TO DOPING VARIATION Suppose that due to a variation in the amount of donor doping in a semiconductor, the electron concentration is nonuniform across the semiconductor, that is, $n = n(x)$. What will be the potential difference between two points in the semiconductors where the electron concentrations are n_1 and n_2? If the donor profile in an *n*-type semiconductor is $N(x) = N_a \exp(-x/b)$, where b is a characteristic of the exponential doping profile, evaluate the built-in field $\mathcal{E}_x$. What is your conclusion? **EXAMPLE 5.14**

SOLUTION

Consider a nonuniformly doped *n*-type semiconductor in which immediately after doping the donor concentration, and hence the electron concentration, decreases toward the right. Initially, the sample is neutral everywhere. The electrons will immediately diffuse from higher- to lower-concentration regions. But this diffusion accumulates *excess* electrons in the right region and exposes the positively charged donors in the left region, as depicted in Figure 5.32. The electric field between the accumulated negative charges and the exposed donors prevents further accumulation. Equilibrium is reached when the diffusion toward the right is just balanced by the drift of electrons toward the left. The total current in the sample must be zero (it is an open circuit),

$$J_e = en\mu_e\mathcal{E}_x + eD_e\frac{dn}{dx} = 0$$

But the field is related to the potential difference by $\mathcal{E}_x = -(dV/dx)$, so

$$-en\mu_e\frac{dV}{dx} + eD_e\frac{dn}{dx} = 0$$

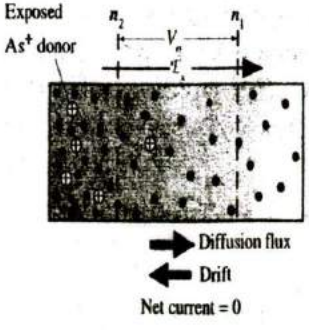


Figure 5.32 Nonuniform doping profile results in electron diffusion toward the less concentrated regions.
This exposes positively charged donors and sets up a built-in field $\mathcal{E}_x$. In the steady state, the diffusion of electrons toward the right is balanced by their drift toward the left.

We can now use the Einstein relation $D_e/\mu_e = kT/e$ to eliminate D_e and μ_e and then cancel dx and integrate the equation,

$$\int_{V_1}^{V_2} dV = \frac{kT}{e}\int_{n_1}^{n_2}\frac{dn}{n}$$

Integrating, we obtain the potential difference between points 1 and 2,

Built-in potential and concentration

$$V_2 - V_1 = \frac{kT}{e}\ln\left(\frac{n_2}{n_1}\right) \qquad [5.41]$$

To find the built-in field, we will assume that (and this is a reasonable assumption) the diffusion of electrons toward the right has not drastically upset the original $n(x) = N_d(x)$ variation because the field builds up quickly to establish equilibrium. Thus

$$n(x) \approx N_d(x) = N_e \exp\left(-\frac{x}{b}\right)$$

Substituting into the equation for $J_e = 0$, and again using the Einstein relation, we obtain $\mathcal{E}_x$ as

Built-in field

$$\mathcal{E}_x = \frac{kT}{be} \qquad [5.42]$$

Note: As a result of the fabrication process, the base region of a bipolar transistor has nonuniform doping, which can be approximated by an exponential $N_d(x)$. The resulting electric field $\mathcal{E}_x$ in Equation 5.42 acts to drift minority carriers faster and therefore speeds up the transistor operation as discussed in Chapter 6.

5.6 CONTINUITY EQUATION[4]

5.6.1 TIME-DEPENDENT CONTINUITY EQUATION

Many semiconductor devices operate on the principle that excess charge carriers are injected into a semiconductor by external means such as illumination or an applied voltage. The injection of carriers upsets the equilibrium concentration. To determine the carrier concentration at any point at any instant we need to solve the **continuity equation,** which is based on accounting for the total charge at that location in the semiconductor. Consider an n-type semiconductor slab as shown in Figure 5.33 in which the hole concentration has been upset along the x axis from its equilibrium value p_{no} by some external means.

Consider an infinitesimally thin elemental volume $A\,\delta x$ as in Figure 5.33 in which the hole concentration is $p_n(x, t)$. The current density at x due to holes flowing into the volume is J_h and that due to holes flowing out at $x + \delta x$ is $J_h + \delta J_h$. There is a change in the hole current density J_h; that is, $J_h(x, t)$ is not uniform along x. (Recall that the total current will also have a component due to electrons.) We assume that $J_h(x, t)$ and $p_n(x, t)$ do not change across the cross section along the y or z directions. If δJ_h is

[4] This section may be skipped without loss of continuity. (No pun intended.)

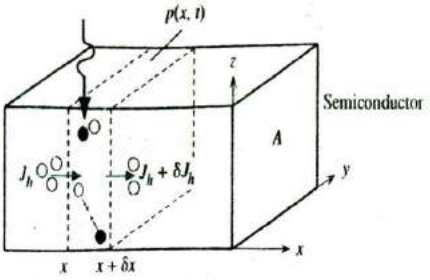


Figure 5.33 Consider an elemental volume $A\,\delta x$ in which the hole concentration is $p(x, t)$.

negative, then the current leaving the volume is less than that entering the volume, which leads to an increase in the hole concentration in $A\,\delta x$. Thus,

$$\frac{1}{A\,\delta x}\left(\frac{-A\,\delta J_h}{e}\right) = \text{Rate of increase in hole concentration due to the change in } J_h \qquad [5.43]$$

The negative sign ensures that negative δJ_h leads to an increase in p_n. Recombination taking place in $A\,\delta x$ removes holes from this volume. In addition, there may also be photogeneration at x at time t. Thus,

The *net* rate of increase in the hole concentration p_n in $A\,\delta x$
= Rate of increase due to decrease in J_h − Rate of recombination + Rate of photogeneration

$$\frac{\partial p_n}{\partial t} = -\frac{1}{e}\left(\frac{\partial J_h}{\partial x}\right) - \frac{p_n - p_{no}}{\tau_h} + G_{ph} \qquad [5.44]$$

Continuity equation for holes

where τ_h is the hole recombination time (lifetime), G_{ph} is the photogeneration rate at x at time t, and we used $\partial J_h/\partial x$ for $\delta J_h/\delta x$ since J_h depends on x and t.

Equation 5.44 is called the **continuity equation** for holes. The current density J_h is given by diffusion and drift components in Equations 5.37 and 5.38. There is a similar expression for electrons as well, but the negative sign multiplying $\partial J_e/\partial x$ is changed to positive (the charge e is negative for electrons).

The solutions of the continuity equation depend on the initial and boundary conditions. Many device scientists and engineers have solved Equation 5.44 for various semiconductor problems to characterize the behavior of devices. In most cases numerical solutions are necessary as analytical solutions are not mathematically tractable. As a simple example, consider uniform illumination of the surface of a semiconductor with suitable electrodes at its end as in Figure 5.28. Photogeneration and current density do not vary with distance along the sample length, so $\partial J_h/\partial x = 0$. If Δp_n is the excess concentration, $\Delta p_n = p_n - p_{no}$, then the time derivative of p_n in Equation 5.44 is the same as Δp_n. Thus, the continuity equation becomes

$$\frac{\partial \Delta p_n}{\partial t} = -\frac{\Delta p_n}{\tau_h} + G_{ph} \qquad [5.45]$$

Continuity equation with uniform photo-generation

which is identical to the semiquantitatively derived Equation 5.27 from which photoconductivity was calculated in Example 5.11.

5.6.2 STEADY-STATE CONTINUITY EQUATION

For certain problems, the continuity equation can be further simplified. Consider, for example, the continuous illumination of one end of an *n*-type semiconductor slab by light that is absorbed in a very small thickness x_o at the surface as depicted in Figure 5.34a. There is no bulk photogeneration, so $G_{ph} = 0$. Suppose we are interested in the **steady-state** behavior; then the time derivative would be zero in Equation 5.44 to give,

Steady-state continuity equation for holes

$$\frac{1}{e}\left(\frac{\partial J_h}{\partial x}\right) = -\frac{p_n - p_{no}}{\tau_h} \qquad [5.46]$$

The hole current density J_h would have diffusion and drift components. If we assume that the electric field is very small, we can use Equation 5.38 with $\mathcal{E} \approx 0$ in Equation 5.46. Further, since the excess concentration $\Delta p_n(x) = p_n(x) - p_{no}$, we obtain,

Steady-state continuity equation with $\mathcal{E} \approx 0$

$$\frac{d^2 \Delta p_n}{dx^2} = \frac{\Delta p_n}{L_h^2} \qquad [5.47]$$

where, by definition, $L_h = \sqrt{D_h \tau_h}$ and is called the **diffusion length of holes.** Equation 5.47 describes the **steady-state** behavior of minority carrier concentration in a semiconductor under time-invariant excitation. When the appropriate boundary conditions are also included, its solution gives the *spatial* dependence of the excess minority carrier concentration $\Delta p_n(x)$.

In Figure 5.34a, both excess electrons and holes are photogenerated at the surface, but the percentage increase in the concentration of holes is much more dramatic since

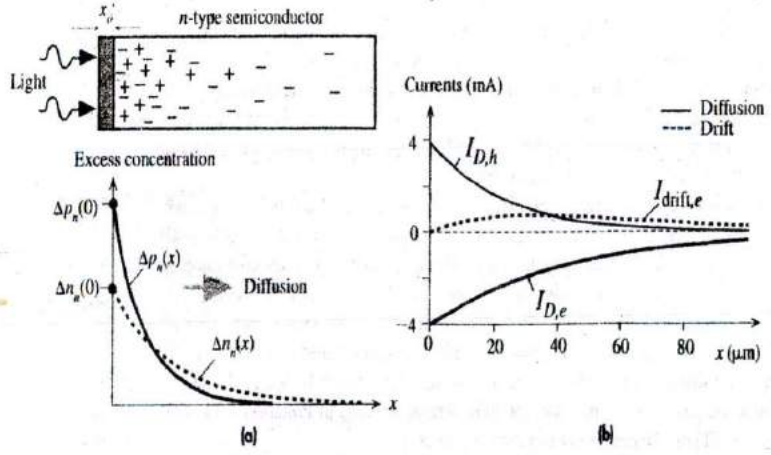


Figure 5.34

(a) Steady-state excess carrier concentration profiles in an *n*-type semiconductor that is continuously illuminated at one end.

(b) Majority and minority carrier current components in open circuit. Total current is zero.

$p_{no} \ll n_{no}$. We will assume **weak injection,** that is, $\Delta p_n \ll n_{no}$. Suppose that illumination is such that it causes the excess hole concentration at $x = 0$ to be $\Delta p_n(0)$. As holes diffuse toward the right, they meet electrons and recombine as a result of which the hole concentration $p_n(x)$ decays with distance into the semiconductor. If the bar is very long, then far away from the injection end we would expect p_n to be equal to the thermal equilibrium concentration p_{no}. The solution of Equation 5.47 with these boundary conditions shows that $\Delta p_n(x)$ decays *exponentially* as

Minority carrier concentration, long bar

$$\Delta p_n(x) = \Delta p_n(0) \exp\left(-\frac{x}{L_h}\right) \qquad [5.48]$$

This decay in the hole concentration results in a hole diffusion current $I_{D,h}(x)$ that has the same spatial dependence. Thus, if A is the cross-sectional area, the hole current is

Hole diffusion current

$$I_h \approx I_{D,h} = -AeD_h \frac{dp_n(x)}{dx} = \frac{AeD_h}{L_h} \Delta p_n(0) \exp\left(-\frac{x}{L_h}\right) \qquad [5.49]$$

We find $\Delta p_n(0)$ as follows. Under steady state, the holes generated per unit time in x_o must be removed by the hole current (at $x = 0$) at the *same* rate. Thus,

$$Ax_o G_{ph} = \frac{1}{e} I_{D,h}(0) = \frac{AD_h}{L_h} \Delta p_n(0)$$

or

$$\Delta p_n(0) = x_o G_{ph} \left(\frac{\tau_h}{D_h}\right)^{1/2} \qquad [5.50]$$

Similarly, electrons photogenerated in x_o diffuse toward the bulk, but their diffusion coefficient D_e and length L_e are larger than those for holes. The excess electron concentration Δn_n decays as

Majority carrier concentration, long bar

$$\Delta n_n(x) = \Delta n_n(0) \exp\left(-\frac{x}{L_e}\right) \qquad [5.51]$$

where $L_e = \sqrt{D_e \tau_h}$ and $\Delta n_n(x)$ decays more slowly than $\Delta p_n(x)$ as $L_e > L_h$. (Note that $\tau_e = \tau_h$.) The electron diffusion current $I_{D,e}$ is

Electron diffusion current

$$I_{D,e} = AeD_e \frac{dn_n(x)}{dx} = -\frac{AeD_e}{L_e} \Delta n_n(0) \exp\left(-\frac{x}{L_e}\right) \qquad [5.52]$$

The field at the surface is zero. Under steady state, the electrons generated per unit time in x_o must be removed by the electron current at the *same* rate. Thus, similarly to Equation 5.50,

$$\Delta n_n(0) = x_o G_{ph} \left(\frac{\tau_h}{D_e}\right)^{1/2} \qquad [5.53]$$

so that

$$\frac{\Delta p_n(0)}{\Delta n_n(0)} = \left(\frac{D_e}{D_h}\right)^{1/2} \qquad [5.54]$$

which is greater than unity for Si.

Table 5.3 Currents in an infinite slab illuminated at one end for weak injection near the surface

Currents at	Minority Diffusion $I_{D,h}$ (mA)	Minority Drift $I_{drift,h}$ (mA)	Majority Diffusion $I_{D,e}$ (mA)	Majority Drift $I_{drift,e}$ (mA)	Field $\mathcal{E}$ (V cm^{-1})
$x = 0$	3.94	0	−3.94	0	0
$x = L_e$	0.70	0.0022	−1.45	0.75	0.035

It is apparent that the hole and electron diffusion currents are in *opposite* directions. At the surface, the electron and hole diffusion currents are equal and opposite, so the total current is zero. As apparent from Equations 5.49 and 5.52, the hole diffusion current decays more rapidly than the electron diffusion current, so there must be some electron drift to keep the total current zero. The electrons are majority carriers which means that even a small field can cause a marked majority carrier drift current. If $I_{drift,e}$ is the electron drift current, then in an open circuit the total current $I = I_{D,h} + I_{D,e} + I_{drift,e} = 0$, so

Electron drift current

$$I_{drift,e} = -I_{D,h} - I_{D,e} \qquad [5.55]$$

The electron drift current increases with distance, so the total current I at every location is zero. It must be emphasized that there must be some field $\mathcal{E}$ in the sample, however small, to provide the necessary drift to balance the currents to zero. The field can be found from $I_{drift,e} \approx Aen_{no}\mu_e\mathcal{E}$, inasmuch as n_{no} does not change significantly (weak injection),

Electric field

$$\mathcal{E} = \frac{I_{drift,e}}{Aen_{no}\mu_e} \qquad [5.56]$$

The hole drift current due to this field is

Hole drift current

$$I_{drift,h} = Ae\mu_h p_n(x)\mathcal{E} \qquad [5.57]$$

and it will be negligibly small as $p_n \ll n_{no}$.

We can use actual values to gauge magnitudes. Suppose that $A = 1$ mm^2 and $N_d = 10^{16}$ cm^{-3} so that $n_{no} = N_d = 10^{16}$ cm^{-3} and $p_{no} = n_i^2/N_d = 1 \times 10^4$ cm^{-3}. The light intensity is adjusted to yield $\Delta p_n(0) = 0.05n_{no} = 5 \times 10^{14}$ cm^{-3}: *weak injection*. Typical values at 300 K for the material properties in this N_d-doped n-type Si would be $\tau_h = 480$ ns, $\mu_e = 1350$ cm^2 V^{-1} s^{-1}, $D_e = 34.9$cm^2 s^{-1}, $L_e = 0.0041$ cm = 41 μm, $\mu_h = 450$ cm^2 V^{-1} s^{-1}, $D_h = 11.6$ cm^2 s^{-1}, $L_h = 0.0024$ cm = 24 μm. We can now calculate each current term using the Equations 5.49, 5.52, 5.55 and 5.57 above as shown in Figure 5.34b. The actual values at two locations, $x = 0$ and $x = L_e = 41$ μm, are shown in Table 5.3.[5]

[5] The reader may have observed that the currents in Table 5.3 do not add exactly to zero. The analysis here is only approximate and, further, it was based on neglecting the hole drift current and taking the field as nearly zero to use Equation 5.47 in deriving the carrier concentration profiles. Note that hole drift current is much smaller than the other current components.

EXAMPLE 5.15

INFINITELY LONG SEMICONDUCTOR ILLUMINATED AT ONE END Find the minority carrier concentration profile $p_n(x)$ in an infinite n-type semiconductor that is illuminated continuously at one end as in Figure 5.34. Assume that photogeneration occurs near the surface. Show that the mean distance diffused by the minority carriers before recombination is L_h.

SOLUTION

Continuous illumination means that we have steady-state conditions and thus Equation 5.47 can be used. The general solution of this second-order differential equation is

$$\Delta p_n(x) = A\exp\left(-\frac{x}{L_h}\right) + B\exp\left(\frac{x}{L_h}\right) \qquad [5.58]$$

where A and B are constants that have to be found from the boundary conditions. For an infinite bar, at $x = \infty$, $\Delta p_n(\infty) = 0$ gives $B = 0$. At $x = 0$, $\Delta p_n = \Delta p_n(0)$ so $A = \Delta p_n(0)$. Thus, the excess (photoinjected) hole concentration at position x is

$$\Delta p_n(x) = \Delta p_n(0)\exp\left(-\frac{x}{L_h}\right) \qquad [5.59]$$

which is shown in Figure 5.34a. To find the mean position of the photoinjected holes, we use the definition of the "mean," that is,

$$\overline{x} = \frac{\int_0^\infty x\,\Delta p_n(x)\,dx}{\int_0^\infty \Delta p_n(x)\,dx}$$

Substituting for $\Delta p_n(x)$ from Equation 5.59 and carrying out the integration gives $\overline{x} = L_h$. We conclude that the **diffusion length** L_h is the average distance diffused by the minority carriers before recombination. As a corollary, we should infer that $1/L_h$ is the mean probability per unit distance that the hole recombines with an electron.

5.7 OPTICAL ABSORPTION

We have already seen that a photon of energy $h\nu$ greater than E_g can be absorbed in a semiconductor, resulting in the excitation of an electron from the valence band to the conduction band, as illustrated in Figure 5.35. The average energy of electrons in the conduction band is $\frac{3}{2}kT$ above E_c (average kinetic energy is $\frac{3}{2}kT$), which means that the electrons are very close to E_c. If the photon energy is much larger than the bandgap energy E_g, then the excited electron is not near E_c and has to lose the extra energy $h\nu - E_g$ to reach thermal equilibrium. The excess energy $h\nu - E_g$ is lost to lattice vibrations as heat as the electron is scattered from one atomic vibration to another. This process is called **thermalization.** If, on the other hand, the photon energy $h\nu$ is less than the bandgap energy, the photon will not be absorbed and we can say that the semiconductor is transparent to wavelengths longer than hc/E_g provided that there are no energy states in the bandgap. There, of course, will be reflections occurring at the air/semiconductor surface due to the change in the refractive index.

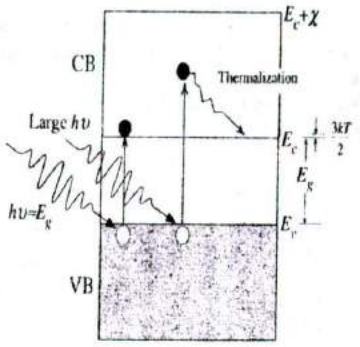


Figure 5.35 Optical absorption generates electron–hole pairs.

Energetic electrons must lose their excess energy to lattice vibrations until their average energy is $\frac{3}{2}kT$ in the CB.

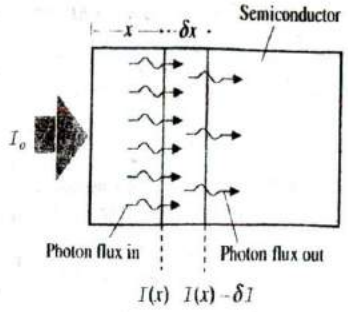


Figure 5.36 Absorption of photons within a small elemental volume of width δx.

Suppose that I_o is the intensity of a beam of photons incident on a semiconductor material. Thus, I_o is the energy incident per unit area per unit time. If Γ_{ph} is the photon flux, then

$$I_o = h\upsilon\Gamma_{ph}$$

When the photon energy is greater than E_g, photons from the incident radiation will be absorbed by the semiconductor. The absorption of photons requires the excitation of valence band electrons, and there are only so many of them with the right energy *per unit volume*. Consequently, absorption depends on the thickness of the semiconductor. Suppose that $I(x)$ is the light intensity at x and δI is the change in the light intensity in the small elemental volume of thickness δx at x due to photon absorption, as illustrated in Figure 5.36. Then δI will depend on the number of photons arriving at this volume $I(x)$ and the thickness δx. Thus

$$\delta I = -\alpha I\,\delta x$$

where α is a proportionality constant that depends on the photon energy and hence wavelength, that is, $\alpha = \alpha(\lambda)$. The negative sign ensures that δI is a reduction. The constant α as defined by this equation is called the **absorption coefficient** of the semiconductor. It is therefore defined by

Definition of absorption coefficient

$$\alpha = -\frac{\delta I}{I\,\delta x} \qquad [5.60]$$

which has the dimensions of length^{-1} (m^{-1}).

When we integrate Equation 5.60 for illumination with constant wavelength light, we get the **Beer–Lambert law,** the transmitted intensity decreases exponentially with the thickness,

Beer–Lambert law

$$I(x) = I_o \exp(-\alpha x) \qquad [5.61]$$

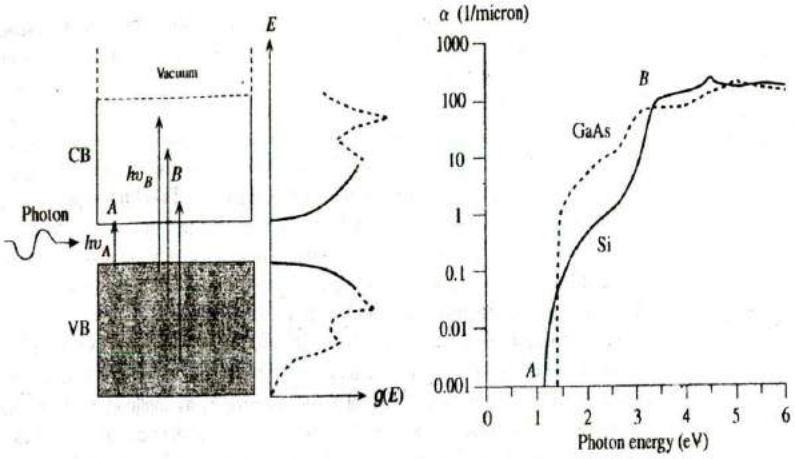


Figure 5.37 The absorption coefficient α depends on the photon energy $h\nu$ and hence on the wavelength. Density of states increases from band edges and usually exhibits peaks and troughs. Generally α increases with the photon energy greater than E_g because more energetic photons can excite electrons from populated regions of the VB to numerous available states deep in the CB.

As apparent from Equation 5.61, over a distance $x = 1/\alpha$, the light intensity falls to a value $0.37I_o$; that is, it decreases by 63 percent. This distance over which 67 percent of the photons are absorbed is called the **penetration depth,** denoted by $\delta = 1/\alpha$.

The absorption coefficient depends on the photon absorption processes occurring in the semiconductor. In the case of **band-to-band (interband) absorption,** α increases rapidly with the photon energy $h\nu$ above E_g as shown for Si ($E_g = 1.1$ eV) and GaAs ($E_g = 1.42$ eV) in Figure 5.37. Notice that α is plotted on a logarithmic scale. The general trend of the α versus $h\nu$ behavior can be intuitively understood from the density of states diagram also shown in the same figure.

Density of states $g(E)$ represents the number of states per unit energy per unit volume. We assume that the VB states are filled and the CB states are empty since the number of electrons in the CB is much smaller than the number of states in this band ($n \ll N_c$). The photon absorption process increases when there are more VB states available as more electrons can be excited. We also need available CB states into which the electrons can be excited, otherwise the electrons cannot find empty states to fill. The probability of photon absorption depends on both the density of VB states *and* the density of CB states. For photons of energy $h\nu_A = E_g$, the absorption can only occur from E_v to E_c where the VB and CB densities of states are low and thus the absorption coefficient is small, which is illustrated as A in Figure 5.37. For photon energies $h\nu_B$, which can take electrons from very roughly the middle region of the VB to the middle of the CB, the densities of states are large and α is also large as indicated by B in Figure 5.37. Furthermore, there are more choices of excitation for the $h\nu_B$ photon as illustrated by the three arrows in the figure. At even higher photon energies,

photon absorption can of course excite electrons from the VB into vacuum. In reality, the density of states $g(E)$ of a real crystalline semiconductor is much more complicated with various sharp peaks and troughs on the density of states function, shown as dashed curves in $g(E)$ in Figure 5.37, particularly away from the band edges. In addition, the absorption process has to satisfy the conservation of momentum and quantum mechanical transition rules which means that certain transitions from the CB to the VB will be more favorable than others. For example, GaAs is a **direct bandgap** semiconductor, so photon absorption can lead directly to the excitation of an electron from the CB to the VB for photon energies just above E_g just as direct recombination of an electron and hole results in photon emission. Si is an **indirect bandgap** semiconductor. Just as direct electron and hole recombination is not possible in silicon, the electron excitation from states near E_v to states near E_c must be accompanied by the emission or absorption of lattice vibrations, and hence the absorption is less efficient; α versus $h\nu$ for GaAs rises more sharply than that for Si above E_g as apparent in Figure 5.37. At sufficiently high photon energies, it is possible to excite electrons directly from the VB to the CB in Si and this gives the sharp rise in α versus $h\nu$ before B in Figure 5.37. (Band-to-band absorption is further discussed in Chapter 9.)

EXAMPLE 5.16 **PHOTOCONDUCTIVITY OF A THIN SLAB** Modify the photoconductivity expression

$$\Delta\sigma = \frac{e\eta I_o \lambda \tau (\mu_e + \mu_h)}{hcD}$$

derived for a direct bandgap semiconductor in Figure 5.28 to take into account that some of the light intensity is transmitted through the material.

SOLUTION

If we assume that all the photons are absorbed (there is no transmitted light intensity), then the photoconductivity expression is

$$\Delta\sigma = \frac{e\eta I_o \lambda \tau (\mu_e + \mu_h)}{hcD}$$

But, in reality, $I_o \exp(-\alpha D)$ is the transmitted intensity through the specimen with thickness D, so absorption is determined by the intensity lost in the material $I_o[1 - \exp(-\alpha D)]$, which means that $\Delta\sigma$ must be accordingly scaled down to

$$\Delta\sigma = \frac{e\eta I_o[1 - \exp(-\alpha D)]\lambda \tau (\mu_e + \mu_h)}{hcD}$$

EXAMPLE 5.17 **PHOTOGENERATION IN GaAs AND THERMALIZATION** Suppose that a GaAs sample is illuminated with a 50 mW HeNe laser beam (wavelength 632.8 nm) on its surface. Calculate how much power is dissipated as heat in the sample during thermalization. Give your answer as mW. The energy bandgap E_g of GaAs is 1.42 eV.

SOLUTION

Suppose P_L is the power in the laser beam; then $P_L = IA$, where I is the intensity of the beam and A is the area of incidence. The photon flux, photons arriving per unit area per unit

time, is

$$\Gamma_{ph} = \frac{I}{h\nu} = \frac{P_L}{Ah\nu}$$

so the number of EHPs generated per unit time is

$$\frac{dN}{dt} = \Gamma_{ph} A = \frac{P_L}{h\nu}$$

These carriers *thermalize*—lose their excess energy as lattice vibrations (heat) via collisions with the lattice—so eventually their average kinetic energy becomes $\frac{3}{2}kT$ above E_g as depicted in Figure 5.35. Remember that we assume that electrons in the CB are nearly free, so they must obey the kinetic theory and hence have an average kinetic energy of $\frac{3}{2}kT$. The average energy of the electron is then $E_g + \frac{3}{2}kT \approx 1.46$ eV. The excess energy

$$\Delta E = h\nu - \left(E_g + \frac{3}{2}kT\right)$$

is lost to the lattice as heat, that is, lattice vibrations. Since each electron loses an amount of energy ΔE as heat, the heat power generated is

$$P_H = \left(\frac{dN}{dt}\right)\Delta E = \left(\frac{P_L}{h\nu}\right)(\Delta E)$$

The incoming photon has an energy $h\nu = hc/\lambda = 1.96$ eV, so

$$P_H = \frac{(50\,\text{mW})(1.96\,\text{eV} - 1.46\,\text{eV})}{1.96\,\text{eV}} = 12.76\,\text{mW}$$

Notice that in this example, and also in Figure 5.35, we have assigned the excess energy $\Delta E = h\nu - E_g - \frac{3}{2}kT$ to the electron rather than share it between the electron and the hole that is photogenerated. This assumption depends on the ratio of the electron and hole effective masses, and hence depends on the semiconductor material. It is approximately true in GaAs because the electron is much lighter than the hole, almost 10 times, and consequently the absorbed photon is able to "impart" a much higher kinetic energy to the electron than to the hole; $h\nu - E_g$ is used in the photogeneration, and the remainder goes to impart kinetic energy to the photogenerated electron hole pair.

5.8 PIEZORESISTIVITY

When a mechanical stress is applied to a semiconductor sample, as shown in Figure 5.38a, it is found that the resistivity of the semiconductor changes by an amount that depends on the stress.[6] **Piezoresistivity** is the change in the resistivity of a semiconductor (indeed, any material), due to an applied stress. **Elastoresistivity** refers to the change in the resistivity due to an induced strain in the substance. Since the application of stress invariably leads to strain, piezoresistivity and elastoresistivity refer to

[6] *Mechanical stress* is defined as the applied force per unit area, $\sigma_m = F/A$, and the resulting strain ε_m is the fractional change in the length of a sample caused by σ_m, $\varepsilon_m = \delta L/L$, where L is the sample length. The two are related through the elastic modulus Y, $\sigma_m = Y\varepsilon_m$. Subscript m is used to distinguish the stress σ_m and strain ε_m from the conductivity σ and permittivity ε.

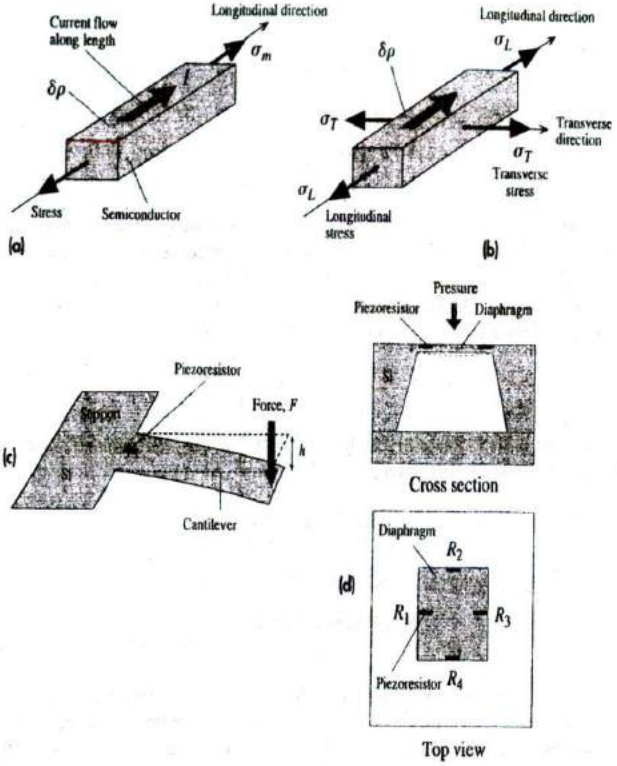


Figure 5.38 Piezoresistivity and its applications.

(a) Stress σ_m along the current (longitudinal) direction changes the resistivity by $\delta\rho$.

(b) Stresses σ_l and σ_t cause a resistivity change.

(c) A force applied to a cantilever bends it. A piezoresistor at the support end (where the stress is large) measures the stress, which is proportional to the force.

(d) A pressure sensor has four piezoresistors R_1, R_2, R_3, R_4 embedded in a diaphragm. The pressure bends the diaphragm, which generates stresses that are sensed by the four piezoresistors.

the same phenomenon. Piezoresistivity is fruitfully utilized in a variety of useful sensor applications such as force, pressure and strain gauges, accelerometers, and microphones.

The change in the resistivity may be due to a change in the concentration of carriers or due to a change in the drift mobility of the carriers, both of which can be modified by a strain in the crystal. Typically, in an extrinsic or doped semiconductor, the concentration of carriers does not change as significantly as the drift mobility; the piezoresistivity is then associated with the change in the mobility. For example, in an n-type Si, the change in the electron mobility μ_e with mechanical strain ε_m, $d\mu_e/d\varepsilon_m$, is of the order of 10^5 cm^2 V^{-1} s^{-1}, so that a strain of 0.015 percent will result in a

change in the mobility that is about 1 percent, and a similar change in the resistivity, which is readily measurable. In this case, the change in the mobility μ_e is due to the induced strain changing the effective mass m_e^* which then modifies μ_e. (Recall that $\mu_e = e\tau/m_e^*$, where τ is the mean scattering time.)

The change in the resistivity $\delta\rho$ has been shown to be proportional to the induced strain in the crystal and hence proportional to the applied stress σ_m. The fractional change $\delta\rho/\rho$ can be written as

$$\frac{\delta\rho}{\rho} = \pi\sigma_m \qquad [5.62]$$

Piezoresistivity

where π is a constant called the **piezoresistive coefficient;** π has the units of 1/stress, *e.g.*, m^2/N or 1/Pa. The piezoresistive coefficient π depends on the type of doping, *p*-or *n*-type; the dopant concentration; the temperature; and the crystallographic direction. A stress along a certain direction in a crystal, for example, along the length of a semiconductor crystal, will change the resistivity not only in the same direction but also in transverse directions. We know from elementary mechanics that a strain in one direction is accompanied by a transverse strain, as implied by the Poisson ratio, so it is not unexpected that a stress in one direction will also modify the resistivity in a transverse direction. Thus, the change in the resistivity of a semiconductor in a "longitudinal" direction, taken as the direction of current flow, is due to stresses in the longitudinal and transverse directions. If σ_L is the stress along a longitudinal direction, the direction of current flow, and σ_T is the stress along a transverse direction, as in Figure 5.38b, then, generally, the fractional change in the resistivity along the current flow direction (longitudinal direction) is given by

$$\frac{\delta\rho}{\rho} = \pi_L\sigma_L + \pi_T\sigma_T \qquad [5.63]$$

Piezoresistivity

where π_L is the piezoresistive coefficient along a longitudinal direction (different for *p*- and *n*-type Si), and π_T is the piezoresistive coefficient in the transverse direction.

The piezoresistive effect is actually more complicated than what we have implied. In reality, we have to consider six types of stresses, three uniaxial stresses along the *x*, *y*, and *z* directions (*e.g.*, trying to pull the crystal along in three independent directions) and three shear stresses (*e.g.*, trying to shear the crystal in three independent ways). In very simple terms, a change in the resistivity $(\delta\rho/\rho)_i$ along a particular direction *i* (an arbitrary direction) can be induced by a stress σ_j along another direction *j* (which may or may not be identical to *i*). The two, $(\delta\rho/\rho)_i$ and σ_j, are then related through a piezoresistivity coefficient denoted by π_{ij}. Consequently, the full description of piezoresistivity involves tensors, and the piezoresistivity coefficients π_{ij} form the elements of this tensor; a treatment beyond the scope of this book. Nonetheless, it is useful to be able to calculate π_L and π_T from various tabulated piezoresistivity coefficients π_{ij}, without having to learn tensors. It turns out that it is sufficient to identify three *principal piezoresistive coefficients* to describe the piezoresistive effect in cubic crystals, which are denoted as π_{11}, π_{12}, and π_{44}. From the latter set we can easily calculate π_L and π_T for a crystallographic direction of interest; the relevant equations can be found in advanced textbooks.

Advances in silicon fabrication technologies and micromachining (ability to fabricate micromechanical structures) have now enabled various piezoresistive silicon microsensors to be developed that have a wide range of useful applications. Figure 5.38c shows a very simple Si microcantilever in which an applied force F to the free end bends the cantilever; the tip of the cantilever is deflected by a distance h. According to elementary mechanics, this deflection induces a maximum stress σ_m that is at the surface, at the support end, of the cantilever. A properly placed piezoresistor at this end can be used to measure this stress σ_m, and hence the deflection or the force. The piezoresistor is implanted by selectively diffusing dopants into the Si cantilever at the support end. Obviously, we need to relate the deflection h of the cantilever tip to the stress σ_m, which is well described in mechanics. In addition, h is proportional to the applied force F through a factor that depends on the elastic modulus and the geometry of the cantilever. Thus, we can measure both the displacement (h) and force (F).

Another useful application is in pressure sensors, which are commercially available. Again, the structure is fabricated from Si. A very thin elastic membrane, called a *diaphragm*, has four piezoresistors embedded, by appropriate dopant diffusion, on its surface as shown in Figure 5.38d. Under pressure, the Si diaphragm deforms elastically, and the stresses that are generated by this deformation cause the resistance of the piezoresistors to change. There are four piezoresistors because the four are connected in a Wheatstone bridge arrangement for better signal detection. The diaphragm area is typically 1 mm × 1 mm, and the thickness is 20 μm. There is no doubt that recent advances in micromachining have made piezoresistivity an important topic for a variety of sensor applications.

EXAMPLE 5.18 **PIEZORESISTIVE STRAIN GAUGE** Suppose that we apply a stress σ_L along the length, taken along the [110] direction, of a p-type silicon crystal sample. We will measure the resistivity along this direction by passing a current along the length and measuring the voltage drop between two fixed points as in Figure 5.38a. The stress σ_L along the length will result in a strain ε_L along the same length given by $\varepsilon_L = \sigma_L/Y$, where Y is the elastic modulus. From Equation 5.63 the change in the resistivity is

$$\frac{\Delta\rho}{\rho} = \pi_L\sigma_L + \pi_T\sigma_T = \pi_L Y\varepsilon_L$$

where we have ignored the presence of any transverse stresses; $\sigma_T \approx 0$. These transverse stresses depend on how the piezoresistor is used, that is, whether it is allowed to contract laterally. If the resistor cannot contract, it must be experiencing a transverse stress. In any event, for the particular direction of interest, [110], the Poisson ratio is very small (less than 0.1), and we can simply neglect any σ_T. Clearly, we can find the strain ε_L from the measurement of $\Delta\rho/\rho$, which is the principle of the strain gauge. The **gauge factor** G of a strain gauge measures the sensitivity of the gauge in terms of the fractional change in the resistance per unit strain,

Semiconductor strain gauge

$$G = \frac{\left(\frac{\Delta R}{R}\right)}{\left(\frac{\Delta L}{L}\right)} \approx \frac{\left(\frac{\Delta\rho}{\rho}\right)}{\varepsilon_L} \approx Y\pi_L$$

where we have assumed that ΔR is dominated by $\Delta\rho$, since the effects from geometric changes in the sample shape can be ignored compared with the piezoresistive effect in semiconductors.

Using typical values for a p-type Si piezoresistor which has a length along [110], $Y \approx 170$ GPa, $\pi_L \approx 72 \times 10^{-11}$ Pa^{-1}, we find $G \approx 122$. This is much greater than $G \approx 1.7$ for metal resistor–based strain gauges. In most metals, the fractional change in the resistance $\Delta R/R$ is due to the geometric effect, the sample becoming elongated and narrower, whereas in semiconductors it is due to the piezoresistive effect.

5.9 SCHOTTKY JUNCTION

5.9.1 SCHOTTKY DIODE

We consider what happens when a metal and an n-type semiconductor are brought into contact. In practice, this process is frequently carried out by the evaporation of a metal onto the surface of a semiconductor crystal in vacuum.

The energy band diagrams for the metal and the semiconductor are shown in Figure 5.39. The work function, denoted as Φ, is the energy difference between the vacuum level and the Fermi level. The vacuum level defines the energy where the electron is free from that particular solid and where the electron has zero KE.

For the metal, the work function Φ_m is the minimum energy required to remove an electron from the solid. In the metal there are electrons at the Fermi level E_{Fm}, but in the

John Bardeen, Walter Schottky, and Walter Brattain. Walter H. Schottky (1886–1976) obtained his PhD from the University of Berlin in 1912. He made many distinct contributions to physical electronics. He invented the screen grid vacuum tube in 1915, and the tetrode vacuum tube in 1919 while at Siemens. The Schottky junction theory was formulated in 1938. He also made distinct contributions to thermal and shot noise in devices. His book *Thermodynamik* was published in 1929 and included an explanation of the Schottky defect (Chapter 1).

1 SOURCE: AIP Emilio Segre Visual Archives, Brattain Collection.

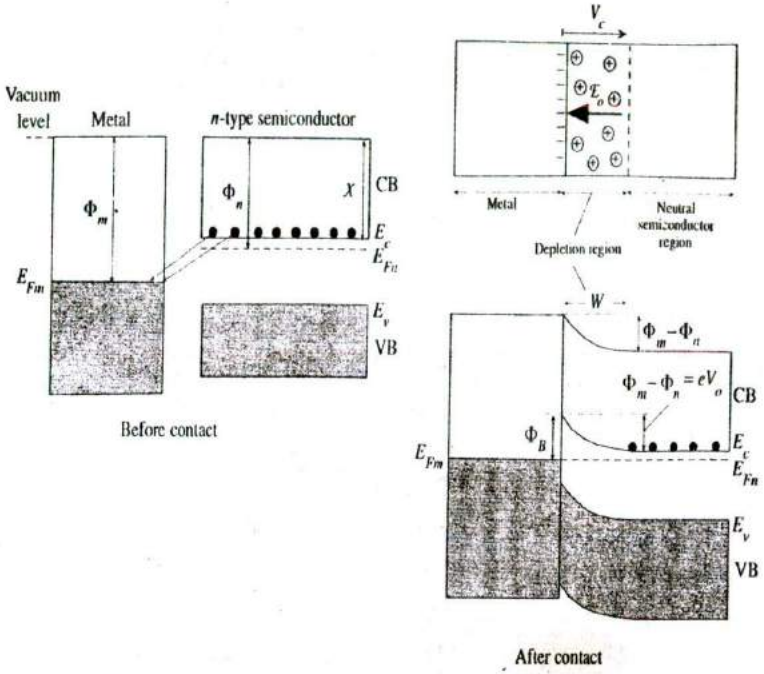


Figure 5.39 Formation of a Schottky junction between a metal and an *n*-type semiconductor when $\Phi_m > \Phi_n$.

semiconductor there are none at E_{Fn}. Nonetheless, the semiconductor work function Φ_n still represents the energy required to remove an electron from the semiconductor. It may be thought that the minimum energy required to remove an electron from the semiconductor is simply the electron affinity χ, but this is not so. Thermal equilibrium requires that only a certain fraction of all the electrons in the semiconductor should be in the CB at a given temperature. When an electron is removed from the conduction band, then thermal equilibrium can be maintained only if an electron is excited from the VB to CB, which involves absorbing heat (energy) from the environment; thus it takes more energy than simply χ. We will not derive the effective thermal energy required to remove an electron but state that, as for a metal, this is equal to Φ_n, even though there are no electrons at E_{Fn}. In fact, the thermionic emission of electrons from a heated semiconductor is also described by the Richardson–Dushman expression in Equation 4.37 but with Φ representing the work function of the semiconductor, Φ_n in the present *n*-type case. (In contrast, the minimum *photon energy* required to remove an electron from a semiconductor above absolute zero would be the electron affinity.)

We assume that $\Phi_m > \Phi_n$, the work function of the metal is greater than that of the semiconductor. When the two solids come into contact, the more energetic electrons in the CB of the semiconductor can readily tunnel into the metal in search of lower empty energy levels (just above E_{Fm}) and accumulate near the surface of the metal, as illustrated in Figure 5.39. Electrons tunneling from the semiconductor leave behind an electron-depleted region of width W in which there are exposed positively charged

donors, in other words, net positive space charge. The contact potential, called the **built-in potential** V_o, therefore develops between the metal and the semiconductor. There is obviously also a **built-in electric field** $\mathcal{E}_o$ from the positive charges to the negative charges on the metal surface. Eventually this built-in potential reaches a value that prevents further accumulation of electrons at the metal surface and an equilibrium is reached. The value of the built-in voltage V_o is the same as that in the metal–metal junction case in Chapter 4, namely, $(\Phi_m - \Phi_n)/e$. The **depletion region** has been depleted of free carriers (electrons) and hence contains the exposed positive donors. This region thus constitutes a **space charge layer** (SCL) in which there is a nonuniform internal field directed from the semiconductor to the metal surface. The maximum value of this built-in field is denoted as $\mathcal{E}_o$ and occurs right at the metal–semiconductor junction (this is where there are a maximum number of field lines from positive to negative charges).

The Fermi level throughout the whole solid, the metal and semiconductor in contact, must be uniform in equilibrium. Otherwise, a change in the Fermi level ΔE_F going from one end to the other end will be available to do external (electrical) work. Thus, E_{Fm} and E_{Fn} line up. The W region, however, has been depleted of electrons, so in this region $E_c - E_{Fn}$ must increase so that n decreases. The bands must bend to increase $E_c - E_{Fn}$ toward the junction, as depicted in Figure 5.39. Far away from the junction, we, of course, still have an n-type semiconductor. The bending is just enough for the vacuum level to be continuous and changing by $\Phi_m - \Phi_n$ from the semiconductor to the metal, as this much energy is needed to take an electron across from the semiconductor to the metal. The PE barrier for electrons moving from the metal to the semiconductor is called the **Schottky barrier height** Φ_B, which is given by

$$\Phi_B = \Phi_m - \chi = eV_o + (E_c - E_{Fn}) \qquad [5.64]$$

which is greater than eV_o.

Under open circuit conditions, there is no net current flowing through the metal–semiconductor junction. The number of electrons thermally emitted over the PE barrier Φ_B from the metal to the semiconductor is equal to the number of electrons thermally emitted over eV_o from the semiconductor to the metal. Emission probability depends on the PE barrier for emission through the Boltzmann factor. There are two current components due to electrons flowing through the junction. The current due to electrons being thermally emitted from the metal to the CB of the semiconductor is

$$J_1 = C_1 \exp\left(-\frac{\Phi_B}{kT}\right) \qquad [5.65]$$

where C_1 is some constant, whereas the current due to electrons being thermally emitted from the CB of the semiconductor to the metal is

$$J_2 = C_2 \exp\left(-\frac{eV_o}{kT}\right) \qquad [5.66]$$

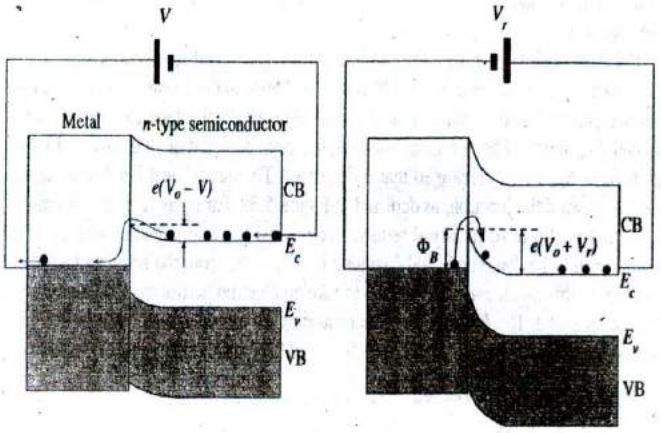
where C_2 is some constant different than C_1.

In equilibrium, that is, open circuit conditions in the dark, the currents are equal but in the reverse directions:

$$J_{\text{open circuit}} = J_2 - J_1 = 0$$

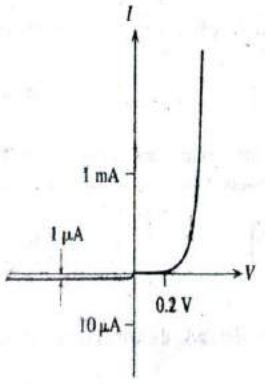
Under forward bias conditions, the semiconductor side is connected to the negative terminal, as depicted schematically in Figure 5.40a. Since the depletion region W has a much larger resistance than the neutral n-region (outside W) and the metal side, nearly all the voltage drop is across the depletion region. The applied bias is in the opposite direction to the built-in voltage V_o. Thus V_o is reduced to $V_o - V$. Φ_B remains unchanged. The semiconductor band diagram outside the depletion region has been effectively shifted up with respect to the metal side by an amount eV because

$$PE = \text{Charge} \times \text{Voltage}$$



(a) Forward-biased Schottky junction. Electrons in the CB of the semiconductor can easily overcome the small *PE* barrier to enter the metal.

(b) Reverse-biased Schottky junction. Electrons in the metal cannot easily overcome the *PE* barrier Φ_B to enter the semiconductor.



(c) I–V characteristics of a Schottky junction exhibits rectifying properties (negative current axis is in microamps).

Figure 5.40 The Schottky junction

The charge is negative but so is the voltage connected to the semiconductor, as shown in Figure 5.40a.

The *PE* barrier for thermal emission of electrons from the semiconductor to the metal is now $e(V_o - V)$. The electrons in the CB can now readily overcome the *PE* barrier to the metal.

The current J_2^{for}, due to the electron emission from the semiconductor to the metal, is now

$$J_2^{for} = C_2 \exp\left[-\frac{e(V_o - V)}{kT}\right] \qquad [5.67]$$

Since Φ_B is the same, J_1 remains unchanged. The net current is then

$$J = J_2^{for} - J_1 = C_2 \exp\left[-\frac{e(V_o - V)}{kT}\right] - C_2 \exp\left(-\frac{eV_o}{kT}\right)$$

or

$$J = C_2 \exp\left(-\frac{eV_o}{kT}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

giving

$$J = J_o\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \qquad [5.68]$$

Schottky junction

where J_o is a constant that depends on the material and surface properties of the two solids. In fact, examination of the above steps shows that J_o is also J_1 in Equation 5.65.

When the Schottky junction is reverse biased, then the positive terminal is connected to the semiconductor, as illustrated in Figure 5.40b. The applied voltage V_r drops across the depletion region since this region has very few carriers and is highly resistive. The built-in voltage V_o thus increases to $V_o + V_r$. Effectively, the semiconductor band diagram is shifted down with respect to the metal side because the charge is negative but the voltage is positive and *PE* = Charge × Voltage. The *PE* barrier for thermal emission of electrons from the CB to the metal becomes $e(V_o + V_r)$, which means that the corresponding current component becomes

$$J_2^{rev} = C_2 \exp\left[-\frac{e(V_o + V_r)}{kT}\right] \ll J_1 \qquad [5.69]$$

Since generally V_o is typically a fraction of a volt and the reverse bias is more than a few volts, $J_2^{rev} \ll J_1$ and the reverse bias current is essentially limited by J_1 only and is very small. Thus, under reverse bias conditions, the current is primarily due to the thermal emission of electrons over the barrier Φ_B from the metal to the CB of the semiconductor as determined by Equation 5.65. Figure 5.40c illustrates the *I–V* characteristics of a typical Schottky junction. The *I–V* characteristics exhibit rectifying properties, and the device is called a **Schottky diode.**

Equation 5.68, which is derived for forward bias conditions, is also valid under reverse bias by making V negative, that is, $V = -V_r$. Furthermore, it turns out to be

applicable not only to Schottky-type metal–semiconductor junctions but also to junctions between a *p*-type and an *n*-type semiconductor, *pn* junctions, as we will show in Chapter 6. Under a forward bias V_f, which is greater than 25 mV at room temperature, the forward current is simply

Schottky junction forward bias

$$J_f = J_o \exp\left(\frac{eV_f}{kT}\right) \qquad V_f > \frac{kT}{e} \tag{5.70}$$

It should be mentioned that it is also possible to obtain a Schottky junction between a metal and a *p*-type semiconductor. This arises when $\Phi_m < \Phi_p$, where Φ_p is the work function for the *p*-type semiconductor.

5.9.2 SCHOTTKY JUNCTION SOLAR CELL

The built-in field in the depletion region of the Schottky junction allows this type of device to function as a photovoltaic device and also as a photodetector. We consider a Schottky device that has a thin metal film (usually Au) deposited onto an *n*-type semiconductor. The energy band diagram is shown in Figure 5.41. The metal is sufficiently thin (~10 nm) to allow light to reach the semiconductor.

For photon energies greater than E_g, EHPs are generated in the depletion region in the semiconductor, as indicated in Figure 5.41. The field in this region separates the EHPs and drifts the electrons toward the semiconductor and holes toward the metal. When an electron reaches the neutral *n*-region, there is now one extra electron there and therefore an additional negative charge. This end therefore becomes more negative with respect to the situation in the dark or the equilibrium situation. When a hole reaches the metal, it recombines with an electron and reduces the effective charge there by one electron, thus making it more positive relative to its dark state. Under open circuit conditions, therefore, a voltage develops across the Schottky junction device with the metal end positive and semiconductor end negative.

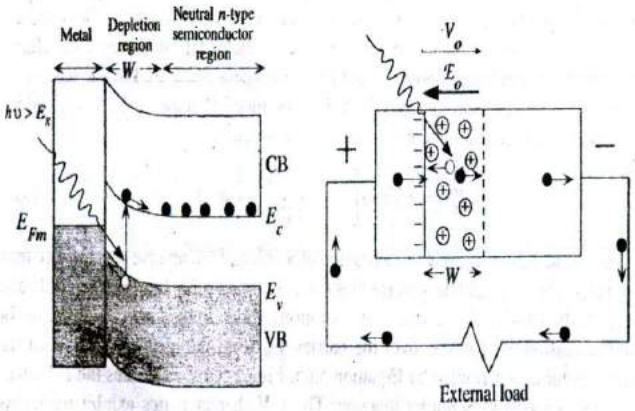


Figure 5.41 The principle of the Schottky junction solar cell.

The photovoltaic explanation in terms of the energy band diagram is simple. At the point of photogeneration, the electron finds itself at a *PE* slope as E_c is decreasing toward the semiconductor, as shown in Figure 5.41. It has no option but to roll down the slope just as a ball that is let go on a slope would roll down the slope to decrease its gravitational *PE*. Recall that there are many more empty states in the CB than electrons, so there is nothing to prevent the electron from rolling down the CB in search of lower energy. When the electron reaches the neutral region (flat E_c region), it upsets the equilibrium there. There is now an additional electron in the CB and this side acquires a negative charge. If we remember that hole energy increases downward on the energy band diagram, then similar arguments also apply to the photogenerated hole in the VB, which rolls down its own *PE* slope to reach the surface of the metal and recombine with an electron there.

If the device is connected to an external load, then the extra electron in the neutral *n*-region is conducted through the external leads, through the load, toward the metal side, where it replenishes the lost electron in the metal. As long as photons are generating EHPs, the flow of electrons around the external circuit will continue and there will be photon energy to electrical energy conversion. Sometimes it is useful to think of the neutral *n*-type semiconductor region as a "conductor," an extension of the external wire (except that the *n*-type semiconductor has a higher resistivity). As soon as the photogenerated electron crosses the depletion region, it reaches a conductor and is conducted around the external circuit to the metal side to replenish the lost electron there.

For photon energies less than E_g, the device can still respond, providing that the $h\nu$ can excite an electron from E_{Fm} in the metal over the *PE* barrier Φ_B into the CB, from where the electron will roll down toward the neutral *n*-region. In this case, $h\nu$ must only be greater than Φ_B.

If the Schottky junction diode is reverse-biased, as shown in Figure 5.42, then the reverse bias V_r increases the built-in potential V_o to $V_o + V_r$ ($V_r \gg V_o$). The internal field increases to substantially high values. This has the advantage of increasing the drift velocity of the EHPs ($v_d = \mu_d \mathcal{E}$) in the depletion region and therefore

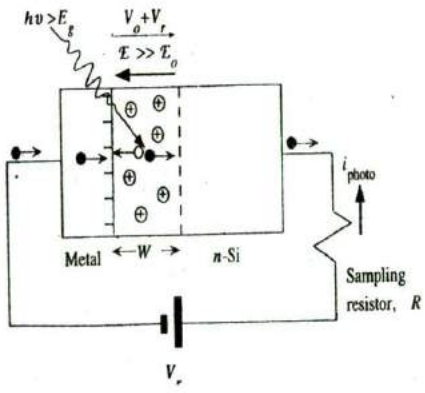


Figure 5.42 Reverse-biased Schottky photodiodes are frequently used as fast photodetectors.

shortening the transit time required to cross the depletion width. The device responds faster and is useful as a fast photodetector. The photocurrent i_{photo} in the external circuit is due to the drift of photogenerated carriers in the depletion region and can be readily measured.

EXAMPLE 5.19 **THE SCHOTTKY DIODE** The reverse saturation current J_o in the Schottky junction, as expressed in Equation 5.68, is the same current that is given by the Richardson–Dushman equation for thermionic emission over a potential barrier $\Phi (= \Phi_B)$ derived in Chapter 4. J_o is given by

Reverse saturation current in Schottky junction

$$J_o = B_t T^2 \exp\left(-\frac{\Phi_B}{kT}\right)$$

where B_t is the effective Richardson constant that depends on the characteristics of the metal–semiconductor junction. B_t for metal–semiconductor junctions, among other factors, depends on the density of states related effective mass of the thermally emitted carriers in the semiconductor. For example, for a metal to n-Si junction, B_t is about 110 A cm^{-2} K^{-2}, and for a metal to p-Si junction, which involves holes, B_t is about 30 A cm^{-2} K^{-2}.

a. Consider a Schottky junction diode between W (tungsten) and n-Si, doped with 10^{16} donors cm^{-3}. The cross-sectional area is 1 mm^2. Given that the electron affinity χ of Si is 4.01 eV and the work function of W is 4.55 eV, what is the theoretical barrier height Φ_B from the metal to the semiconductor?

b. What is the built-in voltage V_o with no applied bias?

c. Given that the experimental barrier height Φ_B is about 0.66 eV, what is the reverse saturation current and the current when there is a forward bias of 0.2 V across the diode?

SOLUTION

a. From Figure 5.39, it is clear that the barrier height Φ_B is

$$\Phi_B = \Phi_m - \chi = 4.55\text{ eV} - 4.01\text{ eV} = 0.54\text{ eV}$$

The experimental value is around 0.66 eV, which is greater than the theoretical value due to various effects at the metal–semiconductor interface arising from dangling bonds, defects, and so forth. For example, dangling bonds give rise to what are called *surface states* within the bandgap of the semiconductor that can capture electrons and modify the Schottky energy band diagram. (The energy band diagram in Figure 5.39 represents an ideal junction with no surface states.) Further, in some cases, such as Pt on n-Si, the experimental value can be lower than the theoretical value.

b. We can find $E_c - E_{Fn}$ in Figure 5.39 from

$$n = N_d = N_c \exp\left(-\frac{E_c - E_{Fn}}{kT}\right)$$

$$10^{16}\text{ cm}^{-3} = (2.8 \times 10^{19}\text{ cm}^{-3}) \exp\left(-\frac{E_c - E_{Fn}}{0.026\text{ eV}}\right)$$

which gives $\Delta E = E_c - E_{Fn} = 0.206$ eV. Thus, the built-in potential V_o can be found from

$$V_o = \frac{\Phi_B}{e} - \frac{E_c - E_{Fn}}{e} = 0.54\text{ V} - 0.206\text{ V} = 0.33\text{ V}$$

c. If A is the cross-sectional area, 0.01 cm^2, taking B_e to be 110 A K^{-2} cm^{-2}, and using the experimental value for the barrier height Φ_B, the saturation current is

$$I_o = AB_eT^2 \exp\left(-\frac{\Phi_B}{kT}\right) = (0.01)(110)(300^2)\exp\left(-\frac{0.66\text{ eV}}{0.026\text{ eV}}\right)$$

$$= 9.36 \times 10^{-7}\text{ A} \quad \text{or} \quad 0.94\,\mu\text{A}$$

When the applied voltage is V_f, the forward current I_f is

$$I_f = I_o\left[\exp\left(\frac{V_f}{kT}\right) - 1\right] = (0.94\,\mu\text{A})\left[\exp\left(\frac{0.2}{0.026}\right) - 1\right] = 2.0\text{ mA}$$

5.10 OHMIC CONTACTS AND THERMOELECTRIC COOLERS

An **ohmic contact** is a junction between a metal and a semiconductor that does not limit the current flow. The current is essentially limited by the resistance of the semiconductor outside the contact region rather than the thermal emission rate of carriers across a potential barrier at the contact. In the Schottky diode, the I–V characteristics were determined by the thermal emission rate of carriers across the contact. It should be mentioned that, contrary to intuition, when we talk about an ohmic contact, we do not generally infer a linear I–V characteristic for the ohmic contact itself. We only imply that the contact does not limit the current flow.

Figure 5.43 shows the formation of an ohmic contact between a metal and an n-type semiconductor. The work function of the metal Φ_m is smaller than the work function Φ_n of the semiconductor. There are more energetic electrons in the metal than

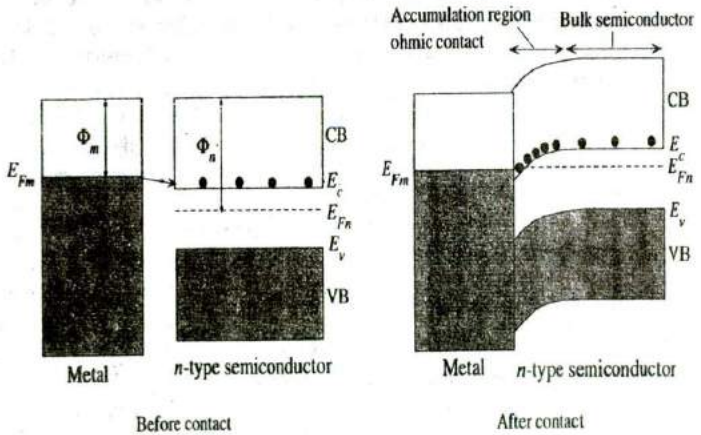


Figure 5.43 When a metal with a smaller work function than an n-type semiconductor is put into contact with the n-type semiconductor, the resulting junction is an ohmic contact in the sense that it does not limit the current flow.

in the CB, which means that the electrons (around E_{Fm}) tunnel into the semiconductor in search of lower energy levels, which they find around E_c, as indicated in Figure 5.43. Consequently, many electrons pile in the CB of the semiconductor near the junction. Equilibrium is reached when the accumulated electrons in the CB of the semiconductor prevent further electrons tunneling from the metal. Put more rigorously, equilibrium is reached when the Fermi level is uniform across the whole system from one end to the other.

The semiconductor region near the junction in which there are excess electrons is called the **accumulation region.** To show the increase in n, we draw the semiconductor energy bands bending downward to decrease $E_c - E_{Fn}$, which increases n. Going from the far end of the metal to the far end of the semiconductor, there are always conduction electrons. In sharp contrast, the depletion region of the Schottky junction separates the conduction electrons in the metal from those in the semiconductor. It can be seen from the contact in Figure 5.43 that the conduction electrons immediately on either side of the junction (at E_{Fm} and E_c) have about the same energy and therefore there is no barrier involved when they cross the junction in either direction under the influence of an applied field.

It is clear that the excess electrons in the accumulation region increase the conductivity of the semiconductor in this region. When a voltage is applied to the structure, the voltage drops across the higher resistance region, which is the bulk semiconductor region. Both the metal and the accumulation region have comparatively high concentrations of electrons compared with the bulk of the semiconductor. The current is therefore determined by the resistance of the bulk region. The current density is then simply $J = \sigma \mathcal{E}$ where σ is the conductivity of the semiconductor in the bulk and $\mathcal{E}$ is the applied field in this region.

One of the interesting and important applications of semiconductors is in **thermoelectric,** or **Peltier,** devices, which enable small volumes to be cooled by direct currents. Whenever a dc current flows through a contact between two dissimilar materials, heat is either released or absorbed in the contact region, depending on the direction of the current. Suppose that there is a dc current flowing from an n-type semiconductor to a metal through an ohmic contact, as depicted in Figure 5.44a. Then electrons are flowing from the metal to the CB of the semiconductor. We only consider the contact region where the Peltier effect occurs. Current is carried by electrons near the Fermi level E_{Fm} in the metal. These electrons then cross over into the CB of the semiconductor and when they reach the end of the contact region, their energy is E_c plus average KE (which is $\frac{3}{2}kT$). There is therefore an increase in the average energy ($PE + KE$) per electron in the contact region. The electron must therefore absorb heat from the environment (lattice vibrations) to gain this energy as it drifts through the junction. Thus, the passage of an electron from the metal to the CB of an n-type semiconductor involves the absorption of heat at the junction.

When the current direction is from the metal to the n-type semiconductor, the electrons flow from the CB of the semiconductor to the Fermi level of the metal as they pass through the contact. Since E_{Fm} is lower than E_c, the passing electron has to lose energy, which it does to lattice vibrations as heat. Thus, the passage of a CB electron from the n-type semiconductor to the metal involves the release of heat at the junction, as indicated in Figure 5.44b.

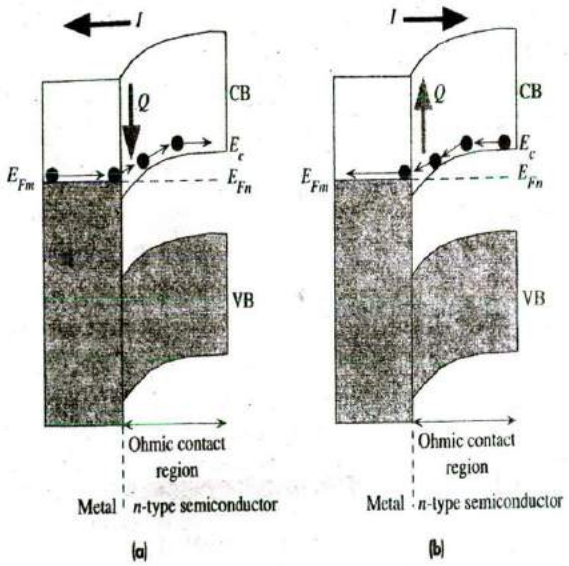


Figure 5.44

(a) Current from an *n*-type semiconductor to the metal results in heat absorption at the junction.

(b) Current from the metal to an *n*-type semiconductor results in heat release at the junction.

It is apparent that depending on the direction of the current flow through a junction between a metal and an *n*-type semiconductor, heat is either absorbed or released at the junction. Although we considered current flow between a metal and an *n*-type semiconductor through an ohmic contact, this thermoelectric effect is a general phenomenon that occurs at a junction between any two dissimilar materials. It is called the **Peltier effect** after its discoverer. In the case of metal–*p*-type semiconductor junctions, heat is absorbed for current flowing from the metal to the *p*-type semiconductor and heat is released in the other direction. Thermoelectric effects occurring at metal–semiconductor junctions are summarized in Figure 5.45. It is important not to confuse the Peltier effect with the Joule heating of the semiconductor and the metal. Joule heating, which we simply call I^2R (or $J^2\rho$) heating, arises from the finite resistivity of the material. It is due to the conduction electrons losing their energy gained from the field to lattice vibrations when they become scattered by such vibrations, as discussed in Chapter 2.

It is self-evident that when a current flows through a semiconductor sample with metal contacts at its ends, as depicted in Figure 5.45, one of the contacts will always absorb heat and the other will always release heat. The contact where heat is absorbed will be cooled and is called the cold junction, whereas the other contact, where heat is released, will warm up and is called the hot junction. One can use the cold junction to

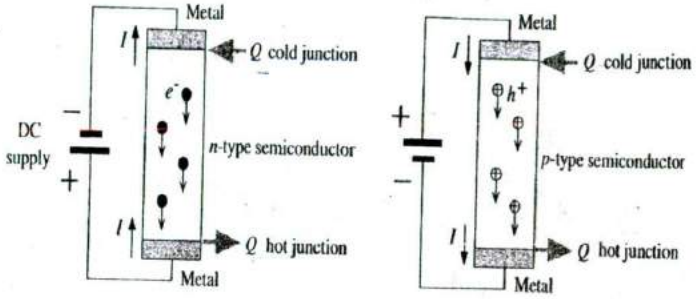


Figure 5.45 When a dc current is passed through a semiconductor to which metal contacts have been made, one junction absorbs heat and cools (the cold junction) and the other releases heat and warms (the hot junction).

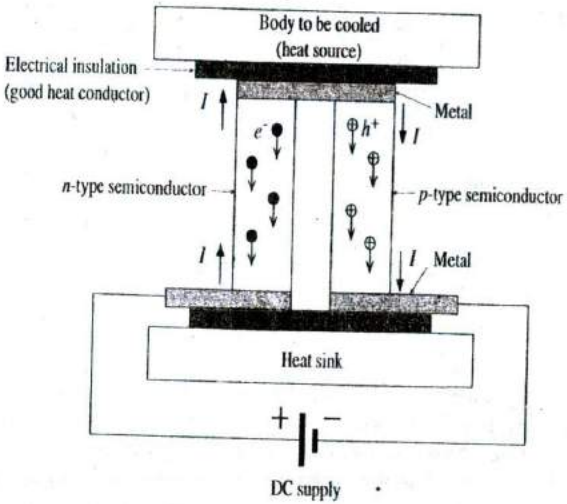


Figure 5.46 Cross section of a typical thermoelectric cooler.

cool another body, providing that the heat generated at the hot junction can be removed from the semiconductor sufficiently quickly to reduce its conduction through the semiconductor to the cold junction. Furthermore, there will always be the Joule heating (I^2R) of the whole semiconductor sample since the bulk will always have a finite resistance.

A simplified schematic diagram of a practical single-element thermoelectric cooling device is shown in Figure 5.46. It uses two semiconductors, one *n*-type and the other *p*-type, each with ohmic contacts. The current direction therefore has opposite thermoelectric effects. On one side, the semiconductors share the same metal

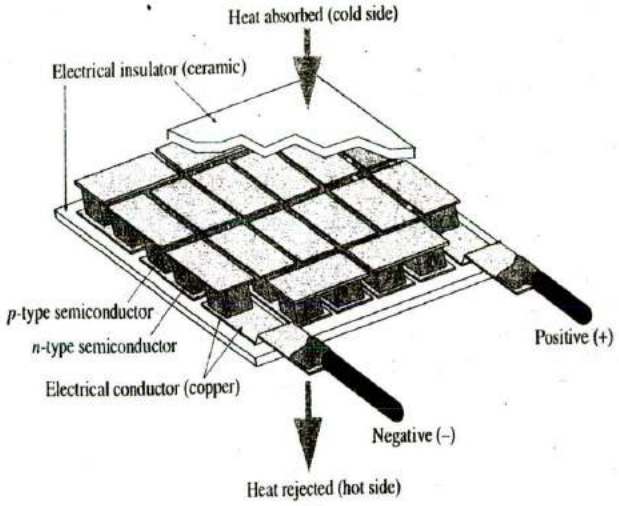


Figure 5.47 Typical structure of a commercial thermoelectric cooler.

electrode. Effectively, the structure is an n-type and a p-type semiconductor connected in series through a common metal electrode. Typically, either Bi_2Te_3, Bi_2Se_3, or Sb_2Te_3 is used as the semiconductor material with copper usually as the metal electrode.

The current flowing through the n-type semiconductor to the common metal electrode causes heat absorption, which cools this junction and hence the metal. The same current then enters the p-type semiconductor and causes heat absorption at this junction, which cools the same metal electrode. Thus the common metal electrode is cooled at both ends. The other ends of the semiconductors are hot junctions. They are connected to a large heat sink to remove the heat and thus prevent heat conduction through the semiconductors toward the cold junctions. The other face of the common metal electrode is in contact, through a thin ceramic plate (electrical insulator but thermal conductor), with the body to be cooled. In commercial Peltier devices, many of these elements are connected in series, as illustrated in Figure 5.47, to increase the cooling efficiency.

EXAMPLE 5.20

THE PELTIER COEFFICIENT Consider the motion of electrons across an ohmic contact between a metal and an n-type semiconductor and hence show that the rate of heat generation Q' at the contact is approximately

$$Q' = \pm \Pi I$$

where Π, called the **Peltier coefficient** between the two materials, is given by

$$\Pi = \frac{1}{e}\left[(E_c - E_{Fn}) + \frac{3}{2}kT\right]$$

where $E_c - E_{Fn}$ is the energy separation of E_c from the Fermi level in the n-type semiconductor. The sign depends on the convention used for heat liberation or absorption.

SOLUTION

We consider Figure 5.44a, which shows only the ohmic contact region between a metal and an n-type semiconductor when a current is passing through it. The majority of the applied voltage drops across the bulk of the semiconductor because the contact region, or the accumulation region, has an accumulation of electrons in the CB. The current is limited by the bulk resistance of the semiconductor. Thus, in the contact region we can take the Fermi level to be almost undisturbed and hence uniform, $E_{Fm} \approx E_{Fn}$. In the bulk of the metal, a conduction electron is at around E_{Fm} (same as E_{Fn}), whereas just at the end of the contact region in the semiconductor it is at E_c plus an average KE of $\frac{3}{2}kT$. The energy difference is the heat absorbed per electron going through the contact region. Since I/e is the rate at which electrons are flowing through the contact,

$$\text{Rate of energy absorption} = \left[\left(E_c + \frac{3}{2}kT\right) - E_{Fm}\right]\left(\frac{I}{e}\right)$$

or

$$Q' = \left[\frac{(E_c - E_{Fn}) + \frac{3}{2}kT}{e}\right] I = \Pi I$$

so the Peltier coefficient is approximately given by the term in the square brackets. A more rigorous analysis gives Π as

$$\Pi = \frac{1}{e}[(E_c - E_{Fn}) + 2kT]$$

ADDITIONAL TOPICS

5.11 DIRECT AND INDIRECT BANDGAP SEMICONDUCTORS

E–k Diagrams We know from quantum mechanics that when the electron is within a potential well of size L, its energy is quantized and given by

$$E_n = \frac{(\hbar k_n)^2}{2m_e}$$

where the wavevector k_n is essentially a quantum number determined by

$$k_n = \frac{n\pi}{L}$$

where $n = 1, 2, 3, \ldots$ The energy increases parabolically with the wavevector k_n. We also know that the electron momentum is given by $\hbar k_n$. This description can be used to represent the behavior of electrons in a metal within which their average

potential energy can be taken to be roughly zero. In other words, we take $V(x) = 0$ within the metal crystal and $V(x)$ to be large [*e.g.*, $V(x) = V_0$] outside so that the electron is contained within the metal. This is the **nearly free electron model** of a metal that has been quite successful in interpreting many of the properties. Indeed, we were able to calculate the density of states $g(E)$ based on the three-dimensional potential well problem. It is quite obvious that this model is too simple since it does not take into account the actual variation of the electron potential energy in the crystal.

The potential energy of the electron depends on its location within the crystal and is periodic due to the regular arrangement of the atoms. How does a periodic potential energy affect the relationship between E and k? It will no longer simply be $E_n = (\hbar k_n)^2/2m_e$.

To find the energy of the electron in a crystal, we need to solve the Schrödinger equation for a periodic potential energy function in three dimensions. We first consider the hypothetical one-dimensional crystal shown in Figure 5.48. The electron potential energy functions for each atom add to give an overall potential energy function $V(x)$, which is clearly periodic in x with the periodicity of the crystal a. Thus,

$$V(x) = V(x + a) = V(x + 2a) = \cdots \qquad [5.71]$$

Periodic potential energy

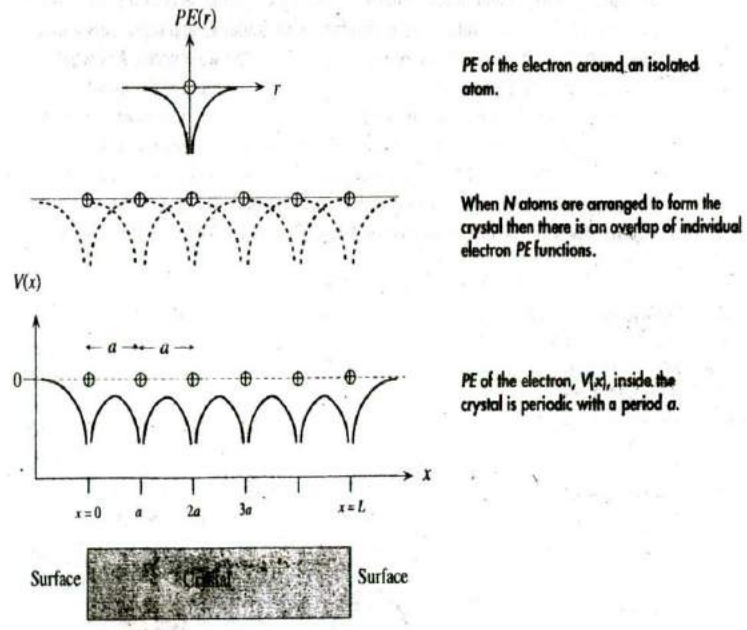


Figure 5.48 The electron potential energy (*PE*), $V(x)$, inside the crystal is periodic with the same periodicity a as that of the crystal. Far away outside the crystal, by choice, $V = 0$ (the electron is free and $PE = 0$).

and so on. Our task is therefore to solve the Schrödinger equation

Schrödinger equation

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}[E - V(x)]\psi = 0 \qquad [5.72]$$

subject to the condition that the potential energy $V(x)$ is periodic in a, that is,

Periodic potential

$$V(x) = V(x + ma) \qquad m = 1, 2, 3, \ldots \qquad [5.73]$$

The solution of Equation 5.72 will give the electron wavefunction in the crystal and hence the electron energy. Since $V(x)$ is periodic, we should expect, by intuition at least, the solution $\psi(x)$ to be periodic. It turns out that the solutions to Equation 5.72, which are called **Bloch wavefunctions,** are of the form

Bloch wavefunction

$$\psi_k(x) = U_k(x)\exp(jkx) \qquad [5.74]$$

where $U_k(x)$ is a periodic function that depends on $V(x)$ and has the same periodicity a as $V(x)$. The term $\exp(jkx)$, of course, represents a traveling wave. We should remember that we have to multiply this by $\exp(-jEt/\hbar)$, where E is the energy, to get the overall wavefunction $\Psi(x, t)$. Thus the electron wavefunction in the crystal is a traveling wave that is modulated by $U_k(x)$.

There are many such Bloch wavefunction solutions to the one-dimensional crystal, each identified with a particular k value, say k_n, which acts as a kind of quantum number. Each $\psi_k(x)$ solution corresponds to a particular k_n and represents a state with an energy E_k. The dependence of the energy E_k on the wavevector k is what we call the E–k diagram. Figure 5.49 shows a typical E–k diagram for the hypothetical one-dimensional solid for k values in the range $-\pi/a$ to $+\pi/a$. Just as $\hbar k$ is the momentum of a free electron, $\hbar k$ for the Bloch electron is the momentum involved in its interaction with external fields, for example, those involved in the photon absorption process. Indeed, the rate of change of $\hbar k$ is the externally applied force F_{ext} on the electron such as that due to an electric field ($F_{ext} = e\mathcal{E}$). Thus, for the electron within

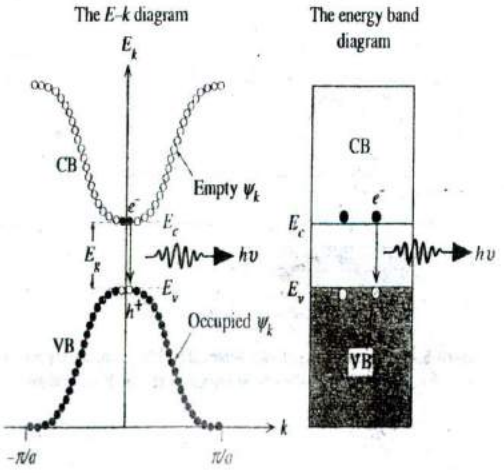


Figure 5.49 The E–k diagram of a direct bandgap semiconductor such as GaAs. The E–k curve consists of many discrete points, each corresponding to a possible state, wavefunction $\psi_k(x)$, that is allowed to exist in the crystal. The points are so close that we normally draw the E–k relationship as a continuous curve. In the energy range E_v to E_c, there are no points [$\psi_k(x)$ solutions].

the crystal,

$$\frac{d(\hbar k)}{dt} = F_{\text{ext}}$$

and consequently we call $\hbar k$ the **crystal momentum** of the electron.[7]

Inasmuch as the momentum of the electron in the x direction in the crystal is given by $\hbar k$, the E–k diagram is an **energy versus crystal momentum plot.** The states $\psi_k(x)$ in the lower E–k curve constitute the wavefunctions for the valence electrons and thus correspond to the states in the VB. Those in the upper E–k curve, on the other hand, correspond to the states in the conduction band (CB) since they have higher energies. All the valence electrons at absolute zero of temperature therefore fill the states, particular k_n values, in the lower E–k diagram.

It should be emphasized that an E–k curve consists of many discrete points, each corresponding to a possible state, wavefunction $\psi_k(x)$, that is allowed to exist in the crystal. The points are so close that we draw the E–k relationship as a continuous curve. It is clear from the E–k diagram that there is a range of energies, from E_v to E_c, for which there are no solutions to the Schrödinger equation and hence there are no $\psi_k(x)$ with energies in E_v to E_c. Furthermore, we also note that the E–k behavior is not a simple parabolic relationship except near the bottom of the CB and the top of the VB.

Above absolute zero of temperature, due to thermal excitation, however, some of the electrons from the top of the valence band will be excited to the bottom of the conduction band. According to the E–k diagram in Figure 5.49, when an electron and hole recombine, the electron simply drops from the bottom of the CB to the top of the VB without any change in its k value, so this transition is quite acceptable in terms of momentum conservation. We should recall that the momentum of the emitted photon is negligible compared with the momentum of the electron. The E–k diagram in Figure 5.49 is therefore for a **direct bandgap semiconductor.**

The simple E–k diagram sketched in Figure 5.49 is for the hypothetical one-dimensional crystal in which each atom simply bonds with two neighbors. In real crystals, we have a three-dimensional arrangement of atoms with $V(x, y, z)$ showing periodicity in more than one direction. The E–k curves are then not as simple as that in Figure 5.49 and often show unusual features. The E–k diagram for GaAs, which is shown in Figure 5.50a, as it turns out, has main features that are quite similar to that sketched in Figure 5.49. GaAs is therefore a direct bandgap semiconductor in which electron–hole pairs can recombine directly and emit a photon. It is quite apparent that light emitting devices use direct bandgap semiconductors to make use of direct recombination.

---

[7] The actual momentum of the electron, however, is not $\hbar k$ because

$$\frac{d(\hbar k)}{dt} \neq F_{\text{external}} + F_{\text{internal}}$$

where $F_{\text{external}} + F_{\text{internal}}$ are all forces acting on the electron. The true momentum p_e satisfies

$$\frac{dp_e}{dt} = F_{\text{external}} + F_{\text{internal}}$$

However, as we are interested in interactions with external forces such as an applied field, we treat $\hbar k$ as if it were the momentum of the electron in the crystal and use the name **crystal momentum.**

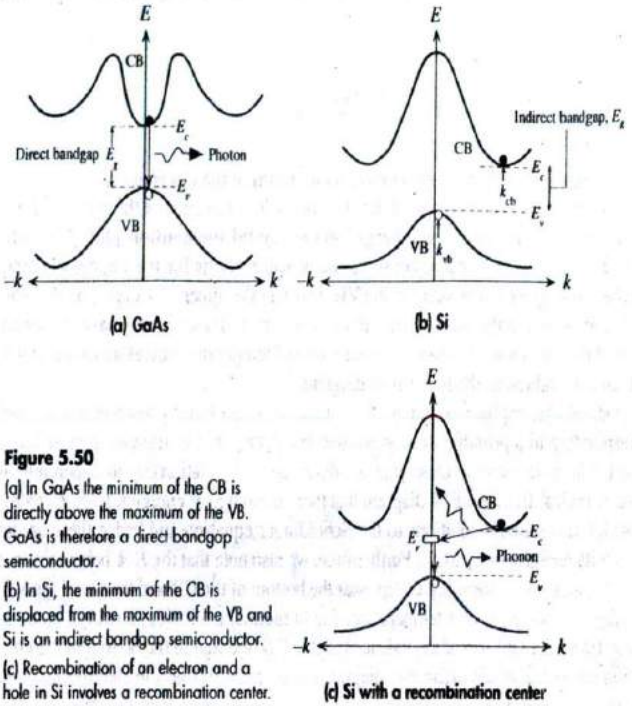


Figure 5.50
(a) In GaAs the minimum of the CB is directly above the maximum of the VB. GaAs is therefore a direct bandgap semiconductor.
(b) In Si, the minimum of the CB is displaced from the maximum of the VB and Si is an indirect bandgap semiconductor.
(c) Recombination of an electron and a hole in Si involves a recombination center.

In the case of Si, the diamond crystal structure leads to an E–k diagram that has the essential features depicted in Figure 5.50b. We notice that the minimum of the CB is not directly above the maximum of the VB. An electron at the bottom of the CB therefore cannot recombine directly with a hole at the top of the VB because, for the electron to fall down to the top of the VB, its momentum must change from k_{cb} to k_{vb}, which is not allowed by the law of conservation of momentum. Thus direct electron–hole recombination does not take place in Si and Ge. The recombination process in these elemental semiconductors occurs via a recombination center at an energy level E_r. The electron is captured by the defect at E_r, from where it can fall down into the top of the VB. The indirect recombination process is illustrated in Figure 5.50c. The energy of the electron is lost by the emission of phonons, that is, lattice vibrations. The E–k diagram in Figure 5.50b for Si is an example of an **indirect bandgap semiconductor.**

In some indirect bandgap semiconductors such as GaP, the recombination of the electron with a hole at certain recombination centers results in photon emission. The E–k diagram is similar to that shown in Figure 5.50c except that the recombination centers at E_r are generated by the purposeful addition of nitrogen impurities to GaP. The electron transition from E_r to E_v involves photon emission.

Electron Motion and Drift We can understand the response of a conduction band electron to an applied external force, for example, an applied field, by examining the E–k diagram. Again, for simplicity, we consider the one-dimensional crystal. The

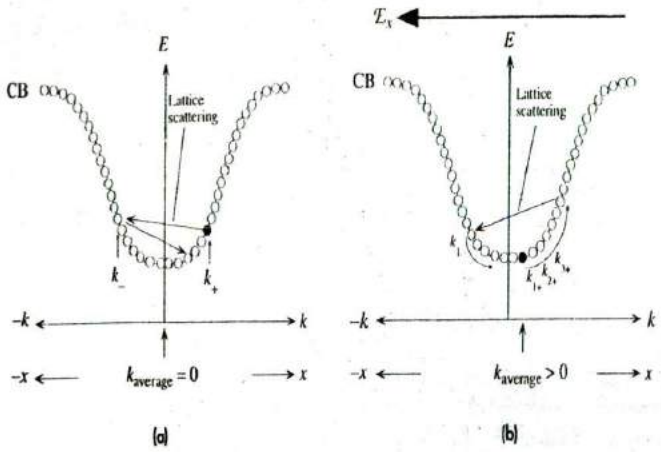


Figure 5.51

(a) In the absence of a field, over a long time, the average of all k values is zero; there is no net momentum in any one particular direction.

(b) In the presence of a field in the $-x$ direction, the electron accelerates in the $+x$ direction increasing its k value along x until it is scattered to a random k value. Over a long time, the average of all k values is along the $+x$ direction. Thus the electron drifts along $+x$.

electron is wandering around the crystal quite randomly due to scattering from lattice vibrations. Thus the electron moves with a certain k value in the $+x$ direction, say k_+, as illustrated in the E–k diagram of Figure 5.51a. When it is scattered by a lattice vibration, its k value changes, perhaps to k_-, which is also shown in Figure 5.51a. This process of k changing randomly from one scattering to another scattering process continues all the time, so over a long time the average value of k is zero; that is, average k_+ is the same as average k_-.

When an electric field is applied, say in the $-x$ direction, then the electron gains momentum in the $+x$ direction from the force of the field $e\mathcal{E}_x$. With time, while the electron is not scattered, it moves up in the E–k diagram from k_{1+} to k_{2+} to k_{3+} and so on until a lattice vibration randomly scatters the electron to say k_{1-} (or to some other random k value) as shown in Figure 5.51b. Over a long time, the average of all k_+ is no longer equal to the average of all k_- and there is a net momentum in the $+x$ direction, which is tantamount to a drift in the same direction.

Effective Mass The usual definition of inertial mass of a particle in classical physics is based on

$$\text{Force} = \text{Mass} \times \text{Acceleration}$$

$$F = ma$$

When we treat the electron as a wave within the semiconductor crystal, we have to determine whether we can still, in some way, use the convenient classical $F = ma$ relation to describe the motion of an electron under an applied force such as $e\mathcal{E}_x$ and, if so, what the apparent mass of the electron in the crystal should be.

We will evaluate the velocity and acceleration of the electron in the CB in response to an electric field $\mathcal{E}_x$ along $-x$ that imposes an external force $F_{ext} = e\mathcal{E}_x$ in the $+x$ direction, as shown in Figure 5.51b. Our treatment will make use of the quantum mechanical E–k diagram.

Since we are treating the electron as a wave, we have to evaluate the group velocity v_g, which, by definition, is $v_g = d\omega/dk$. We know that the time dependence of the wavefunction is $\exp(-jEt/\hbar)$ where the energy $E = \hbar\omega$ (ω is an "angular frequency" associated with the wave motion of the electron). Both E and ω depend on k. Thus, the group velocity is

Electron's group velocity

$$v_g = \frac{1}{\hbar}\frac{dE}{dk} \tag{5.75}$$

Thus the group velocity is determined by the **gradient** of the E–k curve. In the presence of an electric field, the electron experiences a force $F_{ext} = e\mathcal{E}_x$ from which it gains energy and moves up in the E–k diagram until, later on, it collides with a lattice vibration, as shown in Figure 5.51b. During a small time interval δt between collisions, the electron moves a distance $v_g\,\delta t$ and hence gains energy δE, which is

$$\delta E = F_{ext} v_g\,\delta t \tag{5.76}$$

To find the acceleration of the electron and the effective mass, we somehow have to put this equation into a form that looks like $F_{ext} = m_e a$, where a is the acceleration. From Equation 5.76, the relationship between the external force and energy is

$$F_{ext} = \frac{1}{v_g}\frac{dE}{dt} = \hbar\frac{dk}{dt} \tag{5.77}$$

where we used Equation 5.75 for v_g in Equation 5.76. Equation 5.77 is the reason for interpreting $\hbar k$ as the **crystal momentum** inasmuch as the rate of change of $\hbar k$ is the externally applied force.

The acceleration a is defined as dv_g/dt. We can use Equation 5.75,

$$a = \frac{dv_g}{dt} = \frac{d\left[\dfrac{1}{\hbar}\dfrac{dE}{dk}\right]}{dt} = \frac{1}{\hbar}\frac{d^2E}{dk^2}\frac{dk}{dt} \tag{5.78}$$

From Equation 5.78, we can substitute for dk/dt in Equation 5.77, which is then a relationship between F_{ext} and a of the form

External force and acceleration

$$F_{ext} = \frac{\hbar^2}{\left[\dfrac{d^2E}{dk^2}\right]}a \tag{5.79}$$

We know that the response of a free electron to the external force is $F_{ext} = m_e a$, where m_e is its mass in vacuum. Therefore it is quite clear from Equation 5.79 that the **effective mass** of the electron in the crystal is

Effective mass

$$m_e^* = \hbar^2\left[\frac{d^2E}{dk^2}\right]^{-1} \tag{5.80}$$

Thus, the electron responds to an external force and moves as if its mass were given by Equation 5.80. The effective mass obviously depends on the E–k relationship, which in turn depends on the crystal symmetry and the nature of bonding between the atoms. Its value is different for electrons in the CB and for those in the VB, and moreover, it depends on the energy of the electron since it is related to the curvature of the E–k behavior (d^2E/dk^2). Further, it is clear from Equation 5.80 that the effective mass is a quantum mechanical quantity inasmuch as the E–k behavior is a direct consequence of the application of quantum mechanics (the Schrödinger equation) to the electron in the crystal.

It is interesting that, according to Equation 5.80, when the E–k curve is a downward concave as at the top of a band (*e.g.*, Figure 5.49), the effective mass of an electron at these energies in a band is then negative. What does a negative effective mass mean? When the electron moves up on the E–k curve by gaining energy from the field, it actually decelerates, that is, moves more slowly. Its acceleration is therefore in the opposite direction to an electron at the bottom of the band. Electrons in the CB are at the bottom of a band, so their effective masses are positive quantities. At the top of a valence band, however, we have plenty of electrons. These electrons have negative effective masses and under the action of a field, they decelerate. Put differently, they accelerate in the opposite direction to the applied external force F_{ext}. It turns out that we can describe the collective motion of these electrons near the top of a band by considering the motion of a few holes with positive masses.

It should be mentioned that Equation 5.80 defines the meaning of the effective mass in quantum mechanical terms. Its usefulness as a concept lies in the fact that we can measure it experimentally, for example, by cyclotron resonance experiments, and have actual values for it. This means we can simply replace m_e by m_e^* in equations that describe the effect of an external force on electron transport in semiconductors.

Holes To understand the concept of a hole, we consider the E–k curve corresponding to energies in the VB, as shown in Figure 5.52a. If all the states are filled, then there are no empty states for the electrons to move into and consequently an electron cannot gain energy from the field. For each electron moving in the positive x direction with a momentum $\hbar k_+$, there is a corresponding electron with an equal and opposite momentum $\hbar k_-$, so there is no net motion. For example, the electron at b is moving toward the

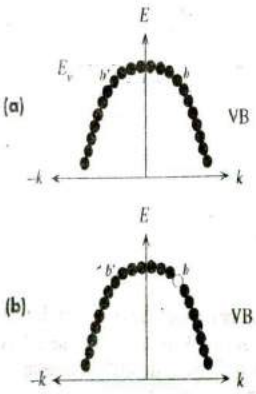


Figure 5.52
(a) In a full valence band, there is no net contribution to the current. There are equal numbers of electrons (e.g., at b and b') with opposite momenta.
(b) If there is an empty state (hole) at b at the top of the band, then the electron at b' contributes to the current.

right with k_{+b}, but its effect is canceled by that at b' moving toward the left with $k_{-b'}$. This cancellation of momenta by electron pairs applies to all the electrons since the VB is assumed to be full. Thus, a full VB cannot contribute to the electric current.

Suppose that one of the states, labeled as b in Figure 5.52b, near the top of the valence band has a missing electron, or a hole, because the electron normally at b has been removed by some means of excitation to the conduction band. It is immediately obvious that the motion of the electron at b' toward the left, that is, $k_{-b'}$, is now *not* canceled, which means that this electron makes a net contribution to the current. We realize that the reason the presence of a hole makes conduction possible is the fact that the momenta of all the VB electrons are canceled except that at b'. It is also clear that in reaching this conclusion, we had to consider all the electrons in the valence band.

Let us maintain strict sign rules so that quantities such as the field ($\mathcal{E}_x$), group velocity (v_g), and acceleration (a) along the $+x$ direction are positive and those along the $-x$ direction are negative. If $\mathcal{E}_x$ is along the $+x$ direction, then the acceleration of a *free* electron from force/mass is $[(-e)(\mathcal{E}_x)]/m_e$, which is negative and along $-x$ as we expect. Similarly, an electron at the bottom of the CB has a positive effective mass and an acceleration that is also negative. Our treatment of conduction in metals by electrons in Chapter 2 inherently assumed that electrons accelerated in the opposite direction to the applied field, that is, positive effective mass.

However, the electrons at the top of the VB have a negative effective mass, which we can write as $-|m_e^*|$. The acceleration a of the electron at b' contributing to the current is

$$a = \frac{-e\mathcal{E}_x}{-|m_e^*|} = \frac{+e\mathcal{E}_x}{+|m_e^*|}$$

which is positive, a along $\mathcal{E}_x$. This means that the acceleration of an electron with a negative effective mass at the top of a VB is equivalent to the acceleration of a positive charge $+e$ with an effective mass $|m_e^*|$. Put differently, we therefore can equivalently describe current conduction by the motion of the hole alone by assigning to it a positive charge and a positive effective mass.

EXAMPLE 5.21 **EFFECTIVE MASS** Show that the effective mass of a free electron is the same as its mass in vacuum.

SOLUTION

The expression for the energy of a free electron is

$$E = \frac{(\hbar k)^2}{2m_e}$$

The effective mass, by definition, is given by

$$m_e^* = \hbar^2 \left[\frac{d^2E}{dk^2}\right]^{-1}$$

Substituting $E = (\hbar k)^2/2m_e$ we get $m_e^* = m_e$. Since the energy of a conduction electron in a metal, within the nearly free electron model, will also have an energy $E = (\hbar k)^2/2m_e$, we can show that the effective mass of the electron in a metal is the same as the mass in vacuum.

EXAMPLE 5.22

CURRENT DUE TO A MISSING ELECTRON IN THE VB First, let us consider a completely full valence band that contains, say, N electrons. $N/2$ of these are moving with momentum in the $+x$, and $N/2$ in the $-x$ direction. Suppose that the crystal is unit volume. An electron with charge $-e$ moving with a group velocity $\mathbf{v}_{gi}$ contributes to the current by an amount $-e\mathbf{v}_{gi}$. We can determine the current density $\mathbf{J}_N$ due to the motion of all the electrons (N of them) in the band,

$$\mathbf{J}_N = -e\sum_{i=1}^{N}\mathbf{v}_{gi} = 0$$

$\mathbf{J}_N$ is zero because for each value of $\mathbf{v}_{gi}$, there is a corresponding velocity equal in magnitude but opposite in direction (b and b' in Figure 5.52a). Our conclusion from this is that the contribution to the current density from a full valence band is nil, as we expect.

Suppose now that the jth electron is missing (b in Figure 5.52b). The net current density is due to $N-1$ electrons in the band, so

$$\mathbf{J}_{N-1} = -e\sum_{i=1, i\neq j}^{N}\mathbf{v}_{gi} \qquad [5.81]$$

where the summation is for $i = 1$ to N and $i \neq j$ (jth electron is missing). We can write the sum as summation to N including the jth electron and minus the missing jth electron contribution,

$$\mathbf{J}_{N-1} = -e\sum_{i=1}^{N}\mathbf{v}_{gi} - (-e\mathbf{v}_{gj})$$

that is,

$$\mathbf{J}_{N-1} = +e\mathbf{v}_{gj} \qquad [5.82]$$

where we used $\mathbf{J}_N = 0$. We see that when there is a missing electron, there is a net current due to that empty state (jth). The current appears as the motion of a charge $+e$ with a velocity $\mathbf{v}_{gj}$, where $\mathbf{v}_{gj}$ is the group velocity of the missing electron. In other words, the current is due to the motion of a positive charge $+e$ at the site of the missing electron at k_j, which is what we call a hole. One should note that Equation 5.81 describes the current by considering the motions of *all* the $N-1$ electrons, whereas Equation 5.82 describes the same current by simply considering the missing electron as if it were a positively charged particle ($+e$) moving with a velocity equal to that of the missing electron. Equation 5.82 is the convenient description universally adopted for a valence band containing missing electrons.

5.12 INDIRECT RECOMBINATION

We consider the recombination of minority carriers in an extrinsic indirect bandgap semiconductor such as Si or Ge. As an example, we consider the recombination of electrons in a p-type semiconductor. In an indirect bandgap semiconductor, the recombination mechanism involves a recombination center, a third body that may be a crystal defect or an impurity, in the recombination process to satisfy the requirements of conservation of momentum. We can view the recombination process as follows. Recombination occurs when an electron is captured by the recombination center at the energy level E_r. As soon as the electron is captured, it will recombine with a hole

because holes are abundant in a p-type semiconductor. In other words, since there are many majority carriers, the limitation on the rate of recombination is the actual capture of the minority carrier by the center. Thus, if τ_r is the electron recombination time, since the electrons will have to be captured by the centers, τ_r is given by

$$\tau_r = \frac{1}{S_r N_r v_{\text{th}}} \qquad [5.83]$$

where S_r is the capture (or recombination) cross section of the center, N_r is the concentration of centers, and v_{th} is the mean speed of the electron that you may take as its effective thermal velocity.

Equation 5.83 is valid under small injection conditions, that is, $p_{p0} \gg n_p$. There is a more general treatment of indirect recombination called the Shockley–Read statistics of indirect recombination and generation, which is treated in more advanced semiconductor physics textbooks. That theory eventually arrives at Equation 5.83 for low-level injection conditions. We derived Equation 5.83 from a purely physical reasoning.

Gold is frequently added to silicon to aid recombination. It is found that the minority carrier recombination time is inversely proportional to the gold concentration, following Equation 5.83.

5.13 AMORPHOUS SEMICONDUCTORS

Up to now we have been dealing with crystalline semiconductors, those crystals that have perfect periodicity and are practically flawless unless purposefully doped for use in device applications. They are used in numerous solid-state devices including large-area solar cells. Today's microprocessor uses a single crystal of silicon that contains millions of transistors; indeed, we are heading for the 1-billion-transistor chip. There are, however, various applications in electronics that require inexpensive large-area devices to be fabricated and hence require a semiconductor material that can be prepared in a large area. In other applications, the semiconductor material is required to be deposited as a film on a flexible substrate for use as a sensor. Best known examples of large-area devices are flat panel displays based on thin-film transistors (TFTs), inexpensive solar cells, photoconductor drums (for printing and photocopying), image sensors, and newly developed X-ray image detectors. Many of these applications typically use hydrogenated amorphous silicon, a-Si:H.

A distinctive property of an electron in a crystalline solid is that its wavefunction is a traveling wave, a Bloch wave, ψ_k, as in Equation 5.74. The Bloch wavefunction is a consequence of the periodicity of an electron's potential energy PE, $V(x)$, within the crystal. One can view the electron's motion as tunneling through the periodic potential energy hills. The wavefunctions ψ_k form **extended states** because they *extend* throughout the whole crystal. The electron belongs to the whole crystal, and there is an equal probability of finding an electron in any unit cell. The wavevector k in this traveling wave ψ_k acts as a quantum number. There are many discrete k_n values, which form a nearly continuous set of k values (see Figure 5.49). We can describe the interaction of the electron with an external force, or with photons and phonons, by assigning a momentum $\hbar k$ to the electron, which is called the electron's crystal momentum.

The electron's wavefunction ψ_k is frequently scattered by lattice vibrations (or by defects or impurities) from one k-value to another, *e.g.*, from ψ_k to $\psi_{k'}$. The scattering of the wavefunction imposes a mean free path ℓ on the electron's motion, that is, a mean distance over which a wave can travel without being scattering. Over the distance ℓ, the wavefunction is coherent, that is, well defined and predictable as a traveling Bloch wave; ℓ is also known as the coherence length of the wavefunction. The mobility is determined by the mean free path ℓ, which at room temperature is typically of the order of several hundreds of mean interatomic separations. The crystal periodicity and the unit cell atomic structure control the types of Bloch wave solutions one can obtain to the Schrödinger equation. The solutions allow the electron energy E to be examined as a function of k (or momentum $\hbar k$) and these $E-k$ diagrams categorize crystalline semiconductors into two classes: direct bandgap (GaAs type) and indirect bandgap (Si type) semiconductors.

Hydrogenated amorphous silicon (a-Si:H) is the noncrystalline form of silicon in which the structure has no long-range order but only short-range order; that is, we can only identify the nearest neighbors of a given atom. Each Si atom has four neighbors as in the crystal, but there is no periodicity or long-range order as illustrated in Figure 1.59. Without the hydrogen, pure a-Si would have dangling bonds. In such a structure sometimes a Si atom would not be able to find a fourth neighboring Si atom to bond with and will be left with a dangling bond as in Figure 1.59b. The hydrogen in the structure (~10 percent) passivates (*i.e.*, neutralizes) the unsatisfied ("dangling") bonds inherent in a noncrystalline structure and so reduces the density of dangling bonds or defects. a-Si:H belongs to a class of solids called **amorphous semiconductors** that do not follow typical crystalline concepts such as Bloch wavefunctions. First, due to the lack of periodicity, we cannot describe the electron as a Bloch wave. Consequently, we cannot use a wavevector k, and hence $\hbar k$, to describe the electron's motion. These semiconductors however do have a short-range order and also possess an energy bandgap that separates a conduction band and a valence band. A window glass has a noncrystalline structure but also has a bandgap, which makes it transparent. Photons with energies less than the bandgap energy can pass through the window glass.

The examination of the structure of a-Si:H in Figure 1.59c should make it apparent that the potential energy $V(x)$ of the electron in this noncrystalline structure fluctuates randomly from site to site. In some cases, the local changes in $V(x)$ can be quite strong, forming effective local *PE* wells (obviously finite wells). Such fluctuations in the *PE* within the solid can capture or trap electrons, that is, localize electrons at certain spatial locations. A localized electron will have a wavefunction that resembles the wavefunction in the hydrogen atom, so the probability of finding the electron is localized to the site. Such locations that can trap electrons, give them localized wavefunctions, are called **localized states.** The amorphous structure also has electrons that possess extended wavefunctions; that is, they belong to the whole solid. These extended wavefunctions are distinctly different than those in the crystal because they have very short coherence lengths due to the random potential fluctuations; the electron is scattered from site to site and hence the mean free path is of the order of a few atomic spacings. The extended wavefunction has random phase fluctuations. Figure 5.53 compares localized and extended wavefunctions in an amorphous semiconductor.

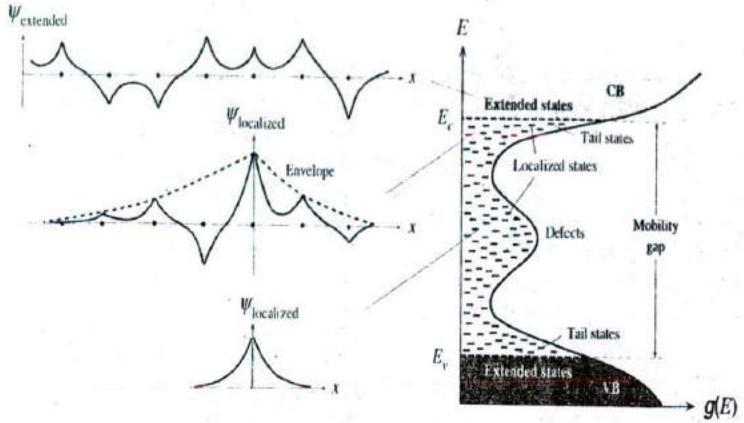


Figure 5.53 Schematic representation of the density of states $g(E)$ versus energy E for an amorphous semiconductor and the associated electron wavefunctions for an electron in the extended and localized states.

Electronic properties of all amorphous semiconductors can be explained in terms of the energy distribution of their density of states (DOS) function, $g(E)$. The DOS function has well-defined energies E_v and E_c that separate extended states from localized states as in Figure 5.53. There is a distribution of localized states, called **tail states** below E_c and above E_v. The usual **bandgap** $E_c - E_v$ is called the **mobility gap.** The reason is that there is a change in the character of charge transport, and hence in the carrier mobility, in going from extended states above E_c to localized states below E_c.

Electron transport above E_c in the conduction band is dominated by scattering from random potential fluctuations arising from the disordered nature of the structure. The electrons are scattered so frequently that their effective mobility is much less than what it is in crystalline Si: μ_e in a-Si:H is typically 5–10 $cm^2\ V^{-1}\ s^{-1}$ whereas it is 1400 $cm^2\ V^{-1}\ s^{-1}$ in a single crystal Si. Electron transport below E_c, on the other hand, requires an electron to jump, or hop, from one localized state to another, aided by thermal vibrations of the lattice, in an analogous way to the diffusion of an interstitial impurity in a crystal. We know from Chapter 1 that the jump or diffusion of the impurity is a thermally activated process because it relies on the thermal vibrations of all the crystal atoms to occasionally give the impurity enough energy to make that jump. The electron's mobility associated with this type of hopping motion among localized states is thermally activated, and its value is small. Thus, there is a change in the electron mobility across E_c, which is called the conduction band **mobility edge.**

The localized states (frequently simply called *traps*) between E_v and E_c have a profound effect on the overall electronic properties. The tail localized states are a direct result of the structural disorder that is inherent in noncrystalline solids, variations in the bond angles and length. Various prominent peaks and features in the DOS within the mobility gap have been associated with possible structural defects, such as under- and overcoordinated atoms in the structure, dangling bonds, and dopants. Electrons that drift in the conduction band can fall into localized states and become immobilized (trapped) for a while. Thus, electron transport in a-Si:H occurs by multiple trapping in

shallow localized states. The effective electron drift mobility in a-Si:H is therefore reduced to $\sim 1\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$. Low drift mobilities obviously prevent the use of amorphous semiconductor materials in high-speed or high-gain electronic applications. Nonetheless, low-speed electronics is just as important as high-speed electronics in the electronics market in such applications as flat panel displays, solar cells, and image sensors. A low-speed flat panel display made from hydrogenated amorphous silicon (a-Si:H) TFTs costs very roughly the same as a high-speed crystalline Si microchip that runs the CPU.

CD Selected Topics and Solved Problems

Selected Topics

Hall Effect in Semiconductors
Transferred Electron Devices: Gunn Effect
Elements of Photoconductivity
Thermoelectric Effects in Semiconductors: Voltage Drift in Semiconductor Devices

Solved Problems

Piezoresistance: Pressure Sensors and Strain Gauges
Hall Effect
Ionization Region in Doped Semiconductors
Compensation Doping of Semiconductors
Electron-Hole Recombination in Semiconductors and Photoconductivity

DEFINING TERMS

Acceptor atoms are dopants that have one less valency than the host atom. They therefore accept electrons from the VB and thereby create holes in the VB, which leads to a $p > n$ and hence to a p-type semiconductor.

Average energy of an electron in the CB is $\frac{3}{2}kT$ as if the electrons were obeying Maxwell–Boltzmann statistics. This is only true for a nondegenerate semiconductor.

Bloch wave refers to an electron wavefunction of the form $\psi_k = U_k(x)\exp(jkx)$, which is a traveling wave that is modulated by a function $U_k(x)$ that has the periodicity of the crystal. The Bloch wavefunction is a consequence of the periodicity of an electron's potential energy within the crystal.

Compensated semiconductor contains both donors and acceptors in the same crystal region that compensate for each other's effects. For example, if there are more donors than acceptors, $N_d > N_a$, then some of the electrons released by donors are captured by acceptors and the net effect is that $N_d - N_a$ number of electrons per unit volume are left in the CB.

Conduction band (CB) is a band of energies for the electron in a semiconductor where it can gain energy from an applied field and drift and thereby contribute to electrical conduction. The electron in the CB behaves as if it were a "free" particle with an effective mass m_e^*.

Degenerate semiconductor has so many dopants that the electron concentration in the CB, or hole concentration in the VB, is comparable with the density of states in the band. Consequently, the Pauli exclusion principle is significant and Fermi–Dirac statistics must be used. The Fermi level is either in the CB for a n^+-type degenerate or in the VB for a p^+-type degenerate semiconductor. The superscript + indicates a heavily doped semiconductor.

Diffusion is a random process by which particles move from high-concentration regions to low-concentration regions.

Donor atoms are dopants that have a valency one more than the host atom. They therefore donate electrons to the CB and thereby create electrons in the CB, which leads to $n > p$ and hence to an n-type semiconductor.

Effective density of states (N_c) at the CB edge is a quantity that represents all the states in the CB per unit volume as if they were all at E_c. Similarly, N_v at the

VB edge is quantity that represents all the states in the VB per unit volume as if they were all at E_v.

Effective mass (m_c^*) of an electron is a quantum mechanical quantity that behaves like the inertial mass in classical mechanics, $F = ma$, in that it measures the object's inertial resistance to acceleration. It relates the acceleration a of an electron in a crystal to the applied external force F_{ext} by $F_{ext} = m_c^* a$. The external force is most commonly the force of an electric field $e\mathcal{E}$ and excludes all internal forces within the crystal.

Einstein relation relates the diffusion coefficient D and the drift mobility μ of a given species of charge carriers through $(D/\mu) = (kT/e)$.

Electron affinity (χ) is the energy required to remove an electron from E_c to the vacuum level.

Energy of the electron in the crystal, whether in the CB or VB, depends on its momentum $\hbar k$ through the E–k behavior determined by the Schrödinger equation. E–k behavior is most conveniently represented graphically through E–k diagrams. For example, for an electron at the bottom of the CB, E increases as $(\hbar k)^2/m_c^*$ where $\hbar k$ is the momentum and m_c^* is the effective mass of the electron, which is determined from the E–k behavior.

Excess carrier concentration is the excess concentration *above* the thermal equilibrium value. Excess carriers are generated by an external excitation such as photogeneration.

Extended state refers to an electron wavefunction ψ_k whose magnitude does not decay with distance; that is, it is extended in the crystal. An extended wavefunction of an electron in a *crystal* is a **Bloch wave,** that is, $\psi_k = U_k(x)\exp(jkx)$, which is a traveling wave that is modulated by a function $U_k(x)$ that has the periodicity of the crystal. There is an equal probability of finding an electron in any unit cell of the crystal. Scattering of an electron in the crystal by lattice vibrations or impurities, etc., corresponds to the electron being scattered from one ψ_k to another $\psi_{k'}$, *i.e.* a change in the wavevector from $\mathbf{k}$ to $\mathbf{k}'$. Valence and conduction bands in a crystal have extended states.

Extrinsic semiconductor is a semiconductor that has been doped so that the concentration of one type of charge carrier far exceeds that of the other. Adding donor impurities releases electrons into the CB and n far exceeds p; thus, the semiconductor becomes n-type.

Fermi energy or level (E_F) may be defined in several equivalent ways. The Fermi level is the energy level corresponding to the energy required to remove an electron from the semiconductor; there need not be any actual electrons at this energy level. The energy needed to remove an electron defines the work function Φ. We can define the Fermi level to be Φ below the vacuum level. E_F can also be defined as that energy value below which all states are full and above which all states are empty at absolute zero of temperature. E_F can also be defined through a difference. A difference in the Fermi energy ΔE_F in a system is the external electrical work done per electron either on the system or by the system such as electrical work done when a charge e moves through an electrostatic PE difference is $e\Delta V$. It can be viewed as a fundamental material property.

Intrinsic carrier concentration (n_i) is the electron concentration in the CB of an intrinsic semiconductor. The hole concentration in the VB is equal to the electron concentration.

Intrinsic semiconductor has an equal number of electrons and holes due to thermal generation across the bandgap E_g. It corresponds to a pure semiconductor crystal in which there are no impurities or crystal defects.

Ionization energy is the energy required to ionize an atom, for example, to remove an electron.

Ionized impurity scattering limited mobility is the mobility of the electrons when their motion is limited by scattering from the ionized impurities in the semiconductor (*e.g.*, donors and acceptors).

k is the wavevector of the electron's wavefunction. In a crystal the electron wavefunction, $\psi_k(x)$ is a *modulated traveling wave* of the form

$$\psi_k(x) = U_k(x)\exp(jkx)$$

where k is the wavevector and $U_k(x)$ is a periodic function that depends on the PE of interaction between the electron and the lattice atoms. k identifies all possible states $\psi_k(x)$ that are allowed to exist in the crystal. $\hbar k$ is called the *crystal momentum* of the electron as its rate of change is the externally applied force to the electron, $d(\hbar k)/dt = F_{external}$.

Lattice-scattering-limited mobility is the mobility of the electrons when their motion is limited by scattering from thermal vibrations of the lattice atoms.

Localized state refers to an electron wavefunction $\psi_{\text{localized}}$ whose magnitude, or the envelope of the wavefunction, decays with distance, which localizes the electron to a spatial region in the semiconductor. For example, a 1*s*-type wavefunction of the form $\psi_{\text{localized}} \propto \exp(-\alpha r)$, where r is the distance measured from some center at $r = 0$, and α is a positive constant, would represent a localized state centered at $r = 0$.

Majority carriers are electrons in an *n*-type and holes in a *p*-type semiconductor.

Mass action law in semiconductor science refers to the law $np = n_i^2$, which is valid under thermal equilibrium conditions and in the absence of external biases and illumination.

Minority carrier diffusion length (L) is the mean distance a minority carrier diffuses before recombination, $L = \sqrt{D\tau}$, where D is the diffusion coefficient and τ is the minority carrier lifetime.

Minority carrier lifetime (τ) is the mean time for a minority carrier to disappear by recombination. $1/\tau$ is the mean probability per unit time that a minority carrier recombines with a majority carrier.

Minority carriers are electrons in a *p*-type and holes in an *n*-type semiconductor.

Nondegenerate semiconductor has electrons in the CB and holes in the VB that obey Boltzmann statistics. Put differently, the electron concentration n in the CB is much less than the effective density of states N_c and similarly $p \ll N_v$. It refers to a semiconductor that has not been heavily doped so that these conditions are maintained; typically, doping concentrations are less than 10^{18} cm^{-3}.

Ohmic contact is a contact that can supply charge carriers to a semiconductor at a rate determined by charge transport through the semiconductor and not by the contact properties itself. Thus the current is limited by the conductivity of the semiconductor and not by the contact.

Peltier effect is the phenomenon of heat absorption or liberation at the contact between two dissimilar materials as a result of a dc current passing through the junction. The rate of heat generation Q' is proportional to the dc current I passing through the contact so that $Q' = +\Pi I$, where Π is called the Peltier coefficient and the sign depends on whether heat is absorbed or released.

Phonon is a quantum of energy associated with the vibrations of the atoms in the crystal, analogous to the photon. A phonon has an energy $\hbar\omega$ where ω is the frequency of the lattice vibration.

Photoconductivity is the change in the conductivity from dark to light, $\sigma_{\text{light}} - \sigma_{\text{dark}}$.

Photogeneration is the excitation of an electron into the CB by the absorption of a photon. If the photon is absorbed by an electron in the VB, then its excitation to the CB will generate an EHP.

Photoinjection is the photogeneration of carriers in the semiconductor by illumination. Photogeneration may be VB to CB excitation, in which case electrons and holes are generated in pairs.

Piezoresistivity is the change in the resistivity of a semiconductor due to an applied mechanical stress σ_m. **Elastoresistivity** refers to the change in the resistivity due to an induced strain in the substance. Application of stress normally leads to strain, so piezoresistivity and elastoresistivity refer to the same phenomenon. In simple terms, the change in the resistivity may be due to a change in the concentration of carriers or due to a change in the drift mobility of the carriers. The fractional change in the resistivity $\delta\rho/\rho$ is proportional to the applied stress σ_m, and the proportionality constant is called the **piezoresistive coefficient** π (1/Pa units), which is a tensor quantity because a stress in one direction in a crystal can alter the resistivity in another direction.

Recombination of an electron–hole pair involves an electron in the CB falling down in energy into an empty state (hole) in the VB to occupy it. The result is the annihilation of an EHP. Recombination is direct when the electron falls directly down into an empty state in the VB as in GaAs. Recombination is indirect if the electron is first captured locally by a defect or an impurity, called a recombination center, and from there it falls down into an empty state (hole) in the VB as in Si and Ge.

Schottky junction is a contact between a metal and a semiconductor that has rectifying properties. For a metal/*n*-type semiconductor junction, electrons on the metal side have to overcome a potential energy barrier Φ_B to enter the conduction band of the semiconductor, whereas the conduction electrons in the semiconductor have to overcome a smaller barrier eV_o to enter the metal. Forward bias decreases eV_o and thereby greatly encourages electron emissions over the barrier $e(V_o - V)$. Under reverse bias, electrons have to overcome Φ_B and the current is very small.

Thermal equilibrium carrier concentrations are those electron and hole concentrations that are solely determined by the statistics of the carriers and the density of states in the band. Thermal equilibrium concentrations obey the mass action law, $np = n_i^2$.

Thermal velocity (v_{th}) of an electron in the CB is its mean (or effective) speed in the semiconductor as it moves around in the crystal. For a nondegenerate semiconductor, it can be obtained simply from $\frac{1}{2}m_e^* v_{th}^2 = \frac{1}{2}kT$

Vacuum level is the energy level where the *PE* of the electron and the *KE* of the electron are both zero. It defines the energy level where the electron is just free from the solid.

Valence band (VB) is a band of energies for the electrons in bonds in a semiconductor. The valence band is made of all those states (wavefunctions) that constitute the bonding between the atoms in the crystal. At absolute zero of temperature, the VB is full of all the bonding electrons of the atoms. When an electron is excited to the CB, this leaves behind an empty state, which is called a hole. It carries a positive charge and behaves as if it were a "free" positively charged entity with an effective mass of m_h^*. It moves around the VB by having a neighboring electron tunnel into the unoccupied state.

Work function (Φ) is the energy required to remove an electron from the solid to the vacuum level.

QUESTIONS AND PROBLEMS

5.1 Bandgap and photodetection

a. Determine the maximum value of the energy gap that a semiconductor, used as a photoconductor, can have if it is to be sensitive to yellow light (600 nm).

b. A photodetector whose area is 5×10^{-2} cm^2 is irradiated with yellow light whose intensity is 2 mW cm^{-2}. Assuming that each photon generates one electron–hole pair, calculate the number of pairs generated per second.

c. From the known energy gap of the semiconductor GaAs ($E_g = 1.42$ eV), calculate the primary wavelength of photons emitted from this crystal as a result of electron–hole recombination.

d. Is the above wavelength visible?

e. Will a silicon photodetector be sensitive to the radiation from a GaAs laser? Why?

5.2 Intrinsic Ge Using the values of the density of states effective masses m_e^* and m_h^* in Table 5.1, calculate the intrinsic concentration in Ge. What is n_i if you use N_c and N_v from Table 5.1? Calculate the intrinsic resistivity of Ge at 300 K.

5.3 Fermi level in intrinsic semiconductors Using the values of the density of states effective masses m_e^* and m_h^* in Table 5.1, find the position of the Fermi energy in intrinsic Si, Ge, and GaAs with respect to the middle of the bandgap ($E_g/2$).

5.4 Extrinsic Si A Si crystal has been doped with P. The donor concentration is 10^{15} cm^{-3}. Find the conductivity and resistivity of the crystal.

5.5 Extrinsic Si Find the concentration of acceptors required for a *p*-Si crystal to have a resistivity of 1 Ω cm.

5.6 Minimum conductivity

a. Consider the conductivity of a semiconductor, $\sigma = en\mu_e + ep\mu_h$. Will doping always increase the conductivity?

b. Show that the minimum conductivity for Si is obtained when it is *p*-type doped such that the hole concentration is

$$p_m = n_i \sqrt{\frac{\mu_e}{\mu_h}}$$

and the corresponding minimum conductivity (maximum resistivity) is

$$\sigma_{min} = 2en_i\sqrt{\mu_e \mu_h}$$

c. Calculate ρ_m and σ_{min} for Si and compare with intrinsic values.

5.7 Extrinsic *p*-Si A Si crystal is to be doped *p*-type with B acceptors. The hole drift mobility μ_h depends on the total concentration of ionized dopants N_{dopant}, in this case acceptors only, as

$$\mu_h \approx 54.3 + \frac{407}{1 + 3.745 \times 10^{-18} N_{dopant}} \quad \text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$$

where N_{dopant} is in cm^{-3}. Find the required concentration of B doping for the resistivity to be 0.1 Ω cm.

5.8 Thermal velocity and mean free path in GaAs Given that the electron effective mass m_e^* for the GaAs is $0.067m_e$, calculate the thermal velocity of the conduction band (CB) electrons. The electron drift mobility μ_e depends on the mean free time τ_e between electron scattering events (between electrons and lattice vibrations). Given $\mu_e = e\tau_e/m_e^*$, and $\mu_e = 8500\ \text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$ for GaAs, calculate τ_e, and hence the mean free path ℓ of CB electrons. How many unit cells is ℓ if the lattice constant a of GaAs is 0.565 nm? Calculate the drift velocity $v_d = \mu_e \mathcal{E}$ of the CB electrons in an applied field $\mathcal{E}$ of $10^4\ \text{V m}^{-1}$. What is your conclusion?

5.9 Compensation doping in Si

a. A Si wafer has been doped *n*-type with 10^{17} As atoms cm^{-3}.

1. Calculate the conductivity of the sample at 27 °C.
2. Where is the Fermi level in this sample at 27 °C with respect to the Fermi level (E_{Fi}) in intrinsic Si?
3. Calculate the conductivity of the sample at 127 °C.

b. The above *n*-type Si sample is further doped with 9×10^{16} boron atoms (*p*-type dopant) per centimeter cubed.

1. Calculate the conductivity of the sample at 27 °C.
2. Where is the Fermi level in this sample with respect to the Fermi level in the sample in (*a*) at 27 °C? Is this an *n*-type or *p*-type Si?

5.10 Temperature dependence of conductivity An *n*-type Si sample has been doped with 10^{15} phosphorus atoms cm^{-3}. The donor energy level for P in Si is 0.045 eV below the conduction band edge energy.

a. Calculate the room temperature conductivity of the sample.

b. Estimate the temperature above which the sample behaves as if intrinsic.

c. Estimate to within 20 percent the lowest temperature above which all the donors are ionized.

d. Sketch schematically the dependence of the electron concentration in the conduction band on the temperature as $\log(n)$ versus $1/T$, and mark the various important regions and critical temperatures. For each region draw an energy band diagram that clearly shows from where the electrons are excited into the conduction band.

e. Sketch schematically the dependence of the conductivity on the temperature as $\log(\sigma)$ versus $1/T$ and mark the various critical temperatures and other relevant information.

***5.11 Ionization at low temperatures in doped semiconductors** Consider an *n*-type semiconductor. The probability that a donor level E_d is occupied by an electron is

Probability of donor occupancy

$$f_d = \frac{1}{1 + \frac{1}{g}\exp\left(\frac{E_d - E_F}{kT}\right)} \qquad [5.84]$$

where k is the Boltzmann constant, T is the temperature, E_f is the Fermi energy, and g is a constant called the degeneracy factor; in Si, $g = 2$ for donors, and for the occupation statistics of acceptors $g = 4$. Show that

Electron concentration in extrinsic semiconductors

$$n^2 + \frac{nN_c}{g\exp\left(\frac{\Delta E}{kT}\right)} - \frac{N_d N_c}{g\exp\left(\frac{\Delta E}{kT}\right)} = 0 \qquad [5.85]$$

where n is the electron concentration in the conduction band, N_c is the effective density of states at the conduction band edge, N_d is the donor concentration, and $\Delta E = E_c - E_d$ is the ionization energy of the donors. Show that Equation 5.85 at low temperatures is equivalent to Equation 5.19. Consider a p-type Si sample that has been doped with 10^{15} gallium (Ga) atoms cm^{-3}. The acceptor energy level for Ga in Si is 0.065 eV above the valence band edge energy, E_v. Estimate the lowest temperature (°C) above which 90 percent of the acceptors are ionized by assuming that the acceptor degeneracy factor $g = 4$.

5.12 **Compensation doping in *n*-type Si** An n-type Si sample has been doped with 1×10^{17} phosphorus (P) atoms cm^{-3}. The drift mobilities of holes and electrons in Si at 300 K depend on the total concentration of dopants N_{dopant} (cm^{-3}) as follows:

Electron drift mobility

$$\mu_e \approx 88 + \frac{1252}{1 + 6.984 \times 10^{-18} N_{dopant}} \quad \text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$$

and

Hole drift mobility

$$\mu_h \approx 54.3 + \frac{407}{1 + 3.745 \times 10^{-18} N_{dopant}} \quad \text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$$

a. Calculate the room temperature conductivity of the sample.

b. Calculate the necessary acceptor doping (*i.e.*, N_a) that is required to make this sample p-type with approximately the same conductivity.

5.13 **GaAs** Ga has a valency of III and As has V. When Ga and As atoms are brought together to form the GaAs crystal, as depicted in Figure 5.54, the three valence electrons in each Ga and the five valence electrons in each As are all shared to form four covalent bonds per atom. In the GaAs crystal with some 10^{23} or so equal numbers of Ga and As atoms, we have an average of four valence electrons per atom, whether Ga or As, so we would expect the bonding to be similar to that in the Si crystal: four bonds per atom. The crystal structure, however, is not that of diamond but rather that of zinc blende (Chapter 1).

a. What is the average number of valence electrons per atom for a pair of Ga and As atoms and in the GaAs crystal?

b. What will happen if Se or Te, from Group VI, are substituted for an As atom in the GaAs crystal?

c. What will happen if Zn or Cd, from Group II, are substituted for a Ga atom in the GaAs crystal?

d. What will happen if Si, from Group IV, is substituted for an As atom in the GaAs crystal?

e. What will happen if Si, from Group IV, is substituted for a Ga atom in the GaAs crystal? What do you think **amphoteric dopant** means?

f. Based on the discussion of GaAs, what do you think the crystal structures of the III–V compound semiconductors AlAs, GaP, InAs, InP, and InSb will be?

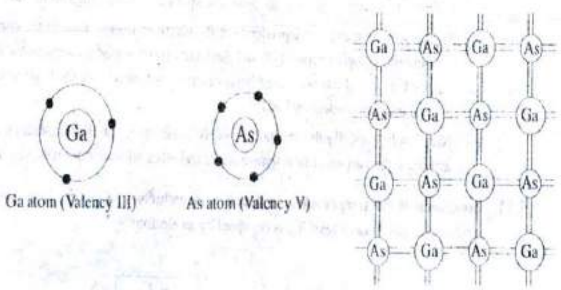


Figure 5.54 The GaAs crystal structure in two dimensions. Average number of valence electrons per atom is four. Each Ga atom covalently bonds with four neighboring As atoms and vice versa.

5.14 **Doped GaAs** Consider the GaAs crystal at 300 K.

a. Calculate the intrinsic conductivity and resistivity.

b. In a sample containing only 10^{15} cm^{-3} ionized donors, where is the Fermi level? What is the conductivity of the sample?

c. In a sample containing 10^{15} cm^{-3} ionized donors and 9×10^{14} cm^{-3} ionized acceptors, what is the free hole concentration?

5.15 **Varshni equation and the change in the bandgap with temperature** The Varshni equation describes the change in the energy bandgap E_g of a semiconductor with temperature T in terms of

$$E_g = E_{go} - \frac{AT^2}{B+T}$$

Varshni equation

where E_{go} is the bandgap at $T = 0$ K, and A and B are material-specific constants. For example, for GaAs, $E_{go} = 1.519$ eV, $A = 5.405 \times 10^{-4}$ eV K^{-1}, $B = 204$ K, so that at $T = 300$ K, $E_g = 1.42$ eV. Show that

$$\frac{dE_g}{dT} = -\frac{AT(T+2B)}{(B+T)^2} = -\frac{(E_{go} - E_g)}{T}\left(\frac{T+2B}{T+B}\right)$$

Bandgap shift with temperature

What is dE_g/dT for GaAs? The Varshni equation can be used to calculate the shift in the peak emission wavelength of a light emitting diode (LED) with temperature or the cutoff wavelength of a detector. If the emitted photon energy from an electron and hole recombination is $h\nu \approx E_g + kT$, find the shift in the emitted wavelength from 27 °C down to −30 °C from a GaAs LED.

5.16 **Degenerate semiconductor** Consider the general exponential expression for the concentration of electrons in the CB,

$$n = N_c \exp\left[-\frac{(E_c - E_F)}{kT}\right]$$

and the mass action law, $np = n_i^2$. What happens when the doping level is such that n approaches N_c and exceeds it? Can you still use the above expressions for n and p?

Consider an n-type Si that has been heavily doped and the electron concentration in the CB is 10^{20} cm^{-3}. Where is the Fermi level? Can you use $np = n_i^2$ to find the hole concentration? What is its resistivity? How does this compare with a typical metal? What use is such a semiconductor?

5.17 **Photoconductivity and speed** Consider two p-type Si samples both doped with 10^{15} B atoms cm^{-3}. Both have identical dimensions of length L (1 mm), width W (1 mm), and depth (thickness) D (0.1 mm). One sample, labeled A, has an electron lifetime of 1 μs whereas the other, labeled B, has an electron lifetime of 5 μs.

a. At time $t = 0$, a laser light of wavelength 750 nm is switched on to illuminate the surface ($L \times W$) of both the samples. The incident laser light intensity on both samples is 10 mW cm^{-2}. At time $t = 50$ μs, the laser is switched off. Sketch the time evolution of the minority carrier concentration for both samples on the same axes.

b. What is the photocurrent (current due to illumination alone) if each sample is connected to a 1 V battery?

***5.18** **Hall effect in semiconductors** The Hall effect in a semiconductor sample involves not only the electron and hole concentrations n and p, respectively, but also the electron and hole drift mobilities μ_e and μ_h. The Hall coefficient of a semiconductor is (see Chapter 2)

$$R_H = \frac{p - nb^2}{e(p+nb)^2} \qquad [5.86]$$

Hall coefficient of a semiconductor

where $b = \mu_e/\mu_h$.

a. Given the mass action law $np = n_i^2$, find n for maximum $|R_H|$ (negative and positive R_H). Assume that the drift mobilities remain relatively unaffected as n changes (due to doping). Given the electron and hole drift mobilities $\mu_e = 1350$ cm^2 V^{-1} s^{-1} and $\mu_h = 450$ cm^2 V^{-1} s^{-1} for silicon, determine n for maximum $|R_H|$ in terms of n_i.

b. Taking $b = 3$, plot R_H as a function of electron concentration n/n_i from 0.01 to 10.

c. Show that, when $n \gg n_i$, $R_H = -1/en$ and when $n \ll n_i$, $R_H = +1/ep$.

5.19 **Hall effect in semiconductors** Most Hall-effect high-sensitivity sensors typically use III–V semiconductors, such as GaAs, InAs, InSb. Hall-effect integrated circuits with integrated amplifiers, on the other hand, use Si. Consider nearly intrinsic samples in which $n \approx p \approx n_i$, and calculate R_H for each using the data in Table 5.4. What is your conclusion? Which sensor would exhibit the worst temperature drift? (Consider the bandgap, and drift in n_i.)

Table 5.4 Hall effect in selected semiconductors

	E_g(eV)	n_i(cm^{-3})	μ_e (cm^2 V^{-1} s^{-1})	μ_h (cm^2 V^{-1} s^{-1})	b	R_H (m^3 A^{-1} s^{-1})
Si	1.10	1×10^{10}	1,350	450	3	−312
GaAs	1.42	2×10^{6}	8,500	400	?	?
InAs	0.36	1×10^{15}	33,000	460	?	?
InSb	0.17	2×10^{16}	78,000	850	?	?

***5.20** **Compound semiconductor devices** Silicon and germanium crystalline semiconductors are what are called elemental Group IV semiconductors. It is possible to have compound semiconductors from atoms in Groups III and V. For example, GaAs is a compound semiconductor that has Ga from Group III and As from Group V, so in the crystalline structure we have an "effective" or "mean" valency of IV per atom and the solid behaves like a semiconductor. Similarly GaSb (gallium antimonide) would be a III–V type semiconductor. Provided we have a stoichiometric compound, the semiconductor will be ideally intrinsic. If, however, there is an excess of Sb atoms in the solid GaSb, then we will have nonstoichiometry and the semiconductor will be extrinsic. In this case, excess Sb atoms will act as donors in the GaSb structure. There are many useful compound semiconductors, the most important of which is GaAs. Some can be doped both *n*- and *p*-type, but many are one type only. For example, ZnO is a II–VI compound semiconductor with a direct bandgap of 3.2 eV, but unfortunately, due to the presence of excess Zn, it is naturally *n*-type and cannot be doped to *p*-type.

a. GaSb (gallium antimonide) is an interesting direct bandgap semiconductor with an energy bandgap $E_g = 0.67$ eV, almost equal to that of germanium. It can be used as an light emitting diode (LED) or laser diode material. What would be the wavelength of emission from a GaSb LED? Will this be visible?

b. Calculate the intrinsic conductivity of GaSb at 300 K taking $N_c = 2.3 \times 10^{19}$ cm^{-3}, $N_v = 6.1 \times 10^{19}$ cm^{-3}, $\mu_e = 5000$ cm^2 V^{-1} s^{-1}, and $\mu_h = 1000$ cm^2 V^{-1} s^{-1}. Compare with the intrinsic conductivity of Ge.

c. Excess Sb atoms will make gallium antimonide nonstoichiometric, that is, $GaSb_{1+\delta}$, which will result in an extrinsic semiconductor. Given that the density of GaSb is 5.4 g cm^{-3}, calculate δ (excess Sb) that will result in GaSb having a conductivity of 100 Ω^{-1} cm^{-1}. Will this be an *n*- or *p*-type semiconductor? You may assume that the drift mobilities are relatively unaffected by the doping.

5.21 **Excess minority carrier concentration** Consider an *n*-type semiconductor and weak injection conditions. Assume that the minority carrier recombination time τ_h is constant (independent of injection—hence the weak injection assumption). The rate of change of the instantaneous hole concentration $\partial p_n/\partial t$ due to recombination is given by

Recombination rate

$$\frac{\partial p_n}{\partial t} = -\frac{p_n}{\tau_h} \qquad [5.87]$$

The net rate of increase (change) in p_n is the sum of the total generation rate G and the rate of change due to recombination, that is,

$$\frac{dp_n}{dt} = G - \frac{p_n}{\tau_h} \quad [5.88]$$

Excess carries under uniform photogeneration and recombination

By separating the generation term G into thermal generation G_o and photogeneration G_{ph} and considering the dark condition as one possible solution, show that

$$\frac{d\Delta p_n}{dt} = G_{ph} - \frac{\Delta p_n}{\tau_h} \quad [5.89]$$

How does your derivation compare with Equation 5.27? What are the assumptions inherent in Equation 5.89?

***5.22 Direct recombination and GaAs** Consider recombination in a direct bandgap p-type semiconductor, *e.g.*, GaAs doped with an acceptor concentration N_a. The recombination involves a direct meeting of an electron–hole pair as depicted in Figure 5.22. Suppose that excess electrons and holes have been injected (*e.g.*, by photoexcitation), and that Δn_p is the excess electron concentration and Δp_p is the excess hole concentration. Assume Δn_p is controlled by recombination and thermal generation only; that is, recombination is the equilibrium storing mechanism. The recombination rate will be proportional to $n_p p_p$, and the thermal generation rate will be proportional to $n_{po} p_{po}$. In the dark, in equilibrium, thermal generation rate is equal to the recombination rate. The latter is proportional to $n_{no} p_{po}$. The rate of change of Δn_p is

$$\frac{\partial \Delta n_p}{\partial t} = -B[n_p p_p - n_{po} p_{po}] \quad [5.90]$$

Recombination rate

where B is a proportionality constant, called the **direct recombination capture coefficient.** The **recombination lifetime** τ_r is *defined* by

$$\frac{\partial \Delta n_p}{\partial t} = -\frac{\Delta n_p}{\tau_r} \quad [5.91]$$

Definition of recombination lifetime

a. Show that for *low-level injection*, $n_{po} \ll \Delta n_p \ll p_{po}$, τ_r is constant and given by

$$\tau_r = \frac{1}{B p_{po}} = \frac{1}{B N_a} \quad [5.92]$$

Low injection recombination time

b. Show that under *high-level injection*, $\Delta n_p \gg p_{po}$,

$$\frac{\partial \Delta n_p}{\partial t} \approx -B \Delta p_p \, \Delta n_p = -B(\Delta n_p)^2 \quad [5.93]$$

High injection

so that the recombination lifetime τ_r is now given by

$$\tau_r = \frac{1}{B \Delta p_p} = \frac{1}{B \Delta n_p} \quad [5.94]$$

High-injection recombination time

that is, the lifetime τ_r is inversely proportional to the injected carrier concentration.

c. Consider what happens in the presence of photogeneration at a rate G_{ph} (electron–hole pairs per unit volume per unit time). Steady state will be reached when the photogeneration rate and recombination rate become equal. That is,

$$G_{ph} = \left(\frac{\partial \Delta n_p}{\partial t}\right)_{\text{recombination}} = B[n_p p_p - n_{po} p_{po}]$$

Steady-state photogeneration rate

A photoconductive film of n-type GaAs doped with 10^{13} cm^{-3} donors is 2 mm long (L), 1 mm wide (W), and 5 μm thick (D). The sample has electrodes attached to its ends (electrode area is therefore 1 mm × 5 μm) which are connected to a 1 V supply through an ammeter. The GaAs photoconductor is uniformly illuminated over the surface area 2 mm × 1 mm with a 1 mW laser

radiation of wavelength $\lambda = 840$ nm (infrared). The recombination coefficient B for GaAs is 7.21×10^{-16} m^3 s^{-1}. At $\lambda = 840$ nm, the absorption coefficient is about 5×10^3 cm^{-1}. Calculate the photocurrent I_{photo} and the electrical power dissipated as Joule heating in the sample. What will be the power dissipated as heat in the sample in an open circuit, where $I = 0$?

5.23 **Piezoresistivity application to deflection and force measurement** Consider the cantilever in Figure 5.38c. Suppose we apply a force F to the free end, which results in a deflection h of the tip of the cantilever from its horizontal equilibrium position. The maximum stress σ_m is induced at the support end of the cantilever, at its surface where the piezoresistor is embedded to measure the stress. When the cantilever is bent, there is a tensile or longitudinal stress σ_L on the surface because the top surface is extended and the bottom surface is contracted. If L, W, and D are respectively the length, width, and thickness of the cantilever, then the relationships between the force F and deflection h, and the maximum stress σ_L are

Cantilever equations

$$\sigma_L(\max) = \frac{3YDh}{2L^2} \quad \text{and} \quad F = \frac{WD^3Y}{4L^3}h$$

where Y is the elastic (Young's) modulus. A particular Si cantilever has a length (L) of 500 μm, width (W) of 100 μm, and thickness (D) of 10 μm. Given $Y = 170$ GPa, and that the piezoresistor embedded in the cantilever is along the [110] direction with $\pi_L \approx 72 \times 10^{-11}$ Pa^{-1}, find the percentage change in the resistance, $\Delta R/R$, of the piezoresistor when the deflection is 0.1 μm. What is the force that would give this deflection? (Neglect the transverse stresses on the piezoresistor.) How does the design choice for the length L of the cantilever depend on whether one is interested in measuring the deflection h or the force F? (Note: σ_L depends on the distance x from the support end; it decreases with x. Assume that the length of the piezoresistor is very short compared with L so that σ_L does not change significantly along its length.)

5.24 **Schottky junction**

a. Consider a Schottky junction diode between Au and n-Si, doped with 10^{16} donors cm^{-3}. The cross-sectional area is 1 mm^2. Given the work function of Au as 5.1 eV, what is the theoretical barrier height Φ_B from the metal to the semiconductor?

b. Given that the experimental barrier height Φ_B is about 0.8 eV, what is the reverse saturation current and the current when there is a forward bias of 0.3 V across the diode? (Use Equation 4.37.)

5.25 **Schottky junction** Consider a Schottky junction diode between Al and n-Si, doped with $5 \cdot 10^{16}$ donors cm^{-3}. The cross-sectional area is 1 mm^2. Given that the electron affinity χ of Si is 4.01 eV and the work function of Al is 4.28 eV, what is the theoretical barrier height Φ_B from the metal to the semiconductor? What is the built-in voltage? If the experimental barrier height Φ_B is about 0.6 eV, what is the reverse saturation current and the current when there is a forward bias of 0.2 V across the diode? Take $B_t = 110$ A cm^{-2} K^{-2}.

5.26 **Schottky and ohmic contacts** Consider an n-type Si sample doped with 10^{16} donors cm^{-3}. The length L is 100 μm; the cross-sectional area A is 10 μm × 10 μm. The two ends of the sample are labeled as B and C. The electron affinity (χ) of Si is 4.01 eV and the work functions Φ of four potential metals for contacts at B and C are listed in Table 5.5.

Table 5.5 Work functions in eV

Cs	Li	Al	Au
1.8	2.5	4.25	5.0

a. Ideally, which metals will result in a Schottky contact?

b. Ideally, which metals will result in an ohmic contact?

c. Sketch the I–V characteristics when both B and C are ohmic contacts. What is the relationship between I and V?

d. Sketch the I–V characteristics when B is ohmic and C is a Schottky junction. What is the relationship between I and V?

e. Sketch the I–V characteristics when both B and C are Schottky contacts. What is the relationship between I and V?

5.27 **Peltier effect and electrical contacts** Consider the Schottky junction and the ohmic contact shown in Figures 5.39 and 5.43 between a metal and n-type semiconductor.

a. Is the Peltier effect similar in both contacts?

b. Is the sign in $Q' = \pm \Pi I$ the same for both contacts?

c. Which junction would you choose for a thermoelectric cooler? Give reasons.

***5.28** **Peltier coolers and figure of merit (FOM)** Consider the thermoelectric effect shown in Figure 5.45 in which a semiconductor has two contacts at its ends and is conducting an electric current I. We assume that the cold junction is at a temperature T_c and the hot junction is at T_h and that there is a temperature difference of $\Delta T = T_h - T_c$ between the two ends of the semiconductor. The current I flowing through the cold junction absorbs Peltier heat at a rate Q'_P, given by

$$Q'_P = \Pi I \qquad [5.95]$$

where Π is the Peltier coefficient for the junction between the metal and semiconductor. The current I flowing through the semiconductor generates heat due to the Joule heating of the semiconductor. The rate of Joule heat generated through the bulk of the semiconductor is

$$Q'_J = \left(\frac{L}{\sigma A}\right) I^2 \qquad [5.96]$$

We assume that half of this heat flows to the cold junction.

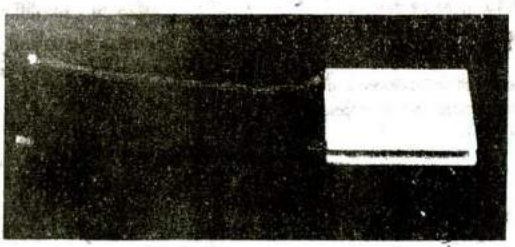
In addition there is heat flow from the hot to the cold junction through the semiconductor, given by the thermal conduction equation

$$Q'_{TC} = \left(\frac{A\kappa}{L}\right) \Delta T \qquad [5.97]$$

The net rate of heat absorption (cooling rate) at the cold junction is then

$$Q'_{\text{net cool}} = Q'_P - \frac{1}{2} Q'_J - Q'_{TC} \qquad [5.98]$$

By substituting from Equations 5.95 to 5.97 into Equation 5.98, obtain the net cooling rate in terms of the current I. Then by differentiating $Q'_{\text{net cool}}$ with respect to current, show that maximum cooling is



A commercial thermoelectric cooler (by Melcor), an example of the Peltier effect. The device area is 5.5 cm x 5.5 cm (approximately 2.2 inches x 2.2 inches). Its maximum current is 14 A; maximum heat pump ability is 67 W; maximum temperature difference between the hot and cold surfaces is 67 °C.

Table 5.6

Material	Π (V)	ρ (Ω m)	κ (W m^{-1}K^{-1})	FOM
n-Bi_2Te_3	6.0×10^{-2}	10^{-5}	1.70	
p-Bi_2Te_3	7.0×10^{-2}	10^{-5}	1.45	
Cu	5.5×10^{-4}	1.7×10^{-8}	390	
W	3.3×10^{-4}	5.5×10^{-8}	167	

obtained when the current is

$$I_m = \left(\frac{A}{L}\right)\Pi\sigma \qquad [5.99]$$

and the maximum cooling rate is

Maximum cooling rate

$$Q'_{\max\,\text{cool}} = \frac{A}{L}\left[\frac{1}{2}\Pi^2\sigma - \kappa\,\Delta T\right] \qquad [5.100]$$

Under steady-state operating conditions, the temperature difference ΔT reaches a steady-state value and the net cooling rate at the junction is then zero (ΔT is constant). From Equation 5.100 show that the maximum temperature difference achievable is

Maximum temperature difference

$$\Delta T_{\max} = \frac{1}{2}\frac{\Pi^2\sigma}{\kappa} \qquad [5.101]$$

The quantity $\Pi^2\sigma/\kappa$ is defined as the **figure of merit** (FOM) for the semiconductor as it determines the maximum ΔT achievable. The same expression also applies to metals, though we will not derive it here.

Use Table 5.6 to determine the FOM for various materials listed therein and discuss the significance of your calculations. Would you recommend a thermoelectric cooler based on a metal-to-metal junction?

***5.29 Seebeck coefficient of semiconductors and thermal drift in semiconductor devices** Consider an n-type semiconductor that has a temperature gradient across it. The right end is hot and the left end is cold, as depicted in Figure 5.55. There are more energetic electrons in the hot region than in the cold region. Consequently, electron diffusion occurs from hot to cold regions, which immediately exposes negatively charged donors in the hot region and therefore builds up an internal field and a built-in voltage, as shown in Figure 5.55. Eventually an equilibrium is reached when the diffusion of electrons is balanced by their drift driven by the built-in field. The net current must be zero. The Seebeck coefficient (or thermoelectric power) S measures

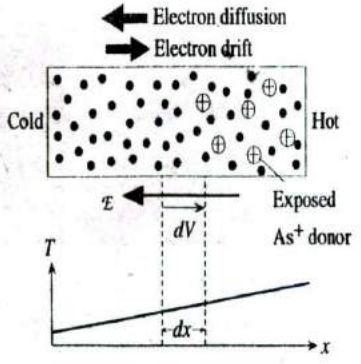


Figure 5.55 In the presence of a temperature gradient, there is an internal field and a voltage difference.

The Seebeck coefficient is defined as dV/dT, the potential difference per unit temperature difference.

this effect in terms of the voltage developed as a result of an applied temperature gradient as

$$S = \frac{dV}{dT} \qquad [5.102]$$

a. How is the Seebeck effect in a *p*-type semiconductor different than that for an *n*-type semiconductor when both are placed in the same temperature gradient in Figure 5.55? Recall that the sign of the Seebeck coefficient is the polarity of the voltage at the cold end with respect to the hot end (see Section 4.8.2).

b. Given that for an *n*-type semiconductor,

Seebeck coefficient n-type semiconductor

$$S_n = -\frac{k}{e}\left[2 + \frac{(E_c - E_F)}{kT}\right] \qquad [5.103]$$

what are typical magnitudes for S_n in Si doped with 10^{14} and 10^{16} donors cm^{-3}? What is the significance of S_n at the semiconductor device level?

c. Consider a *pn* junction Si device that has the *p*-side doped with 10^{18} acceptors cm^{-3} and the *n*-side doped with 10^{14} donors cm^{-3}. Suppose that this *pn* junction forms the input stage of an op amp with a large gain, say 100. What will be the output signal if a small thermal fluctuation gives rise to a 1 °C temperature difference across the *pn* junction?

5.30 Photogeneration and carrier kinetic energies Figure 5.35 shows what happens when a photon with energy $h\nu > E_g$ is absorbed in GaAs to photogenerate an electron and a hole. The figure shows that the electron has a higher kinetic energy (*KE*), which is the excess energy above E_c, than the hole, since the hole is almost at E_v. The reason is that the electron effective mass in GaAs is almost 10 times less than the hole effective mass, so the photogenerated electron has a much higher *KE*. When an electron and hole are photogenerated in a direct bandgap semiconductor, they have the same **k** vector. Energy conservation requires that the photon energy $h\nu$ divides according to

Photogeneration

$$h\nu = E_g + \frac{(\hbar k)^2}{2m_e^*} + \frac{(\hbar k)^2}{2m_h^*}$$

where k is the wavevector of the electron and hole and m_e^* and m_h^* are the effective masses of the electron and hole, respectively.

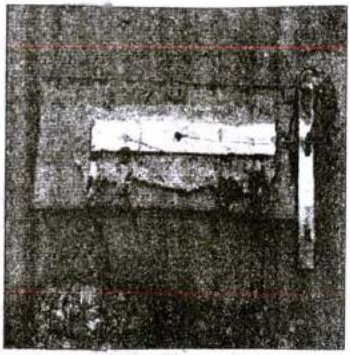
a. What is the ratio of the electron to hole *KE*s right after photogeneration?

b. If the incoming photon has an energy of 2.0 eV, and $E_g = 1.42$ eV for GaAs, calculate the *KE*s of the electron and the hole in eV, and calculate to which energy levels they have been excited with respect to their band edges.

c. Explain why the electron and hole wavevector k should be approximately the same right after photogeneration. Consider k_{photon} for the photon, and the momentum conservation.

William Shockley and his group celebrate Shockley's Nobel prize in 1956. First left, sitting, is G. E. Moore [chairman emeritus of Intel], standing fourth from right is R. N. Noyce, inventor of the integrated circuit, and standing at the extreme right is J. T. Last.

SOURCE: P. K. Bondyopadhyay, "W = Shockley, the Transistor Pioneer—Portrait of an Inventive Genius," *Proceedings IEEE*, vol. 86, no. 1, January 1998, p. 202, figure 16 (Courtesy of IEEE.)



The first monolithic integrated circuit, about the size of a fingertip, was documented and developed at Texas Instruments by Jack Kilby in 1958; he won the 2000 Nobel prize in physics for his contribution to the development of the first integrated circuit. The IC was a chip of a single Ge crystal containing one transistor, one capacitor, and one resistor. Left: Jack Kilby holding his IC (photo, 1998). Right. The photo of the chip.

SOURCE: Courtesy of Texas Instruments.

Robert Noyce and Jean Hoerni (a Swiss physicist) were responsible for the invention of the first planar IC at Fairchild (1961). The planar fabrication process was the key to the success of their IC. The photograph is that of the first logic chip at Fairchild.

SOURCE: Courtesy of Fairchild Semiconductor.

Left to right: Andrew Grove, Robert Noyce (1927–1990), and Gordon Moore, who founded Intel in 1968. Andrew Grove's book *Physics and Technology of Semiconductor Devices* (Wiley, 1967) was one of the classic texts on devices in the sixties and seventies. "Moore's law" that started as a rough rule in 1965 states that the number of transistors in a chip will double every 18 months; Moore updated it in 1995 to every couple of years.

SOURCE: Courtesy of Intel.

CHAPTER

6

Semiconductor Devices

Most diodes are essentially *pn* junctions fabricated by forming a contact between a *p*-type and an *n*-type semiconductor. The junction possesses rectifying properties in that a current in one direction can flow quite easily whereas in the other direction it is limited by a leakage current that is generally very small. A transistor is a three-terminal solid-state device in which a current flowing between two electrodes is controlled by the voltage between the third and one of the other terminals. Transistors are capable of providing current and voltage gains thereby enabling weak signals to be amplified. Transistors can also be used as switches just like electromagnetic relays. Indeed, the whole microcomputer industry is based on transistor switches. The majority of the transistors in microelectronics are of essentially two types: **bipolar junction transistors** (BJTs) and **field effect transistors** (FETs). The appreciation of the underlying principles of the *pn* junction is essential to understanding the operation of not only the bipolar transistor but also a variety of related devices. The central fundamental concept is the **minority carrier injection** as purported by William Shockley in his explanations of the transistor operation. Field effect transistors operate on a totally different principle than BJTs. Their characteristics arise from the effect of the applied field on a conducting channel between two terminals. The last two decades have seen enormous advances and developments in optoelectronic and photonic devices which we now take for granted, the best examples being **light emitting diodes** (LEDs), **semiconductor lasers, photodetectors,** and **solar cells.** Nearly all these devices are based on *pn* junction principles. The present chapter takes the semiconductor concepts developed in Chapter 5 to device level applications, from the basic *pn* junction to heterojunction laser diodes.

6.1 IDEAL *pn* JUNCTION

6.1.1 NO APPLIED BIAS: OPEN CIRCUIT

Consider what happens when one side of a sample of Si is doped *n*-type and the other *p*-type, as shown in Figure 6.1a. We assume that there is an abrupt discontinuity between the *p*- and *n*-regions, which we call the **metallurgical junction** and label as M in Figure 6.1a, where the fixed (immobile) ionized donors and the free electrons (in the conduction band, CB) in the *n*-region and fixed ionized acceptors and holes (in the valence band, VB) in the *p*-region are also shown.

Due to the hole concentration gradient from the *p*-side, where $p = p_{po}$, to the *n*-side, where $p = p_{no}$, holes diffuse toward the right. Similarly the electron concentration

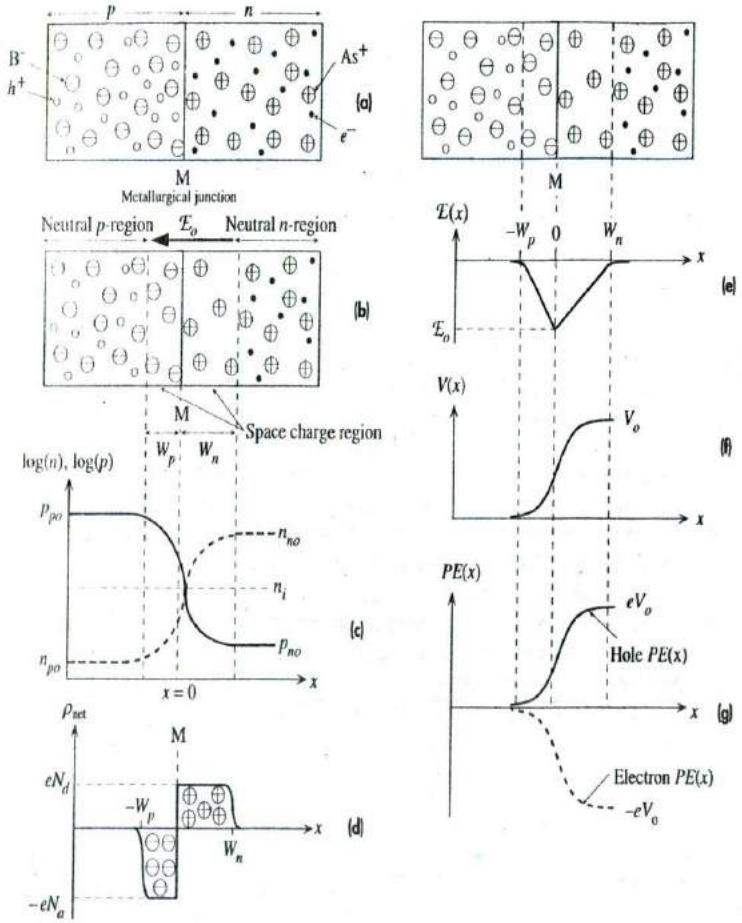


Figure 6.1 Properties of the *pn* junction.

gradient drives the electrons by diffusion toward the left. Holes diffusing and entering the *n*-side recombine with the electrons in the *n*-side near the junction. Similarly, electrons diffusing and entering the *p*-side recombine with holes in the *p*-side near the junction. The junction region consequently becomes depleted of free carriers in comparison with the bulk *p*- and *n*-regions far away from the junction. Note that we must, under equilibrium conditions (*e.g.*, no applied bias or photoexcitation), have $pn = n_i^2$ everywhere. Electrons leaving the *n*-side near the junction M leave behind exposed positively charged donor ions, say As^+, of concentration N_d. Similarly, holes leaving the *p*-region near M expose negatively charged acceptor ions, say B^-, of concentration N_a. There is therefore a **space charge layer** (SCL) around M. Figure 6.1b shows the **depletion region,** or the space charge layer, around M, whereas Figure 6.1c illustrates the hole and electron concentration profiles in which the vertical concentration scale is logarithmic. The depletion region is also called the transition region.

It is clear that there is an internal electric field $\mathcal{E}_o$ from positive ions to negative ions, that is, in the $-x$ direction, that tries to drift the holes back into the *p*-region and electrons back into the *n*-region. This field drives the holes in the opposite direction to their diffusion. As shown in Figure 6.1b, $\mathcal{E}_o$ imposes a drift force on holes in the $-x$ direction, whereas the hole diffusion flux is in the $+x$ direction. A similar situation also applies for electrons with the electric field attempting to drift the electrons against diffusion from the *n*-region to the *p*-region. It is apparent that as more and more holes diffuse toward the right, and electrons toward the left, the internal field around M will increase until eventually an "equilibrium" is reached when the rate of holes diffusing toward the right is just balanced by holes drifting back to the left, driven by the field $\mathcal{E}_o$. The electron diffusion and drift fluxes will also be balanced in equilibrium.

For uniformly doped *p*- and *n*-regions, the net space charge density $\rho_{net}(x)$ across the semiconductor will be as shown in Figure 6.1d. (Why are the edges rounded?) The net space charge density ρ_{net} is negative and equal to $-eN_a$ in the SCL from $x = -W_p$ to $x = 0$ (where we take M to be) and then positive and equal to $+eN_d$ from $x = 0$ to W_n. The total charge on the left-hand side must be equal to that on the right-hand side for overall charge neutrality, so

Depletion widths

$$N_a W_p = N_d W_n \qquad [6.1]$$

In Figure 6.1, we arbitrarily assumed that the donor concentration is less than the acceptor concentration, $N_d < N_a$. From Equation 6.1 this implies that $W_n > W_p$; that is, the depletion region penetrates the *n*-side, the lightly doped side, more than the *p*-side, the heavily doped side. Indeed, if $N_a \gg N_d$, then the depletion region is almost entirely on the *n*-side. We generally indicate heavily doped regions with the plus sign as a superscript, that is, p^+.

The electric field $\mathcal{E}(x)$ and the net space charge density $\rho_{net}(x)$ at a point are related in electrostatics[1] by

Field and net space charge density

$$\frac{d\mathcal{E}}{dx} = \frac{\rho_{net}(x)}{\varepsilon}$$

[1] This is called **Gauss's law in point** form and comes from Gauss's law in electrostatics. Gauss's law is discussed in Section 7.5.

where $\varepsilon = \varepsilon_o\varepsilon_r$ is the permittivity of the medium and ε_o and ε_r are the absolute permittivity and relative permittivity of the semiconductor material. We can thus integrate $\rho_{net}(x)$ across the diode and thus determine the electric field $\mathcal{E}(x)$, that is,

Field in depletion region

$$\mathcal{E}(x) = \frac{1}{\varepsilon}\int_{-W_p}^{x} \rho_{net}(x)\,dx \qquad [6.2]$$

The variation of the electric field across the *pn* junction is shown in Figure 6.1e. The negative field means that it is in the $-x$ direction. Note that $\mathcal{E}(x)$ reaches a maximum value $\mathcal{E}_o$ at the metallurgical junction M.

The potential $V(x)$ at any point x can be found by integrating the electric field since by definition $\mathcal{E} = -dV/dx$. Taking the potential on the *p*-side far away from M as zero (we have no applied voltage), which is an arbitrary reference level, then $V(x)$ increases in the depletion region toward the *n*-side, as indicated in Figure 6.1f. Its functional form can be determined by integrating Equation 6.2, which is, of course, a parabola. Notice that on the *n*-side the potential reaches V_o, which is called the **built-in potential.**

The fact that we are considering an abrupt *pn* junction means that $\rho_{net}(x)$ can simply be described by step functions, as displayed in Figure 6.1d. Using the step form of $\rho_{net}(x)$ in Figure 6.1d in the integration of Equation 6.2 gives the electric field at M as

Built-in field

$$\mathcal{E}_o = -\frac{eN_dW_n}{\varepsilon} = -\frac{eN_aW_p}{\varepsilon} \qquad [6.3]$$

where $\varepsilon = \varepsilon_o\varepsilon_r$. We can integrate the expression for $\mathcal{E}(x)$ in Figure 6.1e to evaluate the potential $V(x)$ and thus find V_o by putting in $x = W_n$. The graphical representation of this integration is the step from Figure 6.1e to f. The result is

Built-in voltage

$$V_o = -\frac{1}{2}\mathcal{E}_oW_o = \frac{eN_aN_dW_o^2}{2\varepsilon(N_a + N_d)} \qquad [6.4]$$

where $W_o = W_n + W_p$ is the total width of the depletion region under a zero applied voltage. If we know W_o, then W_n or W_p follows readily from Equation 6.1. Equation 6.4 is a relationship between the built-in voltage V_o and the depletion region width W_o. If we know V_o, we can calculate W_o.

The simplest way to relate V_o to the doping parameters is to make use of the fact that in the system consisting of *p*- and *n*-type semiconductors joined together, in equilibrium, Boltzmann statistics[2] demands that the concentrations n_1 and n_2 of carriers at potential energies E_1 and E_2 are related by

$$\frac{n_2}{n_1} = \exp\left[-\frac{(E_2 - E_1)}{kT}\right]$$

where $E = qV$, where q is the charge of the carrier. Considering electrons ($q = -e$), we see from Figure 6.1g that $E = 0$ on the *p*-side far away from M where $n = n_{po}$, and

[2] We use Boltzmann statistics, that is, $n(E) \propto \exp[-E/kT]$, because the concentration of electrons in the conduction band, whether on the *n*-side or *p*-side, is never so large that the Pauli exclusion principle becomes important. As long as the carrier concentration in the conduction band is much smaller than N_c, we can use Boltzmann statistics.

$E = -eV_o$ on the n-side away from M where $n = n_{no}$. Thus

$$\frac{n_{po}}{n_{no}} = \exp\left(-\frac{eV_o}{kT}\right) \qquad [6.5a]$$

Boltzmann statistics for electrons

This shows that V_o depends on n_{no} and n_{po} and hence on N_d and N_a. The corresponding equation for hole concentrations is clearly

$$\frac{p_{no}}{p_{po}} = \exp\left(-\frac{eV_o}{kT}\right) \qquad [6.5b]$$

Thus, rearranging Equations 6.5a and b we obtain

$$V_o = \frac{kT}{e}\ln\left(\frac{n_{no}}{n_{po}}\right) \quad \text{and} \quad V_o = \frac{kT}{e}\ln\left(\frac{p_{po}}{p_{no}}\right)$$

We can now write p_{po} and p_{no} in terms of the dopant concentrations inasmuch as $p_{po} = N_a$ and

$$p_{no} = \frac{n_i^2}{n_{no}} = \frac{n_i^2}{N_d}$$

so V_o becomes

$$V_o = \frac{kT}{e}\ln\left(\frac{N_a N_d}{n_i^2}\right) \qquad [6.6]$$

Built-in voltage

Clearly V_o has been conveniently related to the dopant and material properties via N_a, N_d, and n_i^2. The built-in voltage (V_o) is the voltage across a *pn* junction, going from *p*- to *n*-type semiconductor, in an open circuit. It is *not* the voltage across the diode, which is made up of V_o as well as the contact potentials at the metal-to-semiconductor junctions at the electrodes. If we add V_o and the contact potentials at the electroded ends, we will find zero.

Once we know the built-in potential from Equation 6.6, we can then calculate the width of the depletion region from Equation 6.4, namely

$$W_o = \left[\frac{2\varepsilon(N_a + N_d)V_o}{eN_aN_d}\right]^{1/2} \qquad [6.7]$$

Depletion region width

Notice that the depletion width $W_o \propto V_o^{1/2}$. This results in the capacitance of the depletion region being voltage dependent, as we will see in Section 6.3.

EXAMPLE 6.1

THE BUILT-IN POTENTIALS FOR Ge, Si, AND GaAs *pn* JUNCTIONS A *pn* junction diode has a concentration of 10^{16} acceptor atoms cm^{-3} on the *p*-side and a concentration of 10^{17} donor atoms cm^{-3} on the *n*-side. What will be the built-in potential for the semiconductor materials Ge, Si, and GaAs?

SOLUTION

The built-in potential is given by Equation 6.6, which requires the knowledge of the intrinsic concentration for each semiconductor. From Chapter 5 we can tabulate the following

at 300 K:

Semiconductor	E_g (eV)	n_i (cm^{-3})	V_o (V)
Ge	0.7	2.40×10^{13}	0.37
Si	1.1	1.0×10^{10}	0.78
GaAs	1.4	2.1×10^{6}	1.21

Using

$$V_o = \left(\frac{kT}{e}\right) \ln\left(\frac{N_d N_a}{n_i^2}\right)$$

for Si with $N_d = 10^{17}$ cm^{-3} and $N_a = 10^{16}$ cm^{-3}, $kT/e = 0.0259$ V at 300 K, and $n_i = 1.0 \times 10^{10}$ cm^{-3}, we obtain

$$V_o = (0.0259 \text{ V}) \ln\left[\frac{(10^{17})(10^{16})}{(1.0 \times 10^{10})^2}\right] = 0.775 \text{ V}$$

The results for all three semiconductors are summarized in the last column of the table in this example.

EXAMPLE 6.2 **THE p^+n JUNCTION** Consider a p^+n junction, which has a heavily doped p-side relative to the n-side, that is, $N_a \gg N_d$. Since the amount of charge Q on both sides of the metallurgical junction must be the same (so that the junction is overall neutral)

$$Q = eN_a W_p = eN_d W_n$$

it is clear that the depletion region essentially extends into the n-side. According to Equation 6.7, when $N_d \ll N_a$, the width is

$$W_o = \left[\frac{2\varepsilon V_o}{eN_d}\right]^{1/2}$$

What is the depletion width for a pn junction Si diode that has been doped with 10^{18} acceptor atoms cm^{-3} on the p-side and 10^{16} donor atoms cm^{-3} on the n-side?

SOLUTION

To apply the above equation for W_o, we need the built-in potential, which is

$$V_o = \left(\frac{kT}{e}\right) \ln\left(\frac{N_d N_a}{n_i^2}\right) = (0.0259 \text{ V}) \ln\left[\frac{(10^{16})(10^{18})}{(1.0 \times 10^{10})^2}\right] = 0.835 \text{ V}$$

Then with $N_d = 10^{16}$ cm^{-3}, that is, 10^{22} m^{-3}, $V_o = 0.835$ V, and $\varepsilon_r = 11.9$ in the equation for W_o

$$W_o = \left[\frac{2\varepsilon V_o}{eN_d}\right]^{1/2} = \left[\frac{2(11.9)(8.85 \times 10^{-12})(0.835)}{(1.6 \times 10^{-19})(10^{22})}\right]^{1/2}$$

$$= 3.32 \times 10^{-7} \text{ m} \quad \text{or} \quad 0.33 \ \mu\text{m}$$

Nearly all of this region (99 percent of it) is on the n-side.

EXAMPLE 6.3

BUILT-IN VOLTAGE There is a rigorous derivation of the built-in voltage across a *pn* junction. Inasmuch as in equilibrium there is no net current through the *pn* junction, drift of holes due to the built-in field $\mathcal{E}(x)$ must be just balanced by their diffusion due to the concentration gradient dp/dx. We can thus set the total electron and hole current densities (drift + diffusion) through the depletion region to zero. Considering holes alone, from Equation 5.38,

$$J_{\text{hole}}(x) = ep(x)\mu_h \mathcal{E}(x) - eD_h \frac{dp}{dx} = 0$$

The electric field is defined by $\mathcal{E} = -dV/dx$, so substituting we find,

$$-ep\mu_h\, dV - eD_h\, dp = 0$$

We can now use the *Einstein relation* $D_h/\mu_h = kT/e$ to get

$$-ep\, dV - kT\, dp = 0$$

We can integrate this equation. According to Figure 6.1, in the *p*-side, $p = p_{po}$, $V = 0$, and in the *n*-side, $p = p_{no}$, $V = V_o$, thus,

$$\int_0^{V_o} dV + \frac{kT}{e}\int_{p_{po}}^{p_{no}} \frac{dp}{p} = 0$$

that is,

$$V_o + \frac{kT}{e}[\ln(p_{no}) - \ln(p_{po})] = 0$$

giving

$$V_o = \frac{kT}{e}\ln\left(\frac{p_{po}}{p_{no}}\right)$$

which is the same as Equation 6.5b and hence leads to Equation 6.6.

6.1.2 FORWARD BIAS: DIFFUSION CURRENT

Consider what happens when a battery is connected across a *pn* junction so that the positive terminal of the battery is attached to the *p*-side and the negative terminal to the *n*-side. Suppose that the applied voltage is V. It is apparent that the negative polarity of the supply will reduce the potential barrier V_o by V, as shown in Figure 6.2a. The reason for this is that the bulk regions outside the depletion width have high conductivities due to plenty of majority carriers in the bulk, in comparison with the depletion region in which there are mainly immobile ions. Thus, the applied voltage drops mostly across the depletion width W. Consequently, V directly opposes V_o and the potential barrier against diffusion is reduced to $(V_o - V)$, as depicted in Figure 6.2b. This has drastic consequences because the probability that a hole will surmount this potential barrier and diffuse to the right now becomes proportional to $\exp[-e(V_o - V)/kT]$. In other words, the applied voltage effectively reduces the built-in potential and hence the built-in field, which acts against diffusion. Consequently many holes can now diffuse across the depletion region and enter the *n*-side. This results in the **injection of excess minority carriers,** holes, into the *n*-region. Similarly, excess electrons can now diffuse toward the *p*-side and enter this region and thereby become injected minority carriers

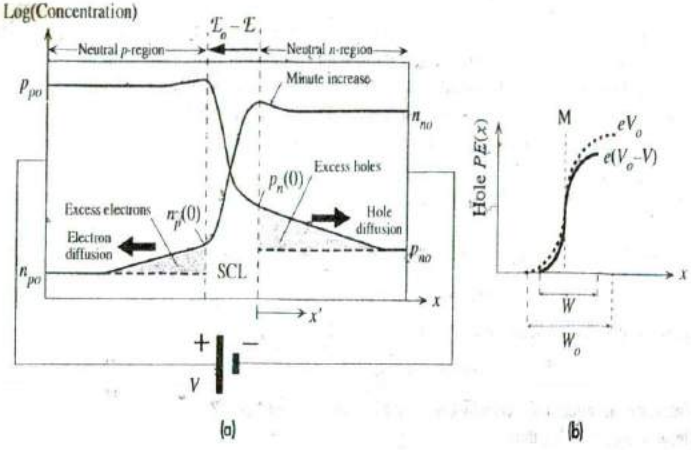


Figure 6.2 Forward-biased *pn* junction and the injection of minority carriers.
(a) Carrier concentration profiles across the device under forward bias.
(b) The hole potential energy with and without an applied bias. *W* is the width of the SCL with forward bias.

The hole concentration

$$p_n(0) = p_n(x' = 0)$$

just outside the depletion region at $x' = 0$ (x' is measured from W_n) is due to the excess of holes diffusing as a result of the reduction in the built-in potential barrier. This concentration $p_n(0)$ is determined by the probability of surmounting the new potential energy barrier $e(V_o - V)$,

$$p_n(0) = p_{po} \exp\left[\frac{e(V_o - V)}{kT}\right] \qquad [6.8]$$

This follows directly from the Boltzmann equation, by virtue of the hole potential energy rising by $e(V_o - V)$ from $x = -W_p$ to $x = W_n$, as indicated in Figure 6.2b, and at the same time the hole concentration falling from p_{po} to $p_n(0)$. By dividing Equation 6.8 by Equation 6.5b, we obtain the effect of the applied voltage directly, which shows how the voltage V determines the amount of excess holes diffusing and arriving at the *n*-region. Equation 6.8 divided by Equation 6.5b is

Law of the junction

$$p_n(0) = p_{no} \exp\left(\frac{eV}{kT}\right) \qquad [6.9]$$

which is called the **law of the junction**. Equation 6.9 is an important equation that we will use again in dealing with *pn* junction devices. It describes the effect of the applied voltage V on the injected minority carrier concentration just outside the depletion region $p_n(0)$. Obviously, with no applied voltage, $V = 0$ and $p_n(0) = p_{no}$, which is exactly what we expect.

Injected holes diffuse in the *n*-region and eventually recombine with electrons in this region as there are many electrons in the *n*-side. Those electrons lost by recombination are readily replenished by the negative terminal of the battery connected to this side. The current due to holes diffusing in the *n*-region can be sustained because more holes can be supplied by the *p*-region, which itself can be replenished by the positive terminal of the battery.

Electrons are similarly injected from the *n*-side to the *p*-side. The electron concentration $n_p(0)$ just outside the depletion region at $x = -W_p$ is given by the equivalent of Equation 6.9 for electrons, that is,

$$n_p(0) = n_{p0} \exp\left(\frac{eV}{kT}\right) \qquad [6.10]$$

Law of the junction

In the *p*-region, the injected electrons diffuse toward the positive terminal looking to be collected. As they diffuse they recombine with some of the many holes in this region. Those holes lost by recombination can be readily replenished by the positive terminal of the battery connected to this side. The current due to the diffusion of electrons in the *p*-side can be maintained by the supply of electrons from the *n*-side, which itself can be replenished by the negative terminal of the battery. It is apparent that an electric current can be maintained through a *pn* junction under forward bias, and that the current flow, surprisingly, seems to be due to the **diffusion of minority carriers.** There is, however, some drift of majority carriers as well.

If the lengths of the *p*- and *n*-regions are longer than the minority carrier diffusion lengths, then we will be justified to expect the hole concentration $p_n(x')$ on the *n*-side to fall exponentially toward the thermal equilibrium value p_{n0}, that is,

$$\Delta p_n(x') = \Delta p_n(0) \exp\left(-\frac{x'}{L_h}\right) \qquad [6.11]$$

Excess minority carrier profile

where

$$\Delta p_n(x') = p_n(x') - p_{n0}$$

Excess minority carrier concentration

is the excess carrier distribution and L_h is the **hole diffusion length,** defined by $L_h = \sqrt{D_h \tau_h}$ in which τ_h is the mean hole recombination lifetime (minority carrier lifetime) in the *n*-region. We base Equation 6.11 on our experience with the minority carrier injection in Chapter 5.[3]

The hole **diffusion current density** $J_{D,\text{hole}}$ is therefore

$$J_{D,\text{hole}} = -eD_h \frac{dp_n(x')}{dx'} = -eD_h \frac{d\Delta p_n(x')}{dx'}$$

that is,

$$J_{D,\text{hole}} = \left(\frac{eD_h}{L_h}\right) \Delta p_n(0) \exp\left(-\frac{x'}{L_h}\right)$$

[3] This is simply the solution of the continuity equation in the absence of an electric field, which is discussed in Chapter 5. Equation 6.11 is identical to Equation 5.48.

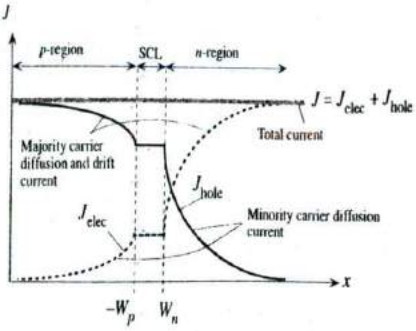


Figure 6.3 The total current anywhere in the device is constant. Just outside the depletion region, it is due to the diffusion of minority carriers.

Although this equation shows that the hole diffusion current depends on location, the total current at any location is the sum of hole and electron contributions, which is independent of x, as indicated in Figure 6.3. The decrease in the minority carrier diffusion current with x' is made up by the increase in the current due to the drift of the majority carriers, as schematically shown in Figure 6.3. The field in the neutral region is not totally zero but a small value, just sufficient to drift the huge number of majority carriers there.

At $x' = 0$, just outside the depletion region, the hole diffusion current is

$$J_{D,\text{hole}} = \left(\frac{eD_h}{L_h}\right)\Delta p_n(0)$$

We can now use the law of the junction to substitute for $\Delta p_n(0)$ in terms of the applied voltage V. Writing

$$\Delta p_n(0) = p_n(0) - p_{no} = p_{no}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

Hole diffusion current in n-side

and substituting in $J_{D,\text{hole}}$, we get

$$J_{D,\text{hole}} = \left(\frac{eD_h p_{no}}{L_h}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

Thermal equilibrium hole concentration p_{no} is related to the donor concentration by

$$p_{no} = \frac{n_i^2}{n_{no}} = \frac{n_i^2}{N_d}$$

Hole diffusion current in n-side

Thus,

$$J_{D,\text{hole}} = \left(\frac{eD_h n_i^2}{L_h N_d}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

There is a similar expression for the electron diffusion current density $J_{D,\text{elec}}$ in the p-region. We will assume (quite reasonably) that the electron and hole currents do not change across the depletion region because, in general, the width of this region is narrow (reality is not quite like the schematic sketches in Figures 6.2 and 6.3). The electron

current at $x = -W_p$ is the same as that at $x = W_n$. The total current density is then simply given by $J_{D,\text{hole}} + J_{D,\text{elec}}$, that is,

$$J = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a}\right) n_i^2 \left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

or

$$J = J_{so}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \qquad [6.12]$$

Ideal diode (Shockley) equation

This is the familiar diode equation with

$$J_{so} = \left[\left(\frac{eD_h}{L_h N_d}\right) + \left(\frac{eD_e}{L_e N_a}\right)\right] n_i^2$$

Reverse saturation current

It is frequently called the **Shockley equation.** The constant J_{so} depends not only on the doping, N_d and N_a, but also on the material via n_i, D_h, D_e, L_h, and L_e. It is known as the **reverse saturation current density,** as explained below. Writing

$$n_i^2 = (N_c N_v) \exp\left(-\frac{eV_g}{kT}\right)$$

Intrinsic concentration

where $V_g = E_g/e$ is the bandgap energy expressed in volts, we can write Equation 6.12 as

$$J = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a}\right)\left[(N_c N_v) \exp\left(-\frac{eV_g}{kT}\right)\right]\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

that is,

$$J = J_1 \exp\left(-\frac{eV_g}{kT}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

or

$$J = J_1 \exp\left[\frac{e(V - V_g)}{kT}\right] \quad \text{for} \quad \frac{eV}{kT} \gg 1 \qquad [6.13]$$

Diode current and bandgap energy

where

$$J_1 = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a}\right)(N_c N_v)$$

is a new constant.

The significance of Equation 6.13 is that it reflects the dependence of *I–V* characteristics on the bandgap (via V_g), as displayed in Figure 6.4 for the three important semiconductors, Ge, Si, and GaAs. Notice that the voltage across the *pn* junction for an appreciable current of say ~0.1 mA is about 0.2 V for Ge, 0.6 V for Si, and 0.9 V for GaAs.

The diode equation, Equation 6.12, was derived by assuming that the lengths of the *p* and *n* regions outside the depletion region are long in comparison with the diffusion lengths L_h and L_e. Suppose that ℓ_p is the length of the *p*-side outside the depletion region

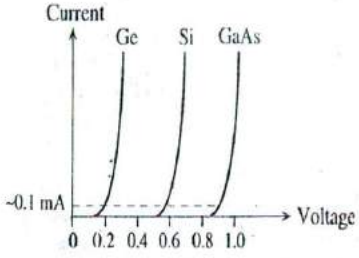


Figure 6.4 Schematic sketch of the *I–V* characteristics of Ge, Si, and GaAs *pn* junctions.

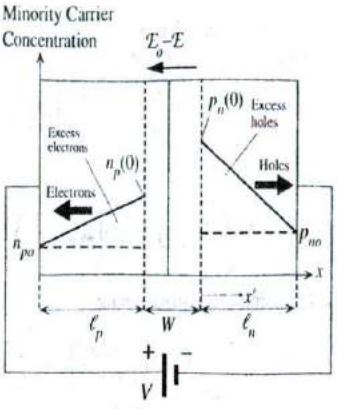


Figure 6.5 Minority carrier injection and diffusion in a short diode.

and ℓ_n is that of the n-side outside the depletion region. If ℓ_p and ℓ_n are shorter than the diffusion lengths L_e and L_h, respectively, then we have what is called a **short diode** and consequently the minority carrier distribution profiles fall almost linearly with distance from the depletion region, as depicted in Figure 6.5. This can be readily proved by solving the continuity equation, but an intuitive explanation makes it clear. At $x' = 0$, the minority carrier concentration is determined by the law of the junction, whereas at the battery terminal there can be no excess carriers as the battery will simply collect these. Since the length of the neutral region is shorter than the diffusion length, there are practically no holes lost by recombination, and therefore the hole flow is expected to be uniform across ℓ_n. This can be so only if the driving force for diffusion, the concentration gradient, is linear.

The excess minority carrier gradient is

$$\frac{d\Delta p_n(x')}{dx'} = -\frac{[p_n(0) - p_{no}]}{\ell_n}$$

The current density $J_{D,\text{hole}}$ due to the injection and diffusion of holes in the n-region as a result of forward bias is

$$J_{D,\text{hole}} = -eD_h \frac{d\Delta p_n(x')}{dx'} = eD_h \frac{[p_n(0) - p_{no}]}{\ell_n}$$

We can now use the law of the junction

$$p_n(0) = p_{no} \exp\left(\frac{eV}{kT}\right)$$

for $p_n(0)$ in the above equation and also obtain a similar equation for electrons diffusing in the p-region and then sum the two for the total current J,

Short diode

$$J = \left(\frac{eD_h}{\ell_n N_d} + \frac{eD_e}{\ell_p N_a}\right) n_i^2 \left[\exp\left(\frac{eV}{kT}\right) - 1\right] \qquad [6.14]$$

It is clear that this expression is identical to that of a long diode, that is, Equation 6.12, if in the latter we replace the diffusion lengths L_h and L_e by the lengths ℓ_n and ℓ_p of the n- and p-regions outside the SCL.

6.1.3 FORWARD BIAS: RECOMBINATION AND TOTAL CURRENT

So far we have assumed that, under a forward bias, the minority carriers diffusing and recombining in the neutral regions are supplied by the external current. However, some of the minority carriers will recombine in the depletion region. The external current must therefore also supply the carriers lost in the recombination process in the SCL. Consider for simplicity a symmetrical *pn* junction as in Figure 6.6 under forward bias. At the metallurgical junction at the center C, the hole and electron concentrations are p_M and n_M and are equal. We can find the SCL recombination current by considering electrons recombining in the p-side in W_p and holes recombining in the n-side in W_n as shown by the shaded areas ABC and BCD, respectively, in Figure 6.6. Suppose that the **mean hole recombination time** in W_n is τ_h and **mean electron recombination time** in W_p is τ_e. The rate at which the electrons in ABC are recombining is the area ABC (nearly all injected electrons) divided by τ_e. The electrons are replenished by the diode current. Similarly, the rate at which holes in BCD are recombining is the area BCD divided by τ_h. Thus, the recombination current density is

$$J_{\text{recom}} = \frac{eABC}{\tau_e} + \frac{eBCD}{\tau_h}$$

We can evaluate the areas ABC and BCD by taking them as triangles, $ABC \approx \frac{1}{2} W_p n_M$, etc., so that

$$J_{\text{recom}} \approx \frac{e\frac{1}{2}W_p n_M}{\tau_e} + \frac{e\frac{1}{2}W_n p_M}{\tau_h}$$

Under steady-state and equilibrium conditions, assuming a nondegenerate semiconductor, we can use Boltzmann statistics to relate these concentrations to the potential

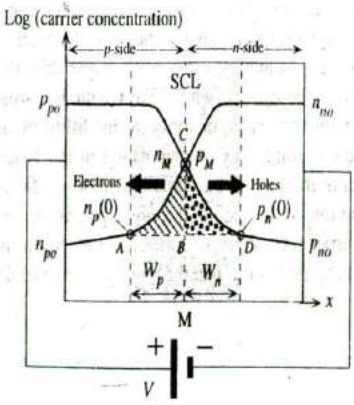


Figure 6.6 Forward-biased *pn* junction and the injection of carriers and their recombination in SCL.

energy. At A, the potential is zero and at M it is $\frac{1}{2}e(V_o - V)$, so

$$\frac{p_M}{p_{po}} = \exp\left[-\frac{e(V_o - V)}{2kT}\right]$$

Since V_o depends on dopant concentrations and n_i as in Equation 6.6 and further $p_{po} = N_a$, we can simplify this equation to

$$p_M = n_i \exp\left(\frac{eV}{2kT}\right)$$

This means that the recombination current for $V > kT/e$ is given by

Recombination current

$$J_{\text{recom}} = \frac{en_i}{2}\left(\frac{W_p}{\tau_e} + \frac{W_n}{\tau_h}\right)\exp\left(\frac{eV}{2kT}\right) \qquad [6.15]$$

From a better quantitative analysis, the expression for the recombination current can be shown to be[4]

Recombination current

$$J_{\text{recom}} = J_{ro}[\exp(eV/2kT) - 1] \qquad [6.16]$$

where J_{ro} is the preexponential constant in Equation 6.15.

Equation 6.15 is the current that supplies the carriers that recombine in the depletion region. The total current into the diode will supply carriers for minority carrier diffusion in the neutral regions and recombination in the space charge layer, so it will be the sum of Equations 6.12 and 6.15.

Total diode current = diffusion + recombination

$$J = J_{so}\exp\left(\frac{eV}{kT}\right) + J_{ro}\exp\left(\frac{eV}{2kT}\right) \qquad \left(V > \frac{kT}{e}\right)$$

This expression is often lumped into a single exponential as

The diode equation

$$J = J_o \exp\left(\frac{eV}{\eta kT}\right) \qquad \left(V > \frac{kT}{e}\right) \qquad [6.17]$$

where J_o is a new constant and η is an **ideality factor,** which is 1 when the current is due to minority carrier diffusion in the neutral regions and 2 when it is due to recombination in the space charge layer. Figure 6.7 shows typical expected I–V characteristics of pn junction Ge, Si, and GaAs diodes. At the highest currents, invariably, the bulk resistances of the neutral regions limit the current (why?). For Ge diodes, typically $\eta = 1$ and the overall I–V characteristics are due to minority carrier diffusion. In the case of GaAs, $\eta \approx 2$ and the current is limited by recombination in the space charge layer. For Si, typically, η changes from 2 to 1 as the current increases, indicating that both processes play an important role. In the case of heavily doped Si diodes, heavy doping leads to short minority carrier recombination times and the current is controlled by recombination in the space charge layer so that the $\eta = 2$ region extends all the way to the onset of bulk resistance limitation.

[4] This is generally proved in advanced texts.

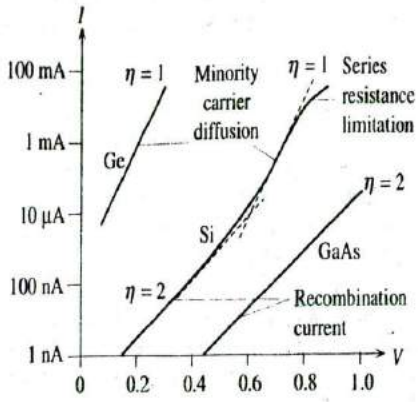


Figure 6.7 Schematic sketch of typical *I–V* characteristics of Ge, Si, and GaAs *pn* junctions as log(*I*) versus *V*.
The slope indicates $e/(\eta kT)$.

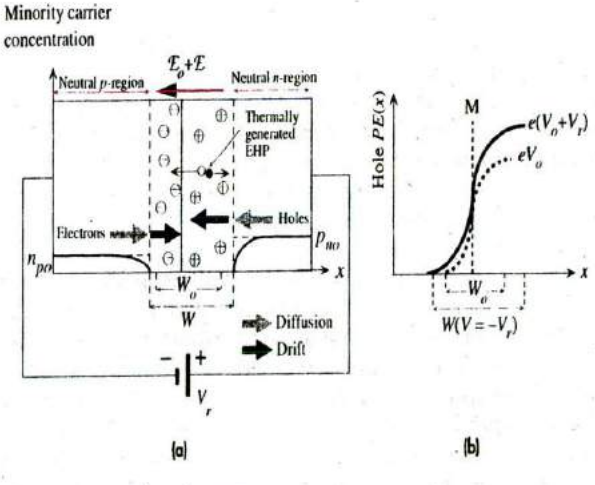


Figure 6.8 Reverse-biased *pn* junction.
(a) Minority carrier profiles and the origin of the reverse current.
(b) Hole *PE* across the junction under reverse bias.

6.1.4 REVERSE BIAS

When a *pn* junction is reverse-biased, as shown in Figure 6.8a, the applied voltage, as before, drops mainly across the depletion region, that is, the space charge layer (SCL), which becomes wider. The negative terminal will attract the holes in the *p*-side to move away from the SCL, which results in more exposed negative acceptor ions and thus a wider SCL. Similarly, the positive terminal will attract electrons away from the SCL, which exposes more positively charged donors. The depletion width on the *n*-side also widens. The movement of electrons in the *n*-region toward the positive battery

terminal cannot be sustained because there is no electron supply to this n-side. The p-side cannot supply electrons to the n-side because it has almost none. However, there is a small reverse current due to two causes.

The applied voltage increases the built-in potential barrier, as depicted in Figure 6.8b. The electric field in the SCL is larger than the built-in internal field $\mathcal{E}_o$. The small number of holes on the n-side near the SCL become extracted and swept by the field across the SCL over to the p-side. This small current can be maintained by the diffusion of holes from the n-side bulk to the SCL boundary.

Assume that the reverse bias $V_r > kT/e = 25$ mV. The hole concentration $p_n(0)$ just outside the SCL is nearly zero by the law of the junction, Equation 6.9, whereas the hole concentration in the bulk (or near the negative terminal) is the equilibrium concentration p_{no}, which is small. There is therefore a small concentration gradient and hence a small hole diffusion current toward the SCL as shown in Figure 6.8a. Similarly, there is a small electron diffusion current from bulk p-side to the SCL. Within the SCL, these carriers are drifted by the field. This minority carrier diffusion current is essentially the Shockley model. The reverse current is given by Equation 6.12 with a negative voltage which leads to a diode current density of $-J_{so}$ called the **reverse saturation current density.** The value of J_{so} depends only on the material via n_i, μ_h, μ_e, dopant concentrations, but not on the voltage ($V_r > kT/e$). Furthermore, as J_{so} depends on n_i^2, it is strongly temperature dependent. In some books it is stated that the causes of reverse current are the thermal generation of minority carriers in the neutral region within a diffusion length to the SCL, the diffusion of these carriers to the SCL, and their subsequent drift through the SCL. This description, in essence, is identical to the Shockley model we just described.

The thermal generation of electron–hole pairs (EHPs) in the SCL, as shown in Figure 6.8a, can also contribute to the observed reverse current since the internal field in this layer will separate the electron and hole and drift them toward the neutral regions. This drift will result in an external current in addition to the reverse current due to the diffusion of minority carriers. The theoretical evaluation of SCL generation current involves an in-depth knowledge of the charge carrier generation processes via recombination centers, which is discussed in advanced texts. Suppose that τ_g is the **mean time to generate an electron–hole pair** by virtue of the thermal vibrations of the lattice; τ_g is also called the **mean thermal generation time.** Given τ_g, the rate of thermal generation per unit volume must be n_i/τ_g because it takes on average τ_g seconds to create n_i number of EHPs per unit volume. Furthermore, since WA, where A is the cross-sectional area, is the volume of the depletion region, the rate of EHP, or charge carrier, generation is $(AWn_i)/\tau_g$. Both holes and electrons drift in the SCL each contributing equally to the current. The observed current density must be $e(Wn_i)/\tau_g$. Therefore the reverse current density component due to thermal generation of EHPs within the SCL should be given by

EHP thermal generation in SCL

$$J_{gen} = \frac{eWn_i}{\tau_g} \qquad [6.18]$$

The reverse bias widens the width W of the depletion layer and hence increases J_{gen}. The total reverse current density J_{rev} is the sum of the diffusion and generation

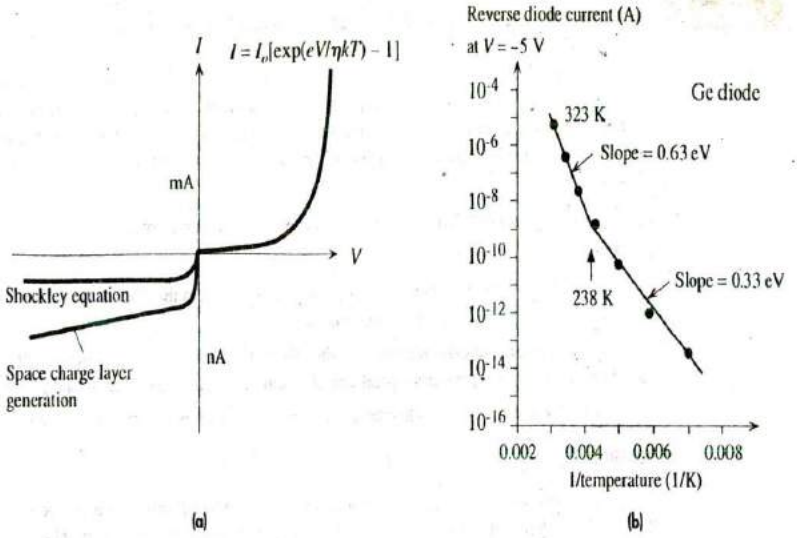


Figure 6.9

(a) Forward and reverse *I–V* characteristics of a *pn* junction (the positive and negative current axes have different scales and hence the discontinuity at the origin).

(b) Reverse diode current in a Ge *pn* junction as a function of temperature in a $\ln(I_{rev})$ versus $1/T$ plot. Above 238 K, I_{rev} is controlled by n_i^2, and below 238 K, it is controlled by n_i. The vertical axis is a logarithmic scale with actual current values.

SOURCE: (b) From D. Scansen and S. O. Kasap, *Cnd. J. Physics*, **70,** 1070, 1992.

components,

$$J_{rev} = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a}\right) n_i^2 + \frac{eWn_i}{\tau_g} \qquad [6.19]$$

Total reverse current

which is shown schematically in Figure 6.9a. The thermal generation component J_{gen} in Equation 6.18 increases with reverse bias V_r because the SCL width W increases with V_r.

The terms in the reverse current in Equation 6.19 are predominantly controlled by n_i^2 and n_i. Their relative importance depends not only on the semiconductor properties but also on the temperature since $n_i \propto \exp(-E_g/2kT)$. Figure 6.9b shows the reverse current I_{rev} in dark in a Ge *pn* junction (a photodiode) plotted as $\ln(I_{rev})$ versus $1/T$ to highlight the two different processes in Equation 6.19. The measurements in Figure 6.9b show that above 238 K, I_{rev} is controlled by n_i^2 because the slope of $\ln(I_{rev})$ versus $1/T$ yields an E_g of approximately 0.63 eV, close to the expected E_g of about 0.66 eV in Ge. Below 238 K, I_{rev} is controlled by n_i because the slope of $\ln(I_{rev})$ versus $1/T$ is equivalent to $E_g/2$ of approximately 0.33 eV. In this range, the reverse current is due to EHP generation in the SCL via defects and impurities (recombination centers).

EXAMPLE 6.4 **FORWARD- AND REVERSE-BIASED Si DIODE** An abrupt Si p^+n junction diode has a cross-sectional area of 1 mm^2, an acceptor concentration of 5×10^{18} boron atoms cm^{-3} on the p-side, and a donor concentration of 10^{16} arsenic atoms cm^{-3} on the n-side. The lifetime of holes in the n-region is 417 ns, whereas that of electrons in the p-region is 5 ns due to a greater concentration of impurities (recombination centers) on that side. Mean thermal generation lifetime (τ_g) is about 1 μs. The lengths of the p- and n-regions are 5 and 100 microns, respectively.

a. Calculate the minority diffusion lengths and determine what type of a diode this is.

b. What is the built-in potential across the junction?

c. What is the current when there is a forward bias of 0.6 V across the diode at 27 °C? Assume that the current is by minority carrier diffusion.

d. Estimate the forward current at 100 °C when the voltage across the diode remains at 0.6 V. Assume that the temperature dependence of n_i dominates over those of D, L, and μ.

e. What is the reverse current when the diode is reverse-biased by a voltage $V_r = 5$ V?

SOLUTION

The general expression for the diffusion length is $L = \sqrt{D\tau}$ where D is the diffusion coefficient and τ is the carrier lifetime. D is related to the carrier mobility μ via the Einstein relationship $D/\mu = kT/e$. We therefore need to know μ to calculate D and hence L. Electrons diffuse in the p-region and holes in the n-region, so we need μ_e in the presence of N_a acceptors and μ_h in the presence of N_d donors. From the drift mobility, μ versus dopant concentration in Figure 5.19, we have the following:

With $N_a = 5 \times 10^{18}$ cm^{-3} $\quad \mu_e \approx 120$ cm^2 V^{-1} s^{-1}

With $N_d = 10^{16}$ cm^{-3} $\quad \mu_h \approx 440$ cm^2 V^{-1} s^{-1}

Thus

$$D_e = \frac{kT\mu_e}{e} \approx (0.0259 \text{ V})(120 \text{ cm}^2 \text{ V}^{-1} \text{s}^{-1}) = 3.10 \text{ cm}^2 \text{ s}^{-1}$$

$$D_h = \frac{kT\mu_h}{e} \approx (0.0259 \text{ V})(440 \text{ cm}^2 \text{ V}^{-1} \text{s}^{-1}) = 11.39 \text{ cm}^2 \text{ s}^{-1}$$

Diffusion lengths are

$$L_e = \sqrt{D_e\tau_e} = \sqrt{[(3.10 \text{ cm}^2 \text{ s}^{-1})(5 \times 10^{-9} \text{ s})]}$$

$$= 1.2 \times 10^{-4} \text{ cm} \quad \text{or} \quad 1.2 \text{ μm} < 5 \text{ μm}$$

$$L_h = \sqrt{D_h\tau_h} = \sqrt{[(11.39 \text{ cm}^2 \text{ s}^{-1})(417 \times 10^{-9} \text{ s})]}$$

$$= 21.8 \times 10^{-4} \text{ cm} \quad \text{or} \quad 21.8 \text{ μm} < 100 \text{ μm}$$

We therefore have a long diode. The built-in potential is

$$V_a = \left(\frac{kT}{e}\right) \ln\left(\frac{N_d N_a}{n_i^2}\right) = (0.0259 \text{ V}) \ln\left[\frac{(5 \times 10^{18} \times 10^{16})}{(1.0 \times 10^{10})^2}\right] = 0.877 \text{ V}$$

To calculate the forward current when $V = 0.6$ V, we need to evaluate both the diffusion and recombination components to the current. It is likely that the diffusion component will exceed the recombination component at this forward bias (this can be easily verified). Assuming

that the forward current is due to minority carrier diffusion in neutral regions,

$$I = I_{so}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \approx I_{so}\exp\left(\frac{eV}{kT}\right) \qquad \text{for } V \gg \frac{kT}{e} \qquad (= 0.0259 \text{ V})$$

where

$$I_{so} = AJ_{so} = Aen_i^2\left[\left(\frac{D_h}{L_h N_d}\right) + \left(\frac{D_e}{L_e N_a}\right)\right] \approx \frac{Aen_i^2 D_h}{L_h N_d}$$

as $N_a \gg N_d$. In other words, the current is mainly due to the diffusion of holes in the *n*-region. Thus,

$$I_{so} = \frac{(0.01 \text{ cm}^2)(1.6 \times 10^{-19} \text{ C})(1.0 \times 10^{10} \text{ cm}^{-3})^2(11.39 \text{ cm}^2 \text{ s}^{-1})}{(21.8 \times 10^{-4} \text{ cm})(10^{16} \text{ cm}^{-3})}$$

$$= 8.36 \times 10^{-14} \text{ A} \quad \text{or} \quad 0.084 \text{ pA}$$

Then the diode current is

$$I \approx I_{so}\exp\left(\frac{eV}{kT}\right) = (8.36 \times 10^{-14} \text{ A})\exp\left[\frac{(0.6 \text{ V})}{(0.0259 \text{ V})}\right]$$

$$= 0.96 \times 10^{-3} \text{ A} \quad \text{or} \quad 0.96 \text{ mA}$$

We note that when a forward bias of 0.6 V is applied, the built-in potential is reduced from 0.877 V to 0.256 V, which encourages minority carrier injection, that is, diffusion of holes from *p*- to *n*-side and electrons from *n*- to *p*-side. To find the current at 100 °C, first we assume that $I_{so} \propto n_i^2$. Then at $T = 273 + 100 = 373$ K, $n_i \approx 1.0 \times 10^{12}$ cm^{-3} (approximately from n_i versus $1/T$ graph in Figure 5.16), so

$$I_{so}(373 \text{ K}) \approx I_{so}(300 \text{ K})\left[\frac{n_i(373 \text{ K})}{n_i(300 \text{ K})}\right]^2$$

$$\approx (8.36 \times 10^{-14})\left(\frac{1.0 \times 10^{12}}{1.0 \times 10^{10}}\right)^2 = 8.36 \times 10^{-10} \text{ A} \quad \text{or} \quad 0.836 \text{ nA}$$

At 100 °C, the forward current with 0.6 V across the diode is

$$I = I_{so}\exp\left(\frac{eV}{kT}\right) = (8.36 \times 10^{-10} \text{ A})\exp\left[\frac{(0.6 \text{ V})(300 \text{ K})}{(0.0259 \text{ V})(373 \text{ K})}\right] = 0.10 \text{ A}$$

When a reverse bias of V_r is applied, the potential difference across the depletion region becomes $V_o + V_r$ and the width W of the depletion region is

$$W = \left[\frac{2\varepsilon(V_o + V_r)}{eN_d}\right]^{1/2} = \left[\frac{2(11.9)(8.85 \times 10^{-12})(0.877 + 5)}{(1.6 \times 10^{-19})(10^{22})}\right]^{1/2}$$

$$= 0.88 \times 10^{-6} \text{ m} \quad \text{or} \quad 0.88 \text{ μm}$$

The thermal generation current with $V_r = 5$ V is

$$I_{gen} = \frac{eAWn_i}{\tau_g} = \frac{(1.6 \times 10^{-19} \text{ C})(0.01 \text{ cm}^2)(0.88 \times 10^{-4} \text{ cm})(1.0 \times 10^{10} \text{ cm}^{-3})}{(10^{-6} \text{ s})}$$

$$= 1.41 \times 10^{-9} \text{ A} \quad \text{or} \quad 1.4 \text{ nA}$$

This thermal generation current is much greater than the reverse saturation current I_{so} (= 0.084 pA). The reverse current is therefore dominated by I_{gen} and it is 1.4 nA.

6.2 *pn* JUNCTION BAND DIAGRAM

6.2.1 OPEN CIRCUIT

Figure 6.10a shows the energy band diagrams for a *p*-type and an *n*-type semiconductor of the same material (same E_g) when the semiconductors are isolated from each other. In the *p*-type material the Fermi level E_{Fp} is Φ_p below the vacuum level and is close to E_v. In the *n*-type material the Fermi level E_{Fn} is Φ_n below the vacuum level and is close to E_c. The separation $E_c - E_{Fn}$ determines the electron concentration n_{no} in the *n*-type and $E_{Fp} - E_v$ determines the hole concentration p_{po}, in the *p*-type semiconductor under thermal equilibrium conditions.

An important property of the Fermi energy E_F is that in a system in equilibrium, the Fermi level must be spatially continuous. A difference in Fermi levels ΔE_F is equivalent to electrical work eV, which is either done on the system or extracted from the system. When the two semiconductors are brought together, as in Figure 6.10b, the Fermi level must be uniform through the two materials and the junction at M, which marks the position of the metallurgical junction. Far away from M, in the bulk of the *n*-type semiconductor, we should still have an *n*-type semiconductor and $E_c - E_{Fn}$ should be the same as before. Similarly, $E_{Fp} - E_v$ far away from M inside the *p*-type material should also be the same as before. These features are sketched in Figure 6.10b keeping E_{Fp} and E_{Fn} the same through the whole system and, of course, keeping the bandgap $E_c - E_v$ the same. Clearly, to draw the energy band diagram, we have to bend the bands E_c and E_v around the junction at M because E_c on the *n*-side is close to E_{Fn} whereas on the *p*-side it is far away from E_{Fp}. How do bands bend and what does it mean?

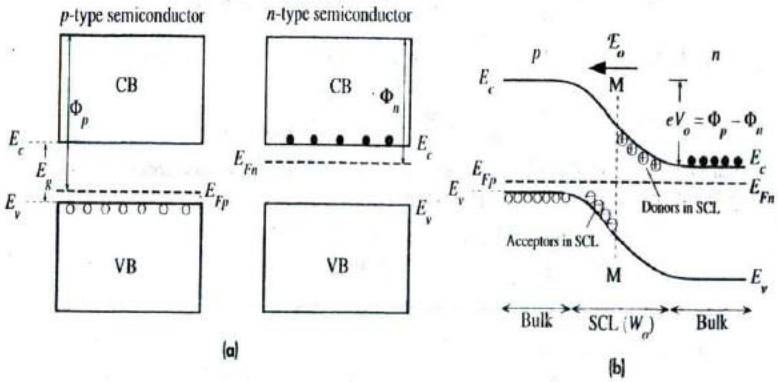


Figure 6.10

(a) Two isolated *p*- and *n*-type semiconductors (same material).

(b) A *pn* junction band diagram when the two semiconductors are in contact. The Fermi level must be uniform in equilibrium. The metallurgical junction is at M. The region around M contains the space charge layer (SCL). On the *n*-side of M, SCL has the exposed positively charged donors, whereas on the *p*-side it has the exposed negatively charged acceptors.

As soon as the two semiconductors are brought together to form the junction, electrons diffuse from the n-side to the p-side and as they do so they deplete the n-side near the junction. Thus E_c must move away from E_{Fn} toward M, which is exactly what is sketched in Figure 6.10b. Holes diffuse from the p-side to the n-side and the loss of holes in the p-type material near the junction means that E_v moves away from E_{Fp} toward M, which is also in the figure.

Furthermore, as electrons and holes diffuse toward each other, most of them recombine and disappear around M, which leads to the formation of a depletion region or the space charge layer, as we saw in Figure 6.1. The electrostatic potential energy (PE) of the electron decreases from 0 inside the p-region to $-eV_o$ inside the n-region, as shown in Figure 6.1g. The total energy of the electron must therefore decrease going from the p- to the n-region by an amount eV_o. In other words, the electron in the n-side at E_c must overcome a PE barrier to go over to E_c in the p-side. This PE barrier is eV_o, where V_o is the built-in potential that we evaluated in Section 6.1. Band bending around M therefore accounts not only for the variation of electron and hole concentrations in this region but also for the effect of the built-in potential (and hence the built-in field as the two are related).

In Figure 6.10b we have also schematically sketched in the positive donor (at E_d) and the negative acceptor (at E_a) charges in the SCL around M to emphasize that there are exposed charges near M. These charges are, of course, immobile and, generally, they are not shown in band diagrams. It should be noted that in the SCL region, marked as W_o, the Fermi level is close to neither E_c nor E_v, compared with the bulk semiconductor regions. This means that both n and p in this zone are much less than their bulk values n_{no} and p_{po}. The metallurgical junction zone has been depleted of carriers compared with the bulk. Any applied voltage must therefore drop across the SCL.

6.2.2 FORWARD AND REVERSE BIAS

The energy band diagram of the pn junction under open circuit conditions is shown in Figure 6.11a. There is no net current, so the diffusion current of electrons from the n- to p-side is balanced by the electron drift current from the p- to n-side driven by the built-in field $\mathcal{E}_o$. Similar arguments apply to holes. The probability that an electron diffuses from E_c in the n-side to E_c in the p-side determines the diffusion current density J_{diff}. The probability of overcoming the PE barrier is proportional to $\exp(-eV_o/kT)$. Therefore, under zero bias,

$$J_{\text{diff}}(0) = B\exp\left(-\frac{eV_o}{kT}\right) \qquad [6.20]$$

$$J_{\text{net}}(0) = J_{\text{diff}}(0) + J_{\text{drift}}(0) = 0 \qquad [6.21]$$

where B is a proportionality constant and $J_{\text{drift}}(0)$ is the current due to the drift of electrons by $\mathcal{E}_o$. Clearly $J_{\text{drift}}(0) = -J_{\text{diff}}(0)$; that is, drift is in the opposite direction to diffusion.

When the pn junction is forward-biased, the majority of the applied voltage drops across the depletion region, so the applied voltage is in opposition to the built-in potential V_o. Figure 6.11b shows the effect of forward bias, which is to reduce the PE

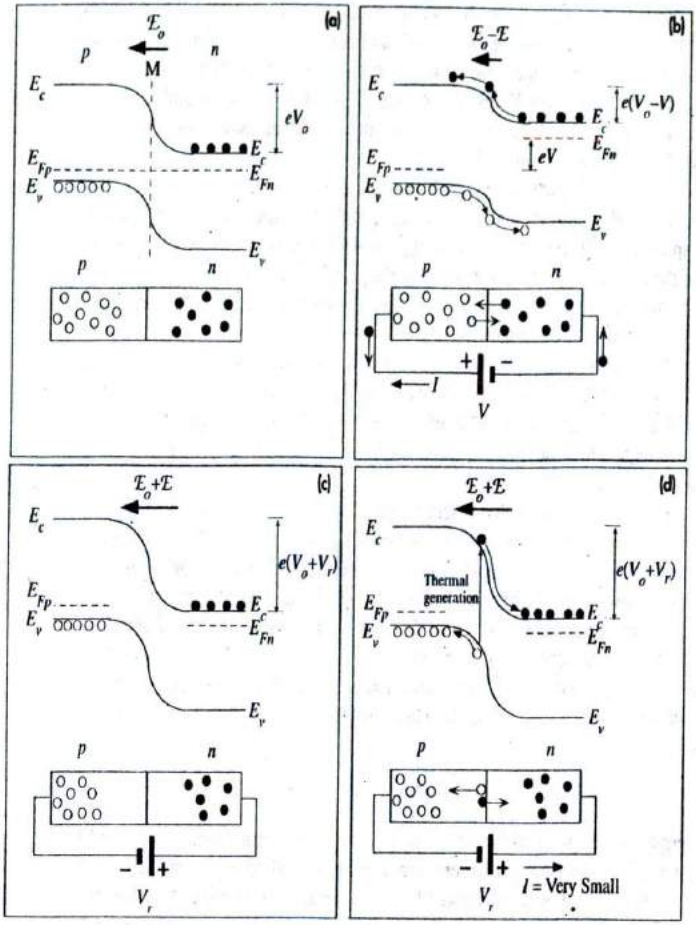


Figure 6.11 Energy band diagrams for a *pn* junction: (a) open circuit, (b) forward bias, (c) reverse bias conditions, (d) thermal generation of electron–hole pairs in the depletion region results in a small reverse current.

barrier from eV_o to $e(V_o - V)$. The electrons at E_c in the *n*-side can now readily overcome the *PE* barrier and diffuse to the *p*-side. The diffusing electrons from the *n*-side can be replenished easily by the negative terminal of the battery connected to this side. Similarly holes can now diffuse from the *p*- to *n*-side. The positive terminal of the battery can replenish those holes diffusing away from the *p*-side. There is therefore a current flow through the junction and around the circuit.

The probability that an electron at E_c in the *n*-side overcomes the new *PE* barrier and diffuses to E_c in the *p*-side is now proportional to $\exp[-e(V_o - V)/kT]$. The latter increases enormously even for small forward voltages. The new diffusion current due

to electrons diffusing from the n- to p-side is

$$J_{\text{diff}}(V) = B \exp\left[-\frac{e(V_o - V)}{kT}\right]$$

There is still a drift current due to electrons being drifted by the new field $\mathcal{E}_o - \mathcal{E}$ ($\mathcal{E}$ is the applied field) in the SCL. This drift current now has the value $J_{\text{drift}}(V)$. The net current is the diode current under forward bias

$$J = J_{\text{diff}}(V) + J_{\text{drift}}(V)$$

$J_{\text{drift}}(V)$ is difficult to evaluate. As a first approximation we can assume that although $\mathcal{E}_o$ has decreased to $\mathcal{E}_o - \mathcal{E}$, there is, however, an increase in the electron concentration in the SCL due to diffusion so that we can approximately take $J_{\text{drift}}(V)$ to remain the same as $J_{\text{drift}}(0)$. Thus

$$J \approx J_{\text{diff}}(V) + J_{\text{drift}}(0) = B \exp\left[-\frac{e(V_o - V)}{kT}\right] - B \exp\left(-\frac{eV_o}{kT}\right)$$

Factoring leads to

$$J \approx B \exp\left(-\frac{eV_o}{kT}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

We should also add to this the hole contribution, which has a similar form with a different constant B. The diode current–voltage relationship then becomes the familiar diode equation,

$$J = J_o\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

pn Junction I–V characteristics

where J_o is a temperature-dependent constant.[5]

When a reverse bias, $V = -V_r$, is applied to the pn junction, the voltage again drops across the SCL. In this case, however, V_r adds to the built-in potential V_o, so the PE barrier becomes $e(V_o + V_r)$, as shown in Figure 6.11c. The field in the SCL at M increases to $\mathcal{E}_o + \mathcal{E}$, where $\mathcal{E}$ is the applied field.

The diffusion current due to electrons diffusing from E_c in the n-side to E_c in the p-side is now almost negligible because it is proportional to $\exp[-e(V_o + V_r)/kT]$, which rapidly becomes very small with V_r. There is, however, a small reverse current arising from the drift component. When an electron–hole pair (EHP) is thermally generated in the SCL, as shown in Figure 6.11d, the field here separates the pair. The electron falls down the PE hill, down to E_c, in the n-side to be collected by the battery. Similarly the hole falls down its own PE hill (energy increases downward for holes) to make it to the p-side. The process of falling down a PE hill is the same process as being driven by a field, in this case by $\mathcal{E}_o + \mathcal{E}$. Under reverse bias conditions, there is therefore a small reverse current that depends on the rate of thermal generation of EHPs in the SCL. An electron in the p-side that is thermally generated within a diffusion length

[5] The derivation is similar to that for the Schottky diode, but there were more assumptions here.

L_e to the SCL can diffuse to the SCL and consequently can become drifted by the field, that is, roll down the PE hill in Figure 6.11d. Such minority carrier thermal generation in neutral regions can also give rise to a small reverse current.

EXAMPLE 6.5 **THE BUILT-IN VOLTAGE V_o FROM THE ENERGY BAND DIAGRAM** The energy band treatment allows a simple way to calculate V_o. When the junction is formed in Figure 6.10 from a to b, E_{Fp} and E_{Fn} must shift and line up. Using the energy band diagrams in this figure and semiconductor equations for n and p, derive an expression for the built-in voltage V_o in terms of the material and doping properties N_d, N_a, and n_i.

SOLUTION

The shift in E_{Fp} and E_{Fn} to line up is clearly $\Phi_p - \Phi_n$, the work function difference. Thus the PE barrier eV_o is $\Phi_p - \Phi_n$. From Figure 6.10, we have

$$eV_o = \Phi_p - \Phi_n = (E_c - E_{Fp}) - (E_c - E_{Fn})$$

But on the p- and n-sides, the electron concentrations in thermal equilibrium are given by

$$n_{po} = N_c \exp\left[-\frac{(E_c - E_{Fp})}{kT}\right]$$

$$n_{no} = N_c \exp\left[-\frac{(E_c - E_{Fn})}{kT}\right]$$

From these equations, we can now substitute for $(E_c - E_{Fp})$ and $(E_c - E_{Fn})$ in the expression for eV_o. The N_c cancel and we obtain

$$eV_o = kT \ln\left(\frac{n_{no}}{n_{po}}\right)$$

Since $n_{po} = n_i^2/N_a$ and $n_{no} = N_d$, we readily obtain the built-in potential V_o.

Built-in voltage

$$V_o = \left(\frac{kT}{e}\right) \ln\left[\frac{(N_a N_d)}{n_i^2}\right]$$

6.3 DEPLETION LAYER CAPACITANCE OF THE *pn* JUNCTION

It is apparent that the depletion region of a *pn* junction has positive and negative charges separated over a distance W similar to a parallel plate capacitor. The stored charge in the depletion region, however, unlike the case of a parallel plate capacitor, does not depend linearly on the voltage. It is useful to define an incremental capacitance that relates the incremental charge stored to an incremental voltage change across the *pn* junction.

The width of the depletion region is given by

Depletion region width

$$W = \left[\frac{2\varepsilon(N_a + N_d)(V_o - V)}{eN_a N_d}\right]^{1/2} \qquad [6.22]$$

where, for forward bias, V is positive, which reduces V_o, and, for reverse bias, V is negative, so V_o is increased. We are interested in obtaining the capacitance of the

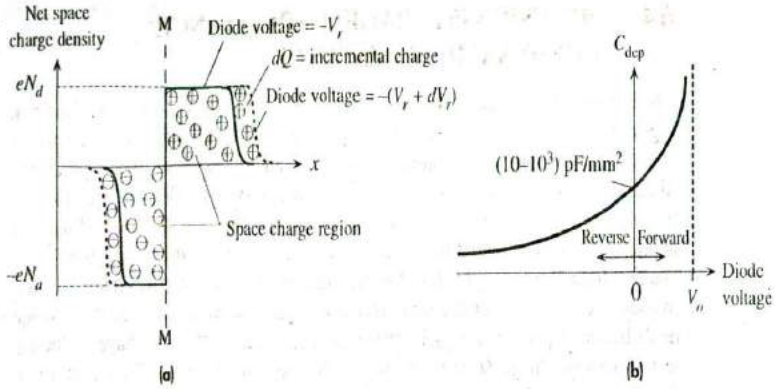


Figure 6.12 The depletion region behaves like a capacitor.

(a) The charge in the depletion region depends on the applied voltage just as in a capacitor. A reverse bias example is shown.

(b) The incremental capacitance of the depletion region increases with forward bias and decreases with reverse bias. Its value is typically in the range of picofarads per mm^2 of device area.

depletion region under dynamic conditions, that is, when V is a function of time. When the applied voltage V changes by dV, to $V + dV$, then W also changes via Equation 6.22, and as a result, the amount of charge in the depletion region becomes $Q + dQ$, as shown in Figure 6.12a for the reverse bias case, that is, $V = -V_r$ and $dV = -dV_r$. The **depletion layer capacitance** C_{dep} is defined by

Definition of depletion layer capacitance

$$C_{dep} = \left| \frac{dQ}{dV} \right| \qquad [6.23]$$

where the amount of charge (on any one side of the depletion layer) is

$$|Q| = eN_d W_n A = eN_a W_p A$$

and $W = W_n + W_p$. We can therefore substitute for W in Equation 6.22 in terms of Q and then differentiate it to obtain dQ/dV. The final result for the depletion capacitance is

Depletion capacitance

$$C_{dep} = \frac{\varepsilon A}{W} = \frac{A}{(V_o - V)^{1/2}} \left[\frac{e\varepsilon (N_a N_d)}{2(N_a + N_d)} \right]^{1/2} \qquad [6.24]$$

We should note that C_{dep} is given by the same expression as that for the parallel plate capacitor, $\varepsilon A / W$, but with W being voltage dependent by virtue of Equation 6.22. The $C_{dep} - V$ behavior is sketched in Figure 6.12b. Notice that C_{dep} decreases with increasing reverse bias, which is expected since the separation of the charges increases via $W \propto (V_o + V_r)^{1/2}$. The capacitance C_{dep} is present under both forward and reverse bias conditions.

The voltage dependence of the depletion capacitance is utilized in **varactor diodes** (varicaps), which are employed as voltage-dependent capacitors in tuning circuits. A varactor diode is reverse biased to prevent conduction, and its depletion capacitance is varied by the magnitude of the reverse bias.

6.4 DIFFUSION (STORAGE) CAPACITANCE AND DYNAMIC RESISTANCE

The diffusion or storage capacitance arises under forward bias only. As shown in Figure 6.2a, when the p^+n junction is forward biased, we have stored a positive charge on the n-side by the continuous injection and diffusion of minority carriers. Similarly, a negative charge has been stored on the p^+-side by electron injection, but the magnitude of this negative charge is small for the p^+n junction. When the applied voltage is increased from V to $V + dV$, as shown in Figure 6.13, then $p_n(0)$ changes from $p_n(0)$ to $p_n'(0)$. If dQ is the additional minority carrier charge injected into the n-side, as a result of a small increase dV in V, then the incremental **storage** or **diffusion capacitance** C_{diff} is defined as $C_{\text{diff}} = dQ/dV$. At voltage V, the injected positive charge Q on the n-side is disappearing by recombination at a rate Q/τ_h, where τ_h is the minority carrier lifetime. The diode current I is therefore Q/τ_h, from which

Injected minority carrier charge

$$Q = \tau_h I = \tau_h I_o \left[\exp\left(\frac{eV}{kT}\right) - 1\right] \qquad [6.25]$$

Thus,

Diffusion capacitance

$$C_{\text{diff}} = \frac{dQ}{dV} = \frac{\tau_h e I}{kT} = \frac{\tau_h I\,(\text{mA})}{25} \qquad [6.26]$$

where we used $e/kT \approx 1/0.025$ at room temperature. Generally the value of the diffusion capacitance, typically in the nanofarads range, far exceeds that of the depletion layer capacitance.

Suppose that the voltage V across the diode is increased by an infinitesimally small amount dV, as shown in an exaggerated way in Figure 6.14. This gives rise to a small increase dI in the diode current. We define the **dynamic** or **incremental resistance** r_d of the diode as dV/dI, so

Dynamic/ incremental resistance

$$r_d = \frac{dV}{dI} = \frac{kT}{eI} = \frac{25}{I\,(\text{mA})} \qquad [6.27]$$

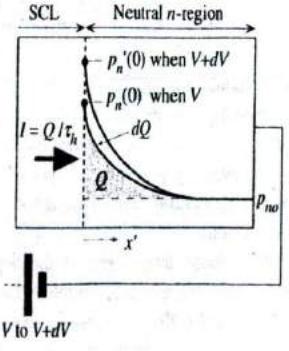


Figure 6.13 Consider the injection of holes into the n-side during forward bias.

Storage or diffusion capacitance arises because when the diode voltage increases from V to $V + dV$, more minority carriers are injected and more minority carrier charge is stored in the n-region.

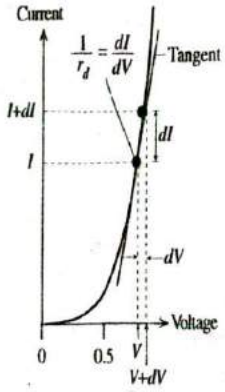


Figure 6.14 The dynamic resistance of the diode is defined as dV/dI, which is the inverse of the tangent at I.

The dynamic resistance is therefore the inverse of the slope of the I–V characteristics at a point and hence depends on the current I. It relates the changes in the diode current and voltage arising from the **diode action** alone, by which we mean the modulation of the rate of minority carrier diffusion by the diode voltage. We could have equivalently defined a dynamic conductance by

$$g_d = \frac{dI}{dV} = \frac{1}{r_d}$$

Dynamic conductance

From Equations 6.26 and 6.27 we have

$$r_d C_{\text{diff}} = \tau_h \qquad [6.28]$$

The dynamic resistance r_d and diffusion capacitance C_{diff} of a diode determine its response to small ac signals under forward bias conditions. By *small* we usually mean voltages smaller than the thermal voltage kT/e or 25 mV at room temperature. For small ac signals we can simply represent a forward-biased diode as a resistance r_d in parallel with a capacitance C_{diff}.

EXAMPLE 6.6

INCREMENTAL RESISTANCE AND CAPACITANCE An abrupt Si p^+n junction diode of cross-sectional area (A) 1 mm^2 with an acceptor concentration of 5×10^{18} boron atoms cm^{-3} on the p-side and a donor concentration of 10^{16} arsenic atoms cm^{-3} on the n-side is forward-biased to carry a current of 5 mA. The lifetime of holes in the n-region is 417 ns, whereas that of electrons in the p-region is 5 ns. What are the small-signal ac resistance, incremental storage, and depletion capacitances of the diode?

SOLUTION

This is the same diode we considered in Example 6.4 for which the built-in potential was 0.877 V and $I_{so} = 0.0836$ pA. The current through the diode is 5 mA. Thus

$$I = I_{so}\exp\left(\frac{eV}{kT}\right) \quad \text{or} \quad V = \left(\frac{kT}{e}\right)\ln\left(\frac{I}{I_{so}}\right) = (0.0259)\ln\left(\frac{5 \times 10^{-3}}{0.0836 \times 10^{-12}}\right) = 0.643 \text{ V}$$

The dynamic diode resistance is given by

$$r_d = \frac{25}{I(\text{mA})} = \frac{25}{5} = 5\ \Omega$$

The depletion capacitance per unit area with $N_a \gg N_d$ is

$$C_{dep} = A\left[\frac{e\varepsilon(N_a N_d)}{2(N_a + N_d)(V_o - V)}\right]^{1/2} \approx A\left[\frac{e\varepsilon N_d}{2(V_o - V)}\right]^{1/2}$$

At $V = 0.643$ V, with $V_o = 0.877$ V, $N_d = 10^{22}$ m^{-3}, $\varepsilon_r = 11.9$, and $A = 10^{-6}$ m^2, the above equation gives

$$C_{dep} = 10^{-6}\left[\frac{(1.6 \times 10^{-19})(11.9)(8.85 \times 10^{-12})(10^{22})}{2(0.877 - 0.643)}\right]^{1/2}$$

$$= 6.0 \times 10^{-10}\ \text{F} \quad \text{or} \quad 600\ \text{pF}$$

The incremental diffusion capacitance C_{diff} due to holes injected and stored in the n-region is

$$C_{diff} = \frac{\tau_h I(\text{mA})}{25} = \frac{(417 \times 10^{-9})(5)}{25} = 8.3 \times 10^{-8}\ \text{F} \quad \text{or} \quad 83\ \text{nF}$$

Clearly the diffusion capacitance (83 nF) that arises during forward bias completely overwhelms the depletion capacitance (600 pF).

We note that there is also a diffusion capacitance due to electrons injected and stored in the p-region. However, electron lifetime in the p-region is very short (here 5 ns), so the value of this capacitance is much smaller than that due to holes in the n-region. In calculating the diffusion capacitance, we normally consider the minority carriers that have the longest recombination lifetime, here τ_h. These are the carriers that take a long time to disappear by recombination when the bias is suddenly switched off.

6.5 REVERSE BREAKDOWN: AVALANCHE AND ZENER BREAKDOWN

The reverse voltage across a *pn* junction cannot be increased without limit. Eventually the *pn* junction breaks down either by the Avalanche or Zener breakdown mechanisms, which lead to large reverse currents, as shown in Figure 6.15. In the $V = -V_{br}$ region, the reverse current increases dramatically with the reverse bias. If unlimited, the large

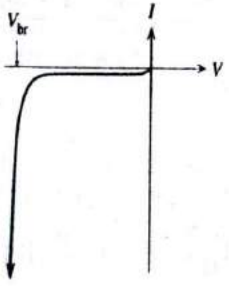


Figure 6.15 Reverse I–V characteristics of a *pn* junction.

reverse current will increase the power dissipated, which in turn raises the temperature of the device, which leads to a further increase in the reverse current and so on. If the temperature does not burn out the device, for example, by melting the contacts, then the breakdown is recoverable. If the current is limited by an external resistance to a value within the power dissipation specifications, then there is no reason why the device cannot operate under breakdown conditions.

6.5.1 AVALANCHE BREAKDOWN

As the reverse bias increases, the field in the SCL can become so large that an electron drifting in this region can gain sufficient kinetic energy to impact on a Si atom and ionize it, or break a Si–Si bond. The phenomenon by which a drifting electron gains sufficient energy from the field to ionize a host crystal atom by bombardment is termed **impact ionization.** The accelerated electron must gain at least an energy equal to E_g as impact ionization breaks a Si–Si bond, which is tantamount to exciting an electron from the valence band to the conduction band. Thus an additional electron–hole pair is created by this process.

Consider what happens when a thermally generated electron just inside the SCL in the p-side is accelerated by the field. The electron accelerates and gains sufficient energy to collide with a host Si atom and release an EHP by impact ionization, as depicted in Figure 6.16. It will lose at least E_g amount of energy, but it can accelerate and head for another ionizing collision further along the depletion region until it reaches the neutral n-region. The EHPs generated by impact ionization themselves can now be accelerated by the field and will themselves give rise to further EHPs by ionizing collisions and so on, leading to an **avalanche effect.** One initial carrier can thus create many carriers in the SCL through an avalanche of impact ionizations.

If the reverse current in the SCL in the absence of impact ionization is I_o, then due to the avalanche of ionizing collisions in the SCL, the reverse current becomes MI_o where M is the multiplication. It is the net number of carriers generated by the avalanche effect per carrier in the SCL. Impact ionization depends strongly on the electric field. Small increases in the reverse bias can lead to dramatic increases in the

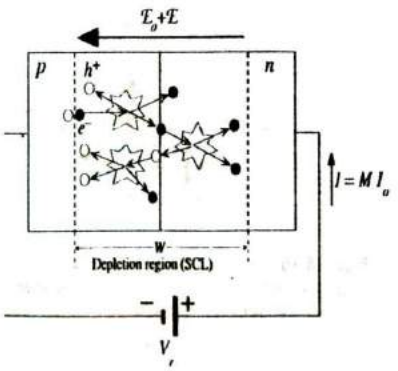


Figure 6.16 Avalanche breakdown by impact ionization.

multiplication process. Typically

$$M = \frac{1}{1 - \left(\frac{V_r}{V_{br}}\right)^n} \qquad [6.29]$$

where V_r is the reverse bias, V_{br} is the breakdown voltage, and n is an index in the range 3 to 5. It is clear that the reverse current MI_o increases sharply with V_r near V_{br}, as depicted in Figure 6.15. Indeed, the voltage across a diode under reverse breakdown remains around V_{br} for very large current variations (several orders of magnitude). If the reverse current under breakdown is limited by an appropriate external resistor R, as shown in Figure 6.17, to prevent destructive power dissipation in the diode, then the voltage across the diode remains approximately at V_{br}. Thus, as long as $V_r > V_{br}$, the diode clamps the voltage between A and B to approximately V_{br}. The reverse current in the circuit is then $(V_r - V_{br})/R$.

Since the electric field in the SCL depends on the width of the depletion region W, which in turn depends on the doping parameters, V_{br} also depends on the doping, as discussed in Example 6.7.

6.5.2 ZENER BREAKDOWN

Heavily doped *pn* junctions have narrow depletion widths, which lead to large electric fields within this region. When a reverse bias is applied to a *pn* junction, the energy band diagram of the *n*-side can be viewed as being lowered with respect to the *p*-side, as depicted in Figure 6.18. For a sufficient reverse bias (typically less than 10 V), E_c

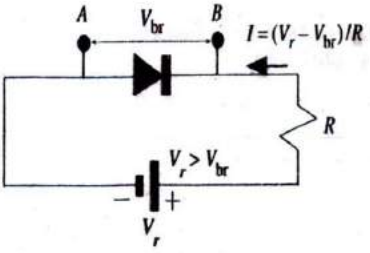


Figure 6.17 If the reverse breakdown current when $V_r > V_{br}$ is limited by an external resistance R to prevent destructive power dissipation, then the diode can be used to clamp the voltage between A and B to remain approximately V_{br}.

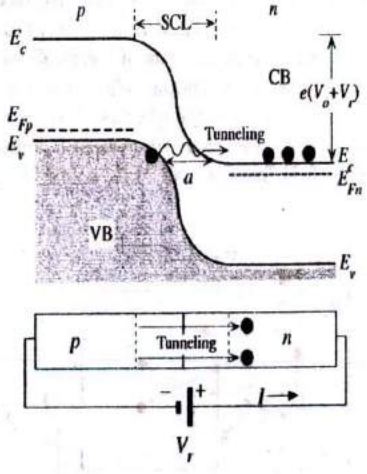


Figure 6.18 Zener breakdown involves electrons tunneling from the VB of *p*-side to the CB of *n*-side when the reverse bias reduces E_c to line up with E_v.

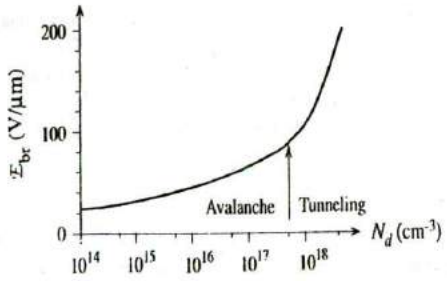


Figure 6.19 The breakdown field $\mathcal{E}_{br}$ in the depletion layer for the onset of reverse breakdown versus doping concentration N_d in the lightly doped region in a one-sided (p^+n or pn^+) abrupt pn junction.

Avalanche and tunneling mechanisms are separated by the arrow.

SOURCE: Data extracted from M. Sze and G. Gibbons, *Solid State Electronics*, 9, no. 831, 1966.

on the n-side may be lowered to be below E_v on the p-side. This means that electrons at the top of the VB in the p-side are now at the same energy level as the empty states in the CB in the n-side. As the separation between the VB and CB narrows, shown as a ($< W$), the electrons easily tunnel from the VB in the p-side to the CB in the n-side, which leads to a current. This process is called the **Zener effect.** As there are many electrons in the VB and many empty states in the CB, the tunneling current can be substantial. The reverse voltage V_r, which starts the tunneling current and hence the Zener breakdown, is clearly that which lowers E_c on the n-side to be below E_v on the p-side and thereby gives a separation that encourages tunneling. In nonquantum mechanical terms, one may intuitively view the Zener effect as the strong electric field in the depletion region ripping out some of those electrons in the Si–Si bonds and thereby releasing them for conduction.

Figure 6.19 shows the dependence of the breakdown field $\mathcal{E}_{br}$ in the depletion region for the onset of avalanche or Zener breakdown in a one-sided (p^+n or pn^+) abrupt junction on the dopant concentration N_d in the lightly doped side. At high fields, the tunneling becomes the dominant reverse breakdown mechanism.

EXAMPLE 6.7

AVALANCHE BREAKDOWN Consider a uniformly doped abrupt p^+n junction ($N_a \gg N_d$) reverse biased by $V = -V_r$.

a. What is the relationship between the depletion width W and the potential difference $(V_o + V_r)$ across W?

b. If avalanche breakdown occurs when the maximum field in the depletion region $\mathcal{E}_a$ reaches the breakdown field $\mathcal{E}_{br}$, show that the breakdown voltage V_{br} ($\gg V_o$) is then given by

$$V_{br} = \frac{\varepsilon \mathcal{E}_{br}^2}{2eN_d}$$

c. An abrupt Si p^+n junction has boron doping of 10^{19} cm^{-3} on the p-side and phosphorus doping of 10^{16} cm^{-3} on the n-side. The dependence of the avalanche breakdown field on the impurity concentration is shown in Figure 6.19.

1. What is the reverse breakdown voltage of this Si diode?
2. Calculate the reverse breakdown voltage when the phosphorus doping is increased to 10^{17} cm^{-3}.

SOLUTION

One can assume that all the applied reverse bias drops across the depletion layer so that the new voltage across W is now $V_o + V_r$. We have to integrate $d\mathcal{E}/dx = \rho_{net}/\varepsilon$ as before across W to find the maximum field. The most important fact to remember here is that the *pn* junction equations relating W, $\mathcal{E}_o$, V_o, N_a, N_d, and so on remain the same but with V_o replaced with $V_o + V_r$ since the applied reverse bias of V_r increases V_o to $V_o + V_r$. Then from Equation 6.4,

$$W^2 = \frac{2\varepsilon(V_o + V_r)(N_a^{-1} + N_d^{-1})}{e} \approx \frac{2\varepsilon(V_o + V_r)}{eN_d}$$

since $N_a \gg N_d$. The maximum field that corresponds to the breakdown field $\mathcal{E}_{br}$ is given by

Maximum field and reverse bias

$$\mathcal{E}_o = -\frac{2(V_o + V_r)}{W}$$

Thus, from these two equations we can eliminate W and obtain $V_{br} = V_r$ as

Breakdown voltage and doping

$$V_{br} = \frac{\varepsilon \mathcal{E}_{br}^2}{2eN_d}$$

Given $N_a \gg N_d$ we have a p^+n junction with $N_d = 10^{16}$ cm^{-3}. The depletion region extends into the *n*-region, so the maximum field actually occurs in the *n*-region. Here the breakdown field $\mathcal{E}_{br}$ depends on the doping level as given in the graph of the critical field at breakdown $\mathcal{E}_{br}$ versus doping concentration N_d in Figure 6.19. Taking $\mathcal{E}_{br} \approx 40$ V/μm or 4.0×10^5 V cm^{-1} at $N_d = 10^{16}$ cm^{-3} and using the above equation for V_{br}, we get $V_{br} = 53$ V.

When $N_d = 10^{17}$ cm^{-3}, $\mathcal{E}_{br}$ from the graph is about 6×10^5 V cm^{-1}, which leads to $V_{br} = 11.8$ V.

6.6 BIPOLAR TRANSISTOR (BJT)

6.6.1 COMMON BASE (CB) DC CHARACTERISTICS

As an example, we will consider the *pnp* bipolar junction transistor (BJT) whose basic structure is shown in Figure 6.20a. The *pnp* transistor has three differently doped semiconductor regions. These regions of different doping occur within the same single crystal by the variation of acceptor and donor concentrations resulting from the fabrication process. The most heavily doped *p*-region (p^+) is called the **emitter.** In contact with this region is the lightly doped *n*-region, which is called the **base.** The next region is the *p*-type doped **collector.** The base region has the most narrow width for reasons discussed below. Although the three regions in Figure 6.20a have identical cross-sectional areas, in practice, due to the fabrication process, the cross-sectional area increases from the emitter to the collector and the collector region has an extended width. For simplicity, we will assume that the cross-sectional area is uniform, as in Figure 6.20a.

The *pnp* BJT connected as shown in Figure 6.20b is said to be operating under normal and active conditions, which means that the base–emitter (BE) junction is forward biased and the base–collector (BC) junction is reverse biased. The circuit in

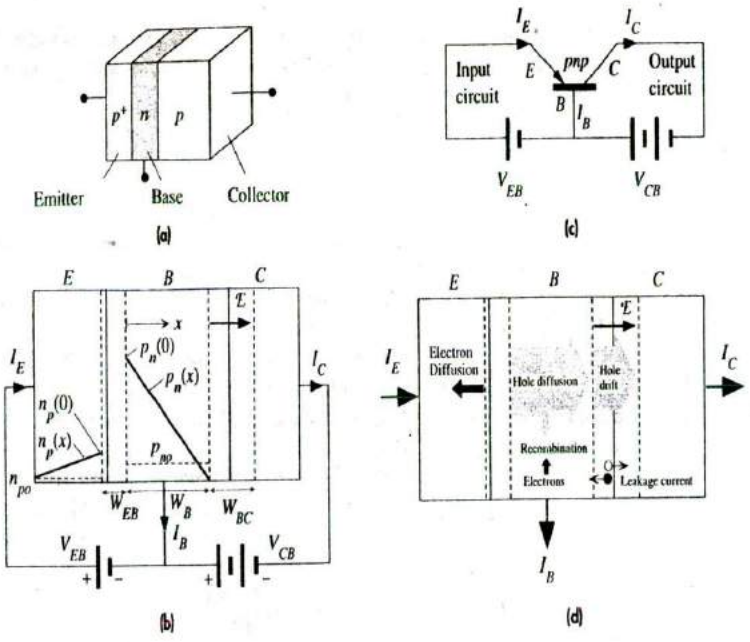


Figure 6.20

(a) A schematic illustration of the *pnp* bipolar transistor with three differently doped regions.

(b) The *pnp* bipolar operated under normal and active conditions.

(c) The CB configuration with input and output circuits identified.

(d) The illustration of various current components under normal and active conditions.

Figure 6.20b, in which the base is common to both the collector and emitter bias voltages, is known as the common base (CB) configuration.[6] Figure 6.20c shows the CB transistor circuit with the BJT represented by its circuit symbol. The arrow identifies the emitter junction and points in the direction of current flow when the EB junction is forward biased. Figure 6.20c also identifies the emitter circuit, where V_{EB} is connected, as the input circuit. The collector circuit, where V_{CB} is connected, is the output circuit.

The base–emitter junction is simply called the **emitter junction** and the base-collector junction is called the **collector junction.** As the emitter is heavily doped, the base–emitter depletion region W_{EB} extends almost entirely into the base. Generally, the base and collector regions have comparable doping, so the base–collector depletion region W_{BC} extends to both sides. The width of the neutral base region outside the depletion regions is labeled as W_B. All these parameters are shown and defined in Figure 6.20b.

[6] CB should not be confused with the conduction band abbreviation.

We should note that all the applied voltages drop across the depletion widths. The applied collector–base voltage V_{CB} reverse biases the BC junction and hence increases the field in the depletion region at the collector junction.

Since the EB junction is forward-biased, minority carriers are then injected into the emitter and base exactly as they are in the forward-biased diode. Holes are injected into the base and electrons into the emitter, as depicted in Figure 6.20d. Hole injection into the base, however, far exceeds the electron injection into the emitter because the emitter is heavily doped. We can then assume that the emitter current is almost entirely due to holes injected from the emitter into the base. Thus, when forward biased, the emitter "emits," that is, injects holes into the base.

Injected holes into the base must diffuse toward the collector junction because there is a hole concentration gradient in the base. Hole concentration $p_n(W_B)$ just outside the depletion region at the collector junction is negligibly small because the increased field sweeps nearly all the holes here across the junction into the collector (the collector junction is reverse biased).

The hole concentration $p_n(0)$ in the base just outside the emitter junction depletion region is given by the law of the junction. Measuring x from this point (Figure 6.20b),

$$p_n(0) = p_{no} \exp\left(\frac{eV_{EB}}{kT}\right) \qquad [6.30]$$

whereas at the collector end, $x = W_B$, $p_n(W_B) \approx 0$.

If no holes are lost by recombination in the base, then all the injected holes diffuse to the collector junction. There is no field in the base to drift the holes. Their motion is by diffusion. When they reach the collector junction, they are quickly swept across into the collector by the internal field $\mathcal{E}$ in W_{BC}. It is apparent that all the injected holes from the emitter become collected by the collector. The collector current is then the same as the emitter current. The only difference is that the emitter current flows across a smaller voltage difference V_{EB}, whereas the collector current flows through a larger voltage difference V_{CB}. This means a *net gain in power* from the emitter circuit to the collector circuit.

Since the current in the base is by diffusion, to evaluate the emitter and collector currents we must know the hole concentration gradient at $x = 0$ and $x = W_B$ and therefore we must know the hole concentration profile $p_n(x)$ across the base.[7] In the first instance, we can approximate the $p_n(x)$ profile in the base as a straight line from $p_n(0)$ to $p_n(W_B) = 0$, as shown in Figure 6.20b. This is only true in the absence of any recombination in the base as in the short diode case. The emitter current is then

$$I_E = -eAD_h\left(\frac{dp_n}{dx}\right)_{x=0} = eAD_h\frac{p_n(0)}{W_B}$$

[7] The actual concentration profile can be calculated by solving the steady-state continuity equation, which can be found in more advanced texts.

We can substitute for $p_n(0)$ from Equation 6.30 to obtain

$$I_E = \frac{eAD_h p_{no}}{W_B} \exp\left(\frac{eV_{EB}}{kT}\right) \quad [6.31]$$

Emitter current

It is apparent that I_E is determined by V_{EB}, the forward bias applied across the EB junction, and the base width W_B. In the absence of recombination, the collector current is the same as the emitter current, $I_C = I_E$. The control of the collector current I_C in the output (collector) circuit by V_{EB} in the input (emitter) circuit is what constitutes the **transistor action.** The common base circuit has a **power gain** because I_C in the output in Figure 6.20c flows around a larger voltage difference V_{CB} compared with I_E in the input, which flows across V_{EB} (about 0.6 V).

The ratio of the collector current I_C to the emitter current I_E is defined as the **CB current gain** or **current transfer ratio** α of the transistor,

$$\alpha = \frac{I_C}{I_E} \quad [6.32]$$

Definition of CB current gain

Typically, α is less than unity, in the range 0.99–0.999, due to two reasons. First is the limitation due to the emitter injection efficiency. When the *BE* junction is forward-biased, holes are injected from the emitter into the base, giving an emitter current $I_{E(\text{hole})}$, and electrons are injected from the base into the emitter, giving an emitter current $I_{E(\text{electron})}$. The total emitter current is, therefore,

$$I_E = I_{E(\text{hole})} + I_{E(\text{electron})}$$

Total emitter current

Only the holes injected into the base are useful in giving a collector current because only they can reach the collector. The emitter injection efficiency is defined as

$$\gamma = \frac{I_{E(\text{hole})}}{I_{E(\text{hole})} + I_{E(\text{electron})}} = \frac{1}{1 + \dfrac{I_{E(\text{electron})}}{I_{E(\text{hole})}}} \quad [6.33]$$

Emitter injection efficiency

Consequently, the collector current, which depends on $I_{E(\text{hole})}$ only, is less than the emitter current. We would like γ to be as close to unity as possible; $I_{E(\text{hole})} \gg I_{E(\text{electron})}$. γ can be readily calculated for the forward-biased *pn* junction current equations as shown in Example 6.9.

Secondly, a small number of the diffusing holes in the narrow base inevitably become lost by recombination with the large number of electrons present in this region as depicted in Figure 6.20d. Thus, a fraction of $I_{E(\text{hole})}$ is lost in the base due to recombination, which further reduces the collector current. We define the **base transport factor** α_T as

$$\alpha_T = \frac{I_C}{I_{E(\text{hole})}} = \frac{I_C}{\gamma I_E} \quad [6.34]$$

Base transport factor

If the emitter were a perfect injector, $I_E = I_{E(\text{hole})}$, then the current gain α would be α_T. If τ_h is the hole (minority carrier) lifetime in the base, then $1/\tau_h$ is the probability per unit time that a hole will recombine and disappear. We also know that in

time t, a particle diffuses a distance x, given by $x = \sqrt{2Dt}$ where D is the diffusion coefficient. The time τ_t it takes for a hole to diffuse across W_B is then given by

Base minority carrier transit time

$$\tau_t = \frac{W_B^2}{2D_h} \quad [6.35]$$

This diffusion time is called the **transit time** of the minority carriers across the base.

The probability of recombination in time τ_t is then τ_t/τ_h. The probability of not recombining and therefore diffusing across is $(1 - \tau_t/\tau_h)$. Since $I_{E(\text{hole})}$ represents the holes entering the base per unit time, $I_{E(\text{hole})}(1 - \tau_t/\tau_h)$ represents the number of holes leaving the base per unit time (without recombining) which is the collector current I_C. Substituting for I_C and $I_{E(\text{hole})}$ in Equation 6.34 gives the base transport factor α_T,

Base transport factor

$$\alpha_T = \frac{I_C}{I_{E(\text{hole})}} = 1 - \frac{\tau_t}{\tau_h} \quad [6.36]$$

Using Equations 6.32, 6.34, and 6.36 we can find the total **CB current gain** α:

CB current gain

$$\alpha = \alpha_T \gamma = \left(1 - \frac{\tau_t}{\tau_h}\right)\gamma \quad [6.37]$$

The recombination of holes with electrons in the base means that the base must be replenished with electrons, which are supplied by the external battery in the form of a small base current I_B, as shown in Figure 6.20d. In addition, the base current also has to supply the electrons injected from the base into the emitter, that is, $I_{E(\text{electron})}$, and shown as electron diffusion in the emitter in Figure 6.20d. The number of holes entering the base per unit time is represented by $I_{E(\text{hole})}$, and the number recombining per unit time is then $I_{E(\text{hole})}(\tau_t/\tau_h)$. Thus, I_B is

Base current

$$I_B = \left(\frac{\tau_t}{\tau_h}\right) I_{E(\text{hole})} + I_{E(\text{electron})} = \gamma \frac{\tau_t}{\tau_h} I_E + (1 - \gamma) I_E \quad [6.38]$$

which further simplifies to $I_E - I_C$; the difference between the emitter current and the collector current is the base current. (This is exactly what we expect from Kirchoff's current law.)

The ratio of the collector current to the base current is defined as the **current gain** β of the transistor.[8] By using Equations 6.32, 6.37, and 6.38, we can relate β to α:

Base-to-collector current gain

$$\beta = \frac{I_C}{I_B} = \frac{\alpha}{1 - \alpha} \approx \frac{\gamma \tau_h}{\tau_t} \quad [6.39]$$

The base–collector junction in Figure 6.20b is reverse biased, which leads to a leakage current into the collector terminal even in the absence of an emitter current. This leakage current is due to thermally generated electron–hole pairs in the depletion region W_{BC} being drifted by the internal field, as schematically illustrated in Figure 6.20d.

[8] β is a useful parameter when the transistor is used in what is called the common emitter (CE) configuration, in which the input current is made to flow into the base of the transistor, and the collector current is made to flow in the output circuit.

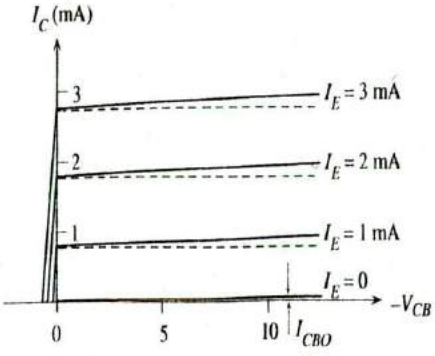


Figure 6.21 DC I–V characteristics of the *pnp* bipolar transistor (exaggerated to highlight various effects).

Suppose that we open circuit the emitter ($I_E = 0$). Then the collector current is simply the leakage current, denoted by I_{CBO}. The base current is then $-I_{CBO}$ (flowing out from the base terminal). In the presence of an emitter current I_E, we have

Active region collector current

$$I_C = \alpha I_E + I_{CBO} \quad [6.40]$$

Active region base current

$$I_B = (1 - \alpha) I_E - I_{CBO} \quad [6.41]$$

Equations 6.40 and 6.41 give the collector and base currents in terms of the input current I_E, which in turn depends on V_{EB}. They only hold when the collector junction is reverse biased and the emitter junction is forward biased, which is defined as the **active region** of the BJT. It should be emphasized that what constitutes the transistor action is the control of I_E, and hence I_C, by V_{EB}.

The dc characteristics of the CB-connected BJT as in Figure 6.20b are normally represented by plotting the collector current I_C as a function of V_{CB} for various fixed values of the emitter current. A typical example of such dc characteristics for a *pnp* transistor is illustrated in Figure 6.21. The following characteristics are apparent. The collector current when $I_E = 0$ is the CB junction leakage current I_{CBO}, typically a fraction of a microampere. As long as the collector is negatively biased with respect to the base, the CB junction is reverse biased and the collector current is given by $I_C = \alpha I_E + I_{CBO}$, which is close to the emitter current when $I_E \gg I_{CBO}$. When the polarity of V_{CB} is changed, the CB junction becomes forward biased. The collector junction is then like a forward biased diode and the collector current is the difference between the forward biased CB junction current and the forward biased EB junction current. As they are in opposite directions, they subtract.

We note that I_C increases slightly with the magnitude of V_{CB} even when I_E is constant. In our treatment I_C did not directly depend on V_{CB}, which simply reverse biased the collector junction to collect the diffusing holes. In our discussions we assumed that the base width W_B does not depend on the applied voltages. This is only approximately true. Suppose that we increase the reverse bias V_{CB} (for example, from -5 to -10 V). Then the base–collector depletion width W_{BC} also increases, as schematically depicted in Figure 6.22. Consequently the base width W_B gets slightly narrower, which leads to a slightly shorter base transit time τ_t. The base transport factor α_T in Equation 6.36 and

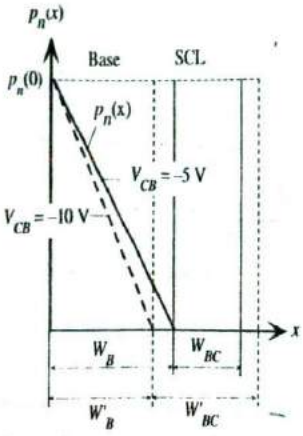


Figure 6.22 The Early effect. When the BC reverse bias increases, the depletion width W_{BC} increases to W'_{BC}, which reduces the base width W_B to W'_B. As $p_n(0)$ is constant (constant V_{EB}), the minority carrier concentration gradient becomes steeper and the collector current, I_C increases.

hence α are then slightly larger, which leads to a small increase in I_C. The modulation of the base width W_B by V_{CB} is not very strong, which means that the slopes of the $I_C - V_{CB}$ lines at a fixed I_E are very small in Figure 6.21. The base width modulation by V_{CB} is called the **Early effect.**

EXAMPLE 6.8 **A *pnp* TRANSISTOR** Consider a *pnp* Si BJT that has the following properties. The emitter region mean acceptor doping is 2×10^{18} cm^{-3}, the base region mean donor doping is 1×10^{16} cm^{-3}, and the collector region mean acceptor doping is 1×10^{16} cm^{-3}. The hole drift mobility in the base is 400 cm^2 V^{-1} s^{-1}, and the electron drift mobility in the emitter is 200 cm^2 V^{-1} s^{-1}. The transistor emitter and base neutral region widths are about 2 μm each when the transistor is under normal operating conditions, that is, when the EB junction is forward-biased and the BC junction is reverse-biased. The effective cross-sectional area of the device is 0.02 mm^2. The hole lifetime in the base is approximately 400 ns. Assume that the emitter has 100 percent injection efficiency, $\gamma = 1$. Calculate the CB current transfer ratio α and the current gain β. What is the emitter–base voltage if the emitter current is 1 mA?

SOLUTION

The hole drift mobility $\mu_h = 400$ cm^2 V^{-1} s^{-1} (minority carriers in the base). From the Einstein relationship we can easily find the diffusion coefficient of holes,

$$D_h = \left(\frac{kT}{e}\right)\mu_h = (0.0259 \text{ V})(400 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}) = 10.36 \text{ cm}^2 \text{ s}^{-1}$$

The minority carrier transit time τ_t across the base is

$$\tau_t = \frac{W_B^2}{2D_h} = \frac{(2 \times 10^{-4} \text{ cm})^2}{2(10.36 \text{ cm}^2 \text{ s}^{-1})} = 1.93 \times 10^{-9} \text{ s} \quad \text{or} \quad 1.93 \text{ ns}$$

The base transport factor and hence the CB current gain is

$$\alpha = \gamma\alpha_B = 1 - \frac{\tau_t}{\tau_h} = 1 - \frac{1.93 \times 10^{-9} \text{ s}}{400 \times 10^{-9} \text{ s}} = 0.99517$$

The current gain β of the transistor is

$$\beta = \frac{\alpha}{1-\alpha} = \frac{0.99517}{1-0.99517} = 206.2$$

The emitter current is due to holes diffusing in the base ($\gamma = 1$),

$$I_E = I_{EO} \exp\left(\frac{eV_{EB}}{kT}\right)$$

where

$$I_{EO} = \frac{eAD_h p_{no}}{W_B} = \frac{eAD_h n_i^2}{N_d W_B}$$

$$= \frac{(1.6 \times 10^{-19}\ \text{C})(0.02 \times 10^{-2}\ \text{cm}^2)(10.36\ \text{cm s}^{-1})(1.0 \times 10^{10}\ \text{cm}^{-3})^2}{(1 \times 10^{16}\ \text{cm}^{-3})(2 \times 10^{-4}\ \text{cm})}$$

$$= 1.66 \times 10^{-14}\ \text{A}$$

Thus,

$$V_{EB} = \frac{kT}{e} \ln\left(\frac{I_E}{I_{EO}}\right) = (0.0259\ \text{V}) \ln\left(\frac{1 \times 10^{-3}\ \text{A}}{1.66 \times 10^{-14}\ \text{A}}\right) = 0.64\ \text{V}$$

The major assumption is $\gamma = 1$, which is generally not true, as shown in Example 6.9. The actual α and hence β will be smaller due to less than 100 percent emitter injection. Note also that W_B is the *neutral region width*, that is, the region of base outside the depletion regions. It is not difficult to calculate the depletion layer widths within the base, which are about 0.2 μm on the emitter side and roughly about 0.7 μm on the collector side, so that the total base width junction to junction is $2 + 0.2 + 0.7 = 2.9$ μm.

The transit time of minority carriers across the base is τ_t. If the input signal changes before the minority carriers have diffused across the base, then the collector current cannot respond to the changes in the input. Thus, if the frequency of the input signal is greater than $1/\tau_t$, the minority carriers will not have time to transit the base and the collector current will remain unmodulated by the input signal. One can set the upper frequency limit at $\sim 1/\tau_t$ which is 518 MHz.

EMITTER INJECTION EFFICIENCY γ **EXAMPLE 6.9**

a. Consider a *pnp* transistor with the parameters as defined in Figure 6.20. Show that the **injection efficiency of the emitter**, defined as

$$\gamma = \frac{\text{Emitter current due to minority carriers injected into the base}}{\text{Total emitter current}}$$

is given by

$$\gamma = \frac{1}{1 + \dfrac{N_d W_B \mu_{e(\text{emitter})}}{N_a W_E \mu_{h(\text{base})}}}$$

b. How would you modify the CB current gain α to include the emitter injection efficiency?

c. Calculate the emitter injection efficiency for the *pnp* transistor in Example 6.8, which has an acceptor doping of $2 \times 10^{18}\ \text{cm}^{-3}$ in the emitter, donor doping of $1 \times 10^{16}\ \text{cm}^{-3}$ in the

base, emitter and base neutral region widths of 2 μm, and a minority carrier lifetime of 400 ns in the base. What are its α and β taking into account the emitter injection efficiency?

SOLUTION

When the BE junction is forward biased, holes are injected into the base, giving an emitter current $I_{E(\text{hole})}$, and electrons are injected into the emitter, giving an emitter current $I_{E(\text{electron})}$. The total emitter current is therefore

$$I_E = I_{E(\text{hole})} + I_{E(\text{electron})}$$

Only the holes injected into the base are useful in giving a collector current because only they can reach the collector. Injection efficiency is defined as

Emitter injection efficiency definition

$$\gamma = \frac{I_{E(\text{hole})}}{I_{E(\text{hole})} + I_{E(\text{electron})}} = \frac{1}{1 + \dfrac{I_{E(\text{electron})}}{I_{E(\text{hole})}}}$$

But, provided that W_E and W_B are shorter than minority carrier diffusion lengths,

$$I_{E(\text{hole})} = \frac{eAD_{h(\text{base})}n_i^2}{N_d W_B}\exp\left(\frac{eV_{EB}}{kT}\right) \quad \text{and} \quad I_{E(\text{electron})} = \frac{eAD_{e(\text{emitter})}n_i^2}{N_a W_E}\exp\left(\frac{eV_{EB}}{kT}\right)$$

When we substitute into the definition of γ and use $D = \mu kT/e$, we obtain

Emitter injection efficiency

$$\gamma = \frac{1}{1 + \dfrac{N_d W_B \mu_{e(\text{emitter})}}{N_a W_E \mu_{h(\text{base})}}}$$

The hole component of the emitter current is given as γI_E. Of this, a fraction $\alpha_T = (1 - \tau_t/\tau_h)$ will give a collector current. Thus, the emitter-to-collector current transfer ratio α, taking into account the emitter injection efficiency, is

Emitter-to-collector current transfer ratio

$$\alpha = \alpha_T \gamma \left(1 - \frac{\tau_t}{\tau_h}\right)$$

In the emitter, $N_{a(\text{emitter})} = 2 \times 10^{18}\ \text{cm}^{-3}$ and $\mu_{e(\text{emitter})} = 200\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$, and in the base, $N_{d(\text{base})} = 1 \times 10^{16}\ \text{cm}^{-3}$ and $\mu_{h(\text{base})} = 400\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$. The emitter injection efficiency is

$$\gamma = \frac{1}{1 + \dfrac{(1 \times 10^{16})(2)(200)}{(2 \times 10^{18})(2)(400)}} = 0.99751$$

The transit time $\tau_t = W_B^2/2D_h = 1.93 \times 10^{-9}$ s (as before), so the overall α is

$$\alpha = 0.99751\left(1 - \frac{1.93 \times 10^{-9}}{400 \times 10^{-9}}\right) = 0.99269$$

and the overall β is

$$\beta = \frac{\alpha}{(1 - \alpha)} = 135.8$$

The same transistor with 100 percent emitter injection in Example 6.8 had a β of 206. It is clear that the emitter injection efficiency γ and the base transport factor α_T have comparable impacts in controlling the overall gain in the example. We neglected the recombination of

electrons and holes in the EB depletion region. In fact, if we were to also consider this recombination component of the emitter current, $I_{E(\text{rec})}$ would have to be even smaller compared with the total I_E, which would make γ and hence β even lower.

6.6.2 COMMON BASE AMPLIFIER

According to Equation 6.31 the emitter current depends exponentially on V_{EB},

$$I_E = I_{EO} \exp\left(\frac{eV_{EB}}{kT}\right) \qquad [6.42]$$

It is therefore apparent that small changes in V_{EB} lead to large changes in I_E. Since $I_C \approx I_E$, we see that small variations in V_{EB} cause large changes in I_C in the collector circuit. This can be fruitfully used to obtain voltage amplification as shown in Figure 6.23. The battery V_{CC}, through R_C, provides a reverse bias for the base–collector junction. The dc voltage V_{EE} forward biases the EB junction, which means that it provides a dc current I_E. The input signal is the ac voltage v_{eb} applied in series with the dc bias voltage V_{EE} to the EB junction. The applied signal v_{eb} modulates the total voltage V_{EB} across the EB junction and hence, by virtue of Equation 6.30, modulates the injected hole concentration $p_n(0)$ up and down about the dc value determined by V_{EE} as depicted in Figure 6.23. This variation in $p_n(0)$ alters the concentration gradient and therefore gives rise to a change in I_E, and hence a nearly identical change in I_C. The change in the collector current can be converted to a voltage change by using a resistor R_C in the collector circuit as shown in Figure 6.23. However, the output is commonly taken between the collector, and the base and this voltage V_{CB} is

$$V_{CB} = -V_{CC} + R_C I_C$$

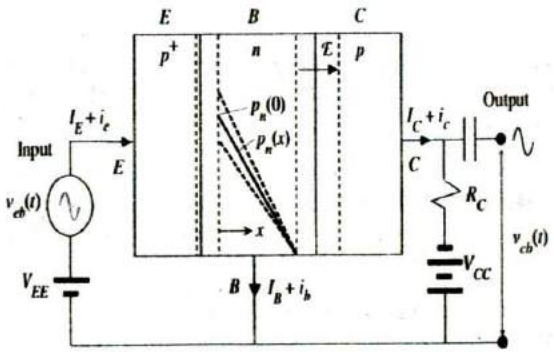


Figure 6.23 A *pnp* transistor operated in the active region in the common base amplifier configuration.

The applied (input) signal v_{eb} modulates the dc voltage across the EB junction and hence modulates the injected hole concentration up and down about the dc value $p_n[0]$. The solid line shows $p_n[x]$ when only the dc bias V_{EE} is present. The dashed lines show how $p_n[x]$ is modulated up and down by the signal v_{eb} superimposed on V_{EE}.

Increasing the emitter–base voltage V_{EB} (by increasing v_{eb}) increases I_C, which increases V_{CB}. Since we are interested in ac signals, that voltage variation across CB is tapped out through a dc blocking capacitor in Figure 6.23.

For simplicity we will assume that changes δV_{EB} and δI_E in the dc values of V_{EB} and I_E are small, which means that δV_{EB} and δI_E can be related by differentiating Equation 6.42. We are hence tacitly assuming an operation under small signals. Further, we will take the changes to represent the ac signal magnitudes, $v_{eb} = \delta V_{EB}$, $i_e = \delta I_E$, $i_c = \delta I_C \approx \delta I_E \approx i_e$, $v_{cb} = \delta V_{CB}$.

The output signal voltage v_{cb} corresponds to the change in V_{CB},

$$v_{cb} = \delta V_{CB} = R_C \delta I_C = R_C \delta I_E$$

The variation in the emitter current δI_E depends on the variation δV_{EB} in V_{EB}, which can be determined by differentiating Equation 6.42,

$$\frac{\delta I_E}{\delta V_{EB}} = \frac{e}{kT} I_E$$

By definition, δV_{EB} is the input signal v_{eb}. The change δI_E in I_E is the input signal current (i_e) flowing into the emitter as a result of δV_{EB}. Therefore the quantity $\delta V_{EB}/\delta I_E$ represents an input resistance r_e seen by the source v_{eb}.

Input resistance

$$r_e = \frac{\delta V_{EB}}{\delta I_E} = \frac{kT}{eI_E} = \frac{25}{I_E(\text{mA})} \qquad [6.43]$$

The output signal is then

$$v_{cb} = R_C \delta I_E = R_C \frac{v_{eb}}{r_e}$$

so the voltage amplification is

CB voltage gain

$$A_V = \frac{v_{cb}}{v_{eb}} = \frac{R_C}{r_e} \qquad [6.44]$$

To obtain a voltage gain we obviously need $R_C > r_e$, which is invariably the case by the appropriate choice of I_E, hence r_e, and R_C. For example, when the BJT is biased so that I_E is 10 mA and r_e is 2.5 Ω, and if R_C is chosen to be 50 Ω, then the gain is 20.

EXAMPLE 6.10 **A COMMON BASE AMPLIFIER** Consider a *pnp* Si BJT that has been connected as in Figure 6.23. The BJT has a $\beta = 135$ and has been biased to operate with a 5 mA collector current. What is the small-signal input resistance? What is the required R_C that will provide a voltage gain of 20? What is the base current? What should be the V_{CC} in Figure 6.23? Suppose $V_{CC} = -6$ V, what is the largest swing in the output voltage V_{CB} in Figure 6.23 as the input signal is increased and decreased about the bias point V_{EE}, taken as 0.65 V?

SOLUTION

The emitter and collector currents are approximately the same. From Equation 6.43,

$$r_e = \frac{25}{I_E\,(\text{mA})} = \frac{25}{5} = 5\ \Omega$$

The voltage gain A_V from Equation 6.44 is

$$A_V = \frac{R_C}{r_e} \quad \text{or} \quad 20 = \frac{R_C}{5\ \Omega}$$

so a gain of 20 requires $R_C = 100\ \Omega$.

$$\text{Base current } I_B = \frac{I_C}{\beta} = \frac{5\text{ mA}}{135} = 0.037\text{ mA} \quad \text{or} \quad 37\ \mu\text{A}$$

There is a dc voltage across R_C given by $I_C R_C = (0.005\text{ A})(100\ \Omega) = 0.5$ V. V_{CC} has to provide the latter voltage across R_C and also a sufficient voltage to keep the BC junction reverse biased at all times under normal operation. Let us set $V_{CC} = -6$ V. Thus, in the absence of any input signal v_{eb}, V_{CB} is set to $-6\text{ V} + 0.5\text{ V} = -5.5$ V. As we increase the signal v_{eb}, V_{EB} and hence I_C increase until the point C becomes nearly zero,[9] that is, $V_{CB} = 0$, which occurs when I_C is maximum at $I_{C\max} = |V_{CC}|/R_C$ or 60 mA. As v_{eb} decreases, so does V_{EB} and hence I_C. Eventually I_C will simply become zero, and point C will be at -6 V, so $V_{CB} = V_{CC}$. Thus, V_{CB} can only swing from -5.5 V to 0 V (for increasing input until $I_C = I_{C\max}$), or from -5.5 to -6 V (for decreasing input until $I_C = 0$).

6.6.3 COMMON EMITTER (CE) DC CHARACTERISTICS

An *npn* bipolar transistor when connected in the common emitter (CE) configuration has the emitter common to both the input and output circuits, as shown in Figure 6.24a. The dc voltage V_{BE} forward biases the BE junction and thereby injects electrons as minority carriers into the base. These electrons diffuse to the collector junction where the field $\mathcal{E}$ sweeps them into the collector to constitute the collector current I_C. V_{BE} controls the current I_E and hence I_B and I_C. The advantage of the CE configuration is that the **input current** is the current flowing between the ac source and the base, which is the base current I_B. This current is much smaller than the emitter current by about a factor of β. The output current is the current flowing between V_{CE} and the collector, which is I_C. In the CE configuration, the dc voltage V_{CE} must be greater than V_{BE} to reverse bias the collector junction and collect the diffusing electrons in the base.

The dc characteristics of the BJT in the CE configuration are normally given as I_C versus V_{CE} for various values of fixed base currents I_B, as shown in Figure 6.24b. The characteristics can be readily understood by Equations 6.40 and 6.41. We should note that, in practice, we are essentially adjusting V_{BE} to obtain the desired I_B because, by Equation 6.41,

$$I_B = (1 - \alpha)I_E - I_{CBO}$$

and I_E depends on V_{BE} via Equation 6.42.

Increasing I_B requires increasing V_{BE}, which increases I_C. Using Equations 6.40 and 6.41, we can obtain I_C in terms of I_B alone,

$$I_C = \beta I_B + \frac{1}{(1 - \alpha)} I_{CBO}$$

[9] Various saturation effects are ignored in this approximate discussion.

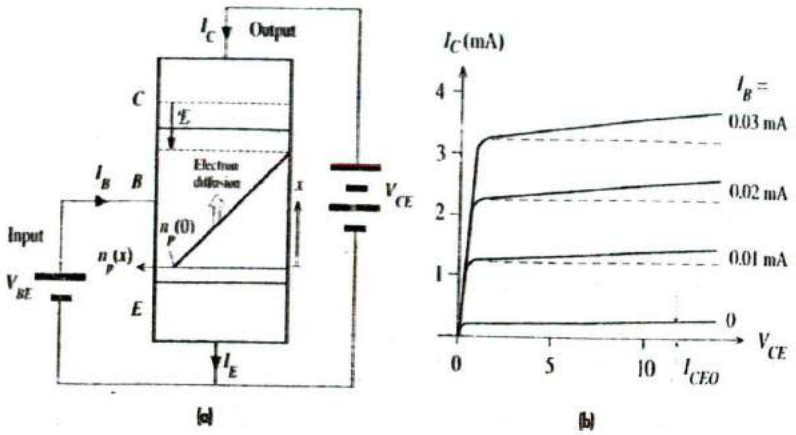


Figure 6.24

(a) An *npn* transistor operated in the active region in the common emitter configuration. The input current is the current that flows between V_{BE} and the base which is I_B.

(b) DC I-V characteristics of the *npn* bipolar transistor in the CE configuration. [Exaggerated to highlight various effects.]

Active region collector current

or

$$I_C = \beta I_B + I_{CEO} \qquad [6.45]$$

where

$$I_{CEO} = \frac{I_{CBO}}{(1-\alpha)} \approx \beta I_{CBO}$$

is the leakage current into the collector when the base is open circuited. This is much larger in the CE circuit than in the CB configuration.

Even when I_B is kept constant, I_C still exhibits a small increase with V_{CE}, which, according to Equation 6.45, indicates an increase in the current gain β with V_{CE}. This is due to the Early effect or modulation of the base width by V_{CB}, shown in Figure 6.22. Increasing V_{CE} increases V_{CB}, which increases W_{BC}, reduces W_B, and hence shortens τ_t. The resulting effect is a larger β ($\approx \tau_b/\tau_t$).

When V_{CE} is less than V_{BE}, the collector junction becomes forward biased and Equation 6.45 is not valid. The collector current is then the difference between forward currents of emitter and collector junctions. The transistor operating in this region is said to be **saturated.**

6.6.4 LOW-FREQUENCY SMALL-SIGNAL MODEL

The *npn* bipolar transistor in the CE (common emitter) amplifier configuration is shown in Figure 6.25. The input circuit has a dc bias V_{BB} to forward bias the base-emitter (BE) junction and the output circuit has a dc voltage V_{CC} (larger than V_{BB}) to reverse bias the base-collector (BC) junction through a collector resistor R_C.

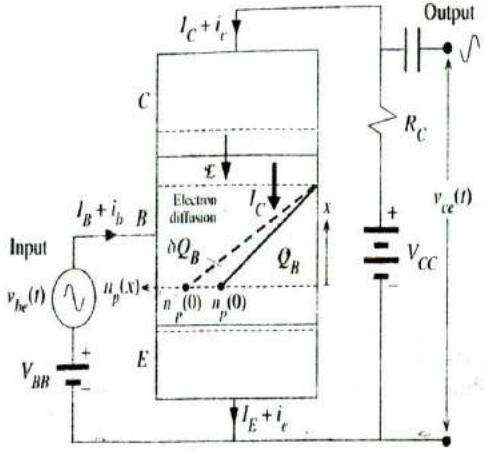


Figure 6.25 An *npn* transistor operated in the active region in the common emitter amplifier configuration.
The applied signal v_{be} modulates the dc voltage across the BE junction and hence modulates the injected electron concentration up and down about the dc value $n_p[0]$. The solid line shows $n_p[x]$ when only the dc bias V_{BB} is present. The dashed line shows how $n_p[x]$ is modulated up by a positive small signal v_{be} superimposed on V_{BB}.

The actual reverse bias voltage across the BC junction is $V_{CE} - V_{BE}$, where V_{CE} is

$$V_{CE} = V_{CC} - I_C R_C$$

An input signal in the form of a small ac signal v_{be} is applied in series with the bias voltage V_{BB} and modulates the voltage V_{BE} across the BE junction about its dc value V_{BB}. The varying voltage across the BE modulates $n_p(0)$ up and down about its dc value, which leads to a varying emitter current and hence to an almost identically varying collector current in the output circuit. The variation in the collector current is converted to an output voltage signal by the collector resistance R_C. Note that increasing V_{BE} increases I_C, which leads to a decrease in V_{CE}. Thus, the output voltage is 180° out of phase with the input voltage.

Since the BE junction is forward-biased, the relationship between I_E and V_{BE} is exponential,

$$I_E = I_{EO} \exp\left(\frac{eV_{BE}}{kT}\right) \qquad [6.46]$$

Emitter current and V_{BE}

where I_{EO} is a constant. We can differentiate this expression to relate small variations in I_E and V_{BE} as in the presence of small signals superimposed on dc values. For small signals, we have $v_{be} = \delta V_{BE}$, $i_b = \delta I_B$, $i_c = \delta I_E$, $i_c = \delta I_C$. Then from Equation 6.45 we see that $\delta I_C = \beta \delta I_B$, so $i_c = \beta i_b$. Since $\alpha \approx 1$, $i_c \approx i_e$.

What is the advantage of the CE circuit over the common base (CB) configuration? First, the input current is the base current, which is about a factor of β smaller than the emitter current. The ac input resistance of the CE circuit is therefore a factor of β higher than that of the CB circuit. This means that the amplifier does not load the ac source; the input resistance of the amplifier is much greater than the internal (or output) resistance of the ac source at the input. The small-signal input resistance r_{be} is

$$r_{be} = \frac{v_{be}}{i_b} = \frac{\delta V_{BE}}{\delta I_B} \approx \beta \frac{\delta V_{BE}}{\delta I_E} = \frac{\beta kT}{eI_E} \approx \frac{\beta 25}{I_C(\text{mA})} \qquad [6.47]$$

Input resistance

where we differentiated Equation 6.46.

The output ac signal v_{ce} develops across the CE and is tapped out through a capacitor. Since $V_{CE} = V_{CC} - I_C R_C$, as I_C increases, V_{CE} decreases. Thus,

$$v_{ce} = \delta V_{CE} = -R_C \, \delta I_C = -R_C i_c$$

The voltage amplification is

Voltage gain

$$A_V = \frac{v_{ce}}{v_{be}} = \frac{-R_C i_c}{r_{be} i_b} = \frac{-R_C \beta}{r_{be}} \approx -\frac{R_C I_C(\text{mA})}{25} \qquad [6.48]$$

which is the same as that in the CB configuration. However, in the CE configuration the output to input current ratio $i_c / i_b = \beta$, whereas this is almost unity in the CB configuration. Consequently, the CE configuration provides a greater power amplification, which is the second advantage of the CE circuit.

The input signal v_{be} gives rise to an output current i_c. This input voltage to output current conversion is defined in a parameter called the **mutual conductance,** or **transconductance,** g_m.

Transconductance

$$g_m = \frac{i_c}{v_{be}} \approx \frac{\delta I_E}{\delta V_{BE}} = \frac{I_E(\text{mA})}{25} = \frac{1}{r_e} \qquad [6.49]$$

The voltage amplification of the CE amplifier is then

Voltage gain

$$A_V = -g_m R_C \qquad [6.50]$$

We generally find it convenient to use a small-signal equivalent circuit for the low-frequency behavior of a BJT in the CE configuration. Between the base and emitter, the applied ac source voltage v_s sees only an input resistance of r_{be}, as shown in Figure 6.26. To underline the importance of the transistor input resistance, the output (or the internal) resistance R_s of the ac source is also shown. In the output circuit there is a voltage-controlled current source i_c which generates a current of $g_m v_{be}$. The current i_c passes through the load (or collector) resistance R_C across which the voltage signal develops. As we are only interested in ac signals, the batteries are taken as a short-circuit path for the ac current, which means that the internal resistances of the batteries are taken as zero. This model, of course, is valid only under normal and active operating conditions and small signals about dc values, and at low frequencies.

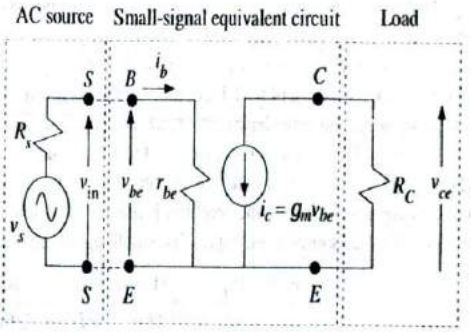
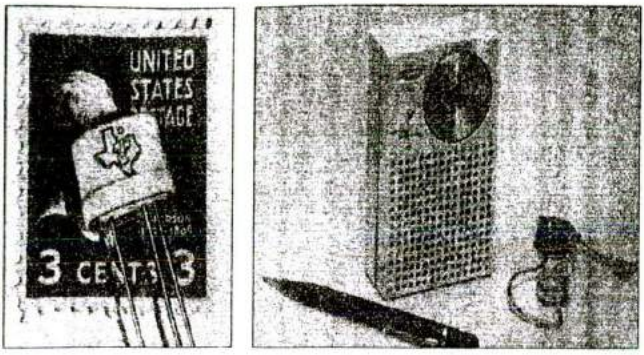


Figure 6.26 Low-frequency small-signal simplified equivalent circuit of the bipolar transistor in the CE configuration with a load resistor R_C in the collector circuit.



Left: The first commercial Si transistor from Texas Instruments (1954). Right: The first transistor pocket radio (1954). It had four Ge *npn* transistors.

SOURCE: Courtesy of Texas Instruments.

The bipolar transistor general dc current equation $I_C = \beta I_B$, where $\beta \approx \tau_h/\tau_t$ is a material-dependent constant, implies that the ac small-signal collector current is

$$\delta I_C = \beta \delta I_B \quad \text{or} \quad i_c = \beta i_b$$

Thus the CE dc and ac small-signal current gains are the same. This is a reasonable approximation in the low-frequency range, typically at frequencies below $1/\tau_h$. It is useful to have a relationship between β, g_m, and r_{be}. Using Equations 6.47 and 6.49, we have

$$\beta = g_m r_{be} \qquad \textbf{[6.51]}$$

β at low frequencies

In transistor data books, the dc current gain I_C/I_B is denoted as h_{FE} whereas the small-signal ac current gain i_c/i_b is denoted as h_{fe}. Except at high frequencies, $h_{fe} \approx h_{FE}$.

EXAMPLE 6.11

CE LOW-FREQUENCY SMALL-SIGNAL EQUIVALENT CIRCUIT Consider a BJT with a β of 100, used in a CE amplifier in which the collector current is 2.5 mA and R_C is 1 kΩ. If the ac source has an rms voltage of 1 mV and an output resistance R_s of 50 Ω, what is the rms output voltage? What is the input and output power and the overall power amplification?

SOLUTION

As the collector current is 2.5 mA, the input resistance and the transconductance are

$$r_{be} = \frac{\beta 25}{I_C(\text{mA})} = \frac{(100)(25)}{2.5} = 1000\ \Omega$$

and

$$g_m = \frac{I_C(\text{mA})}{25} = \frac{2.5}{25} = 0.1\ \text{A/V}$$

The *magnitude* of the voltage gain of the BJT small-signal equivalent circuit is

$$A_V = \frac{v_{ce}}{v_{be}} = g_m R_C = (0.1)(1000) = 100$$

When the ac source is connected to the B and E terminals (Figure 6.26), the input resistance r_{be} of the BJT loads the ac source, so v_{be} across BE is

$$v_{be} = v_s \frac{r_{be}}{(r_{be} + R_s)} = (1 \text{ mV}) \frac{1000 \ \Omega}{(1000 \ \Omega + 50 \ \Omega)} = 0.952 \text{ mV}$$

The output voltage (rms) is, therefore,

$$v_{ce} = A_V v_{be} = 100(0.952 \text{ mV}) = 95.2 \text{ mV}$$

The loading effect makes the output less than 100 mV. To reduce the loading of the ac source, we need to increase r_{be}, *i.e.*, reduce the collector current, but that also reduces the gain. So to keep the gain the same, we need to reduce I_C and increase R_C. However, R_C cannot be increased indefinitely because R_C itself is loaded by the input of the next stage and, in addition, there is an incremental resistance between the collector and emitter terminals (typically ~100 kΩ) that shunts R_C (not shown in Figure 6.26).

The power amplification of the CE BJT itself is

$$A_P = \frac{i_c v_{ce}}{i_b v_{be}} = \beta A_V = (100)(100) = 10{,}000$$

The input power into the BE terminals is

$$P_{in} = v_{be} i_b = \frac{v_{be}^2}{r_{be}} = \frac{(0.952 \times 10^{-3} \text{ V})^2}{1000 \ \Omega} = 9.06 \times 10^{-10} \text{ W} \quad \text{or} \quad 0.906 \text{ nW}$$

The output power is

$$P_{out} = P_{in} A_P = (9.06 \times 10^{-10})(10{,}000) = 9.06 \times 10^{-6} \text{ W} \quad \text{or} \quad 9.06 \ \mu\text{W}$$

6.7 JUNCTION FIELD EFFECT TRANSISTOR (JFET)

6.7.1 GENERAL PRINCIPLES

The basic structure of the junction field effect transistor (JFET) with an n-type channel (n-channel) is depicted in Figure 6.27a. An n-type semiconductor slab is provided with contacts at its ends to pass current through it. These terminals are called **source** (S) and **drain** (D). Two of the opposite faces of the n-type semiconductor are heavily p-type doped to some small depth so that an n-type channel is formed between the source and drain terminals, as shown in Figure 6.27a. The two p^+ regions are normally electrically connected and are called the **gate** (G). As the gate is heavily doped, the depletion layers extend almost entirely into the n-channel, as shown in Figure 6.27. For simplicity we will assume that the two gate regions are identical (both p^+ type) and that the doping in the n-type semiconductor is uniform. We will define the n-channel to be the region of conducting n-type material contained between the two depletion layers.

The basic and idealized symmetric structure in Figure 6.27a is useful in explaining the principle of operation as discussed later but does not truly represent

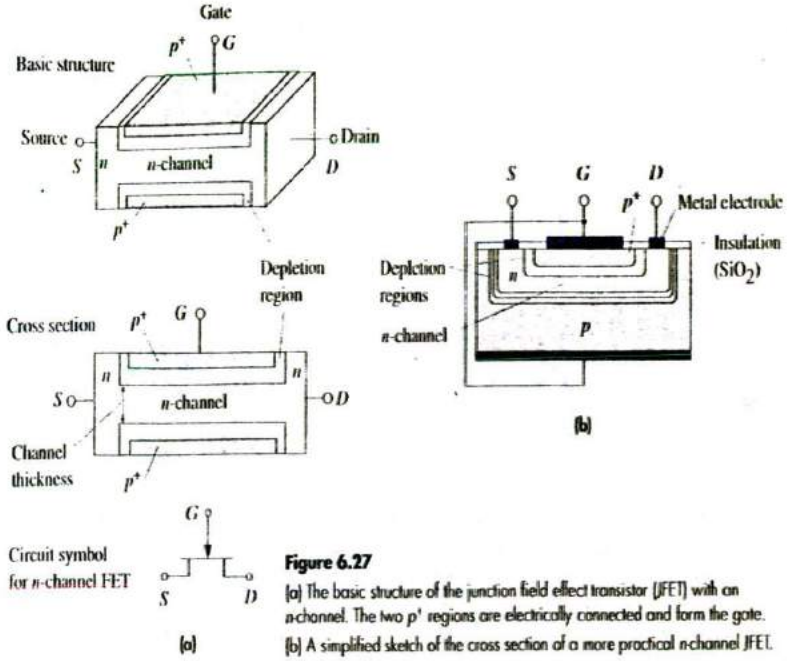


Figure 6.27

(a) The basic structure of the junction field effect transistor (JFET) with an n-channel. The two p^+ regions are electrically connected and form the gate. (b) A simplified sketch of the cross section of a more practical n-channel JFET.

the structure of a typical practical device. A simplified schematic sketch of the cross section of a more practical device (as, for example, fabricated by the planar technology) is shown in Figure 6.27b where it is apparent that the two gate regions do not have identical doping and that, except for one of the gates, all contacts are on one surface.

We first consider the behavior of the JFET with the gate and source shorted ($V_{GS} = 0$), as shown in Figure 6.28a. The resistance between S and D is essentially the resistance of the conducting n-channel between A and B. R_{AB}. When a positive voltage is applied to D with respect to S ($V_{DS} > 0$), then a current flows from D to S, which is called the **drain current** I_D. There is a voltage drop along the channel, between A and B, as indicated in Figure 6.28a. The voltage in the n-channel is zero at A and V_{DS} at B. As the voltage along the n-channel is positive, the p^+n junctions between the gates and the n-channel become progressively more reverse-biased from A to B. Consequently the depletion layers extend more into the channel and thereby decrease the thickness of the conducting channel from A to B.

Increasing V_{DS} increases the widths of the depletion layers, which penetrate more into the channel and hence result in more channel narrowing toward the drain. The resistance of the n-channel R_{AB} therefore increases with V_{DS}. The drain current therefore does not increase linearly with V_{DS} but falls below it because

$$I_D = \frac{V_{DS}}{R_{AB}}$$

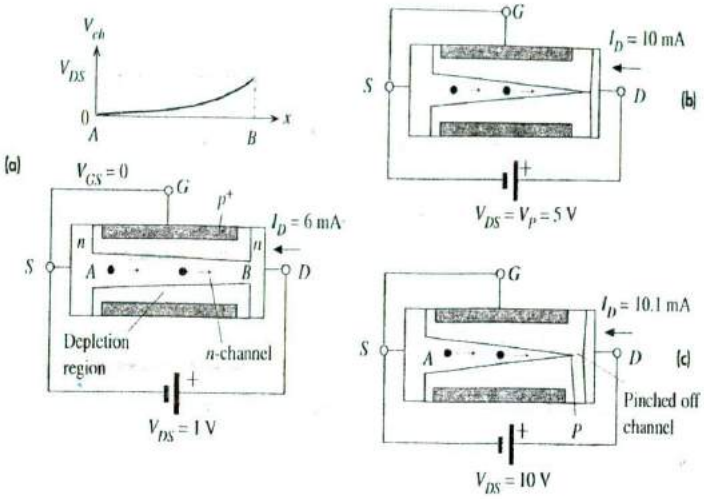


Figure 6.28

(a) The gate and source are shorted ($V_{GS} = 0$) and V_{DS} is small.

(b) V_{DS} has increased to a value that allows the two depletion layers to just touch, when $V_{DS} = V_P (= 5\text{ V})$ and the p^+n junction voltage at the drain end, $V_{GD} = -V_{DS} = -V_P = -5$ V.

(c) V_{DS} is large ($V_{DS} > V_P$), so a short length of the channel is pinched off.

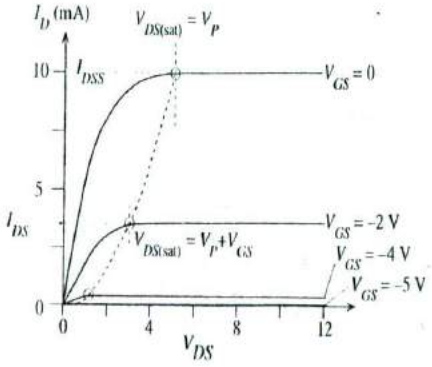


Figure 6.29 Typical I_D versus V_{DS} characteristics of a JFET for various fixed gate voltages V_{GS}.

and R_{AB} increases with V_{DS}. Thus I_D versus V_{DS} exhibits a sublinear behavior, as shown in the $V_{DS} < 5$ V region in Figure 6.29.

As V_{DS} increases further, the depletion layers extend more into the channel and eventually, when $V_{DS} = V_P (= 5\text{ V})$, the two depletion layers around B meet at point P at the drain end of the channel, as depicted in Figure 6.28b. The channel is then said to be "pinched off" by the two depletion layers. The voltage V_P is called the **pinch-off voltage.** It is equal to the magnitude of reverse bias needed across the p^+n junctions to

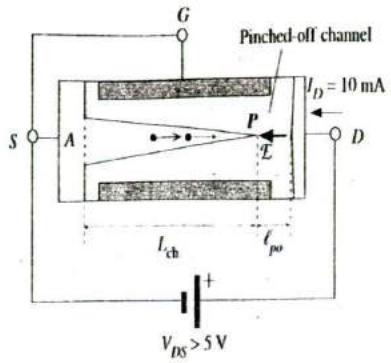


Figure 6.30 The pinched-off channel and conduction for $V_{DS} > V_P$ (= 5 V).

make them just touch at the drain end. Since the actual bias voltage across the p^+n junctions at the drain end (B) is V_{GD}, the pinch-off occurs whenever

$$V_{GD} = -V_P \qquad [6.52]$$

Pinch-off condition

In the present case, gate to source is shorted, $V_{GS} = 0$, so $V_{GD} = -V_{DS}$ and pinch-off occurs when $V_{DS} = V_P$ (5 V). The drain current from pinch-off onwards, as shown in Figure 6.29, does not increase significantly with V_{DS} for reasons given below. Beyond $V_{DS} = V_P$, there is a short pinched-off channel of length ℓ_{po}.

The pinched-off channel is a reverse-biased depletion region that separates the drain from the n-channel, as depicted in Figure 6.30. There is a very strong electric field $\mathcal{E}$ in this pinched-off region in the D to S direction. This field is the vector sum of the fields from positive donors to negative acceptors in the depletion regions of the channel and the gate on the drain side. Electrons in the n-channel drift toward P, and when they arrive at P, they are swept across the pinched-off channel by $\mathcal{E}$. This process is similar to minority carriers in the base of a BJT reaching the collector junction depletion region, where the internal field there sweeps them across the depletion layer into the collector. Consequently the drain current is actually determined by the resistance of the conducting n-channel over L_{ch} from A to P in Figure 6.30 and not by the pinched-off channel.

As V_{DS} increases, most of the additional voltage simply drops across ℓ_{po} as this region is depleted of carriers and hence highly resistive. Point P, where the depletion layers first meet, moves slightly toward A, thereby slightly reducing the channel length L_{ch}. Point P must still be at a potential V_P because it is this potential that just makes the depletion layers touch. Thus the voltage drop across L_{ch} remains as V_P. Beyond pinch-off then

$$I_D = \frac{V_P}{R_{AP}} \qquad (V_{DS} > V_P)$$

Since R_{AP} is determined by L_{ch}, which decreases slightly with V_{DS}, I_D increases slightly with V_{DS}. In many cases, I_D is conveniently taken to be saturated at a value I_{DSS} for $V_{DS} > V_P$. Typical I_D versus V_{DS} behavior is shown in Figure 6.29.

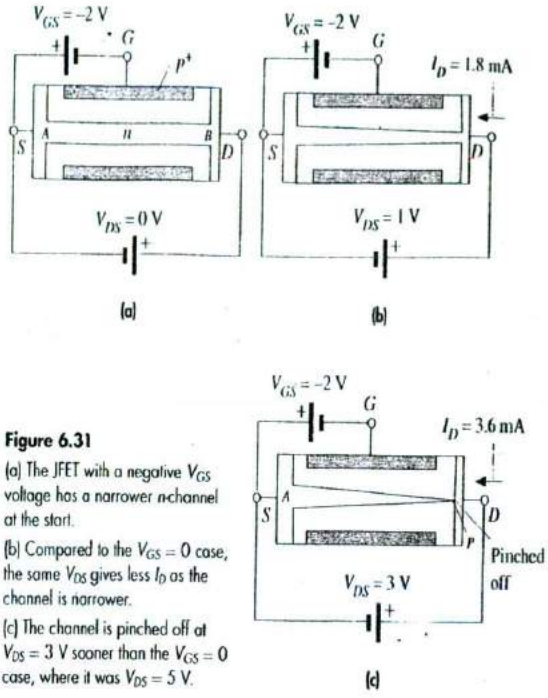


Figure 6.31

(a) The JFET with a negative V_{GS} voltage has a narrower n-channel at the start.

(b) Compared to the $V_{GS} = 0$ case, the same V_{DS} gives less I_D as the channel is narrower.

(c) The channel is pinched off at $V_{DS} = 3$ V sooner than the $V_{GS} = 0$ case, where it was $V_{DS} = 5$ V.

We now consider what happens when a negative voltage, say $V_{GS} = -2$ V, is applied to the gate with respect to the source, as shown in Figure 6.31a with $V_{DS} = 0$. The p^+n junctions are now reverse-biased from the start, the channel is narrower, and the channel resistance is now larger than in the $V_{GS} = 0$ case. The drain current that flows when a small V_{DS} is applied, as in Figure 6.31b, is now smaller than in the $V_{GS} = 0$ case as apparent in Figure 6.29. The p^+n junctions are now progressively more reverse-biased from V_{GS} at the source end to $V_{GD} = V_{GS} - V_{DS}$ at the drain end. We therefore need a smaller V_{DS} (= 3 V) to pinch off the channel, as shown in Figure 6.31c. When $V_{DS} = 3$ V, the G to D voltage V_{GD} across the p^+n junctions at the drain end is -5 V, which is $-V_P$, so the channel becomes pinched off. Beyond pinch-off, I_D is nearly saturated just as in the $V_{GS} = 0$ case, but its magnitude is obviously smaller as the thickness of the channel at A is smaller; compare Figures 6.28 and 6.31. In the presence of V_{GS}, the pinch-off occurs at $V_{DS} = V_{DS(sat)}$, and from Equation 6.52.

Pinch-off condition

$$V_{DS(sat)} = V_P + V_{GS} \qquad [6.53]$$

where V_{GS} is a negative voltage (reducing V_P). Beyond pinch-off when $V_{DS} > V_{DS(sat)}$, the point P where the channel is *just pinched* still remains at potential $V_{DS(sat)}$, given by Equation 6.53.

For $V_{DS} > V_{DS(sat)}$, I_D becomes nearly saturated at a value denoted as I_{DS}, which is indicated in Figure 6.29. When G and S are shorted ($V_{GS} = 0$), I_{DS} is called I_{DSS} (which

stands for I_{DS} with shorted gate to source). Beyond pinch-off, with negative V_{GS}, I_{DS} is

$$I_D \approx I_{DS} \approx \frac{V_{DS(sat)}}{R_{AP}(V_{GS})} = \frac{V_P + V_{GS}}{R_{AP}(V_{GS})} \qquad V_{DS} > V_{DS(sat)} \qquad [6.54]$$

where R_{AP} (V_{GS}) is the effective resistance of the conducting n-channel from A to P (Figure 6.31b), which depends on the channel thickness and hence on V_{GS}. The resistance increases with more negative gate voltage as this increases the reverse bias across the p^+n junctions, which leads to the narrowing of the channel. For example, when $V_{GS} = -4$ V, the channel thickness at A becomes narrower than in the case with $V_{GS} = -2$ V, thereby increasing the resistance, R_{AP}, of the conducting channel and therefore decreasing I_{DS}. Further, there is also a reduction in the drain current by virtue of $V_{DS(sat)}$ decreasing with negative V_{GS}, as apparent in Equation 6.54. Figure 6.29 shows the effect of the gate voltage on the I_D versus V_{DS} behavior. The two effects, that from $V_{DS(sat)}$ and that from R_{AP} (V_{GS}) in Equation 6.54, lead to I_{DS} almost decreasing parabolically with $-V_{GS}$.

When the gate voltage is such that $V_{GS} = -V_P$ (= −5 V) with the source and drain shorted ($V_{DS} = 0$), then the two depletion layers touch over the entire channel length and the whole channel is closed, as illustrated in Figure 6.32. The channel is said to be off. The only drain current that flows when a V_{DS} is applied is due to the thermally generated carriers in the depletion layers. This current is very small.

Figure 6.29 summarizes the full I_D versus V_{DS} characteristics of the n-channel JFET at various gate voltages V_{GS}. It is apparent that I_{DS} is relatively independent of V_{DS} and that it is controlled by the gate voltage V_{GS}, as expected by Equation 6.54. This is analogous to the BJT in which the collector current I_C is controlled by the base–emitter bias voltage V_{BE}. Figure 6.33a shows the dependence of I_{DS} on the gate voltage V_{GS}. The transistor action is the control of the drain current I_{DS}, in the drain–source (output) circuit by the voltage V_{GS} in the gate–source (input circuit), as shown in Figure 6.33b. This control is only possible if $V_{DS} > V_{DS(sat)}$. When $V_{GS} = -V_P$, the drain current is nearly zero because the channel has been totally pinched off. This gate–source voltage is denoted by $V_{GS(off)}$ as the drain current has been switched off. Furthermore, we should note that as V_{GS} reverse biases the p^+n junction, the current into the gate I_G is the reverse leakage current of these junctions. It is usually very small. In some JFETs, I_G is as low as a fraction of a nanoampere. We should also note that the circuit symbol for the JFET, as shown in Figure 6.27a, has an arrow to identify the gate and the pn junction direction.

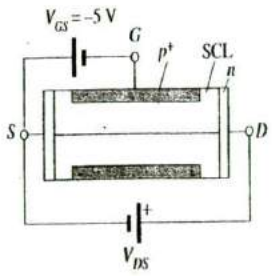


Figure 6.32 When $V_{GS} = -5$ V, the depletion layers close the whole channel from the start, at $V_{DS} = 0$. As V_{DS} is increased, there is a very small drain current, which is the small reverse leakage current due to thermal generation of carriers in the depletion layers.

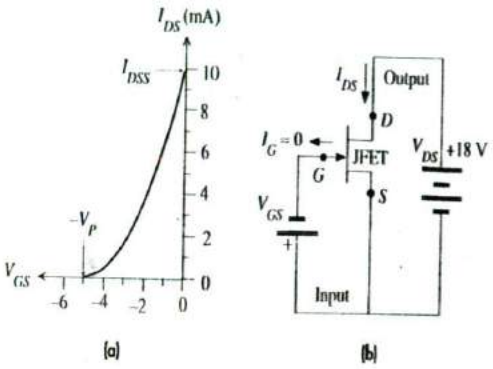


Figure 6.33

(a) Typical I_{DS} versus V_{GS} characteristics of a JFET.

(b) The dc circuit where V_{GS} in the gate–source circuit (input) controls the drain current I_{DS} in the drain–source (output) circuit in which V_{DS} is kept constant and large ($V_{DS} > V_P$).

Is there a convenient relationship between I_{DS} and V_{GS}? If we calculate the effective resistance R_{AP} of the n-channel between A and P, we can obtain its dependence on the channel thickness, and thus on the widths of the depletion layers and hence on V_{GS}. We can then find I_{DS} from Equation 6.54. It turns out that a simple parabolic dependence seems to represent the data reasonably well,

Beyond pinch-off

$$I_{DS} = I_{DSS}\left[1 - \left(\frac{V_{GS}}{V_{GS(off)}}\right)\right]^2 \qquad [6.55]$$

where I_{DSS} is the drain current when $V_{GS} = 0$ (Figure 6.33) and $V_{GS(off)}$ is defined as $-V_P$, that is, that gate–source voltage that just pinches off the channel. The pinch-off voltage V_P here is a positive quantity because it was introduced through $V_{DS(sat)}$. $V_{GS(off)}$ however is negative, $-V_P$. We should note two important facts about the JFET. Its name originates from the effect that modulating the electric field in the reverse-biased depletion layers (by changing V_{GS}) varies the depletion layer penetration into the channel and hence the resistance of the channel. The transistor action hence can be thought of as being based on a **field effect.** Since there is a p^+n junction between the gate and the channel, the name has become JFET. This junction in reverse bias provides the isolation between the gate and channel.

Secondly, the region beyond pinch-off, where Equations 6.54 and 6.55 hold, is commonly called the **current saturation region,** as well as **constant current region** and **pentode region.** The term **saturation** should not be confused with similar terms used for saturation effects in bipolar transistors. A saturated BJT cannot be used as an amplifier, but JFETs are invariably used as amplifiers in the saturated current region.

6.7.2 JFET AMPLIFIER

The transistor action in the JFET is the control of I_{DS} by V_{GS}, as shown in Figure 6.33. The input circuit is therefore the gate–source circuit containing V_{GS} and the output circuit is the drain–source circuit in which the drain current I_{DS} flows. The JFET is almost never used with the pn junction between the gate and channel forward-biased ($V_{GS} > 0$) as this would lead to a very large gate current and near shorting of the gate to source

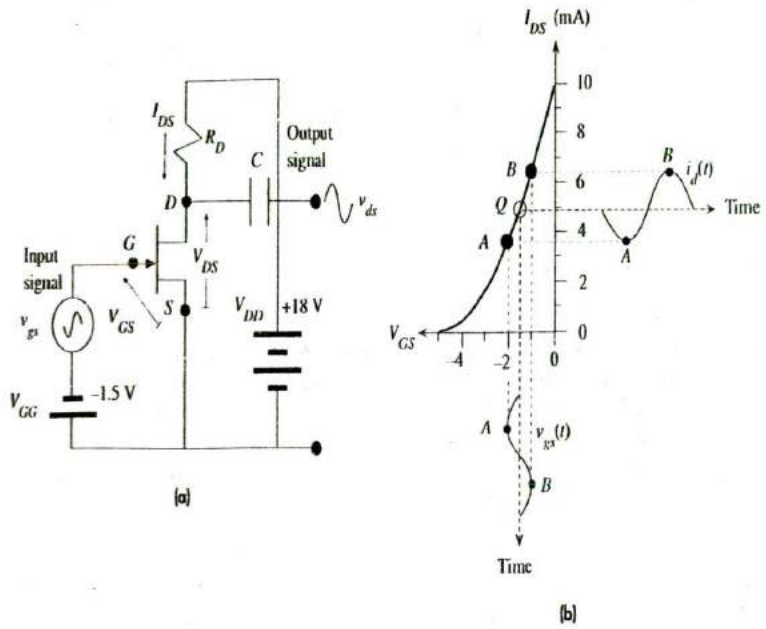


Figure 6.34

(a) Common source (CS) ac amplifier using a JFET.

(b) Explanation of how I_D is modulated by the signal v_{gs} in series with the dc bias voltage V_{GG}.

voltage. With V_{GS} limited to negative voltages, the maximum current in the output circuit can only be I_{DSS}, as shown in Figure 6.33a. The maximum input voltage V_{GS} should therefore give an I_{DS} less than I_{DSS}.

Figure 6.34a shows a simplified illustration of a typical JFET voltage amplifier. As the source is common to both the input and output circuits, this is called a **common source (CS) amplifier.** The input signal is the ac source v_{gs} connected in series with a negative dc bias voltage V_{GG} of -1.5 V in the GS circuit. First we will find out what happens when there is no ac signal in the circuit ($v_{gs} = 0$). The dc supply (-1.5 V) in the input provides a negative dc voltage to the gate and therefore gives a dc current I_{DS} in the output circuit (less than I_{DSS}). Figure 6.34b shows that when $V_{GS} = -1.5$ V, point Q on the I_{DS} versus V_{GS} characteristics gives $I_{DS} = 4.9$ mA. Point Q, which determines the dc operation, is called the **quiescent point.**

The ac source v_{gs} is connected in series with the negative dc bias voltage V_{GS}. It therefore modulates V_{GS} up and down about -1.5 V with time, as shown in Figure 6.34b. Suppose that v_{gs} varies sinusoidally between -0.5 V and $+0.5$ V. Then, as shown in Figure 6.34b when v_{gs} is -0.5 V (point A), $V_{GS} = -2.0$ V and the drain current is given by point A on the I_{DS}–V_{GS} curve and is about 3.6 mA. When v_{gs} is $+0.5$ V (point B), then $V_{GS} = -1.0$ V and the drain current is given by point B on the I_{DS}–V_{GS} curve and is about 6.4 mA. The input variation from -0.5 V to $+0.5$ V has thus been

Table 6.1 Voltage and current in the common source amplifier of Figure 6.34a

v_{gs} (V)	V_{GS} (V)	I_{DS} (mA)	i_d (mA)	$V_{DS} = V_{DD} - I_{DS}R_D$	v_{ds} (V)	Voltage Gain	Comment
0	−1.5	4.9	0	8.2	0		dc conditions, point Q
−0.5	−2.0	3.6	−1.3	10.8	+2.6	−5.2	Point A
+0.5	−1.0	6.4	+1.5	5.2	−3.0	−6	Point B

NOTE: $V_{DD} = 18$ V and $R_D = 2000\ \Omega$.

converted to a drain current variation from 3.6 mA to 6.4 mA as indicated in Figure 6.34b. We could have just as easily calculated the drain current from Equation 6.55. Table 6.1 summarizes what happens to the drain current as the ac input voltage is varied about zero.

The change in the drain current with respect to its dc value is the output signal current denoted as i_d. Thus at A,

$$i_d = 3.6 - 4.9 = -1.3 \text{ mA}$$

and at B,

$$i_d = 6.4 - 4.9 = 1.5 \text{ mA}$$

The variation in the output current is not quite symmetric as that in the input signal v_{gs} because the I_{DS}–V_{GS} relationship, Equation 6.55, is not linear.

The drain current variations in the DS circuit are converted to voltage variations by the resistance R_D. The voltage across DS is

$$V_{DS} = V_{DD} - I_{DS}R_D \qquad [6.56]$$

where V_{DD} is the bias battery voltage in the DS circuit. Thus, variations in I_{DS} result in variations in V_{DS} that are in the opposite direction or 180° out of phase. The ac output voltage between D and S is tapped out through a capacitor C, as shown in Figure 6.34a. The capacitor C simply blocks the dc. Suppose that $R_D = 2000\ \Omega$ and $V_{DD} = 18$ V, then using Equation 6.56 we can calculate the dc value of V_{DS} and also the minimum and maximum values of V_{DS}, as shown in Table 6.1.

It is apparent that as v_{gs} varies from −0.5 V, at A, to +0.5 V, at B, V_{DS} varies from 10.8 V to 5.2 V, respectively. The change in V_{DS} with respect to dc is what constitutes the output signal v_{ds}, as only the ac is tapped out. From Equation 6.56, the change in V_{DS} is related to the change in I_{DS} by

$$v_{ds} = -R_D i_d \qquad [6.57]$$

Thus the output, v_{ds}, changes from −3.0 V to 2.6 V. The peak-to-peak voltage amplification is

$$A_{V(\text{pk-pk})} = \frac{\Delta V_{DS}}{\Delta V_{GS}} = \frac{v_{ds(\text{pk-pk})}}{v_{gs(\text{pk-pk})}} = \frac{-3\text{ V} - (2.6\text{ V})}{0.5\text{ V} - (-0.5\text{ V})} = -5.6$$

The negative sign represents the fact that the output and input voltages are out of phase by 180°. This can also be seen from Table 6.1 where a negative v_{gs} results in a positive v_{ds}. Even though the ac input signal v_{gs} is symmetric about zero, ±0.5 V, the ac output signal v_{ds} is not symmetric, which is due to the I_{DS} versus V_{GS} curve being nonlinear, and thus varies between −3.0 V and 2.6 V. If we were to calculate the voltage amplification for the most negative input signal, we would find −5.2, whereas for the most positive input signal, it would be −6. The peak-to-peak voltage amplification, which was −5.6, represents a mean gain taking both negative and positive input signals into account.

The amplification can of course be increased by increasing R_D, but we must maintain V_{DS} at all times above $V_{DS(sat)}$ (beyond pinch-off) to ensure that the drain current I_{DS} in the output circuit is only controlled by V_{GS} in the input circuit.

When the signals are small about dc values, we can use differentials to represent small signals. For example, $v_{gs} = \delta V_{GS}$, $i_d = \delta I_{DS}$, $v_{ds} = \delta V_{DS}$, and so on. The variation δI_{DS} due to δV_{GS} about the dc value may be used to define a **mutual transconductance** g_m (sometimes denoted as g_{fs}) for the JFET,

$$g_m = \frac{dI_{DS}}{dV_{GS}} \approx \frac{\delta I_{DS}}{\delta V_{GS}} = \frac{i_d}{v_{gs}}$$

Definition of JFET transconductance

This transconductance can be found by differentiating Equation 6.55,

$$g_m = \frac{dI_{DS}}{dV_{GS}} = -\frac{2I_{DSS}}{V_{GS(off)}}\left[1 - \left(\frac{V_{GS}}{V_{GS(off)}}\right)\right] = -\frac{2|I_{DSS}I_{DS}|^{1/2}}{V_{GS(off)}} \qquad [6.58]$$

JFET transconductance

The output signal current is

$$i_d = g_m v_{gs}$$

so using Equation 6.57, the small-signal voltage amplification is

$$A_V = \frac{v_{ds}}{v_{gs}} = \frac{-R_D(g_m v_{gs})}{v_{gs}} = -g_m R_D \qquad [6.59]$$

Small-signal voltage gain

Equation 6.59 is only valid under small-signal conditions in which the variations about the dc values are small compared with the dc values themselves. The negative sign indicates that v_{ds} and v_{gs} are 180° out of phase.

EXAMPLE 6.12

THE JFET AMPLIFIER Consider the *n*-channel JFET common source amplifier shown in Figure 6.34a. The JFET has an I_{DSS} of 10 mA and a pinch-off voltage V_P of 5 V as in Figure 6.34b. Suppose that the gate dc bias voltage supply $V_{GG} = -1.5$ V, the drain circuit supply $V_{DD} = 18$ V, and $R_D = 2000\ \Omega$. What is the voltage amplification for small signals? How does this compare with the peak-to-peak amplification of −5.6 found for an input signal that had a peak-to-peak value of 1 V?

SOLUTION

We first calculate the operating conditions at the bias point with no ac signals. This corresponds to point Q in Figure 6.34b. The dc bias voltage V_{GS} across the gate to source is −1.5 V. The

resulting dc drain current I_{DS} can be calculated from Equation 6.55 with $V_{GS(off)} = -V_P = -5$ V:

$$I_{DS} = I_{DSS}\left[1 - \left(\frac{V_{GS}}{V_{GS(off)}}\right)\right]^2 = (10 \text{ mA})\left[1 - \left(\frac{-1.5}{-5}\right)\right]^2 = 4.9 \text{ mA}$$

The transconductance at this dc current (at Q) is given by Equation 6.58,

$$g_m = -\frac{2(I_{DSS}I_{DS})^{1/2}}{V_{GS(off)}} = -\frac{2[(10 \times 10^{-3})(4.9 \times 10^{-3})]^{1/2}}{-5} = 2.8 \times 10^{-3} \text{ A/V}$$

The voltage amplification of small signals about point Q is

$$A_V = -g_m R_D = -(2.8 \times 10^{-3})(2000) = -5.6$$

This turns out to be the same as the peak-to-peak voltage amplification we calculated in Table 6.1. When the input ac signal v_{gs} varies between −0.5 and +0.5 V, as in Table 6.1, the output signal is not symmetric. It varies between −3 V and 2.8 V, so the voltage gain depends on the input signal. The amplifier is then said to exhibit **nonlinearity.**

6.8 METAL-OXIDE-SEMICONDUCTOR FIELD EFFECT TRANSISTOR (MOSFET)

6.8.1 FIELD EFFECT AND INVERSION

The metal-oxide-semiconductor field effect transistor is based on the effect of a field penetrating into a semiconductor. Its operation can be understood by first considering a parallel plate capacitor with metal electrodes and a vacuum as insulation in between, as shown in Figure 6.35a. When a voltage V is applied between the plates, charges $+Q$ and $-Q$ (where $Q = CV$) appear on the plates and there is an electric field given by $\mathcal{E} = V/L$. The origins of these charges are the conduction electrons for $-Q$ and exposed positively charged metal ions for $+Q$. Metallic bonding is based on all the valence electrons forming a sea of conduction electrons and permeating the space between metal ions that are fixed at crystal lattice sites. Since the electrons are mobile, they are readily displaced by the field. Thus in the lower plate $\mathcal{E}$ displaces some of the conduction electrons to the surface to form $-Q$. In the top plate $\mathcal{E}$ displaces some electrons from the surface into the bulk to expose positively charged metal ions to form $+Q$.

Suppose that the plate area is 1 cm^2 and spacing is 0.1 μm and that we apply 2 V across it. The capacitance C is 8.85 nF and the magnitude of charge Q on each plate is 1.77×10^{-8} C, which corresponds to 1.1×10^{11} electrons. A typical metal such as copper has something like 1.9×10^{15} atoms per cm^2 on the surface. Thus, there will be that number of positive metal ions and electrons on the surface (assuming one conduction electron per atom). The charges $+Q$ and $-Q$ can therefore be generated by the electrons and metal ions at the surface alone. For example, if one in every 1.7×10^4 electrons on the surface moves one atomic spacing (~0.3 nm) into the bulk, then the surface will have a charge of $+Q$ due to exposed positive metal ions. It is clear that, for all practical purposes, the electric field does not penetrate into the metal and terminates at the metal surface.

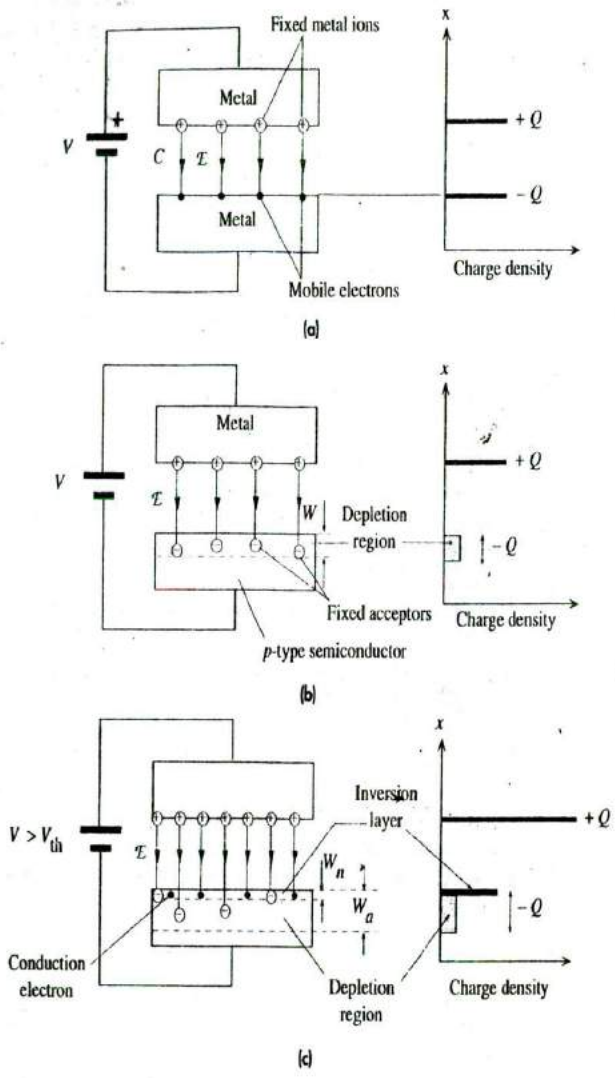


Figure 6.35 The field effect.

(a) In a metal-air-metal capacitor, all the charges reside on the surface.

(b) Illustration of field penetration into a p-type semiconductor.

(c) As the field increases, eventually when $V > V_t$, an inversion layer is created near the surface in which there are conduction electrons.

The same is not true when one of the electrodes is a semiconductor, as shown in Figure 6.35b where the structure now is of the metal-insulator-semiconductor type. Suppose that we replace the lower metal in Figure 6.35a with a p-type semiconductor with an acceptor concentration of $10^{15}\ \text{cm}^{-3}$. The number of acceptor atoms on the surface[10] is $1 \times 10^{10}\ \text{cm}^{-2}$. We may assume that at room temperature all the acceptors are ionized and thus negatively charged. It is immediately apparent that we do not have a sufficient number of negative acceptors at the surface to generate the charge $-Q$. We must therefore also expose negative acceptors in the bulk, which means that the field must penetrate into the semiconductor. Holes in the surface region of the semiconductor become repelled toward the bulk and thereby expose more negative acceptors. We can estimate the width W into which the field penetrates since the total negative charge exposed $eAWN_a$ must be Q. We find that W is of the order of 1 μm, which is something like 4000 atomic layers. Our conclusion is that the field penetrates into a semiconductor by an amount that depends on the doping concentration.

The penetrating field into the semiconductor drifts away most of the holes in this region and thereby exposes negatively charged acceptors to make up the charge $-Q$. The region into which the field penetrates has lost holes and is therefore depleted of its equilibrium concentration of holes. We refer to this region as a **depletion layer.** As long as $p > n$ even though $p \ll N_a$, this still has p-type characteristics as holes are in the majority.

If the voltage increases further, $-Q$ also increases, as the field becomes stronger and penetrates more into the semiconductor but eventually it becomes more difficult to make up the charge $-Q$ by simply extending the depletion layer width W into the bulk. It becomes possible (and more favorable) to attract conduction electrons into the depletion layer and form a thin electron layer of width W_n near the surface. The charge $-Q$ is now made up of the fixed negative charge of acceptors in W_a and of conduction electrons in W_n, as shown in Figure 6.35c. Further increases in the voltage do not change the width W_a of the depletion layer but simply increase the electron concentration in W_n. Where do these electrons come from as the semiconductor is doped p-type? Some are attracted into the depletion layer from the bulk, where they were minority carriers. But most are thermally generated by the breaking of Si–Si bonds (*i.e.*, across the bandgap) in the depleted layer. Thermal generation in the depletion layer generates electron–hole pairs that become separated by the field. The holes are then drifted by the field into the bulk and the electrons toward the surface. Recombination of the thermally generated electrons and holes with other carriers is greatly reduced because the depletion layer has so few carriers. Since the electron concentration in the electron layer exceeds the hole concentration and this layer is within a normally p-type semiconductor, we call this an **inversion layer.**

It is now apparent that increasing the field in the metal-insulator-semiconductor device first creates a depletion layer and then an inversion layer at the surface when the voltage exceeds some threshold value V_{th}. This is the basic principle of the field effect device. As long as $V > V_{th}$, any increase in the field and hence $-Q$ leads to more electrons in the inversion layer, whereas the width of the depletion layer W_a and hence the quantity

[10] Surface concentration of atoms (atoms per unit area) can be found from $n_{surf} \approx (n_{bulk})^{2/3}$.

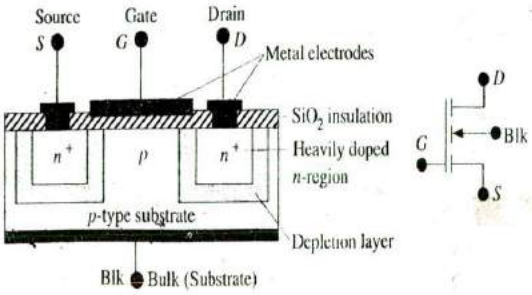


Figure 6.36 The basic structure of the enhancement MOSFET and its circuit symbol.

of fixed negative charge remain constant. The insulator between the metal and the semiconductor, that is, a vacuum in Figure 6.35, is typically SiO_2 in many devices.

6.8.2 ENHANCEMENT MOSFET

Figure 6.36 shows the basic structure of an enhancement n-channel MOSFET device (NMOSFET). A metal-insulator-semiconductor structure is formed between a p-type Si substrate and an aluminum electrode, which is called the gate (G). The insulator is the SiO_2 oxide grown during fabrication. There are two n^+ doped regions at the ends of the MOS device that form the source (S) and drain (D). A metal contact is also made to the p-type Si substrate (or the bulk), which in many devices is connected to the source terminal as shown in Figure 6.36. Further, many MOSFETs have a degenerately doped polycrystalline Si material as the gate that serves the same function as the metal electrode.

With no voltage applied to the gate, S to D is an n^+pn^+ structure that is always reverse-biased whatever the polarity of the source to drain voltage. However, if the substrate (bulk) is connected to the source, a negative V_{DS} will forward bias the n^+p junction between the drain and the substrate. As the n-channel MOSFET device is not normally used with a negative V_{DS}, we will not consider this polarity.

When a positive voltage less than V_{th} is applied to the gate, $V_{GS} < V_{th}$, as shown in Figure 6.37a, the p-type semiconductor under the gate develops a depletion layer as a result of the expulsion of holes into the bulk, just as in Figure 6.35b. Since S and D are isolated by a low-conductivity p-doped region that has a depletion layer from S to D, no current can flow for any positive V_{DS}.

With $V_{DS} = 0$, as soon as V_{GS} is increased beyond the threshold voltage V_{th}, an n-channel inversion layer is formed within the depletion layer under the gate and immediately below the surface, as shown in Figure 6.37b. This n-channel links the two n^+ regions of source and drain. We then have a continuous n-type material with electrons as mobile carriers between the source and drain. When a small V_{DS} is applied, a drain current I_D flows that is limited by the resistance of the n-channel $R_{n\text{-ch}}$:

$$I_D = \frac{V_{DS}}{R_{n\text{-ch}}} \qquad [6.60]$$

Thus, I_D initially increases with V_{DS} almost linearly, as shown in Figure 6.37b.

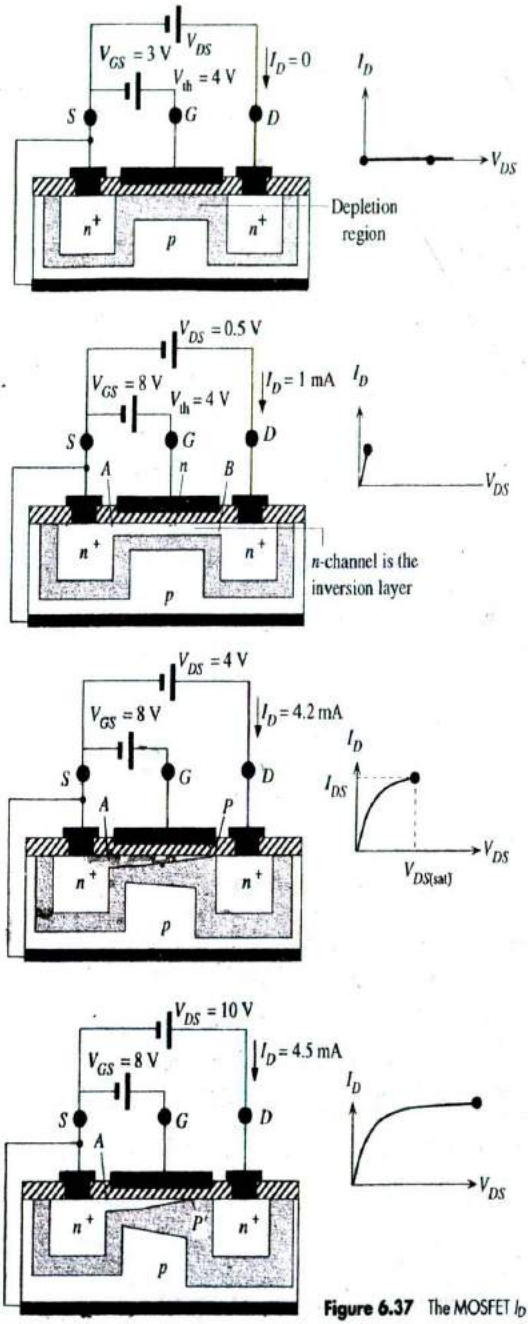


Figure 6.37 The MOSFET I_D versus V_{DS} characteristics.

The voltage variation along the channel is from zero at A (source end) to V_{DS} at B (drain end). The gate to the n-channel voltage is then V_{GS} at A and $V_{GD} = V_{GS} - V_{DS}$ at B. Thus point A depends only on V_{GS} and remains undisturbed by V_{DS}. As V_{DS} increases, the voltage at B (V_{GD}) decreases and thereby causes less inversion. This means that the channel gets narrower from A to B and its resistance $R_{n\text{-ch}}$ increases with V_{DS}. I_D versus V_{DS} then falls increasingly below the $I_D \propto V_{DS}$ line. Eventually when the gate to n-channel voltage at B decreases to just below V_{th}, the inversion layer at B disappears and a depletion layer is exposed, as illustrated in Figure 6.37c. The n-channel becomes pinched off at this point P. This occurs when $V_{DS} = V_{DS(sat)}$, satisfying

$$V_{GD} = V_{GS} - V_{DS(sat)} = V_{th} \qquad [6.61]$$

It is apparent that the whole process of the narrowing of the n-channel and its eventual pinch-off is similar to the operation of the n-channel JFET. When the drifting electrons in the n-channel reach P, the large electric field within the very narrow depletion layer at P sweeps the electrons across into the n^+ drain. The current is limited by the supply of electrons from the n-channel to the depletion layer at P, which means that it is limited by the effective resistance of the n-channel between A and P.

When V_{DS} exceeds $V_{DS(sat)}$, the additional V_{DS} drops mainly across the highly resistive depletion layer at P, which extends slightly to P' toward A, as shown in Figure 6.37d. At P', the gate to channel voltage must still be just V_{th} as this is the voltage required to just pinch off the channel and just eliminate inversion. The widening of the depletion layer (from B to P') at the drain end with V_{DS}, however, is small compared with the channel length AB. The resistance of the channel from A to P' does not change significantly with increasing V_{DS}, which means that the drain current is then nearly saturated at I_{DS},

$$i_D \approx I_{DS} \approx \frac{V_{DS(sat)}}{R_{AP'n\text{-ch}}} \qquad V_{DS} > V_{DS(sat)} \qquad [6.62]$$

As $V_{DS(sat)}$ depends on V_{GS}, so does I_{DS}. The overall I_{DS} versus V_{DS} characteristics for various fixed gate voltages V_{GS} of a typical enhancement MOSFET is shown in Figure 6.38a. It can be seen that there is only a slight increase in I_{DS} with V_{DS} beyond $V_{DS(sat)}$. The I_{DS} versus V_{GS} when $V_{DS} > V_{DS(sat)}$ characteristics are shown in Figure 6.38b. It is apparent that as long as $V_{DS} > V_{DS(sat)}$, the saturated drain current I_{DS} in the source–drain (or output) circuit is almost totally controlled by the gate voltage V_{GS} in the source–gate (or input) circuit. This is what constitutes the MOSFET action. Variations in V_{GS} then lead to variations in the drain current I_{DS} (just as in the JFET), which forms the basis of the MOSFET amplifier. The term *enhancement* refers to the fact that a gate voltage exceeding V_{th} is required to enhance a conducting channel between the source and drain. This contrasts with the JFET where the gate voltage depletes the channel and decreases the drain current.

The experimental relationship between I_{DS} and V_{GS} (when $V_{DS} > V_{DS(sat)}$) has been found to be best described by a parabolic equation similar to that for the JFET, except that now V_{GS} enhances the channel when $V_{GS} > V_{th}$ so I_{DS} exists only when $V_{GS} > V_{th}$.

Enhancement NMOSFET

$$I_{DS} = K(V_{GS} - V_{th})^2 \qquad [6.63]$$

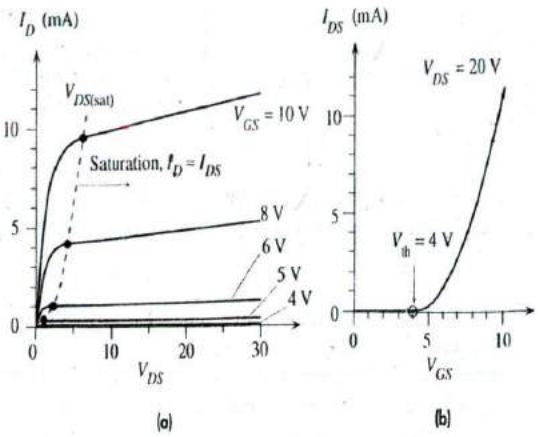


Figure 6.38
(a) Typical I_D versus V_{DS} characteristics of an enhancement MOSFET (V_{th} = 4 V) for various fixed gate voltages V_{GS}. (b) Dependence of I_{DS} on V_{GS} at a given V_{DS} [> $V_{DS(sat)}$].

where K is a constant. For an ideal MOSFET, it can be expressed as

Enhancement NMOSFET constant

$$K = \frac{Z\mu_e\varepsilon}{2Lt_{ox}}$$

where μ_e is the electron drift mobility in the channel, L and Z are the length and width of the gate controlling the channel, and ε and t_{ox} are the permittivity ($\varepsilon_r\varepsilon_0$) and thickness of the oxide insulation under the gate. According to Equation 6.63, I_{DS} is independent of V_{DS}. The shallow slopes of the I_D versus V_{DS} lines beyond $V_{DS(sat)}$ in Figure 6.38a can be accounted for by writing Equation 6.63 as

Enhancement NMOSFET

$$I_{DS} = K(V_{GS} - V_{th})^2(1 + \lambda V_{DS}) \qquad [6.64]$$

where λ is a constant that is typically 0.01 V^{-1}. If we extend the I_{DS} versus V_{DS} lines, they intersect the $-V_{DS}$ axis at $1/\lambda$, which is called the **Early voltage.** It should be apparent that I_{DSS}, which is I_{DS} with the gate and source shorted ($V_{GS} = 0$), is zero and is not a useful quantity in describing the behavior of the enhancement MOSFET.

EXAMPLE 6.13 **THE ENHANCEMENT NMOSFET** A particular enhancement NMOS transistor has a gate with a width (Z) of 50 μm, length (L) of 10 μm, and SiO_2 thickness of 450 Å. The relative permittivity of SiO_2 is 3.9. The p-type bulk is doped with 10^{16} acceptors cm^{-3}. Its threshold voltage is 4 V. Estimate the drain current when V_{GS} = 8 V and V_{DS} = 20 V, given λ = 0.01. Due to the strong scattering of electrons near the crystal surface assume that the electron drift mobility μ_e in the channel is half the drift mobility in the bulk.

SOLUTION

Since $V_{DS} > V_{th}$, we can assume that the drain current is saturated and we can use the I_{DS} versus V_{GS} relationship in Equation 6.64,

$$I_{DS} = K(V_{GS} - V_{th})^2(1 + \lambda V_{DS})$$

where

$$K = \frac{Z\mu_e\varepsilon}{2Lt_{ox}}$$

The electron mobility in the bulk when $N_a = 10^{16}$ cm^{-3} is 1300 cm^2 V^{-1} s^{-1} (Chapter 5). Thus

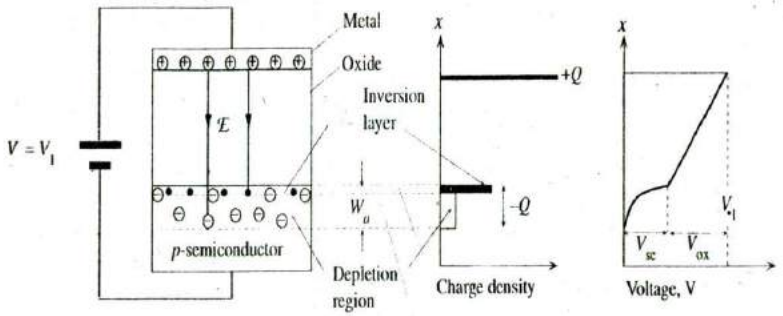
$$K = \frac{Z\mu_e \varepsilon_i \varepsilon_0}{2L t_{ox}} = \frac{(50 \times 10^{-6})\left(\frac{1}{2} \times 1300 \times 10^{-4}\right)(3.9 \times 8.85 \times 10^{-12})}{2(10 \times 10^{-6})(450 \times 10^{-10})} = 0.000125$$

When $V_{GS} = 8$ V and $V_{DS} = 20$ V, with $\lambda = 0.01$, we have

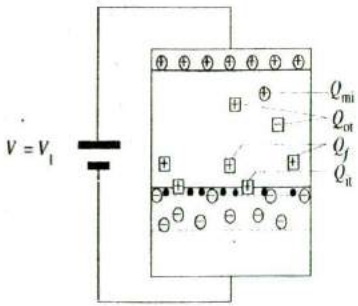
$$I_{DS} = 0.000125(8 - 4)^2[1 + (0.01)(20)] = 0.0024 \text{ A} \quad \text{or} \quad 2.4 \text{ mA}$$

6.8.3 THRESHOLD VOLTAGE

The threshold voltage is an important parameter in MOSFET devices. Its control in device fabrication is therefore essential. Figure 6.39a shows an idealized MOS structure where all the electric field lines from the metal pass through the oxide and penetrate the *p*-type semiconductor. The charge $-Q$ is made up of fixed negative acceptors in a surface region of W_a and of conduction electrons in the inversion layer at the surface, as shown in Figure 6.39a. The voltage drop across the MOS structure, however,



(a)



(b)

Figure 6.39

(a) The threshold voltage and the ideal MOS structure.

(b) In practice, there are several charges in the oxide and at the oxide–semiconductor interface that affect the threshold voltage: Q_{mi} = mobile ionic charge (e.g., Na^+), Q_{ot} = trapped oxide charge, Q_f = fixed oxide charge, and Q_{it} = charge trapped at the interface.

is not uniform. As the field penetrates the semiconductor, there is a voltage drop V_{sc} across the field penetration region of the semiconductor by virtue of $\mathcal{E} = -dV/dx$, as shown in Figure 6.39a. The field terminates on both electrons in the inversion layer and acceptors in W_d, so within the semiconductor $\mathcal{E}$ is not uniform and therefore the voltage drop is not constant. But the field in the oxide is uniform, as we assumed there were no charges inside the oxide. The voltage drop across the oxide is constant and is V_{ox}, as shown in Figure 6.39a. As the applied voltage is V_1, we must have $V_{sc} + V_{ox} = V_1$. The actual voltage drop V_{sc} across the semiconductor determines the condition for inversion. We can show this as follows. If the acceptor doping concentration is 10^{16} cm^{-3}, then the Fermi level E_F in the bulk of the p-type semiconductor must be 0.347 eV below E_{Fi} in intrinsic Si. To make the surface n-type we need to shift E_F at the surface to go just above E_{Fi}. Thus we need to shift E_F from bulk to surface by at least 0.347 eV. We have to bend the energy band by 0.347 eV at the surface. Since the voltage drop across the semiconductor is V_{sc} and the corresponding electrostatic PE change is eV_{sc}, this must be 0.347 eV or $V_{sc} = 0.347$ V. The gate voltage for the start of inversion will then be $V_{ox} + 0.347$ V. By inversion, however, we generally infer that the electron concentration at the surface is comparable to the hole concentration in the bulk. This means that we actually have to shift E_F above E_{Fi} by another 0.347 eV, so the gate threshold voltage V_{th} must be $V_{ox} + 0.694$ V.

In practice there are a number of other important effects that must be considered in evaluating the threshold voltage. Invariably there are charges both within the oxide and at the oxide–semiconductor interface that alter the field penetration into the semiconductor and hence the threshold voltage needed at the gate to cause inversion. Some of these are depicted in Figure 6.39b and can be qualitatively summarized as follows.

There may be some mobile ions within the SiO_2, such as alkaline ions (Na^+, K^+), which are denoted as Q_{mi} in Figure 6.39b. These may be introduced unintentionally, for example, during cleaning and etching processes in the fabrication. In addition there may be various trapped (immobile) charges within the oxide Q_{ot} due to structural defects, for example, an interstitial Si^+. Frequently these oxide trapped charges are created as a result of radiation damage (irradiation by X-rays or other high-energy beams). They can be reduced by annealing the device.

A significant number of fixed positive charges (Q_f) exist in the oxide region close to the interface. They are believed to originate from the nonstoichiometry of the oxide near the oxide–semiconductor interface. They are generally attributed to positively charged Si^+ ions. During the oxidation process, a Si atom is removed from the Si surface to react with the oxygen diffusing in through the oxide. When the oxidation process is stopped suddenly, there are unfulfilled Si ions in this region. Q_f depends on the crystal orientation and on the oxidation and annealing processes. The semiconductor to oxide interface itself is a sudden change in the structure from crystalline Si to amorphous oxide. The semiconductor surface itself will have various defects, as discussed in Chapter 1. There is some inevitable mismatch between the two structures at the interface, and consequently there are broken bonds, dangling bonds, point defects such as vacancies and Si^+, and other defects at this interface that trap charges (*e.g.*, holes). All these interface charges are represented as Q_{it} in Figure 6.39b. Q_{it} depends not only on the crystal orientation but also on the chemical composition of the interface. Both Q_f and Q_{it} overall represent a positive charge that effectively reduces the

gate voltage needed for inversion. They are smaller for the (100) surface than the (111) surface, so (100) is the preferred surface for the Si MOS device.

In addition to various charges in the oxide and at the interface shown in Figure 6.39b, there will also be a voltage difference, denoted as V_{FB}, between the semiconductor surface and the metal surface, even in the absence of an applied voltage. V_{FB} arises from the work function difference between the metal and the p-type semiconductor, as discussed in Chapter 4. The metal work function is generally smaller than the semiconductor work function, which means that the semiconductor surface will have an accumulation of electrons and the metal surface will have positive charges (exposed metal ions). The gate voltage needed for inversion will therefore also depend on V_{FB}. Since V_{FB} is normally positive and Q_f and Q_{it} are also positive, there may already be an inversion layer formed at the semiconductor surface even without a positive gate voltage. The fabrication of an enhancement MOSFET then requires special fabrication procedures, such as ion implantation, to obtain a positive and predictable V_{th}.

The simplest way to control the threshold gate voltage is to provide a separate electrode to the bulk of an enhancement MOSFET, as shown in Figure 6.36, and to apply a bias voltage to the bulk with respect to the source to obtain the desired V_{th} between the gate and source. This technique has the disadvantage of requiring an additional bias supply for the bulk and also adjusting the bulk to source voltage almost individually for each MOSFET.

6.8.4 ION IMPLANTED MOS TRANSISTORS AND POLY-SI GATES

The most accurate method of controlling the threshold voltage is by ion implantation, as the number of ions that are implanted into a device and their location can be closely controlled. Furthermore, ion implantation can also provide a self-alignment of the edges of the gate electrode with the source and drain regions. In the case of an n-channel enhancement MOSFET, it is generally desirable to keep the p-type doping in the bulk low to avoid small V_{DS} for reverse breakdown between the drain and the bulk (see Figure 6.36). Consequently, the surface, in practice, already has an inversion layer (without any gate voltage) due to various fixed positive charges residing in the oxide and at the interface, as shown in Figure 6.39b (positive Q_f and Q_{it} and V_{FB}). It then becomes necessary to implant the surface region under the gate with boron acceptors to remove the electrons and restore this region to a p-type behavior.

The ion implantation process is carried out in a vacuum where the required impurity ions are generated and then accelerated toward the device. The energy of the arriving ions and hence their penetration into the device can be readily controlled. Typically, the device is implanted with B acceptors under the gate oxide, as shown in Figure 6.40. The distribution of implanted acceptors as a function of distance into the device from the surface of the oxide is also shown in the figure. The position of the peak depends on the energy of the ions and hence on the accelerating voltage. The peak of the concentration of implanted acceptors is made to occur just below the surface of the semiconductor. Since ion implantation involves the impact of energetic ions with the crystal structure, it results in the inevitable generation of various defects within the implanted region. The defects are almost totally eliminated by annealing the device at an

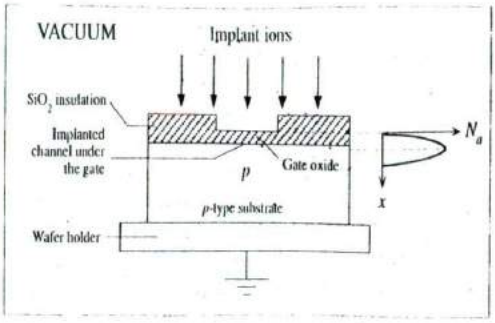


Figure 6.40 Schematic illustration of ion implantation for the control of V_{th}.

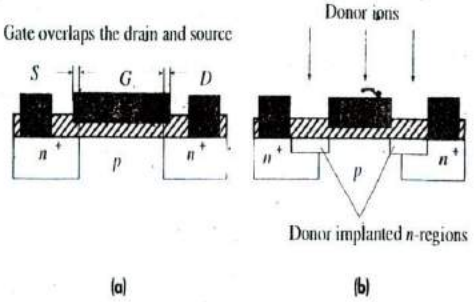


Figure 6.41

(a) There is an overlap of the gate electrode with the source and drain regions and hence additional capacitance between the gate and drain.

(b) n^+-type ion implantation extends the drain and source to line up with the gate.

elevated temperature. Annealing also broadens the acceptor implanted region as a result of increased diffusion of implanted acceptors.

Ion implantation also has the advantage of providing self-alignment of the drain and source with the edges of the gate electrode. In a MOS transistor, it is important that the gate electrode extends all the way from the source to the drain regions so that the channel formed under the gate can link the two regions; otherwise, an incomplete channel will be formed. To avoid the possibility of forming an incomplete channel, it is necessary to allow for some overlap, as shown in Figure 6.41a, between the gate and source and drain regions because of various tolerances and variations involved in the fabrication of a MOSFET by conventional masking and diffusional techniques. The overlap, however, results in additional capacitances between the gate and source and the gate and drain and adversely affects the high-frequency (or transient) response of the device. It is therefore desirable to align the edges of the gate electrode with the source and drain regions. Suppose that the gate electrode is made narrower so that it does not extend all the way between the source and drain regions, as shown in Figure 6.41b. If the device is now ion implanted with donors, then donor ions passing through the thin oxide will extend the n^+ regions up to the edges of the gate and thereby align the drain and source with the edges of the gate. The thick metal gate is practically impervious to the arriving donor ions.

Another method of controlling V_{th} is to use silicon instead of Al for the gate electrode. This technique is called **silicon gate technology.** Typically, the silicon for the

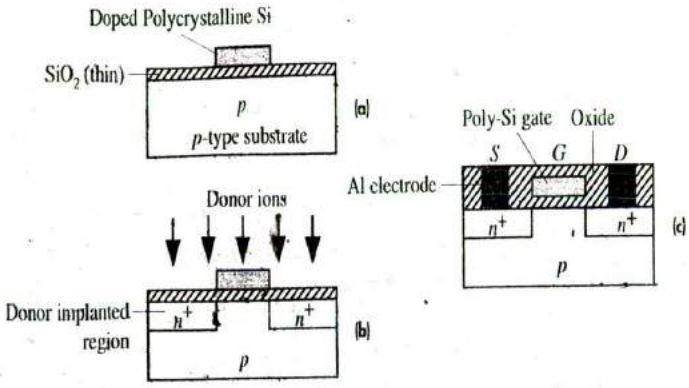


Figure 6.42 The poly-Si gate technology.
(a) Poly-Si is deposited onto the oxide, and the areas outside the gate dimensions are etched away.
(b) The poly-Si gate acts as a mask during ion implantation of donors to form the n^+ source and drain regions.
(c) A simplified schematic sketch of the final poly-Si MOS transistor.

gate is vacuum deposited (*e.g.*, by chemical vapor deposition using silane gas) onto the oxide, as shown in Figure 6.42. As the oxide is noncrystalline, the Si gate is polycrystalline (rather than a single crystal) and is therefore called a **poly-Si gate.** Normally it is heavily doped to ensure that it has sufficiently low resistivity to avoid *RC* time constant limitations in charging and discharging the gate capacitance during transient or ac operations. The advantage of the poly-Si gate is that its work function depends on the doping (type and concentration) and can be controlled so that V_{FB} and hence V_{th} can also be controlled. There are also additional advantages in using the poly-Si gate. For example, it can be raised to high temperatures (Al melts at 660 °C). It can be used as a mask over the gate region of the semiconductor during the formation of the source and drain regions. If ion implantation is used to deposit donors into the semiconductor, then the n^+ source and drain regions are self-aligned with the poly-Si gate, as shown in Figure 6.42.

6.9 LIGHT EMITTING DIODES (LED)

6.9.1 LED PRINCIPLES

A light emitting diode (LED) is essentially a *pn* junction diode typically made from a direct bandgap semiconductor, for example, GaAs, in which the electron–hole pair (EHP) recombination results in the emission of a photon. The emitted photon energy $h\nu$ is approximately equal to the bandgap energy E_g. Figure 6.43a shows the energy band diagram of an unbiased pn^+ junction device in which the *n*-side is more heavily doped than the *p*-side. The Fermi level E_F is uniform through the device, which is a requirement of equilibrium with no applied bias. The depletion region extends mainly into the *p*-side. There is a *PE* barrier eV_o from E_c on the *n*-side to E_c on the *p*-side

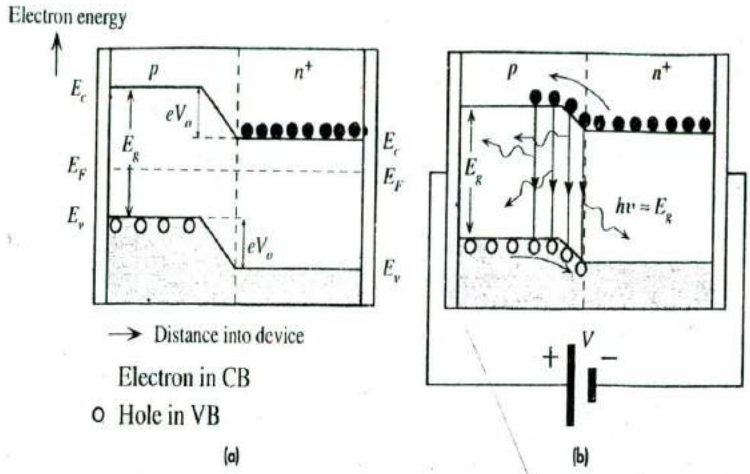


Figure 6.43 Energy band diagram of a *pn* (heavily *n*-type doped) junction.
(a) No bias voltage.
(b) With forward bias *V*. Recombination around the junction and within the diffusion length of the electrons in the *p*-side leads to photon emission.

where V_o is the *built-in voltage*. The *PE* barrier eV_o prevents the diffusion of electrons from the *n*-side to the *p*-side.

When a forward bias *V* is applied, the built-in potential V_o is reduced to $V_o - V$, which then allows the electrons from the n^+-side to diffuse, that is, become injected, into the *p*-side as depicted in Figure 6.43b. The hole injection component from *p* into the n^+-side is much smaller than the electron injection component from the n^+-side to the *p*-side. The recombination of injected electrons in the depletion region and within a volume extending over the electron diffusion length L_e in the *p*-side leads to photon emission. The phenomenon of light emission from the EHP recombination as a result of minority carrier injection is called **injection electroluminescence.** Due to the statistical nature of the recombination process between electrons and holes, the emitted photons are in random directions; they result from spontaneous emission processes. The LED structure has to be such that the emitted photons can escape the device without being reabsorbed by the semiconductor material. This means the *p*-side has to be sufficiently narrow or we have to use *heterostructure* devices as discussed below.

One very simple LED structure is shown in Figure 6.44. First a doped semiconductor layer is grown on a suitable substrate (GaAs or GaP). The growth is done **epitaxially**; that is, the crystal of the new layer is grown to follow the structure of the substrate crystal. The **substrate** is essentially a sufficiently thick crystal that serves as a mechanical support for the *pn* junction device (the doped layers) and can be of different crystal. The pn^+ junction is formed by growing another epitaxial layer but doped *p*-type. Those photons that are emitted toward the *n*-side become either absorbed or reflected back at the substrate interface depending on the substrate thickness and the exact structure of the LED. If the epitaxial layer and the substrate crystals have different

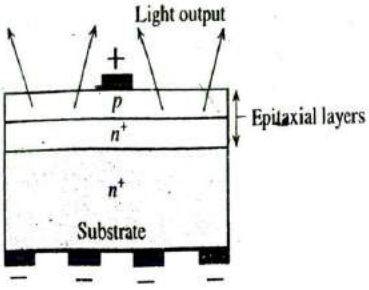


Figure 6.44 A schematic illustration of one possible LED device structure. First an n^+ layer is epitaxially grown on a substrate. A thin p layer is then epitaxially grown on the first layer.

crystal lattice parameters, then there is a lattice mismatch between the two crystal structures. This causes lattice strain in the LED layer and hence leads to crystal defects. Such crystal defects encourage radiationless EHP recombinations. That is, a defect acts as a recombination center. Such defects are reduced by lattice matching the LED epitaxial layer to the substrate crystal. It is therefore important to lattice match the LED layer to the substrate crystal. For example, one of the AlGaAs alloys is a direct bandgap semiconductor that has a bandgap in the red-emission region. It can be grown on GaAs substrates with excellent lattice match which results in high-efficiency LED devices.

There are various direct bandgap semiconductor materials that can be readily doped to make commercial pn junction LEDs which emit radiation in the red and infrared range of wavelengths. An important class of commercial semiconductor materials that covers the visible spectrum is the **III–V ternary alloys** based on alloying GaAs and GaP and denoted as $GaAs_{1-y}P_y$. In this compound, As and P atoms from Group V are distributed randomly at normal As sites in the GaAs crystal structure. When $y < 0.45$, the alloy $GaAs_{1-y}P_y$ is a direct bandgap semiconductor and hence the EHP recombination process is direct as depicted in Figure 6.45a. The rate of recombination is directly proportional to the product of electron and hole concentrations. The emitted wavelengths range from about 630 nm, red, for $y = 0.45$ ($GaAs_{0.55}P_{0.45}$) to 870 nm for $y = 0$ (GaAs).

$GaAs_{1-y}P_y$ alloys (which include GaP) with $y > 0.45$ are indirect bandgap semiconductors. The EHP recombination processes occur through recombination centers and involve lattice vibrations rather than photon emission. However, if we add

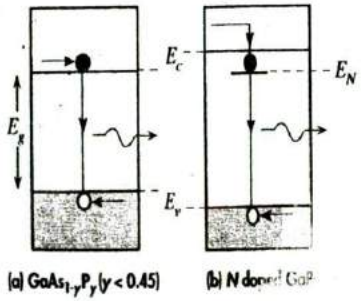


Figure 6.45
(a) Photon emission in a direct bandgap semiconductor.
(b) GaP is an indirect bandgap semiconductor. When it is doped with nitrogen, there is an electron recombination center at E_N. Direct recombination between a captured electron at E_N and a hole emits a photon.

Table 6.2 Selected LED semiconductor materials

Semiconductor Active Layer	Structure	D or I	λ (nm)	$\eta_{external}$ (%)	Comments
GaAs	DH	D	870–900	10	Infrared (IR)
$Al_xGa_{1-x}As$ ($0 < x < 0.4$)	DH	D	640–870	3–20	Red to IR
$In_{1-x}Ga_xAs_yP_{1-y}$ ($y \approx 2.20x$, $0 < x < 0.47$)	DH	D	1–1.6 μm	>10	LEDs in communications
$In_{0.49}Al_xGa_{0.51-x}P$	DH	D	590–630	>10	Amber, green, red; high luminous intensity
InGaN/GaN quantum well	QW	D	450–530	5–20	Blue to green
$GaAs_{1-y}P_y$ ($y < 0.45$)	HJ	D	630–870	< 1	Red to IR
$GaAs_{1-y}P_y$ ($y > 0.45$) (N or Zn, O doping)	HJ	I	560–700	< 1	Red, orange, yellow
SiC	HJ	I	460–470	0.02	Blue, low efficiency
GaP (Zn)	HJ	I	700	2–3	Red
GaP (N)	HJ	I	565	< 1	Green

NOTE: Optical communication channels are at 850 nm (local network) and at 1.3 and 1.55 μm (long distance). D = direct bandgap, I = indirect bandgap. $\eta_{external}$ is typical and may vary substantially depending on the device structure. DH = double heterostructure, HJ = homojunction, QW = quantum well.

isoelectronic impurities such as nitrogen (in the same Group V as P) into the semiconductor crystal, then some of these N atoms substitute for P atoms. Since N and P have the same valency, N atoms substituting for P atoms form the same number of bonds and do not act as donors or acceptors. The electronic cores of N and P, however, are different. The positive nucleus of N is less shielded by electrons compared with that of the P atom. This means that a conduction electron in the neighborhood of a N atom will be attracted and may become captured at this site. N atoms therefore introduce localized energy levels, or electron traps, E_N near the conduction band (CB) edge as depicted in Figure 6.45b. When a conduction electron is captured at E_N, it can attract a hole (in the valence band) in its vicinity by Coulombic attraction and eventually recombine with it directly and emit a photon. The emitted photon energy is only slightly less than E_g as E_N is typically close to E_c. As the recombination process depends on N doping, it is not as efficient as direct recombination. Thus, the efficiency of LEDs from N doped indirect bandgap $GaAs_{1-y}P_y$ semiconductors is less than those from direct bandgap semiconductors. Nitrogen doped indirect bandgap $GaAs_{1-y}P_y$ alloys are widely used in inexpensive green, yellow, and orange LEDs.

The **external efficiency** $\eta_{external}$ of an LED quantifies the efficiency of conversion of electric energy into an emitted external optical energy. It incorporates the internal efficiency of the radiative recombination process and the subsequent efficiency of photon extraction from the device. The input of electric power into an LED is simply the diode current and diode voltage product (IV). If P_{out} is the optical power emitted by the device, then

External efficiency

$$\eta_{external} = \frac{P_{out}(\text{optical})}{IV} \times 100\% \qquad [6.65]$$

and some typical values are listed in Table 6.2. For indirect bandgap semiconductors, $\eta_{external}$ are generally less than 1 percent, whereas for direct bandgap semiconductors with the right device structure, $\eta_{external}$ can be substantial.

6.9.2 HETEROJUNCTION HIGH-INTENSITY LEDS

A *pn* junction between two differently doped semiconductors that are of the same material, that is, the same bandgap E_g, is called a **homojunction.** A junction between two different bandgap semiconductors is called a **heterojunction.** A semiconductor device structure that has junctions between different bandgap materials is called a **heterostructure device.**

LED constructions for increasing the intensity of the output light make use of the double heterostructure. Figure 6.46a shows a **double-heterostructure** (DH) device based on two junctions between different semiconductor materials with different bandgaps. In this case the semiconductors are AlGaAs with $E_g \approx 2$ eV and GaAs with $E_g \approx 1.4$ eV. The double heterostructure in Figure 6.46a has an n^+p heterojunction between n^+-AlGaAs and p-GaAs. There is another heterojunction between p-GaAs and p-AlGaAs. The p-GaAs region is a thin layer, typically a fraction of a micron, and it is lightly doped.

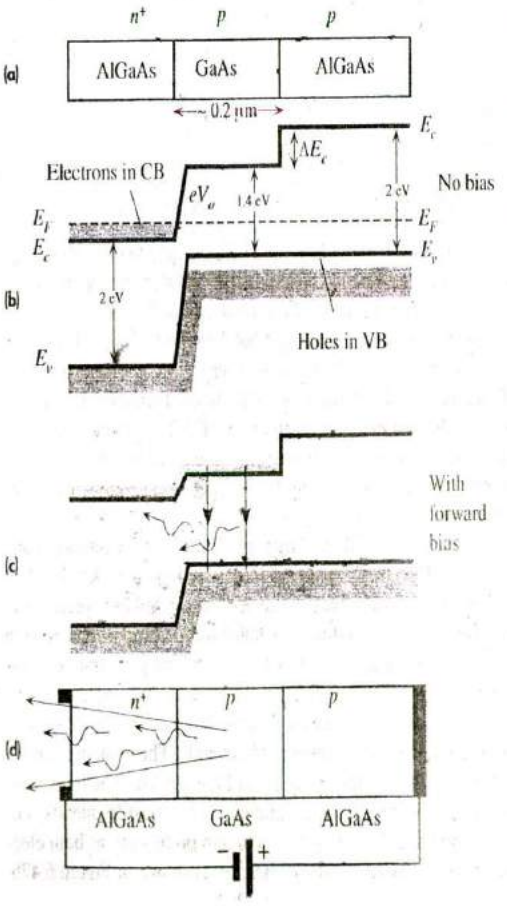


Figure 6.46

(a) A double heterostructure diode has two junctions which are between two different bandgap semiconductors (GaAs and AlGaAs).

(b) A simplified energy band diagram with exaggerated features. E_F must be uniform.

(c) Forward-biased simplified energy band diagram.

(d) Forward-biased LED. Schematic illustration of photons escaping reabsorption in the AlGaAs layer and being emitted from the device.

The simplified energy band diagram for the whole device in the absence of an applied voltage is shown in Figure 6.46b. The Fermi level E_F is continuous throughout the whole structure. There is a potential energy barrier eV_0 for electrons in the CB of n^+-AlGaAs against diffusion into p-GaAs. There is a bandgap change at the junction between p-GaAs and p-AlGaAs which results in a step change ΔE_c in E_c between the two conduction bands of p-GaAs and p-AlGaAs. This ΔE_c is effectively a *potential energy barrier* that prevents any electrons in the CB in p-GaAs passing to the CB of p-AlGaAs. (There is also a step change ΔE_v in E_v, but this is small and is not shown.)

When a forward bias is applied, most of this voltage drops between the n^+-AlGaAs and p-GaAs and reduces the potential energy barrier eV_0, just as in the normal pn junction. This allows electrons in the CB of n^+-AlGaAs to be injected into p-GaAs as shown in Figure 6.46c. These electrons, however, are *confined* to the CB of p-GaAs since there is a barrier ΔE_c between p-GaAs and p-AlGaAs. The wide bandgap AlGaAs layers therefore act as **confining layers** that restrict injected electrons to the p-GaAs layer. The recombination of injected electrons and the holes already present in this p-GaAs layer results in spontaneous photon emission. Since the bandgap E_g of AlGaAs is greater than GaAs, the emitted photons do not get reabsorbed as they escape the active region and can reach the surface of the device as depicted in Figure 6.46d. Since light is also not absorbed in p-AlGaAs, it can be reflected to increase the light output.

6.9.3 LED CHARACTERISTICS

The energy of an emitted photon from an LED is not simply equal to the bandgap energy E_g because electrons in the conduction band are distributed in energy and so are the holes in the valence band (VB). Figure 6.47a and b illustrate the energy band diagram and the energy distributions of electrons and holes in the CB and VB, respectively. The electron concentration as a function of energy in the CB is given by $g(E)f(E)$ where $g(E)$ is the density of states and $f(E)$ is the Fermi–Dirac function (probability of finding an electron in a state with energy E). The product $g(E)f(E)$ represents the electron concentration per unit energy or the concentration in energy and is plotted along the horizontal axis in Figure 6.47b. There is a similar energy distribution for holes in the VB.

The electron concentration in the CB as a function of energy is asymmetrical and has a peak at $\frac{1}{2}kT$ above E_c. The energy spread of these electrons is typically $\sim 2kT$ from E_c as shown in Figure 6.47b. The hole concentration is similarly spread from E_v in the valence band. Recall the rate of direct recombination is proportional to both the electron and hole concentrations at the energies involved. The transition which is identified as 1 in Figure 6.47a involves the direct recombination of an electron at E_c and a hole at E_v. But the carrier concentrations near the band edges are very small and hence this type of recombination does not occur frequently. The relative intensity of light at this photon energy $h\nu_1$ is small as shown in Figure 6.47c. The transitions that involve the largest electron and hole concentrations occur most frequently. For example, the transition 2 in Figure 6.47a has the maximum probability as both electron and hole concentrations are largest at these energies as shown in Figure 6.47b.

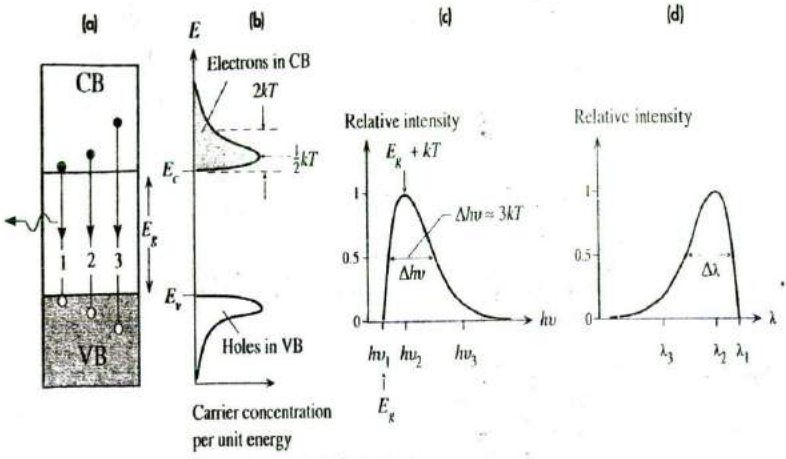


Figure 6.47

(a) Energy band diagram with possible recombination paths.

(b) Energy distribution of electrons in the CB and holes in the VB. The highest electron concentration is $\frac{1}{2}kT$ above E_c.

(c) The relative light intensity as a function of photon energy based on (b).

(d) Relative intensity as a function of wavelength in the output spectrum based on (b) and (c).

The relative intensity of light corresponding to this transition energy $h\nu_2$ is then maximum, or close to maximum, as indicated in Figure 6.47c.[11] The transitions marked as 3 in Figure 6.47a that emit relatively high energy photons $h\nu_3$ involve energetic electrons and holes whose concentrations are small as apparent in Figure 6.47b. Thus, the light intensity at these relatively high photon energies is small. The fall in light intensity with photon energy is shown in Figure 6.47c. The relative light intensity versus photon energy characteristic of the output spectrum is shown in Figure 6.47c and represents an important LED characteristic. Given the spectrum in Figure 6.47c we can also obtain the relative light intensity versus wavelength characteristic as shown in Figure 6.47d since $\lambda = c/\nu$. The **linewidth** of the output spectrum, $\Delta\nu$ or $\Delta\lambda$, is defined as the width between half-intensity points as shown in Figure 6.47c and d.

The wavelength for the peak intensity and the linewidth $\Delta\lambda$ of the emitted spectrum are obviously related to the energy distributions of the electrons and holes in the conduction and valence bands and therefore to the density of states in these bands. The photon energy for the peak emission is roughly $E_g + kT$ inasmuch as it corresponds to peak-to-peak transitions in the energy distributions of the electrons and holes in Figure 6.47b. The linewidth $\Delta(h\nu)$ of the output radiation between the half intensity points is approximately $3kT$ as shown in Figure 6.47c. It is relatively straightforward to calculate the corresponding spectral linewidth $\Delta\lambda$ in terms of wavelength as explained in Example 6.14.

[11] The intensity is not necessarily maximum when both the electron and hole concentrations are maximum, but it will be close.

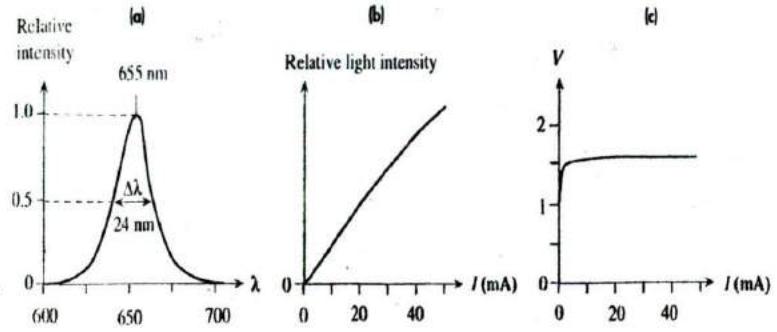


Figure 6.48

(a) A typical output spectrum from a red GaAsP LED.

(b) Typical output light power versus forward current.

(c) Typical *I–V* characteristics of a red LED. The turn-on voltage is around 1.5 V.

The output spectrum, or the relative intensity versus wavelength characteristics, from an LED depends not only on the semiconductor material but also on the structure of the *pn* junction diode, including the dopant concentration levels. The spectrum in Figure 6.47d represents an idealized spectrum without including the effects of heavy doping on the energy bands and the reabsorption of some of the photons.

Typical characteristics of a red LED (655 nm), as an example, are shown in Figure 6.48a to c. The output spectrum in Figure 6.48a exhibits less asymmetry than the idealized spectrum in Figure 6.47d. The width of the spectrum is about 24 nm, which corresponds to a width of about $2.7kT$ in the energy distribution of the emitted photons. As the LED current increases so does the injected minority carrier concentration, and thus the rate of recombination and hence the output light intensity. The increase in the output light power is not however linear with the LED current as apparent in Figure 6.48b. At high current levels, a strong injection of minority carriers leads to the recombination time depending on the injected carrier concentration and hence on the current itself; this leads to a nonlinear recombination rate with current. Typical current–voltage characteristics are shown in Figure 6.48c where it can be seen that the **turn-on**, or **cut-in, voltage** is about 1.5 V from which point the current increases very steeply with voltage. The turn-on voltage depends on the semiconductor and generally increases with the energy bandgap E_g. For example, typically, for a blue LED it is about 3.5–4.5 V, for a yellow LED it is about 2 V, and for a GaAs infrared LED it is around 1 V.

EXAMPLE 6.14 **SPECTRAL LINEWIDTH OF LEDS** We know that a spread in the output wavelengths is related to a spread in the emitted photon energies as depicted in Figure 6.47. The emitted photon energy $E_{ph} = hc/\lambda$ and the spread in the photon energies $\Delta E_{ph} = \Delta(hc/\lambda) \approx 3kT$ between the half-intensity points as shown in Figure 6.47c. Show that the corresponding **linewidth** $\Delta\lambda$ between the ***half-intensity points*** in the output spectrum is

LED spectral linewidth

$$\Delta\lambda = \lambda^2 \frac{3kT}{hc} \qquad [6.66]$$

What is the spectral linewidth of an optical communications LED operating at 1550 nm and at 300 K?

SOLUTION

First consider the relationship between the photon frequency ν and λ,

$$\lambda = \frac{c}{\nu} = \frac{hc}{h\nu}$$

in which $h\nu$ is the photon energy. We can differentiate this,

$$\frac{d\lambda}{d(h\nu)} = -\frac{hc}{(h\nu)^2} = -\frac{\lambda^2}{hc}$$

The negative sign implies that increasing the photon energy decreases the wavelength. We are only interested in changes or spreads; thus $\Delta\lambda/\Delta(h\nu) \approx |d\lambda/d(h\nu)|$,

$$\Delta\lambda = \frac{\lambda^2}{hc}\Delta(h\nu) = \frac{\lambda^2}{hc}3kT$$

where we used $\Delta(h\nu) = 3kT$, and obtained Equation 6.66. We can substitute $\lambda = 1550$ nm and $T = 300$ K to calculate the linewidth of the 1550 nm LED:

$$\Delta\lambda = \lambda^2\frac{3kT}{hc} = (1550 \times 10^{-9})^2\frac{3(1.38 \times 10^{-23})(300)}{(6.626 \times 10^{-34})(3 \times 10^8)}$$

$$= 1.50 \times 10^{-7}\text{ m} \quad \text{or} \quad 150\text{ nm}$$

The spectral linewidth of an LED output is due to the spread in the photon energies, which is fundamentally about $3kT$. The only option for decreasing $\Delta\lambda$ at a given wavelength is to reduce the temperature. The output spectrum of a laser, on the other hand, has a much narrower linewidth. A single-mode laser can have an output linewidth less than 1 nm.

6.10 SOLAR CELLS

6.10.1 PHOTOVOLTAIC DEVICE PRINCIPLES

A simplified schematic diagram of a typical solar cell is shown in Figure 6.49. Consider a *pn* junction with a very narrow and more heavily doped *n*-region. The illumination is through the thin *n*-side. The depletion region (W) or the space charge layer (SCL) extends primarily into the *p*-side. There is a built-in field $\mathcal{E}_o$ in this depletion layer. The electrodes attached to the *n*-side must allow illumination to enter the device and at the same time result in a small series resistance. They are deposited on the *n*-side to form an array of **finger electrodes** on the surface as depicted in Figure 6.50. A thin **antireflection coating** on the surface (not shown in the figure) reduces reflections and allows more light to enter the device.

As the *n*-side is very narrow, most of the photons are absorbed within the depletion region (W) and within the neutral *p*-side (ℓ_p) and photogenerate EHPs in these regions. EHPs photogenerated in the depletion region are immediately separated by the built-in field $\mathcal{E}_o$ which drifts them apart. The electron drifts and reaches the neutral n^+-side whereupon it makes this region negative by an amount of charge $-e$. Similarly,

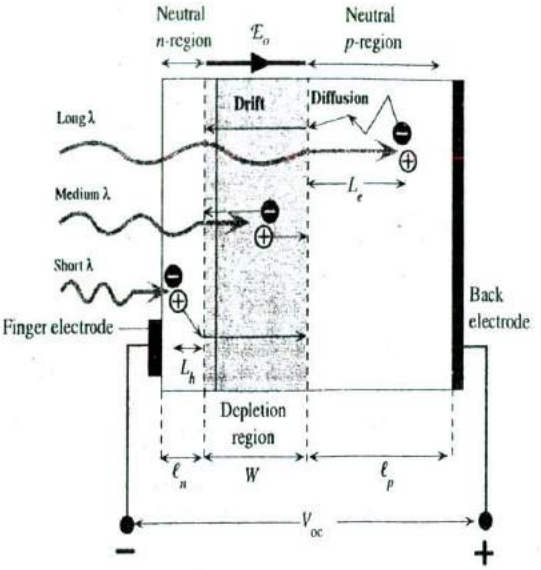


Figure 6.49 The principle of operation of the solar cell (exaggerated features to highlight principles).

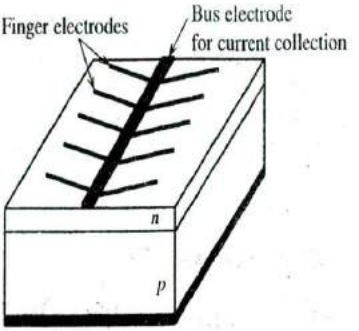


Figure 6.50 Finger electrodes on the surface of a solar cell reduce the series resistance.

the hole drifts and reaches the neutral p-side and thereby makes this side positive. Consequently an **open circuit voltage** develops between the terminals of the device with the p-side positive with respect to the n-side. If an external load is connected, then the excess electron in the n-side can travel around the external circuit, do work, and reach the p-side to recombine with the excess hole there. It is important to realize that without the internal field $\mathcal{E}_o$ it is not possible to drift apart the photogenerated EHPs and accumulate excess electrons on the n-side and excess holes on the p-side.

The EHPs photogenerated by long-wavelength photons that are absorbed in the neutral p-side diffuse around in this region as there is no electric field. If the recombination lifetime of the electron is τ_e, it diffuses a mean distance $L_e = \sqrt{2D_e\tau_e}$ where D_e is its diffusion coefficient in the p-side. Those electrons within a distance L_e to the depletion region can readily diffuse and reach this region whereupon they become drifted

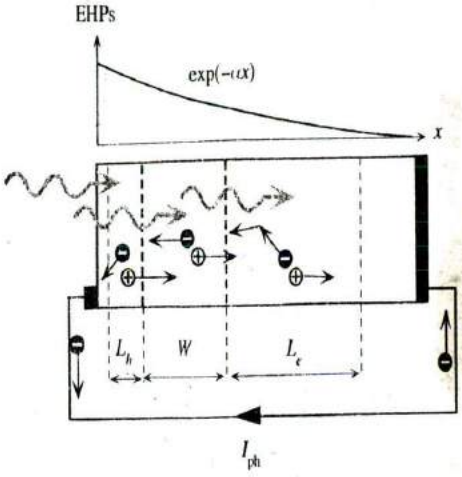


Figure 6.51 Photogenerated carriers within the volume $L_h + W + L_e$ give rise to a photocurrent I_{ph}.
The variation in the photogenerated EHP concentration with distance is also shown where α is the absorption coefficient at the wavelength of interest.

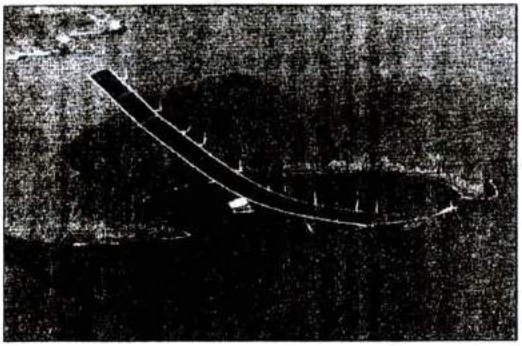
by $\mathcal{E}_o$ to the n-side as shown in Figure 6.49. Consequently only those EHPs photogenerated within the minority carrier diffusion length L_e to the depletion layer can contribute to the photovoltaic effect. Again the importance of the built-in field $\mathcal{E}_o$ is apparent. Once an electron diffuses to the depletion region, it is swept over to the n-side by $\mathcal{E}_o$ to give an additional negative charge there. Holes left behind in the p-side contribute a net positive charge to this region. Those photogenerated EHPs further away from the depletion region than L_e are lost by recombination. It is therefore important to have the minority carrier diffusion length L_e be as long as possible. This is the reason for choosing this side of a Si pn junction to be p-type which makes electrons the minority carriers; the electron diffusion length in Si is longer than the hole diffusion length. The same ideas also apply to EHPs photogenerated by short-wavelength photons absorbed in the n-side. Those holes photogenerated within a diffusion length L_h can reach the depletion layer and become swept across to the p-side. The photogeneration of EHPs that contributes to the photovoltaic effect therefore occurs in a volume covering $L_h + W + L_e$. If the terminals of the device are shorted, as in Figure 6.51, then the excess electron in the n-side can flow through the external circuit to neutralize the excess hole in the p-side. This current due to the flow of the photogenerated carriers is called the **photocurrent.**

Under a steady-state operation, there can be no net current through an *open circuit* solar cell. This means the photocurrent inside the device due to the flow of photogenerated carriers must be exactly balanced by a flow of carriers in the opposite direction. The latter carriers are minority carriers that become injected by the appearance of the photovoltaic voltage across the pn junction as in a normal diode. This is not shown in Figure 6.49.

EHPs photogenerated by energetic photons absorbed in the n-side near the surface region or outside the diffusion length L_h to the depletion layer are lost by recombination as the lifetime in the n-side is generally very short (due to heavy doping). The n-side is therefore made very thin, typically less than 0.2 μm. Indeed, the length ℓ_n of

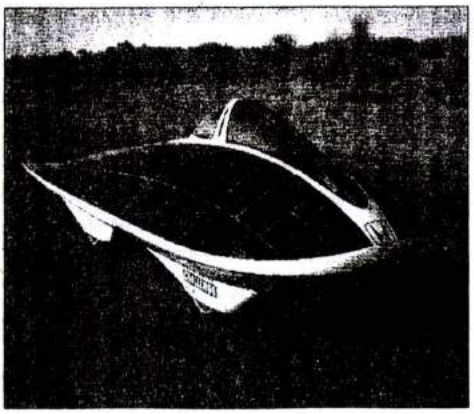
Solar cell inventors at Bell Labs (left to right): Gerald Pearson, Daryl Chapin, and Calvin Fuller. They are checking a Si solar cell sample for the amount of voltage produced (1954).

SOURCE: Courtesy of Bell Labs, Lucent Technologies.



Helios is a solar cell–powered airplane that is remotely piloted. It has been able to fly as high as about 30 km during the day. Its wingspan is 9 m. It has fuel cells to fly at night.

SOURCE: Courtesy of NASA, Dryden Flight Center.



pn Junction Si solar cells at work. Honda's two-seated Dream car is powered by photovoltaics. The Honda Dream was first to finish 3,010 km in four days in the 1996 World Solar Challenge.

SOURCE: Courtesy of Centre for Photovoltaic Engineering, University of New South Wales, Sydney, Australia.

the n-side may be shorter than the hole diffusion length L_h. The EHPs photogenerated very near the surface of the n-side, however, disappear by recombination due to various surface defects acting as recombination centers as discussed below.

At long wavelengths, around 1–1.2 μm, the absorption coefficient α of Si is small and the *absorption depth* $(1/\alpha)$ is typically greater than 100 μm. To capture these long-wavelength photons, we therefore need a thick p-side and at the same time a long minority carrier diffusion length L_e. Typically the p-side is 200–500 μm and L_e tends to be shorter than this.

Crystalline silicon has a bandgap of 1.1 eV which corresponds to a threshold wavelength of 1.1 μm. The incident energy in the wavelength region greater than 1.1 μm is then wasted; this is not a negligible amount (~25 percent). The worst part of the efficiency limitation however comes from the high-energy photons becoming absorbed near the crystal surface and being lost by recombination in the surface region. Crystal surfaces and interfaces contain a high concentration of recombination centers which facilitate the recombination of photogenerated EHPs near the surface. Losses due to EHP recombinations near or at the surface can be as high as 40 percent. These combined effects bring the efficiency down to about 45 percent. In addition, the antireflection coating is not perfect, which reduces the total collected photons by a factor of about 0.8–0.9. When we also include the limitations of the photovoltaic action itself (discussed below), the upper limit to a photovoltaic device that uses a single crystal of Si is about 24–26 percent at room temperature.

Consider an ideal *pn* junction photovoltaic device connected to a resistive load R as shown in Figure 6.52a. Note that I and V in the figure define the convention for the direction of positive current and positive voltage. If the load is a short circuit, then the only current in the circuit is that generated by the incident light. This is the photocurrent I_{ph} shown in Figure 6.52b which depends on the number of EHPs photogenerated within the volume enclosing the depletion region (W) and the diffusion lengths to the depletion region (Figure 6.51). The greater is the light intensity, the

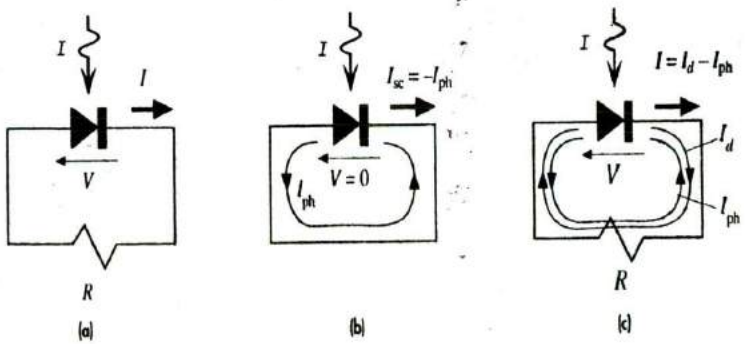


Figure 6.52

(a) The solar cell connected to an external load R and the convention for the definitions of positive voltage and positive current.

(b) The solar cell in short circuit. The current is the photocurrent I_{ph}.

(c) The solar cell driving an external load R. There is a voltage V and current I in the circuit.

higher is the photogeneration rate and the larger is I_{ph}. If I is the light intensity, then the **short circuit current** is

Short circuit solar cell current in light

$$I_{sc} = -I_{ph} = -KI \qquad [6.67]$$

where K is a constant that depends on the particular device. The photocurrent does not depend on the voltage across the pn junction because there is always some internal field to drift the photogenerated EHP. We exclude the secondary effect of the voltage modulating the width of the depletion region. The photocurrent I_{ph} therefore flows even when there is not a voltage across the device.

If R is not a short circuit, then a positive voltage V appears across the pn junction as a result of the current passing through it as shown in Figure 6.52c. This voltage reduces the built-in potential of the pn junction and hence leads to minority carrier injection and diffusion just as it would in a normal diode. Thus, in addition to I_{ph} there is also a forward diode current I_d in the circuit as shown in Figure 6.52c which arises from the voltage developed across R. Since I_d is due to the normal pn junction behavior, it is given by the diode characteristics,

$$I_d = I_o\left[\exp\left(\frac{eV}{\eta kT}\right) - 1\right]$$

where I_o is the "reverse saturation current" and η is the ideality factor ($\eta = 1 - 2$). In an open circuit, the net current is zero. This means that the photocurrent I_{ph} develops just enough photovoltaic voltage V_{oc} to generate a diode current $I_d = I_{ph}$.

Thus the **total current** through the solar cell, as shown in Figure 6.52c, is

Solar cell I–V

$$I = -I_{ph} + I_o\left[\exp\left(\frac{eV}{\eta kT}\right) - 1\right] \qquad [6.68]$$

The overall I–V characteristics of a typical Si solar cell are shown in Figure 6.53. It can be seen that it corresponds to the normal dark characteristics being shifted down

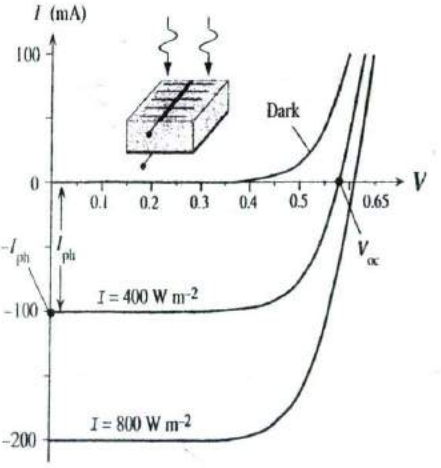


Figure 6.53 Typical I–V characteristics of a Si solar cell.
The short circuit current is I_{ph} and the open circuit voltage is V_{oc}. The I–V curves for positive current require an external bias voltage. Photovoltaic operation is always in the negative current region.

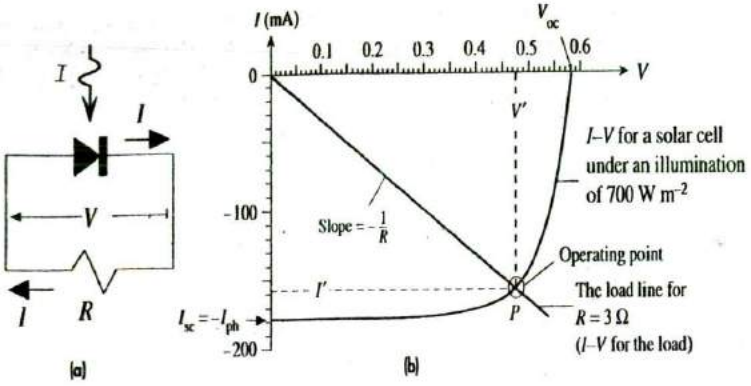


Figure 6.54

(a) When a solar cell drives a load R, R has the same voltage as the solar cell but the current through it is in the opposite direction to the convention that current flows from high to low potential.

(b) The current I' and voltage V' in the circuit of (a) can be found from a load line construction. Point P is the operating point (I', V'). The load line is for $R = 3\ \Omega$.

by the photocurrent I_{ph}, which depends on the light intensity I. The open circuit output voltage V_{oc}, of the solar cell is given by the point where the I–V curve cuts the V axis ($I = 0$). It is apparent that although it depends on the light intensity, its value typically lies in the range 0.5–0.7 V.

Equation 6.68 gives the I–V characteristics of the solar cell. When the solar cell is connected to a load as in Figure 6.54a, the load has the same voltage as the solar cell and carries the same current. But the current I through R is now in the opposite direction to the convention that current flows from high to low potential. Thus, as shown in Figure 6.54a,

$$I = -\frac{V}{R} \qquad [6.69]$$ *The load line*

The actual current I' and voltage V' in the circuit must satisfy both the I–V characteristics of the solar cell, Equation 6.68, and that of the load, Equation 6.69. We can find I' and V' by solving these two equations simultaneously or using a graphical solution. I' and V' in the solar cell circuit are most easily found by using a **load line construction.** The I–V characteristics of the load in Equation 6.69 is a straight line with a negative slope $-1/R$. This is called the **load line** and is shown in Figure 6.54b along with the I–V characteristics of the solar cell under a given intensity of illumination. The load line cuts the solar cell characteristic at P where the load and the solar cell have the same current and voltage I' and V'. Point P therefore satisfies both Equations 6.68 and 6.69 and thus represents the **operating point of the circuit.**

The **power delivered** to the load is $P_{out} = I'V'$, which is the area of the rectangle bound by the I and V axes and the dashed lines shown in Figure 6.54b. Maximum power is delivered to the load when this rectangular area is maximized (by changing R or the intensity of illumination), when $I' = I_m$ and $V' = V_m$. Since the maximum

possible current is I_{sc} and the maximum possible voltage is V_{oc}. $I_{sc}V_{oc}$ represents the desirable goal in power delivery for a given solar cell. Therefore it makes sense to compare the maximum power output $I_m V_m$ with $I_{sc}V_{oc}$. The **fill factor** FF, which is a figure of merit for the solar cell, is defined as

Definition of fill factor

$$\text{FF} = \frac{I_m V_m}{I_{sc} V_{oc}} \qquad [6.70]$$

The FF is a measure of the closeness of the solar cell I–V curve to the rectangular shape (the ideal shape). It is clearly advantageous to have the FF as close to unity as possible, but the exponential *pn* junction properties prevent this. Typically FF values are in the range 70–85 percent and depend on the device material and structure.

EXAMPLE 6.15 **A SOLAR CELL DRIVING A RESISTIVE LOAD** Consider the solar cell in Figure 6.54 that is driving a load of 3 Ω. This cell has an area of 3 cm × 3 cm and is illuminated with light of intensity 700 W m^{-2}. Find the current and voltage in the circuit. Find the power delivered to the load, the efficiency of the solar cell in this circuit, and the fill factor of the solar cell.

SOLUTION

The I–V characteristic of the load in Figure 6.54a, is the load line in Equation 6.69; that is, $I = -V/(3\,\Omega)$. The line is drawn in Figure 6.54b with a slope $1/(3\,\Omega)$. It cuts the I–V characteristics of the solar cell at $I' = 157$ mA and $V' = 0.475$ V as apparent in Figure 6.54b, which are the current and voltage, respectively, in the photovoltaic circuit of Figure 6.54a. The power delivered to the load is

$$P_{out} = I'V' = (157 \times 10^{-3})(0.475\ \text{V}) = 0.0746\ \text{W} \quad \text{or} \quad 74.6\ \text{mW}$$

The input of sunlight power is

$$P_{in} = (\text{Light intensity})(\text{Surface area}) = (700\ \text{W m}^{-2})(0.03\ \text{m})^2 = 0.63\ \text{W}$$

The efficiency is

$$\eta_{\text{photovoltaic}} = (100\%)\frac{P_{out}}{P_{in}} = (100\%)\frac{(0.0746\ \text{W})}{(0.63\ \text{W})} = 11.8\%$$

This will increase if the load is adjusted to extract the maximum power from the solar cell, but the increase will be small as the rectangular area $I'V'$ in Figure 6.54b is already quite close to the maximum.

The fill factor can also be calculated since point P in Figure 6.54b is close to the optimum operation, maximum output power, in which the rectangular area $I'V'$ is maximum:

$$\text{FF} = \frac{I_m V_m}{I_{sc} V_{oc}} \approx \frac{I'V'}{I_{sc} V_{oc}} = \frac{(157\ \text{mA})(0.475\ \text{V})}{(178\ \text{mA})(0.58\ \text{V})} = 0.722 \quad \text{or} \quad 72\%$$

EXAMPLE 6.16 **OPEN CIRCUIT VOLTAGE AND ILLUMINATION** A solar cell under an illumination of 500 W m^{-2} has a short circuit current I_{sc} of 150 mA and an open circuit output voltage V_{oc} of 0.530 V. What are the short circuit current and open circuit voltage when the light intensity is doubled? Assume $\eta = 1.5$, a typical value for various Si *pn* junctions.

SOLUTION

The general I–V characteristic under illumination is given by Equation 6.68. Setting $I = 0$ for open circuit,

Open circuit condition

$$I = -I_{ph} + I_o\left[\exp\left(\frac{eV_{oc}}{\eta kT}\right) - 1\right] = 0$$

Assuming that $V_{oc} \gg \eta kT/e$, rearranging the above equation we can find V_{oc},

Open circuit output voltage

$$V_{oc} = \frac{\eta kT}{e}\ln\left(\frac{I_{ph}}{I_o}\right)$$

The photocurrent I_{ph} depends on the light intensity I via $I_{ph} = KI$, where K is a constant. Thus, at a given temperature, the change in V_{oc} is

Open circuit voltage and light intensity

$$V_{oc2} - V_{oc1} = \frac{\eta kT}{e}\ln\left(\frac{I_{ph2}}{I_{ph1}}\right) = \frac{\eta kT}{e}\ln\left(\frac{I_2}{I_1}\right)$$

The short circuit current is the photocurrent, so at double the intensity this is

$$I_{sc2} = I_{sc1}\left(\frac{I_2}{I_1}\right) = (150\text{ mA})(2) = 300\text{ mA}$$

Assuming $\eta = 1.5$, the new open circuit voltage is

$$V_{oc2} = V_{oc1} + \frac{\eta kT}{e}\ln\left(\frac{I_2}{I_1}\right) = 0.530\text{ V} + (1.5)(0.026)\ln(2) = 0.557\text{ V}$$

This is a 5 percent increase compared with the 100 percent increase in illumination and the short circuit current.

6.10.2 SERIES AND SHUNT RESISTANCE

Practical solar cells can deviate substantially from the ideal *pn* junction solar cell behavior depicted in Figure 6.53 due to a number of reasons. Consider an illuminated *pn* junction driving a load resistance R_L and assume that photogeneration takes place in the depletion region. As shown in Figure 6.55, the photogenerated electrons have to traverse a surface semiconductor region to reach the nearest finger electrode. All these electron paths in the n-layer surface region to finger electrodes introduce an **effective series resistance** R_s into the photovoltaic circuit. If the finger electrodes are thin, then the resistance of the electrodes themselves will further increase R_s. There is also a series resistance due to the neutral p-region, but this is generally small compared with the resistance of the electron paths to the finger electrodes.

Figure 6.56a shows the equivalent circuit of an ideal *pn* junction solar cell. The photogeneration process is represented by a *constant current generator* I_{ph}, which generates a current that is proportional to the light intensity. The flow of photogenerated carriers across the junction gives rise to a photovoltaic voltage difference V across the junction, and this voltage leads to the normal diode current $I_d = I_o[\exp(eV/\eta kT) - 1]$. This diode current I_d is represented by an ideal *pn* junction diode in the circuit as shown in Figure 6.56a. As apparent, I_{ph} and I_d are in opposite directions (I_{ph} is "up"

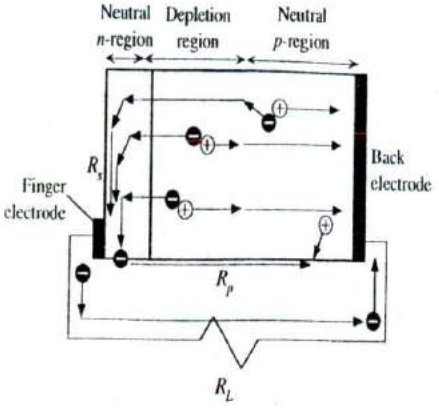


Figure 6.55 Series and shunt resistances and various fates of photogenerated EHPs.

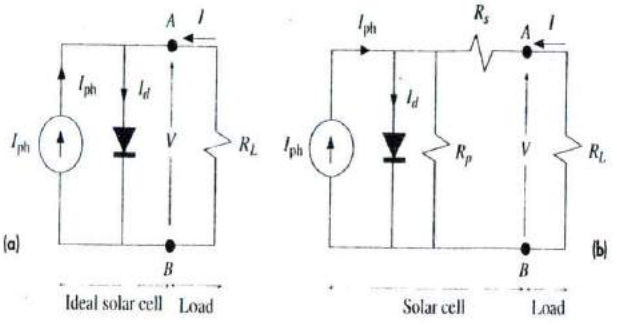


Figure 6.56 The equivalent circuit of a solar cell.
(a) Ideal *pn* junction solar cell.
(b) Parallel and series resistances R_s and R_p.

and I_d is "down"), so in an open circuit the photovoltaic voltage is such that I_{ph} and I_d have the same magnitude and cancel each other. By convention, positive current I at the output terminal is normally taken to flow into the terminal and is given by Equation 6.68. (In reality, of course, the solar cell current is negative, as in Figure 6.53, which represents a current that is flowing out into the load.)

Figure 6.56b shows the equivalent circuit of a more practical solar cell. The **series resistance** R_s in Figure 6.56b gives rise to a voltage drop and therefore prevents the ideal photovoltaic voltage from developing at the output between A and B when a current is drawn. A fraction (usually small) of the photogenerated carriers can also flow through the crystal surfaces (edges of the device) or through *grain boundaries in polycrystalline devices* instead of flowing though the external load R_L. These effects that prevent photogenerated carriers from flowing in the external circuit can be represented by an effective internal **shunt or parallel resistance** R_p that diverts the photocurrent away from the load R_L. Typically R_p is less important than R_s in overall

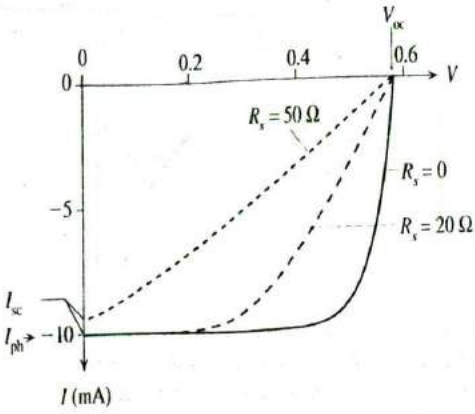


Figure 6.57 The series resistance broadens the *I-V* curve and reduces the maximum available power and hence the overall efficiency of the solar cell.
The example is a Si solar cell with $\eta \approx 1.5$ and $I_0 \approx 3 \times 10^{-6}$ mA. Illumination is such that the photocurrent $I_{ph} = 10$ mA.

device behavior, unless the device is highly polycrystalline and the current component flowing through grain boundaries is not negligible.

The series resistance R_s can significantly deteriorate the solar cell performance as illustrated in Figure 6.57 where $R_s = 0$ is the best solar cell case. It is apparent that the available maximum output power decreases with the series resistance which therefore reduces the cell efficiency. Notice also that when R_s is sufficiently large, it limits the short circuit current. Similarly, low shunt resistance values, due to extensive defects in the material, also reduce the efficiency. The difference is that although R_s does not affect the open circuit voltage V_{oc}, low R_p leads to a reduced V_{oc}.

6.10.3 SOLAR CELL MATERIALS, DEVICES, AND EFFICIENCIES

Most solar cells use crystalline silicon because silicon-based semiconductor fabrication is now a mature technology that enables cost-effective devices to be manufactured. Typical Si-based solar cell efficiencies range from about 18 percent for polycrystalline to 22–24 percent in high-efficiency single-crystal devices that have special structures to absorb as many of the incident photons as possible. Solar cells fabricated by making a *pn* junction in the same crystal are called *homojunctions*. The best Si homojunction solar cell efficiencies are about 24 percent for expensive single-crystal passivated emitter rear locally diffused (PERL) cells.[12] The PERL and similar cells have a textured surface that is an array of "inverted pyramids" etched into the surface to capture as much of the incoming light as possible as depicted in Figure 6.58. Normal reflections from a flat crystal surface lead to a loss of light, whereas reflections inside the pyramid allow a second or even a third chance for absorption. Further, after refraction, photons would be entering the semiconductor at oblique angles which means that they will be absorbed in the useful photogeneration volume, that is, within the electron diffusion length of the depletion layer as shown in Figure 6.58.

[12] Much of the pioneering work for high-efficiency PERL solar cells was done by Martin Green and coworkers at the University of New South Wales.

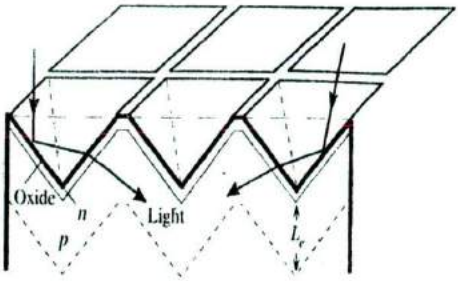


Figure 6.58 An inverted pyramid textured surface substantially reduces reflection losses and increases absorption probability in the device.

Table 6.3 summarizes some typical characteristics of various solar cells. GaAs and Si solar cells have comparable efficiencies though theoretically GaAs with a higher bandgap is supposed to have a better efficiency. The largest factors reducing the efficiency of a Si solar cell are the unabsorbed photons with $h\nu < E_g$ and short wavelength photons absorbed near the surface. Both these factors are improved if tandem cell structures or heterojunctions are used.

There are a number of III–V semiconductor alloys that can be prepared with different bandgaps but with the same lattice constant. Heterojunctions (junctions between different materials) from these semiconductors have negligible interface defects. AlGaAs has a wider bandgap than GaAs and would allow most solar photons to pass through. If we use a thin AlGaAs layer on a GaAs pn junction, as shown in Figure 6.59, then this layer passivates the surface defects normally present in a homojunction GaAs cell. The AlGaAs window layer therefore overcomes the surface recombination limitation and improves the cell efficiency (such cells have efficiencies of about 24 percent).

Table 6.3 Typical characteristics of various solar cells at room temperature under AM1.5 illumination of 1000 W m^{-2}

Semiconductor	E_g (eV)	V_{oc} (V)	J_{sc} (mA cm^{-2})	FF	η (%)	Comments
Si, single crystal	1.1	0.5–0.7	42	0.7–0.8	16–24	Single crystal, PERL
Si, polycrystalline	1.1	0.5–0.65	38	0.7–0.8	12–19	
Amorphous Si:Ge:H film					8–13	Amorphous film with tandem structure, convenient large-area fabrication
GaAs, single crystal	1.42	1.02	28	0.85	24–25	
GaAlAs/GaAs, tandem		1.03	27.9	0.864	24.8	Different bandgap materials in tandem increases absorption efficiency
GaInP/GaAs, tandem		2.5	14	0.86	25–30	Different bandgap materials in tandem increases absorption efficiency
CdTe, thin film	1.5	0.84	26	0.75	15–16	
InP, single crystal	1.34	0.87	29	0.85	21–22	
$CuInSe_2$	1.0				12–13	

NOTE: AM1.5 refers to a solar illumination of "Air Mass 1.5," which represents solar radiation falling on the Earth's surface with a total intensity (or irradiance) of 1000 W m^{-2}. AM1.5 is widely used for comparing solar cells.

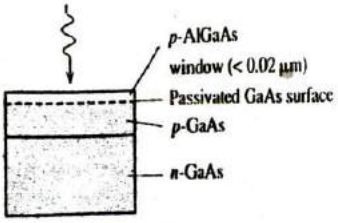


Figure 6.59 AlGaAs window layer on GaAs passivates the surface states and thereby increases the photogeneration efficiency.

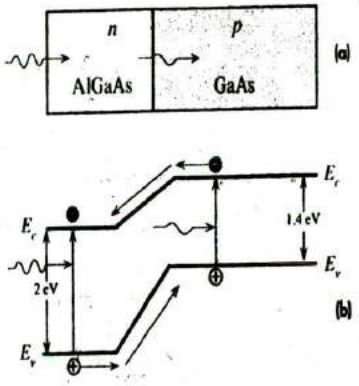


Figure 6.60 A heterojunction solar cell between two different bandgap semiconductors (GaAs and AlGaAs).

Heterojunctions between different bandgap III–V semiconductors that are lattice matched offer the potential of developing high-efficiency solar cells. The simplest single heterojunction example, shown in Figure 6.60, consists of a *pn* junction using a wider bandgap *n*-AlGaAs with *p*-GaAs. Energetic photons ($h\nu > 2$ eV) are absorbed in AlGaAs, whereas those with energies less than 2 eV but greater than 1.4 eV are absorbed in the GaAs layer. In more sophisticated cells, the bandgap of AlGaAs is graded slowly from the surface by varying the composition of the AlGaAs layer.

Tandem or cascaded cells use two or more cells in tandem or in cascade to increase the absorbed photons from the incident light as illustrated in Figure 6.61. The first cell is made from a wider bandgap (E_{g1}) material and only absorbs photons with $h\nu > E_{g1}$. The second cell with bandgap E_{g2} absorbs photons that pass the first cell and have $h\nu > E_{g2}$. The whole structure can be grown within a single crystal by using lattice-matched crystalline layers leading to a monolithic tandem cell. If, in addition, light concentrators are also used, the efficiency can be further increased. For example, a GaAs–GaSb tandem cell operating under a 100-sun condition, that is, 100 times that of ordinary sunlight, have exhibited an efficiency of about 34 percent. Tandem cells have been used in thin-film a-Si:H (hydrogenated amorphous Si) *pin* (*p*-type, intrinsic, and *n*-type structure) solar cells to obtain efficiencies up to about 12 percent. These tandem cells have a-Si:H and a-Si:Ge:H cells and are easily fabricated in large areas.

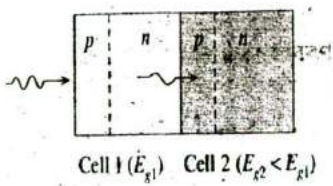


Figure 6.61 A tandem cell. Cell 1 has a wider bandgap and absorbs energetic photons with $h\nu > E_{g1}$. Cell 2 absorbs photons that pass through cell 1 and have $h\nu > E_{g2}$.

ADDITIONAL TOPICS

6.11 *pin* DIODES, PHOTODIODES, AND SOLAR CELLS

The *pin* Si diode is a device that has a structure with three distinct layers: a heavily doped thin p^+-type layer, a relatively thick intrinsic (i-Si) layer, and a heavily doped thin n^+-type layer, as shown in Figure 6.62a. For simplicity we will assume that the i-layer is truly intrinsic, or at least doped so lightly compared with p^+ and n^+ layers that it behaves almost as if intrinsic. The intrinsic layer is much wider than the p^+ and n^+ regions, typically 5–50 μm depending on the particular application. When the structure is first formed, holes diffuse from the p^+-side and electrons from the n^+-side into the i-Si layer where they recombine and disappear. This leaves behind a thin layer of exposed negatively charged acceptor ions in the p^+-side and a thin layer of exposed positively charged donor ions in the n^+-side as shown in Figure 6.22b. The two charges are separated by the i-Si layer of thickness W. There is a uniform built-in field $\mathcal{E}_o$ in the i-Si layer from the exposed positive ions to the exposed negative ions as illustrated in Figure 6.22c. (Since there is no net space charge in the i-layer, from $d\mathcal{E}/dx = \rho/\varepsilon_o\varepsilon_r = 0$, the field must be uniform.) In contrast, the built-in field in the depletion layer of a *pn* junction is not uniform. With no applied bias, the equilibrium is maintained by the built-in field $\mathcal{E}_o$ which prevents further diffusion of majority carriers from the p^+ and n^+ layers into the i-Si layer. A hole that manages to diffuse from the p^+-side into the i-layer is drifted back by $\mathcal{E}_o$, so the net current is zero. As in the *pn* junction, there is also a built-in potential V_o from the edge of the p^+-side depletion region to the edge of the n^+-side depletion region. V_o (like $\mathcal{E}_o$) provides a potential barrier against further net diffusion of holes and electrons into the i-layer and maintains the equilibrium in the open circuit (net current being zero) as in the *pn* junction. It is apparent from Figure 6.62c that, in the absence of an applied voltage, $\mathcal{E}_o = V_o/W$.

One of the distinct advantages of *pin* diodes is that the depletion layer capacitance is very small and independent of the voltage. The separation of two very thin layers of negative and positive charges by a fixed distance, width W of the i-Si layer, is the same as that in a parallel plate capacitor. The **junction or depletion layer capacitance** of the *pin* diode is simply given by

Junction capacitance of pin

$$C_{dep} = \frac{\varepsilon_o\varepsilon_r A}{W} \qquad [6.70]$$

where A is the cross-sectional area and $\varepsilon_o\varepsilon_r$ is the permittivity of the semiconductor (Si), respectively. Further, since the width W of the i-Si layer is fixed by the structure, the junction capacitance does not depend on the applied voltage in contrast to that of the *pn* junction. C_{dep} is typically of the order of a picofarad in fast *pin* photodiodes, so with a 50 Ω resistor, the RC_{dep} time constant is about 50 ps.

When a reverse bias voltage V_r is applied across the *pin* device, it drops almost entirely across the width of the i-Si layer. The depletion layer widths of the thin sheets of acceptor and donor charges in the p^+ and n^+ sides are negligible compared with W. The reverse bias V_r increases the built-in voltage to $V_o + V_r$ as shown in Figure 6.62d. The field $\mathcal{E}$ in the i-Si layer is still uniform and increases to

Reverse-biased pin

$$\mathcal{E} = \mathcal{E}_o + \frac{V_r}{W} \approx \frac{V_r}{W} \qquad (V_r \gg V_o) \qquad [6.71]$$

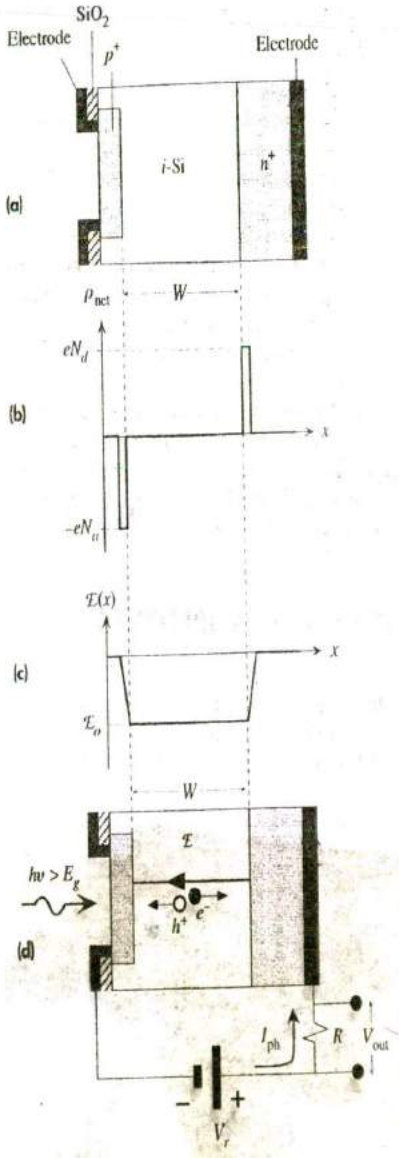


Figure 6.62
(a) The schematic structure of an idealized *pin* photodiode.
(b) The net space charge density across the photodiode.
(c) The built-in field across the diode.
(d) The *pin* photodiode in photodetection is reverse-biased.

Since the width of the *i*-layer in a *pin* device is typically much larger than the depletion layer width in an ordinary *pn* junction, the *pin* devices usually have higher breakdown voltages, which makes them useful where high breakdown voltages are required.

In *pin* photodetectors, the *pin* structure is designed so that photon absorption occurs primarily over the *i*-Si layer. The photogenerated electron–hole pairs (EHPs) in the *i*-Si layer are then separated by the field $\mathcal{E}$ and drifted toward the n^+ and p^+ sides,

respectively, as illustrated in Figure 6.62d. While the photogenerated carriers are drifting through the *i*-Si layer, they give rise to an external photocurrent which is easily detected as a voltage across a small sampling resistor R in Figure 6.62d (or detected by a current-to-voltage converter). The response time of the *pin* photodiode is determined by the transit times of the photogenerated carriers across the width W of the *i*-Si layer. Increasing W allows more photons to be absorbed, which increases the output signal per input light intensity, but it slows down the speed of response because carrier transit times become longer.

The simple *pn* junction photodiode has two major drawbacks. Its junction or depletion layer capacitance is not sufficiently small to allow photodetection at high modulation frequencies. This is an RC time constant limitation. Secondly, its depletion layer is at most a few microns. This means that at long wavelengths where the penetration depth is greater than the depletion layer width, the majority of photons are absorbed outside the depletion layer where there is no field to separate the EHPs and drift them. The photodetector efficiency is correspondingly low at these long wavelengths. These problems are substantially reduced in the *pin* photodiode.[13] The *pin* photovoltaic devices, such as a-Si:H solar cells, are designed to have the photogeneration occur in the *i*-layer as in the case of photodetectors. Obviously, there is no external applied bias, and the built-in field $\mathcal{E}_o$ separates the EHPs and drives the photocurrent.

6.12 SEMICONDUCTOR OPTICAL AMPLIFIERS AND LASERS

All practical semiconductor laser diodes are double heterostructures (DH) whose energy band diagrams are similar to the LED diagram in Figure 6.46. The energy band diagram of a forward biased DH laser diode is shown in Figure 6.63a and b.

Izuo Hayashi and Morton Panish at Bell Labs (1971) were able to design the first semiconductor laser that operated continuously at room temperature. (Notice the similarity of the energy band diagram on the chalkboard with that in Figure 6.63.)

SOURCE: Courtesy of Bell Labs, Lucent Technologies.

[13] The *pin* photodiode was invented by J. Nishizawa and his research group in Japan in 1950.

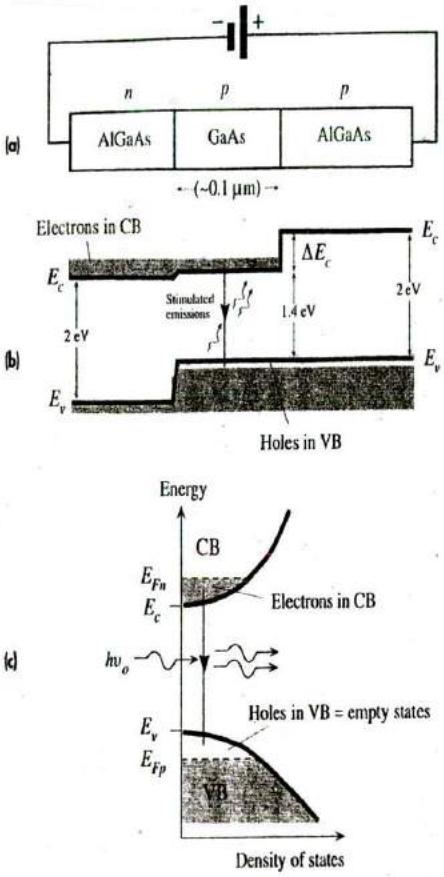


Figure 6.63

(a) A double heterostructure diode has two junctions which are between two different bandgap semiconductors (GaAs and AlGaAs).

(b) Simplified energy band diagram under a large forward bias. Lasing recombination takes place in the p-GaAs layer, the *active layer*.

(c) The density of states and energy distribution of electrons and holes in the conduction and valence bands in the active layer.

In this case the semiconductors are AlGaAs with $E_g \approx 2$ eV and GaAs with $E_g \approx$ 1.4 eV. The p-GaAs region is a thin layer, typically 0.1–0.2 μm, and constitutes the **active layer** in which stimulated emissions take place. Both p-GaAs and p-AlGaAs are heavily p-type doped and are degenerate with the Fermi level E_{Fp} in the valence band. When a sufficiently large forward bias is applied, E_c of n-AlGaAs moves very close to the E_c of p-GaAs which leads to a large injection of electrons in the CB of n-AlGaAs into p-GaAs as shown in Figure 6.63b. In fact, with a sufficient large forward bias, E_c of AlGaAs can be moved above the E_c of GaAs, which causes an enormous electron injection from n-AlGaAs into the CB of p-GaAs. These injected electrons, however, are *confined* to the CB of p-GaAs since there is a barrier ΔE_c between p-GaAs and p-AlGaAs due to the change in the bandgap.

The p-GaAs layer is degenerately doped. Thus, the top of its valence band (VB) is full of holes, or it has all the electronic states *empty* above the Fermi level E_{Fp}

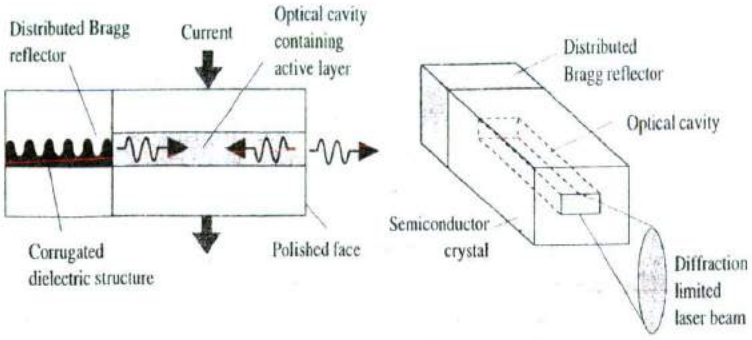


Figure 6.64 Semiconductor lasers have an optical cavity to build up the required electromagnetic oscillations. In this example, one end of the cavity has a Bragg distributed reflector, a reflection grating, that reflects only certain wavelengths back into the cavity.

in this layer. The large forward bias injects a very large concentration of electrons from n-AlGaAs into the conduction band of p-GaAs. Consequently, as shown in Figure 6.63c, there is a large concentration of electrons in the CB and totally empty states at the top of the VB, which means that there is a *population inversion*. An incoming photon with an energy $h\nu_o$ just above E_g can stimulate a conduction electron in the p-GaAs layer to fall down from the CB to the VB and emit a photon by *stimulated emission* as depicted in Figure 6.63c. Such a transition is a photon-stimulated electron–hole recombination, or a lasing recombination. Thus, an avalanche of stimulated emissions in the active layer provides an **optical amplification** of photons with $h\nu_o$ in this layer. The amplification depends on the extent of population inversion and hence on the diode forward current. The device operates as a **semiconductor optical amplifier** which amplifies an optical signal that is passed through the active layer. There is a threshold current below which there is no stimulated emission and no optical amplification.

To construct a **semiconductor laser** with a self-sustained lasing emission we have to incorporate the active layer into an *optical cavity* just as in the case of the HeNe laser in Chapter 3. The optical cavity with reflecting ends, reflects the coherent photons back and forward and encourages their constructive interference within the cavity as depicted in Figure 6.64. This leads to a buildup of high-energy electromagnetic oscillations in the cavity. Some of this electromagnetic energy in the cavity is tapped out as output radiation by having one end of the cavity as partially reflecting. For example, one type of optical cavity, as shown in Figure 6.64, has a special reflector, called a **Bragg distributed reflector** (BDR), at one end to reflect only certain wavelengths back into the cavity.[14] A BDR is a periodic corrugated

[14] Partial reflections of waves from the corrugations in the DBR can interfere constructively and constitute a reflected wave only for certain wavelengths, called *Bragg wavelengths*, that are related to the periodicity of the corrugations. A DBR acts like a reflection grating in optics.

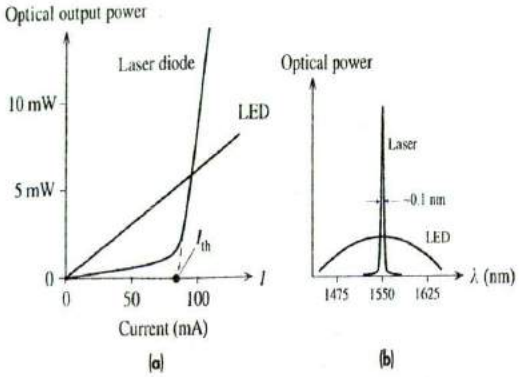


Figure 6.65
(a) Typical optical power output versus forward current for a laser diode and an LED.
(b) Comparison of spectral output characteristics.

structure, like a reflection grating, etched in a semiconductor that reflects only certain wavelengths that are related to the corrugation periodicity. This Bragg reflector has a corrugation periodicity such that it reflects only one desirable wavelength that falls within the optical gain of the active region. This wavelength selective reflection leads to only one possible electromagnetic radiation mode existing in the cavity, which leads to a very narrow output spectrum: a *single-mode output*, that is, only one peak in the output spectrum shown in Figure 3.43. Semiconductor lasers that operate with only one mode in the radiation output are called **single-mode** or **single-frequency lasers;** the spectral linewidth of a single-mode laser output is typically ~0.1 nm, which should be compared with an LED spectral width of 150 nm operating at a 1550 nm emission.

The double heterostructure has further advantages. Wider bandgap semiconductors generally have lower refractive indices, which means AlGaAs has a lower refractive index than that of GaAs. The change in the refractive index defines an optical dielectric waveguide that confines the photons to the active region of the optical cavity and thereby reduces photon losses and increases the photon concentration. This increase in the photon concentration increases the rate of stimulated emissions and the efficiency of the laser.

To achieve the necessary stimulated emissions from a laser diode and build up the necessary optical oscillations in the cavity (to overcome all the optical losses) the current must exceed a certain **threshold current** I_{th} as shown in Figure 6.65a. The optical power output at a current I is then very roughly proportional to $I - I_{th}$. There is still some weak optical power output below I_{th}, but this is simply due to spontaneous recombinations of injected electrons and holes in the active layer; the laser diode behaves like a "poor" LED below I_{th}. The output light from an LED however increases almost in proportion to the diode current. Figure 6.65b compares the output spectrum from the two devices. Remember that the output light from the laser diode is *coherent radiation*, whereas that from an LED is a stream of incoherent photons.

CD Selected Topics and Solved Problems

Selected Topics

The *pn* Junction: Diffusion or Drift? Fick or Ohm?
Shot Noise Generated by the *pn* Junction
Voltage Drift in Semiconductor Devices due to Thermoelectric Effects
Transistor Switches: Why the Saturated Collector–Emitter Voltage is 0.2 V
Semiconductor Device Fabrication (Overview)
Photolithography and Minimum Line Width in Semiconductor Fabrication
Depletion MOSFET Fundamentals
High-Frequency Small-Signal BJT Model

Solved Problems

pn Junction: The Shockley Model
Recombination Current and I–V Characteristics of a *pn* Junction Diode
Design of a *pn* Junction Diode
Bipolar Junction Transistors at Low Frequencies: Principles and Solved Problems
BJT and Nonuniform Base Doping Effect
Junction Field Effect Transistor (JFET)
Enhancement MOSFET and CS Amplifier
LED Emission Wavelength and Temperature

DEFINING TERMS

Accumulation occurs when an applied voltage to the gate (or metal electrode) of a MOS device causes the semiconductor under the oxide to have a greater number of majority carriers than the equilibrium value. Majority carriers have been accumulated at the surface of the semiconductor under the oxide.

Active device is a device that exhibits gain (current or voltage, or both) and has a directional electronic function. Transistors are active devices, whereas resistors, capacitors, and inductors are passive devices.

Antireflection coating reduces light reflection from a surface.

Avalanche breakdown is the enormous increase in the reverse current in a *pn* junction when the applied reverse field is sufficiently high to cause the generation of electron–hole pairs by impact ionization in the space charge layer.

Base width modulation (the Early effect) is the modulation of the base width by the voltage appearing across the base–collector junction. An increase in the base to collector voltage increases the collector junction depletion layer width, which results in the narrowing of the base width.

Bipolar junction transistor (BJT) is a transistor whose normal operation is based on the injection of carriers from the emitter into the base region, where they become minority carriers, and their subsequent diffusion to the collector, where they give rise to a collector current. The voltage between the base and the emitter controls the collector current.

Built-in field is the internal electric field in the depletion region of a *pn* junction that is maximum at the metallurgical junction. It is due to exposed negative acceptors on the *p*-side and positive donors on the *n*-side of the junction.

Built-in voltage (V_o) is the voltage across a *pn* junction, going from a *p*- to *n*-type semiconductor, in an open circuit.

Channel is the conducting strip between the source and drain regions of a MOSFET.

Chip is a piece (or a volume) of a semiconductor crystal that contains many integrated active and passive components to implement a circuit.

Collector junction is the metallurgical junction between the base and the collector of a bipolar transistor.

Critical electric field is the field in the space charge (or depletion) region at reverse breakdown (avalanche or Zener).

Depletion layer (or **space charge layer, SCL**) is a region around the metallurgical junction where recombination of electrons and holes has depleted this region of its large number of equilibrium majority carriers.

Depletion (space charge) layer capacitance is the incremental capacitance (dQ/dV) due to the change in the exposed dopant charges in the depletion layer as a result of the change in the voltage across the *pn* junction.

Diffusion is the flow of particles of a given species from high- to low-concentration regions by virtue of their random thermal motions.

Diffusion (storage) capacitance is the *pn* junction capacitance due to the diffusion and storage of minority carriers in the neutral regions when a forward bias is applied.

Dynamic (incremental) resistance r_d of a diode is the change in the voltage across the diode per unit change in the current through the diode $r_d = dV/dI$. It is the low-frequency ac resistance of the diode. *Dynamic conductance* g_d is the reciprocal dynamic resistance: $g_d = 1/r_d$.

Emitter junction is the metallurgical junction between the emitter and the base.

Enhancement MOSFET is a MOSFET device that needs a gate to source voltage above the threshold voltage to form a conducting channel between the source and the drain. In the absence of a gate voltage, there is no conduction between the source and drain. In its usual mode of operation, the gate voltage enhances the conductance of the source to drain inversion layer and increases the drain current.

Epitaxial layer is a thin layer of crystal that has been grown on the surface of another crystal which is usually a substrate, a mechanical support for the new crystal layer. The atoms of the new layer bond to follow the crystal pattern of the substrate, so the crystal structure of the epitaxial layer is matched with the crystal structure of the substrate.

External quantum efficiency is the optical power emitted from a light emitting device per unit electric input power.

Field effect transistor (FET) is a transistor whose normal operation is based on controlling the conductance of a channel between two electrodes by the application of an external field. The effect of the applied field is to control the current flow. The current is due to majority carrier drift from the source to the drain and is controlled by the voltage applied to the gate.

Fill factor (FF) is a figure of merit for a solar cell that represents, as a percentage, the maximum power $I_m V_m$ available to an external load as a fraction of the *ideal* theoretical power determined by the product of the short circuit current I_{sc} and the open circuit voltage V_{oc}: $\text{FF} = (I_m V_m)/(I_{sc} V_{oc})$.

Forward bias is the application of an external voltage to a *pn* junction such that the positive terminal is connected to the *p*-side and the negative to the *n*-side. The applied voltage reduces the built-in potential.

Heterojunction is a junction between different semiconductor materials, for example, between GaAs and AlGaAs ternary alloy. There may or may not be a change in the doping.

Homojunction is a junction between differently doped regions of the same semiconducting material, for example, a *pn* junction in the same silicon crystal; there is no change in the bandgap energy E_g.

Impact ionization is the process by which a high electric field accelerates a free charge carrier (electron in the CB), which then impacts with a Si–Si bond to generate a free electron–hole pair. The impact excites an electron from E_v to E_c.

Integrated circuit (IC) is a chip of a semiconductor crystal in which many active and passive components have been miniaturized and integrated together to form a sophisticated circuit.

Inversion occurs when an applied voltage to the gate (or metal electrode) of a MOS device causes the semiconductor under the oxide to develop a conducting layer (or a channel) at the surface of the semiconductor. The conducting layer has opposite polarity carriers to the bulk semiconductor and hence is termed an inversion layer.

Ion implantation is a process that is used to bombard a sample in a vacuum with ions of a given species of

atom. First the dopant atoms are ionized in a vacuum and then accelerated by applying voltage differences to impinge on a sample to be doped. The sample is grounded to neutralize the implanted ions.

Isoelectronic impurity atom has the same valency as the host atom.

Law of the junction relates the injected minority carrier concentration just outside the depletion layer to the applied voltage. For holes in the n-side, it is

$$p_n(0) = p_{no} \exp\left(\frac{eV}{kT}\right)$$

where $p_n(0)$ is the hole concentration just outside the depletion layer.

Linewidth is the width of the intensity versus wavelength spectrum, usually between the half-intensity points, emitted from a light emitting device.

Long diode is a pn junction with neutral regions longer than the minority carrier diffusion lengths.

Metallurgical junction is where there is an effective junction between the p-type and n-type doped regions in the crystal. It is where the donor and acceptor concentrations are equal or where there is a transition from n- to p-type doping.

Metal-oxide-semiconductor transistor (MOST) is a field effect transistor in which the conductance between the source and drain is controlled by the voltage supplied to the gate electrode, which is insulated from the channel by an oxide layer.

Minority carrier injection is the flow of electrons into the p-side and holes into the n-side of a pn junction when a voltage is applied to reduce the built-in voltage across the junction.

MOS is short for a metal-insulator-semiconductor structure in which the insulator is typically silicon oxide. It can also be a different type of dielectric; for example, it can be the nitride Si_3N_4.

NMOS is an enhancement type n-channel MOSFET.

Passive device or component is a device that exhibits no gain and no directional function. Resistors, capacitors, and inductors are passive components.

Photocurrent is the current generated by a light-receiving device when it is illuminated.

Pinch-off voltage is the gate to source voltage needed to just pinch off the conducting channel between the source and drain with no source to drain voltage applied. It is also the source to drain voltage that just pinches off the channel when the gate and source are shorted. Beyond pinch-off, the drain current is almost constant and controlled by V_{GS}.

PMOS is an enhancement type p-channel MOSFET.

Poly-Si gate is short for a polycrystalline and highly doped Si gate.

Recombination current flows under forward bias to replenish the carriers recombining in the space charge (depletion) layer. Typically, it is described by $I = I_{ro}[\exp(eV/2kT) - 1]$.

Reverse bias is the application of an external voltage to a pn junction such that the positive terminal is connected to the n-side and the negative to the p-side. The applied voltage increases the built-in potential.

Reverse saturation current is the reverse current that would flow in a reverse-biased ideal pn junction obeying the Shockley equation.

Shockley diode equation relates the diode current to the diode voltage through $I = I_o[\exp(eV/kT) - 1]$. It is based on the injection and diffusion of injected minority carriers by the application of a forward bias.

Short diode is a pn junction in which the neutral regions are shorter than the minority carrier diffusion lengths.

Small-signal equivalent circuit of a transistor replaces the transistor with an equivalent circuit that consists of resistances, capacitances, and dependent sources (current or voltage). The equivalent circuit represents the transistor behavior under small-signal ac conditions. The batteries are replaced with short circuits (or their internal resistances). Small signals imply small variations about dc values.

Substrate is a single mechanical support that carries active and passive devices. For example, in integrated circuit technology, typically, many integrated circuits are fabricated on a single silicon crystal wafer that serves as the substrate.

Thermal generation current is the current that flows in a reverse-biased pn junction as a result of the thermal

generation of electron–hole pairs in the depletion layer that become separated and swept across by the built-in field.

Threshold voltage is the gate voltage needed to establish a conducting channel between the source and drain of an enhancement MOST (metal-oxide-semiconductor transistor).

Transistor is a three-terminal solid-state device in which a current flowing between two electrodes is controlled by the voltage between the third and one of the other terminals or by a current flowing into the third terminal.

Turn-on, or **cut-in,** voltage of a diode is the voltage beyond which there is a substantial increase in the current. The turn-on voltage of a Si diode is about 0.6 V whereas it is about 1 V for a GaAs LED. The turn-on voltage of a *pn* junction diode depends on the bandgap of the semiconductor and the device structure.

Zener breakdown is the enormous increase in the reverse current in a *pn* junction when the applied voltage is sufficient to cause the tunneling of electrons from the valence band in the *p*-side to the conduction band in the *n*-side. Zener breakdown occurs in *pn* junctions that are heavily doped on both sides so that the depletion layer width is narrow.

QUESTIONS AND PROBLEMS

6.1 **The *pn* junction** Consider an abrupt Si pn^+ junction that has 10^{15} acceptors cm^{-3} on the *p*-side and 10^{19} donors on the *n*-side. The minority carrier recombination times are $\tau_e = 490$ ns for electrons in the *p*-side and $\tau_h = 2.5$ ns for holes in the *n*-side. The cross-sectional area is 1 mm^2. Assuming a long diode, calculate the current *I* through the diode at room temperature when the voltage *V* across it is 0.6 V. What are V/I and the incremental resistance (r_d) of the diode and why are they different?

***6.2** **The Si *pn* junction** Consider a long *pn* junction diode with an acceptor doping N_a of 10^{18} cm^{-3} on the *p*-side and donor concentration of N_d on the *n*-side. The diode is forward-biased and has a voltage of 0.6 V across it. The diode cross-sectional area is 1 mm^2. The minority carrier recombination time τ depends on the dopant concentration $N_{dopant}(cm^{-3})$ through the following approximate relation

$$\tau = \frac{5 \times 10^{-7}}{(1 + 2 \times 10^{-17} N_{dopant})}$$

a. Suppose that $N_d = 10^{15}$ cm^{-3}. Then the depletion layer extends essentially into the *n*-side and we have to consider minority carrier recombination time τ_h in this region. Calculate the diffusion and recombination contributions to the total diode current. What is your conclusion?

b. Suppose that $N_d = N_a = 10^{18}$ cm^{-3}. Then *W* extends equally to both sides and, further, $\tau_e = \tau_h$. Calculate the diffusion and recombination contributions to the diode current. What is your conclusion?

6.3 **Junction capacitance of a *pn* junction** The capacitance (*C*) of a reverse-biased abrupt Si p^+n junction has been measured as a function of the reverse bias voltage V_r as listed in Table 6.4. The *pn* junction cross-sectional area is 500 μm × 500 μm. By plotting $1/C^2$ versus V_r, obtain the built-in potential V_o and the donor concentration N_d in the *n*-region. What is N_d?

Table 6.4 Capacitance at various values of reverse bias (V_r)

V_r (V)	1	2	3	5	10	15	20
C (pF)	38.3	30.7	26.4	21.3	15.6	12.9	11.3

6.4 **Temperature dependence of diode properties**

a. Consider the reverse current in a *pn* junction. Show that

$$\frac{\delta I_{rev}}{I_{rev}} \approx \left(\frac{E_g}{\eta kT}\right)\frac{\delta T}{T}$$

where $\eta = 2$ for Si and GaAs, in which thermal generation in the depletion layer dominates the reverse current, and $\eta = 1$ for Ge, in which the reverse current is due to minority carrier diffusion to the depletion layer. It is assumed that $E_g \gg kT$ at room temperature. Order the semiconductors Ge, Si, and GaAs according to the sensitivity of the reverse current to temperature.

b. Consider a forward-biased *pn* junction carrying a *constant* current I. Show that the change in the voltage across the *pn* junction per unit change in the temperature is given by

$$\frac{dV}{dT} = -\left(\frac{V_g - V}{T}\right)$$

where $V_g = E_g/e$ is the energy gap expressed in volts. Calculate typical values for dV/dT for Ge, Si, and GaAs assuming that, typically, $V = 0.2$ V for Ge, 0.6 V for Si, and 0.9 V for GaAs. What is your conclusion? Can one assume that, typically, $dV/dT \approx -2$ mV °C^{-1} for these diodes?

6.5 **Avalanche breakdown** Consider a Si p^+n junction diode that is required to have an avalanche breakdown voltage of 25 V. Given the breakdown field $\mathcal{E}_{br}$ in Figure 6.19, what should be the donor doping concentration?

6.6 **Design of a *pn* junction diode** Design an abrupt Si pn^+ junction that has a reverse breakdown voltage of 100 V and provides a current of 10 mA when the voltage across it is 0.6 V. Assume that, if N_{dopant} is in cm^{-3}, the minority carrier recombination time is given by

$$\tau = \frac{5 \times 10^{-7}}{(1 + 2 \times 10^{-17} N_{dopant})}$$

Mention any assumptions made.

6.7 **Minority carrier profiles (the hyperbolic functions)** Consider a *pnp* BJT under normal operating conditions in which the EB junction is forward-biased and the BC junction is reverse-biased. The field in the neutral base region outside the depletion layers can be assumed to be negligibly small. The continuity equation for holes $p_n(x)$ in the *n*-type base region is

$$D_h \frac{d^2 p_n}{dx^2} - \frac{p_n - p_{no}}{\tau_h} = 0 \qquad [6.71]$$

where $p_n(x)$ is the hole concentration at x from just outside the depletion region and p_{no} and τ_h are the equilibrium hole concentration and hole recombination lifetime in the base.

a. What are the boundary conditions at $x = 0$ and $x = W_B$, just outside the collector region depletion layer? (Consider the law of the junction.)

b. Show that the following expression for $p_n(x)$ is a solution of the continuity equation

$$p_n(x) = p_{no}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]\left[\frac{\sinh\left(\frac{W_B - x}{L_h}\right)}{\sinh\left(\frac{W_B}{L_h}\right)}\right] + p_{no}\left[1 - \frac{\sinh\left(\frac{x}{L_h}\right)}{\sinh\left(\frac{W_B}{L_h}\right)}\right] \qquad [6.72]$$

where $V = V_{EB}$ and $L_h = \sqrt{D_h \tau_h}$.

c. Show that Equation 6.72 satisfies the boundary conditions.

***6.8** **The *pnp* bipolar transistor** Consider a *pnp* transistor in a common base configuration and under normal operating conditions. The emitter–base junction is forward-biased and the base–collector junction is reverse-biased. The emitter, base, and collector dopant concentrations are $N_{a(E)}$, $N_{d(B)}$,

and $N_{a(C)}$, respectively, where $N_{a(E)} \gg N_{d(B)} \geq N_{a(C)}$. For simplicity, assume uniform doping in all the regions. The base and emitter widths are W_B and W_E, respectively, both much shorter than the minority carrier diffusion lengths, L_h and L_e. The minority carrier lifetime in the base is the hole recombination time τ_h. The minority carrier mobility in the base and emitter are denoted by μ_h and μ_e, respectively.

The minority carrier concentration profile in the base can be represented by Equation 6.72.

a. Assuming that the emitter injection efficiency is unity show that

1. $I_E \approx \dfrac{eAD_h n_i^2 \coth\left(\dfrac{W_B}{L_h}\right)}{L_h N_{d(B)}} \exp\left(\dfrac{eV_{EB}}{kT}\right)$

2. $I_C \approx \dfrac{eAD_h n_i^2 \operatorname{cosech}\left(\dfrac{W_B}{L_h}\right)}{L_h N_{d(B)}} \exp\left(\dfrac{eV_{EB}}{kT}\right)$

3. $\alpha \approx \operatorname{sech}\left(\dfrac{W_B}{L_h}\right)$

4. $\beta \approx \dfrac{\tau_h}{\tau_t}$ where $\tau_t = \dfrac{W_B^2}{2D_h}$ is the base transit time.

b. Consider the total emitter current I_E through the EB junction, which has diffusion and recombination components as follows:

$$I_E = I_{E(so)} \exp\left(\frac{eV_{EB}}{kT}\right) + I_{E(ro)} \exp\left(\frac{eV_{EB}}{2kT}\right)$$

Only the hole component of the diffusion current (first term) can contribute to the collector current. Show that when $N_{a(E)} \gg N_{d(B)}$, the emitter injection efficiency γ is given by

$$\gamma \approx \left[1 + \frac{I_{E(ro)}}{I_{E(so)}} \exp\left(-\frac{eV_{EB}}{2kT}\right)\right]^{-1}$$

How does $\gamma < 1$ modify the expressions derived in part (*a*)? What is your conclusion (consider small and large emitter currents, or $V_{EB} = 0.4$ and 0.7 V)?

6.9 **Characteristics of an *npn* Si BJT** Consider an idealized silicon *npn* bipolar transistor with the properties in Table 6.5. Assume uniform doping in each region. The emitter and base widths are between metallurgical junctions (not neutral regions). The cross-sectional area is 100 μm × 100 μm. The transistor is biased to operate in the normal active mode. The base–emitter forward bias voltage is 0.6 V and the reverse bias base–collector voltage is 18 V.

Table 6.5 Properties of an *npn* BJT

Emitter Width	Emitter Doping	Hole Lifetime in Emitter	Base Width	Base Doping	Electron Lifetime in Base	Collector Doping
10 μm	1×10^{18} cm^{-3}	10 ns	5 μm	1×10^{16} cm^{-3}	200 ns	1×10^{16} cm^{-3}

a. Calculate the depletion layer width extending from the collector into the base and also from the emitter into the base. What is the width of the neutral base region?

b. Calculate α and hence β for this transistor, assuming unity emitter injection efficiency. How do α and β change with V_{CB}?

c. What is the emitter injection efficiency and what are α and β, taking into account that the emitter injection efficiency is not unity?

d. What are the emitter, collector, and base currents?

e. What is the collector current when $V_{CB} = 19$ V but $V_{EB} = 0.6$ V? What is the incremental collector output resistance defined as $\Delta V_{CB}/\Delta I_C$?

***6.10 Bandgap narrowing and emitter injection efficiency** Heavy doping in semiconductors leads to what is called *bandgap narrowing* which is an effective narrowing of the bandgap E_g. If ΔE_g is the reduction in the bandgap, then for an n-type semiconductor, according to Lanyon and Tuft (1979),

Bandgap narrowing

$$\Delta E_g(\text{meV}) = 22.5\left(\frac{n}{10^{18}}\right)^{1/2}$$

where n (in cm^{-3}) is the concentration of majority carriers which is equal to the dopant concentration if they are all ionized (for example, at room temperature). The new effective intrinsic concentration n_{ieff} due to the reduced bandgap is given by

Bandgap narrowing

$$n_{ieff}^2 = N_c N_v \exp\left[-\frac{(E_g - \Delta E_g)}{kT}\right] = n_i^2 \exp\left(\frac{\Delta E_g}{kT}\right)$$

where n_i is the intrinsic concentration in the absence of emitter bandgap narrowing.

Mass action law with bandgap narrowing

The equilibrium electron and hole concentrations n_{no} and p_{no}, respectively, obey

$$n_{no}p_{no} = n_{ieff}^2$$

where $n_{no} = N_d$ since nearly all donors would be ionized at room temperature.

Consider a Si npn bipolar transistor operating under normal active conditions with the base-emitter forward biased, and the base-collector reverse biased. The transistor has narrow emitter and base regions. The emitter neutral region width W_E is 1 μm, and the donor doping is 10^{19} cm^{-3}. The width W_B of the neutral base region is 1 μm, and the acceptor doping is 10^{17} cm^{-3}. Assume that W_E and W_B are less than the minority carrier diffusion lengths in the emitter and the base.

a. Obtain an expression for the emitter injection efficiency taking into account the emitter bandgap narrowing effect above.

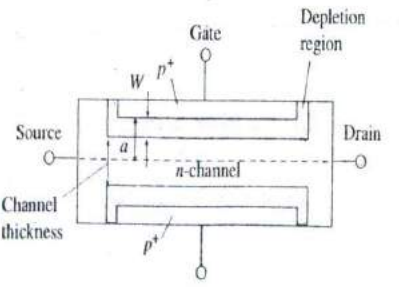
b. Calculate the emitter injection efficiency with and without the emitter bandgap narrowing.

c. Calculate the common emitter current gain β with and without the emitter bandgap narrowing effect given a perfect base transport factor ($\alpha_T = 1$).

6.11 The JFET pinch-off voltage Consider the symmetric n-channel JFET shown in Figure 6.66. The width of each depletion region extending into the n-channel is W. The thickness, or depth, of the channel, defined between the two metallurgical junctions, is $2a$. Assuming an abrupt pn junction and $V_{DS} = 0$, show that when the gate to source voltage is $-V_p$ the channel is pinched off where

$$V_p = \frac{a^2 e N_d}{2\varepsilon} - V_o$$

Figure 6.66 A symmetric JFET.



where V_o is the built-in potential between p^+n junction and N_d is the donor concentration of the channel.

Calculate the pinch-off voltage of a JFET that has an acceptor concentration of 10^{19} cm^{-3} in the p^+ gate, a channel donor doping of 10^{16} cm^{-3}, and a channel thickness (depth) $2a$ of 2 μm.

6.12 **The JFET** Consider an n-channel JFET that has a symmetric p^+n gate–channel structure as shown in Figures 6.27a and 6.66. Let L be the gate length, Z the gate width, and $2a$ the channel thickness. The pinch-off voltage is given by Question 6.11. The drain saturation current I_{DSS} is the drain current when $V_{GS} = 0$. This occurs when $V_{DS} = V_{DS(sat)} = V_P$ (Figure 6.29), so $I_{DSS} = V_P G_{ch}$, where G_{ch} is the conductance of the channel between the source and the pinched-off point (Figure 6.30). Taking into account the shape of the channel at pinch-off, if G_{ch} is about one-third of the conductance of the free or unmodulated (rectangular) channel, show that

$$I_{DSS} = V_P\left[\frac{1}{3}\frac{(e\mu_e N_d)(2a)Z}{L}\right]$$

A particular n-channel JFET with a symmetric p^+n gate–channel structure has a pinch-off voltage of 3.9 V and an I_{DSS} of 5.5 mA. If the gate and channel dopant concentrations are $N_a = 10^{19}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}, respectively, find the channel thickness $2a$ and Z/L. If $L = 10$ μm, what is Z? What is the gate–source capacitance when the JFET has no voltage supplies connected to it?

6.13 **The JFET amplifier** Consider an n-channel JFET that has a pinch-off voltage (V_P) of 5 V and $I_{DSS} = 10$ mA. It is used in a common source configuration as in Figure 6.34a in which the gate to source bias voltage (V_{GS}) is −1.5 V. Suppose that $V_{DD} = 25$ V.

a. If a small-signal voltage gain of 10 is needed, what should be the drain resistance (R_D)? What is V_{DS}?

b. If an ac signal of 3 V peak-to-peak is applied to the gate in series with the dc bias voltage, what will be the ac output voltage peak-to-peak? What is the voltage gain for positive and negative input signals? What is your conclusion?

6.14 **The enhancement NMOSFET amplifier** Consider an n-channel Si enhancement NMOS transistor that has a gate width (Z) of 150 μm, channel length (L) of 10 μm, and oxide thickness (t_{ox}) of 500 Å. The channel has $\mu_e = 700$ cm^2 V^{-1}s^{-1} and the threshold voltage (V_{th}) is 2 V ($\varepsilon_r = 3.9$ for SiO_2).

a. Calculate the drain current when $V_{GS} = 5$ V and $V_{DS} = 5$ V and assuming $\lambda = 0.01$.

b. What is the small-signal voltage gain if the NMOSFET is connected as a common source amplifier, as shown in Figure 6.67, with a drain resistance R_D of 2.2 kΩ, the gate biased at 5 V with respect to

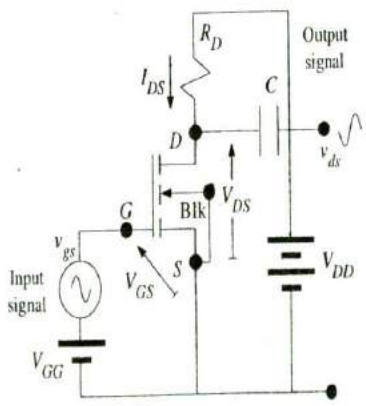


Figure 6.67 NMOSFET amplifier.

source ($V_{GG} = 5$ V) and V_{DD} is such that $V_{DS} = 5$ V? What is V_{DD}? What will happen if the drain supply is smaller?

c. Estimate the most positive and negative input signal voltages that can be amplified if V_{DD} is fixed at the above value in part (*b*).

d. What factors will lead to a higher voltage amplification?

***6.15 Ultimate limits to device performance**

a. Consider the speed of operation of an *n*-channel FET-type device. The time required for an electron to transit from the source to the drain is $\tau_t = L/v_d$, where L is the channel length and v_d is the drift velocity. This transit time can be shortened by shortening L and increasing v_d. As the field increase, the drift velocity eventually saturates at about $v_{dsat} = 10^5$ m s^{-1} when the field in the channel is equal to $\mathcal{E}_t \approx 10^6$ V m^{-1}. A short τ_t requires a field that is at least $\mathcal{E}_t$.

1. What is the change in the *PE* of an electron when it traverses the channel length L from source to drain if the voltage difference is V_{DS}?
2. This energy must be greater than the energy due to thermal fluctuations, which is of the order of kT. Otherwise, electrons would be brought in and out of the drain due to thermal fluctuations. Given the minimum field and V_{DS}, what is the minimum channel length and hence the minimum transit time?

b. Heisenberg's uncertainty principle relates the energy and the time duration in which that energy is possessed through a relationship of the form (Chapter 3) $\Delta E\, \Delta t > h$. Given that during the transit of the electron from the source to the drain its energy changes by eV_{DS}, what is the shortest transit time τ satisfying Heisenberg's uncertainty principle? How does it compare with your calculation in part (*a*)?

c. How does electron tunneling limit the thickness of the gate oxide and the channel length in a MOSFET? What would be typical distances for tunneling to be effective? (Consider Example 3.10.)

6.16 Energy distribution of electrons in the conduction band of a semiconductor and LED emission spectrum

a. Consider the energy distribution of electrons $n_E(E)$ in the conduction band (CB). Assuming that the density of state $g_{cb}(E) \propto (E - E_c)^{1/2}$ and using Boltzmann statistics $f(E) \approx \exp[-(E - E_F)/kT]$, show that the energy distribution of the electrons in the CB can be written as

$$n_x(x) = Cx^{1/2}\exp(-x)$$

where $x = (E - E_c)/kT$ is electron energy in terms of kT measured from E_c, and C is a temperature-dependent constant (independent of E).

b. Setting arbitrarily $C = 1$, plot n_x versus x. Where is the maximum, and what is the full width at half maximum (FWHM), *i.e.*, between half maximum points?

c. Show that the average electron energy in the CB is $\frac{3}{2}kT$, by using the definition of the average,

$$x_{\text{average}} = \frac{\int_0^\infty x n_x\, dx}{\int_0^\infty n_x\, dx}$$

where the integration is from $x = 0$ (E_c) to say $x = 10$ (far away from E_c where $n_x \to 0$). You need to use a numerical integration.

d. Show that the maximum in the energy distribution is at $x = \frac{1}{2}$ or at $E_{\max} = \frac{1}{2}kT$ above E_c.

e. Consider the recombination of electrons and holes in GaAs. The recombination involves the emission of a photon. Given that both electron and hole concentrations have energy distributions in the conduction and valence bands, respectively, sketch schematically the expected light

intensity emitted from electron and hole recombinations against the photon energy. What is your conclusion?

6.17 LED output spectrum Given that the width of the relative light intensity between half-intensity points versus photon energy spectrum of an LED is typically $\sim 3kT$, what is the linewidth $\Delta\lambda$ in the output spectrum in terms of the peak emission wavelength? Calculate the spectral linewidth $\Delta\lambda$ of the output radiation from a green LED emitting at 570 nm at 300 K.

6.18 LED output wavelength variations Show that the change in the emitted wavelength λ with temperature T from an LED is approximately given by

$$\frac{d\lambda}{dT} \approx -\frac{hc}{E_g^2}\left(\frac{dE_g}{dT}\right)$$

where E_g is the bandgap. Consider a GaAs LED. The bandgap of GaAs at 300 K is 1.42 eV which changes (decreases) with temperature as $dE_g/dT = -4.5 \times 10^{-4}$ eV K^{-1}. What is the change in the emitted wavelength if the temperature change is 10 °C?

6.19 Linewidth of direct recombination LEDs Experiments carried out on various direct bandgap semiconductor LEDs give the output spectral linewidth (between half-intensity points) listed in Table 6.6. Since wavelength $\lambda = hc/E_{ph}$, where $E_{ph} = h\nu$ is the photon energy, we know that the spread in the wavelength is related to a spread in the photon energy,

$$\Delta\lambda \approx \frac{hc}{E_{ph}^2}\Delta E_{ph}$$

Suppose that we write $E_{ph} = hc/\lambda$ and $\Delta E_{ph} = \Delta(h\nu) \approx nkT$ where n is a numerical constant. Show that,

$$\Delta\lambda = \frac{nkT}{hc}\lambda^2$$

LED output spectrum linewidth

and by appropriately plotting the data in Table 6.6 find n.

Table 6.6 Linewidth $\Delta\lambda_{1/2}$ between half-points in the output spectrum (intensity versus wavelength) of GaAs and AlGaAs LEDs

	Peak wavelength of emission λ (nm)							
	650	**810**	**820**	**890**	**950**	**1150**	**1270**	**1500**
$\Delta\lambda_{1/2}$ (nm)	22	36	40	50	55	90	110	150
Material (direct E_g)	AlGaAs	AlGaAs	AlGaAs	GaAs	GaAs	InGaAsP	InGaAsP	InGaAsP

6.20 AlGaAs LED emitter An AlGaAs LED emitter for use in a local optical fiber network has the output spectrum shown in Figure 6.68. It is designed for peak emission at 820 nm at 25 °C.

a. What is the linewidth $\Delta\lambda$ between half power points at temperatures −40 °C, 25 °C, and 85 °C? Given these three temperatures, plot $\Delta\lambda$ and T (in K) and find the empirical relationship between $\Delta\lambda$ and T. How does this compare with $\Delta(h\nu) \approx 2.5kT$ to $3kT$?

b. Why does the peak emission wavelength increase with temperature?

c. What is the bandgap of AlGaAs in this LED?

d. The bandgap E_g of the ternary alloys $Al_xGa_{1-x}As$ follows the empirical expression

$$E_g(\text{eV}) = 1.424 + 1.266x + 0.266x^2$$

What is the composition of the AlGaAs in this LED?

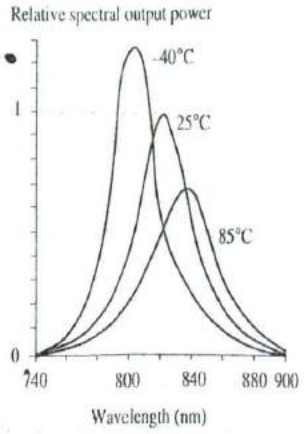


Figure 6.68 The output spectrum from an AlGaAs LED.
Values are normalized to peak emission at 25 °C.

6.21 Solar cell driving a load

a. A Si solar cell of area 2.5 cm × 2.5 cm is connected to drive a load R as in Figure 6.54a. It has the I–V characteristics in Figure 6.53. Suppose that the load is 2 Ω and it is used under a light intensity of 800 W m^{-2}. What are the current and voltage in the circuit? What is the power delivered to the load? What is the efficiency of the solar cell in this circuit?

b. What should the load be to obtain maximum power transfer from the solar cell to the load at 800 W m^{-2} illumination? What is this load at 400 W m^{-2}?

c. Consider using a number of such solar cells to drive a calculator that needs a minimum of 3 V and draws 50 mA at 3–4 V. It is to be used at a light intensity of about 400 W m^{-2}. How many solar cells would you need and how would you connect them?

6.22 Open circuit voltage A solar cell under an illumination of 1000 W m^{-2} has a short circuit current I_{sc} of 50 mA and an open circuit output voltage V_{oc} of 0.65 V. What are the short circuit current and open circuit voltages when the light intensity is halved?

6.23 Maximum power from a solar cell Suppose that the power delivered by a solar cell, $P = IV$, is maximum when $I = I_m$ and $V = V_m$. Suppose that we define normalized voltage and current for maximum power as

Normalized solar cell voltage and current

$$v = \frac{V_m}{\eta V_T} \quad \text{and} \quad i = \frac{I_m}{I_{sc}}$$

where η is the ideality factor, $V_T = kT/e$ is called the thermal voltage (0.026 V at 300 K), and $I_{sc} = -I_{ph}$. Suppose that $v_{oc} = V_{oc}/(\eta V_T)$ is the normalized open circuit voltage. Under illumination with the solar cell delivering power with $V > \eta V_T$,

Power delivered by solar cell

$$P = IV = \left[-I_{ph} + I_o \exp\left(\frac{V}{\eta V_T}\right)\right] V$$

One can differentiate $P = IV$ with respect to V, set it to zero for maximum power, and find expressions for I_m and V_m for maximum power. One can then use the open circuit condition ($I = 0$) to relate V_{oc} to I_o. Show that maximum power occurs when

Maximum power delivery

$$v = v_{oc} - \ln(v + 1) \quad \text{and} \quad i = 1 - \exp[-(v_{oc} - v)]$$

Consider a solar cell with $\eta = 1.5$, $V_{oc} = 0.60$ V, and $I_{ph} = 35$ mA, with an area of 1 cm^2. Find i and v, and hence the current I_m and voltage V_m for maximum power. (Note: Solve the first equation numerically or graphically to find $v \approx 12.76$.) What is the fill factor?

6.24 **Series resistance** The series resistance causes a voltage drop when a current is drawn from a solar cell. By convention, the positive current is taken to flow into the device. (If calculations yield a negative value, it means that, physically, the current is flowing out, which is the actual case under illumination.) If V is the actual voltage across the solar cell output (accessed by the user), then the voltage across the diode is $V - IR_s$. The solar cell equation becomes

$$I = -I_{ph} + I_d = -I_{ph} + I_o \exp\left(\frac{e(V - IR_s)}{\eta kT}\right)$$

Solar cell with series resistance

Plot I versus V for a Si solar cell that has $\eta = 1.5$ and $I_o = 3 \times 10^{-6}$ mA, for an illumination such that $I_{ph} = 10$ mA for $R_s = 0$, 20 and 50 Ω. What is your conclusion?

6.25 **Shunt resistance** Consider the shunt resistance R_p of a solar cell. Whenever there is a voltage V at the terminals of the solar cell, the shunt resistance draws a current V/R_p. Thus, the total current as seen at the terminals (and flowing in by convention) is

$$I = -I_{ph} + I_d + \frac{V}{R_p} = -I_{ph} + I_o \exp\left(\frac{eV}{\eta kT}\right) + \frac{V}{R_p} = 0$$

Solar cell with shunt resistance

Plot I versus V for a polycrystalline Si solar cell that has $\eta = 1.5$ and $I_o = 3 \times 10^{-6}$ mA, for an illumination such that $I_{ph} = 10$ mA. Use $R_p = \infty$, 1000, 100 Ω. What is your conclusion?

***6.26** **Series connected solar cells** Consider two identical solar cells connected in series. There are two R_s in series and two pn junctions in series. If I is the total current through the devices, then the voltage across one pn junction is $V_d = \frac{1}{2}[V - I(2R_s)]$ so that the current I flowing into the combined solar cells is

$$I \approx -I_{ph} + I_o \exp\left[\frac{V - I(2R_s)}{2\eta V_T}\right] \qquad V_d > \eta\left(\frac{kT}{e}\right)$$

Two solar cells in series

where $V_T = kT/e$ is the thermal voltage. Rearranging, for two cells in series,

$$V = 2\eta V_T \ln\left(\frac{I + I_{ph}}{I_o}\right) + 2R_s I$$

Two solar cells in series

whereas for one cell,

$$V = \eta V_T \ln\left(\frac{I + I_{ph}}{I_o}\right) + R_s I$$

One solar cell

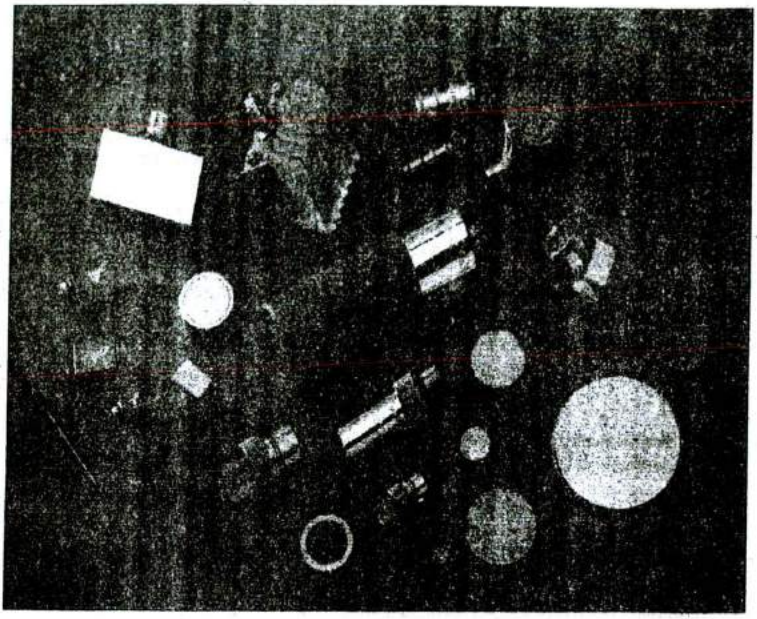
Suppose that the cells have the properties $I_o = 25 \times 10^{-6}$ mA, $\eta = 1.5$, $R_s = 20$ Ω, and both are subjected to the same illumination so that $I_{ph} = 10$ mA. Plot the individual I–V characteristics and the I–V characteristics of the two cells in series. Find the maximum power that can be delivered by one cell and two cells in series. Find the corresponding voltage and current at the maximum power point.

6.27 **A solar cell used in Eskimo Point** The intensity of light arriving at a point on Earth, where the solar latitude is α can be approximated by the Meinel and Meinel equation:

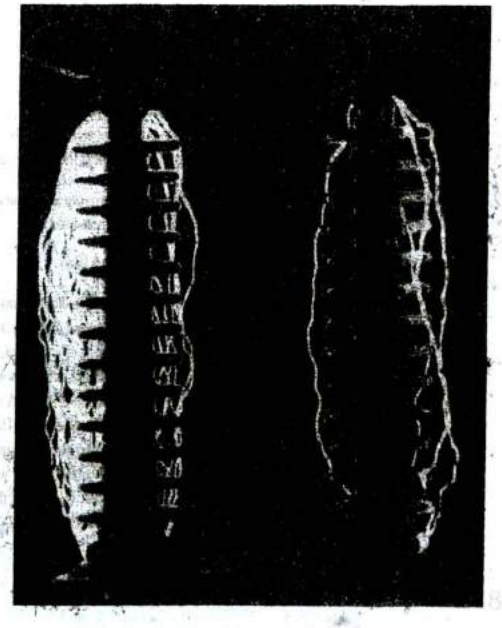
$$I = 1.353(0.7)^{(\text{cosec}\,\alpha)^{0.678}} \quad \text{kW m}^{-2}$$

where $\text{cosec}\,\alpha = 1/(\sin\alpha)$. The solar latitude α is the angle between the sun's rays and the horizon. Around September 23 and March 22, the sun's rays arrive parallel to the plane of the equator. What is the maximum power available for a photovoltaic device panel of area 1 m^2 if its efficiency of conversion is .J percent?

A manufacturer's characterization tests on a particular Si pn junction solar cell at 27 °C specifies an open circuit output voltage of 0.45 V and a short circuit current of 400 mA when illuminated directly with a light of intensity 1 kW m^{-2}. The fill factor for the solar cell is 0.73. This solar cell is to be used in a portable equipment application near Eskimo Point (Canada) at a geographical latitude (ϕ) of 63°. Calculate the open circuit output voltage and the maximum available power when the solar cell is used at noon on September 23 when the temperature is around −10 °C. What is the maximum current this solar cell can supply to an electronic equipment? What is your conclusion? (Note: $\alpha + \phi = \pi/2$, and assume $\eta = 1$, and that $I_o \propto n_i^2$)



A selection of ultrasonic transducers (piezoelectric effect devices).
SOURCE: Courtesy of Valpey Fisher.



An HV capacitor bushing being subjected to mains-frequency overvoltage. The photo is one of prolonged exposure, recording multiple surface flashovers.
SOURCE: Courtesy of Dr. Simon Rowland, UMIST, England.