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 3*s* and 3*p* 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. 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

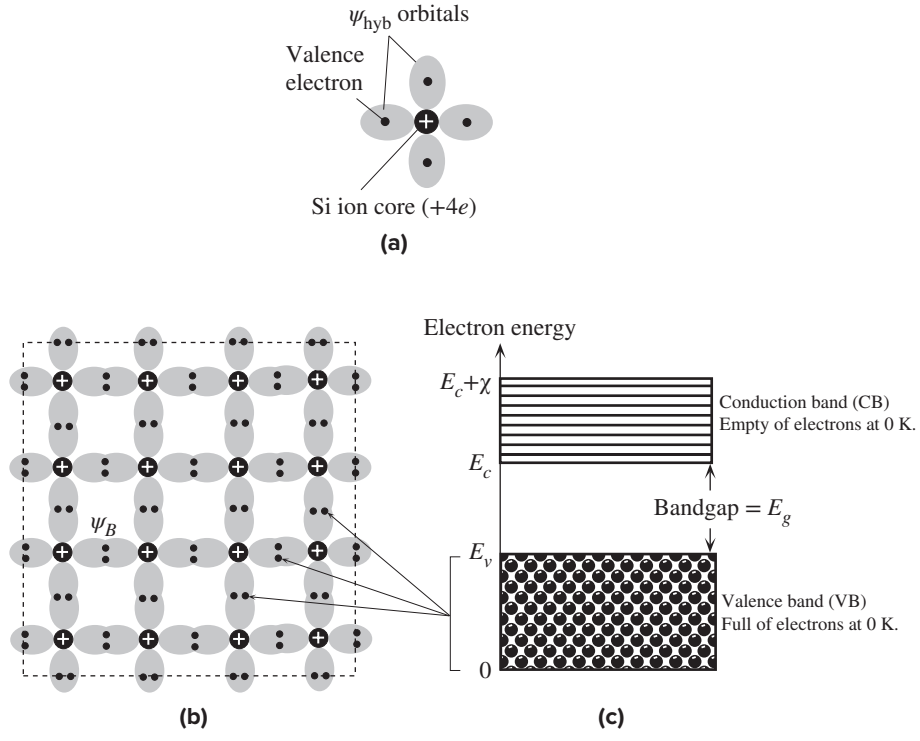


Figure 5.1 (a) A simplified two-dimensional illustration of a Si atom with four hybrid orbitals ψ_{hyb}. Each orbital has one electron. (b) A simplified two-dimensional view of a region of the Si crystal showing covalent bonds. (c) The energy band diagram at absolute zero of temperature.

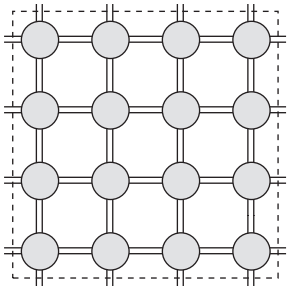


Figure 5.2 A two-dimensional pictorial view of the Si crystal showing covalent bonds as two lines where each line is a valence electron.

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.

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

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

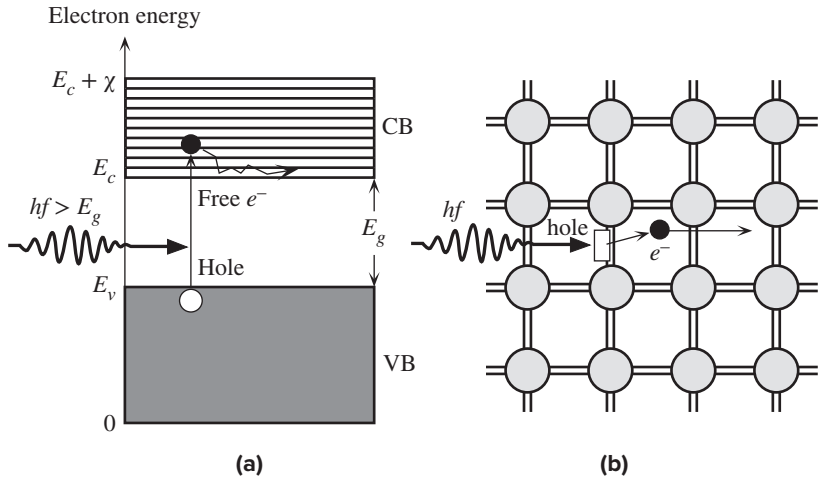


Figure 5.3 (a) A photon with an energy greater than E_g can excite an electron from the VB to the CB. (b) When a photon breaks a Si–Si bond, a free electron and a hole in the Si–Si bond are created.

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 $hf > E_g$ 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 $hf > E_g$ 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 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

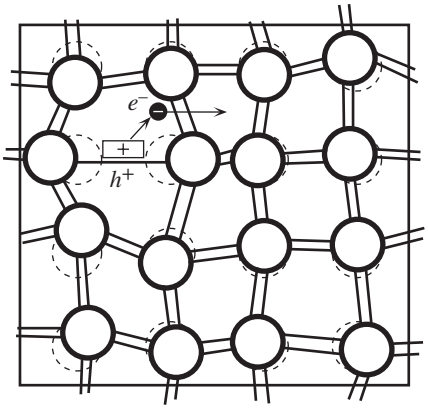


Figure 5.4 Thermal vibrations of atoms can break bonds and thereby create electron–hole pairs.

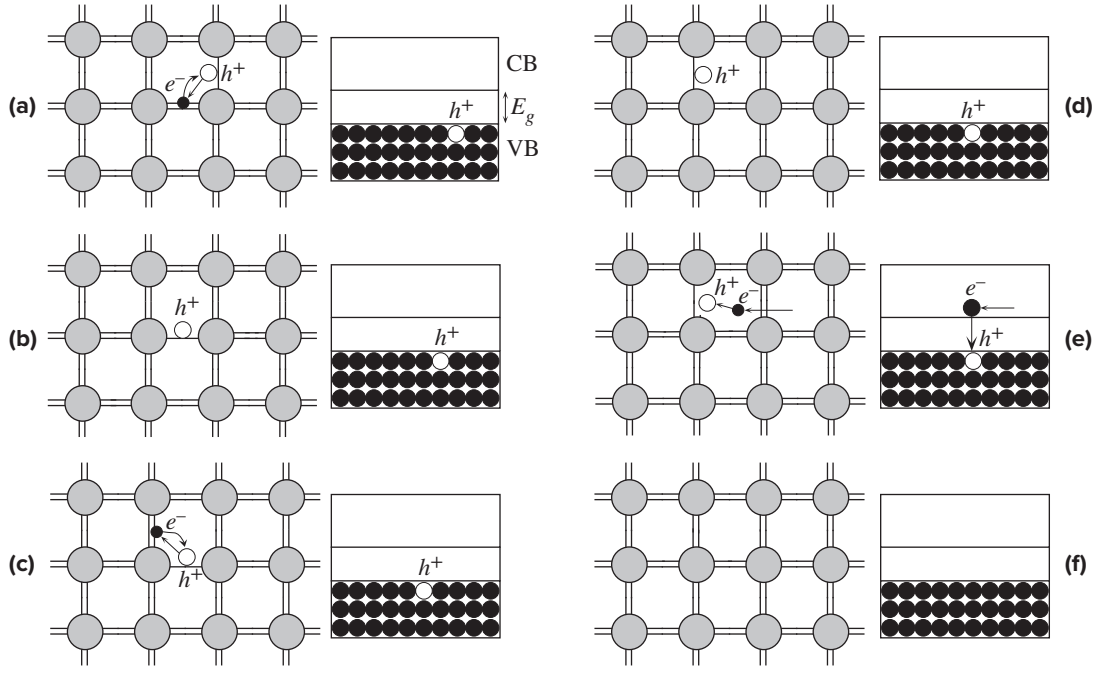


Figure 5.5 A pictorial illustration of a hole in the valence band wandering around the crystal due to the tunneling of electrons from neighboring bonds.

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 5.13). 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.

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

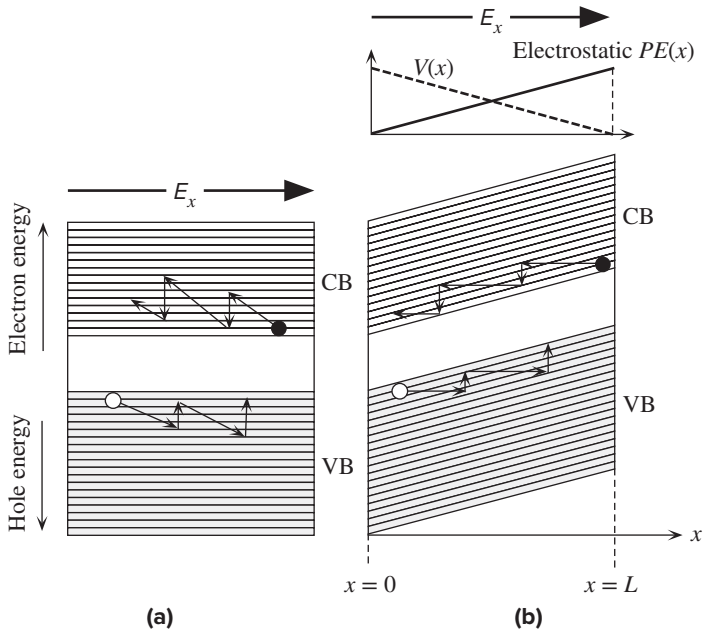


Figure 5.6 When an electric field is applied, electrons in the CB and holes in the VB can drift and contribute to the conductivity. (a) A simplified illustration of drift in E_x. (b) Applied field bends the energy bands since the electrostatic PE of the electron is $-eV(x)$ and $V(x)$ decreases in the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

additional electrostatic PE contribution that is not constant in an applied electric field. A uniform electric field E_x implies a linearly decreasing potential $V(x)$, by virtue of $(dV/dx) = -E_x$, 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 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 $\mu_e E_x$ 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.

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 E_x. Thus,

Electron and hole drift velocities

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

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

Drift mobility and scattering time

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

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.*, $-eE_x$) by $F_{\text{ext}} = m_e^* a$ just as we do for the electron in vacuum by $F_{\text{ext}} = m_e a$. In applying the $F_{\text{ext}} = m_e^* a$ 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.

Conductivity of a semiconductor

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

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

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 of the CB gives the electron concentration n, number of electrons per unit volume, in the CB. In other words,

$$n = \int_{E_c}^{\text{Top of CB}} n_E(E)dE = \int_{E_c}^{\text{Top of CB}} g_{cb}(E)f(E)dE$$

Boltzmann tail of Fermi–Dirac distribution

We will assume that $(E_c - F_F) \gg kT$ (*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.

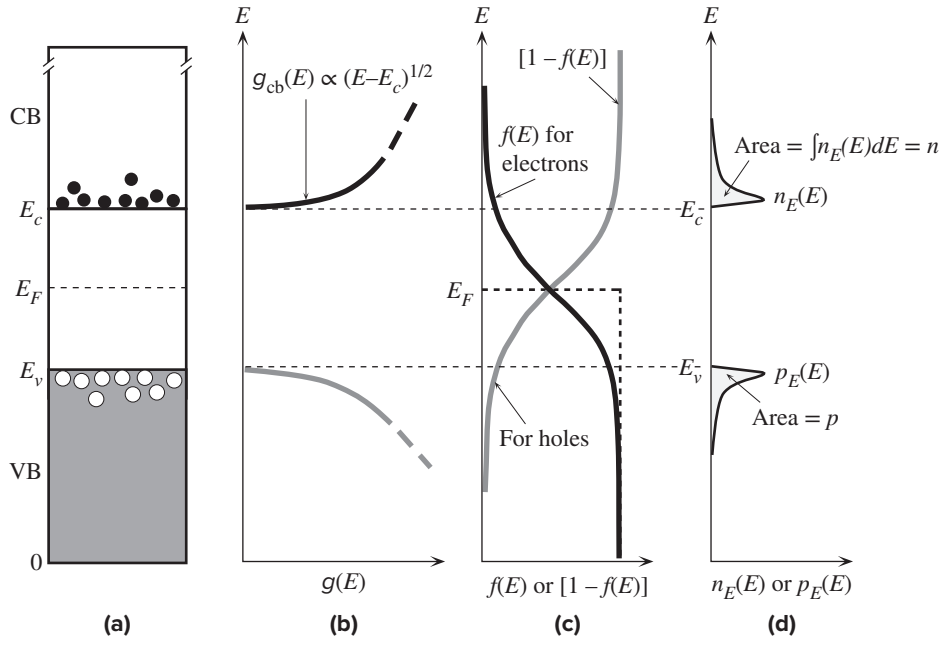


Figure 5.7 (a) Energy band diagram. (b) Density of states (number of states per unit energy per unit volume). (c) Fermi–Dirac probability function (probability of occupancy of a state). (d) The product of $g(E)$ and $f(E)$ is the energy density of electrons in the CB (number of electrons per unit energy per unit volume). The area under $n_E(E)$ versus E is the electron concentration.

Further, we will take the upper limit to be $E = \infty$ 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

$$n = N_c \exp\left[-\frac{(E_c - E_F)}{kT}\right] \quad [5.6]$$

where

Effective density of states at CB edge

$$N_c = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{3/2} \quad [5.7]$$

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 constant, that is, independent

of the Fermi energy, and is called the **effective density of states at the CB edge.** Notice that N_c depends on the effective mass[2] and has a small temperature dependence as apparent from Equation 5.7. 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

$$p = N_v \exp\left[-\frac{(E_F - E_v)}{kT}\right] \quad [5.8]$$

where N_v is the effective density of states at the VB edge and is given by

Effective density of states at VB edge

$$N_v = 2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{3/2} \quad [5.9]$$

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

$$np = N_c N_v \exp\left(-\frac{E_g}{kT}\right) \quad [5.10]$$

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, $N_c N_v \exp(-E_g/kT)$, behaves as if it were a constant for a given material at a given temperature; it depends on the bandgap E_g but not on the position of the Fermi level. In the special case of an intrinsic

[2] The effective mass in Equation 5.7 is called the *density of states effective mass*, and is not the same as that used in describing the electron drift mobility.

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

$$np = n_i^2 = N_c N_v \exp\left(-\frac{E_g}{kT}\right) \quad [5.11a]$$

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 constant at a given temperature and depends only on the semiconductor material. Equation 5.11a is called the **mass action law.** 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. From Equation 5.11a,

$$n_i = (N_c N_v)^{1/2} \exp\left(-\frac{E_g}{2kT}\right) \quad [5.11b]$$

Intrinsic concentration

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 5.11b 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 5.11a.

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

$$E_{Fi} = E_v + \frac{1}{2}E_g - \frac{1}{2}kT \ln\left(\frac{N_c}{N_v}\right) \quad [5.12]$$

Furthermore, substituting the proper expressions for N_c and N_v we get

Fermi energy in intrinsic semiconductor

$$E_{Fi} = E_v + \frac{1}{2}E_g - \frac{3}{4}kT \ln\left(\frac{m_e^*}{m_h^*}\right) \quad [5.13]$$

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

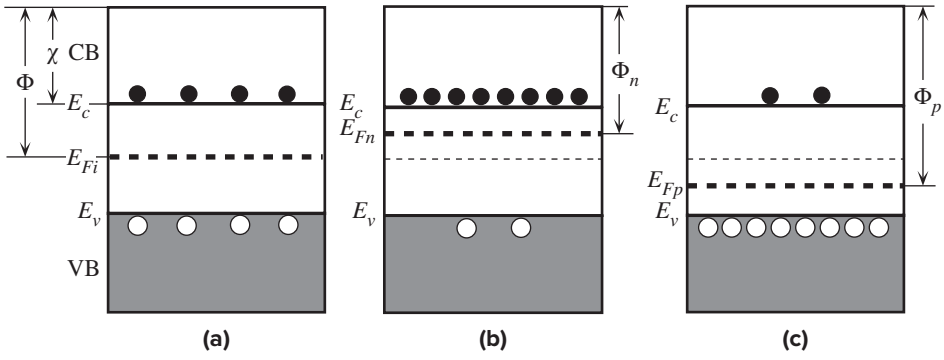


Figure 5.8 Energy band diagrams for (a) intrinsic, (b) *n*-type, and (c) *p*-type semiconductors. In all cases, $np = n_i^2$.

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.

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

$$\overline{E}_{CB} = \frac{1}{n}\int_{CB} E g_{cb}(E) f(E)\ dE$$

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

$$\overline{E}_{CB} = E_c + \frac{3}{2}kT \qquad [5.14]$$

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.

Table 5.1 Selected typical properties of Ge, Si, InP, 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}	2.3×10^{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}	1400	450	0.26*a* 1.08*b*	0.38*a* 0.60*b*	11.9
InP	1.34	4.50	5.2×10^{17}	1.1×10^{19}	1.3×10^{7}	4600	190	0.079*a,b*	0.46*a* 0.58*b*	12.6
GaAs	1.42	4.07	4.4×10^{17}	7.7×10^{18}	2.1×10^{6}	8800	400	0.067*a,b*	0.40*a* 0.50*b*	13.0

NOTE: Ge and Si are indirect whereas InP and GaAs are direct bandgap semiconductors. 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 1.45×10^{10} cm^{-3} and is therefore the most widely used value of n_i for Si, though the correct value is actually 1.0×10^{10} cm^{-3}. (Green, M.A., *Journal of Applied Physics,* 67, 2944, 1990.) (Data combined from various sources.)

This is just like the average kinetic energy of gas atoms (such as He atoms) in a tank assuming that the atoms (or the "particles") 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 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 compares some of the properties of the important semiconductors, Ge, Si, InP, 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 1400 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

$$N_c = 2\left[\frac{2\pi(1.08 \times 9.1 \times 10^{-31}\ \text{kg})(1.38 \times 10^{-23}\ \text{J K}^{-1})(300\ \text{K})}{(6.63 \times 10^{-34}\ \text{J s})^2}\right]^{3/2}$$
$$= 2.81 \times 10^{25}\ \text{m}^{-3} \quad \text{or} \quad 2.81 \times 10^{19}\ \text{cm}^{-3}$$

and

$$N_v = 2\left[\frac{2\pi(0.60 \times 9.1 \times 10^{-31}\ \text{kg})(1.38 \times 10^{-23}\ \text{J K}^{-1})(300\ \text{K})}{(6.63 \times 10^{-34}\ \text{J s})^2}\right]^{3/2}$$
$$= 1.16 \times 10^{25}\ \text{m}^{-3} \quad \text{or} \quad 1.16 \times 10^{19}\ \text{cm}^{-3}$$

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,

$$\sigma = (1.6 \times 10^{-19}\ \text{C})(1.0 \times 10^{10}\ \text{cm}^{-3})(1400 + 450\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1})$$
$$= 3.0 \times 10^{-6}\ \Omega^{-1}\ \text{cm}^{-1}$$

The resistivity is

$$\rho = \frac{1}{\sigma} = 3.3 \times 10^5\ \Omega\ \text{cm}$$

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[3] $n_i = 1.0 \times 10^{10}\ \text{cm}^{-3}$, 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

Suppose that v_{th} is the root mean square velocity of the electron in the CB, then the average KE, $\frac{1}{2}m_e^* v_{\text{th}}^2$, of this electron from Equation 5.14 is $\frac{3}{2}kT$. Thus,

$$v_{\text{th}} = \left(\frac{3kT}{m_e^*}\right)^{1/2} = \left[\frac{(3 \times 1.38 \times 10^{-23} \times 300)}{(0.26 \times 9.1 \times 10^{-31})}\right]^{1/2} = 2.3 \times 10^5\ \text{m s}^{-1}$$

---

[3] 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).

The above velocity v_{th} is called the **thermal velocity,** and it is roughly the same as the mean speed of the electron in the CB. (See Example 1.11.)

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_{th} of the electron and the scattering cross section of the thermal vibrations, that is,

$$\tau \propto \frac{1}{v_{th}(\pi a^2)}$$

where a is the amplitude of the atomic thermal vibrations. But, $v_{th} \propto T^{1/2}$ 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.

EXAMPLE 5.3 **MEAN FREE PATH OF ELECTRONS IN THE CB** Consider the motion of electrons in the CB of an undoped GaAs crystal. What is the mean free path of an average electron in the conduction band? How does this compare with the mean free path of a conduction electron in copper which has a drift mobility of 32 $cm^2\ V^{-1}\ s^{-1}$ and a mean free path of 40 nm. What is your conclusion?

SOLUTION

The drift mobility of electrons in a semiconductor is controlled by various scattering mechanisms that limit the mean scattering time or the free time of an electron in the CB. If τ is the mean scattering time for electrons in the CB, then, from Chapter 2, drift mobility $\mu_e = e\tau/m_e^*$, where m_e^* is the effective mass of the electron in the crystal. Thus,

$$\tau = \frac{\mu_e m_e^*}{e} = \frac{(1400 \times 10^{-4}\ m^2\ V^{-1}\ s^{-1})(0.26 \times 9.11 \times 10^{-31}\ kg)}{(1.602 \times 10^{-19}\ C)} = 2.1 \times 10^{-13}\ s$$

We know from Example 5.1 that the mean velocity or the thermal velocity v_{th} of electrons in the CB is approximately $2.3 \times 10^5\ m\ s^{-1}$. The mean free path $\ell = v_{th}\tau = (2.3 \times 10^5\ m\ s^{-1})(2.1 \times 10^{-13}\ s) = 48 \times 10^{-9}$ m or 48 nm. The conduction electrons in copper have significantly lower drift mobility but their mean free path is almost the same as a conduction electron in Si. Recall from Chapter 4 that conduction electrons in a metal follow Fermi–Dirac statistics and their mean speed is very much larger than the thermal velocity of electrons in Si.

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

$$E_b = -E_1 = \frac{m_e e^4}{8\varepsilon_o^2 h^2} = 13.6 \text{ eV}$$

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

$$E_b^{\text{Si}} = \frac{m_e^* e^4}{8\varepsilon_r^2\varepsilon_o^2 h^2} = (13.6 \text{ eV})\left(\frac{m_e^*}{m_e}\right)\left(\frac{1}{\varepsilon_r^2}\right) \quad [5.15]$$

With $\varepsilon_r = 11.9$ and $m_e^* \approx \frac{1}{3}m_e$ for silicon, we find $E_b^{\text{Si}} = 0.032$ eV, which is comparable with the average thermal energy of atomic vibrations at room temperature, $\sim 3kT$ ($\sim$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,

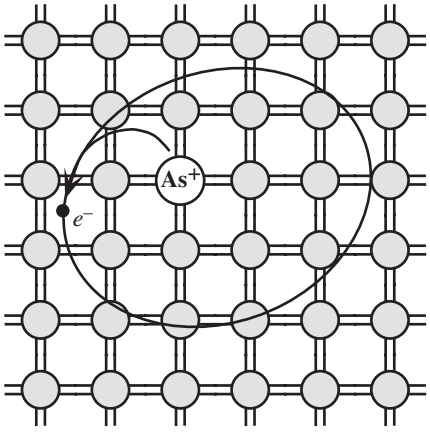


Figure 5.9 Arsenic-doped Si crystal.
The four valence electrons of As allow it to bond just like Si, but the fifth electron is left orbiting the As site. The energy required to release the free fifth electron into the CB is very small.

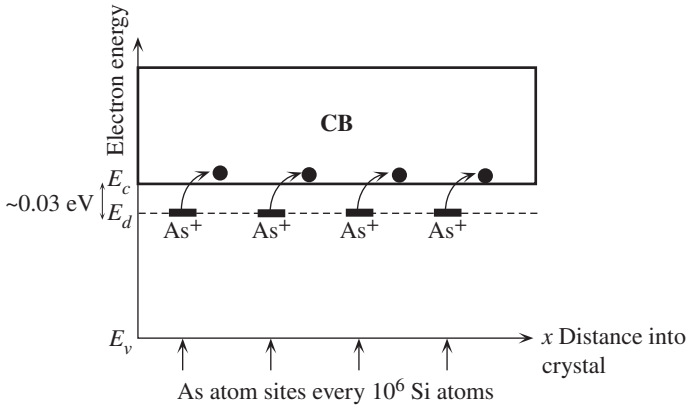


Figure 5.10 Energy band diagram for an *n*-type Si doped with 1 ppm As.
There are donor energy levels just below E_c around As^+ sites.

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 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

$$f_d(E_d) = \frac{1}{1 + \frac{1}{2}\exp\left[\frac{(E_d - E_F)}{kT}\right]} \qquad [5.17]$$

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[4] (once the donor has been occupied,

[4] The proof can be found in advanced solid-state physics texts.

a second electron cannot enter this site). Thus, the concentration of ionized donors at a temperature T is given by

$$\begin{aligned} N_d^+ &= N_d \times (\text{probability of not finding an electron at } E_d) \\ &= N_d[1 - f_d(E_d)] \\ &= \frac{N_d}{1 + 2\exp\left[\dfrac{(E_F - E_d)}{kT}\right]} \end{aligned} \qquad [5.18]$$

Ionized donor concentration

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

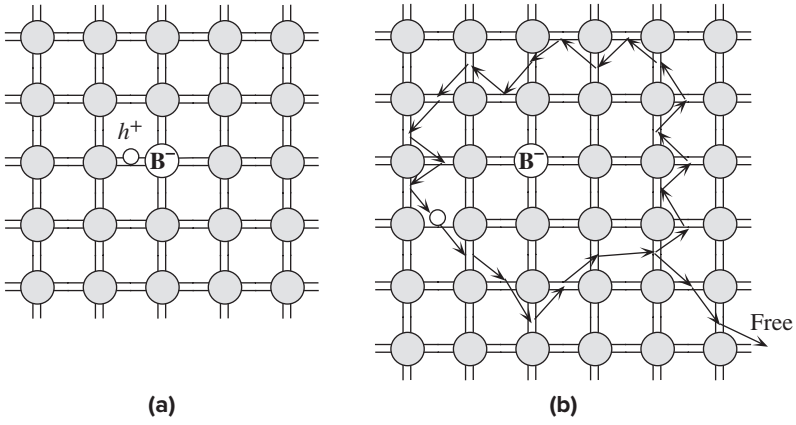


Figure 5.11 Boron-doped Si crystal.

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.

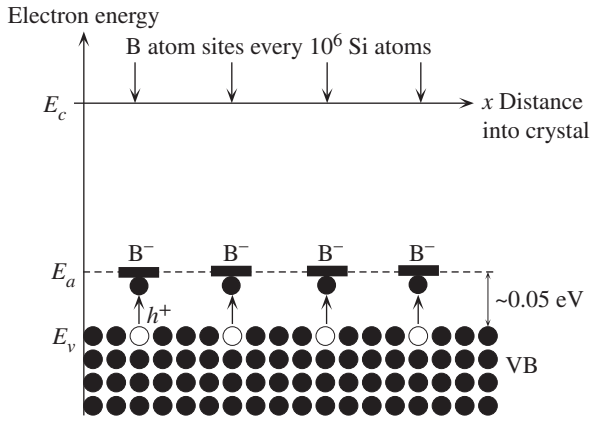


Figure 5.12 Energy band diagram for a *p*-type Si doped with 1 ppm B. There are acceptor energy levels E_a just above E_v around B^- sites. These acceptor levels accept electrons from the VB and therefore create holes in the VB.

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

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 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.

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 5.9.

EXAMPLE 5.4

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)? Note that the atomic concentration in silicon is 5×10^{22} cm^{-3}, $n_i = 1.0 \times 10^{10}$ cm^{-3}, $\mu_e = 1400$ 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

$$\sigma = (1.6 \times 10^{-19}\ \text{C})(1.0 \times 10^{10}\ \text{cm}^{-3})(1400 + 450\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1})$$
$$= 2.96 \times 10^{-6}\ \Omega^{-1}\ \text{cm}^{-1}$$

Since $L = 1$ cm and $A = 1$ cm^2, the resistance is

$$R = \frac{L}{\sigma A} = \frac{1}{\sigma} = 3.47 \times 10^5\ \Omega \qquad \text{or} \qquad 347\ \text{k}\Omega$$

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,

$$\sigma = en\mu_e = (1.6 \times 10^{-19}\ \text{C})(5 \times 10^{13}\ \text{cm}^{-3})(1400\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1})$$
$$= 1.12 \times 10^{-2}\ \Omega^{-1}\ \text{cm}^{-1}$$

Further,

$$R = \frac{L}{\sigma A} = \frac{1}{\sigma} = 89.3\ \Omega$$

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 $N_a = 5 \times 10^{13}$ 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 5.5 **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 $n_i = 1.0 \times 10^{10}$ 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 5.6

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 $N_d = 10^{16}$ cm^{-3}, and since $N_d \gg n_i$ ($=1.0 \times 10^{10}$ 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

$$E_{Fp} - E_{Fi} = -kT \ln\left(\frac{p}{n_i}\right) = -(0.0259 \text{ eV}) \ln\left(\frac{1.9 \times 10^{17}}{1.0 \times 10^{10}}\right)$$
$$= -0.43 \text{ eV}$$

EXAMPLE 5.7 **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 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

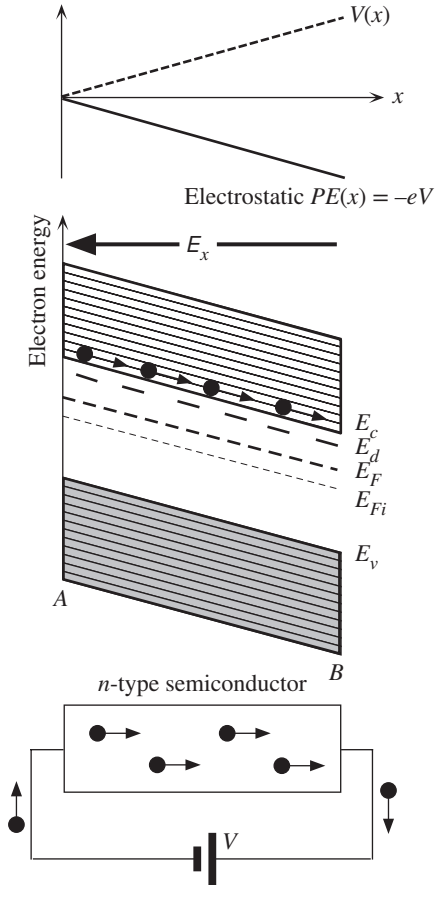


Figure 5.13 Energy band diagram of an *n*-type semiconductor connected to a voltage supply of V volts.

The whole energy diagram tilts because the electron now also has an electrostatic potential energy.

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

$$n = \left(\frac{1}{2}N_c N_d\right)^{1/2} \exp\left(-\frac{\Delta E}{2kT}\right) \qquad [5.19]$$

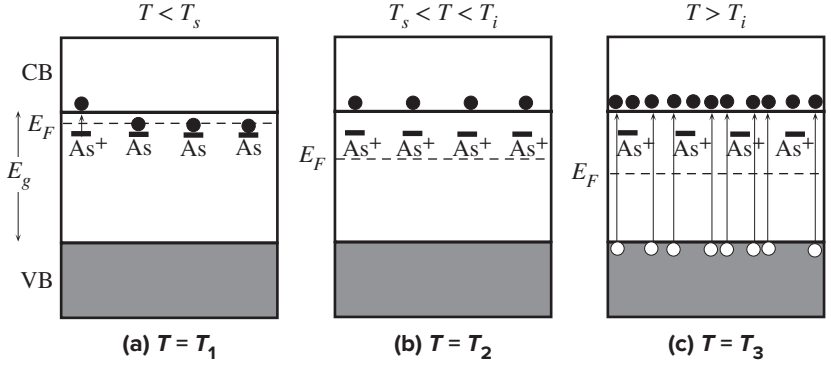


Figure 5.14 (a) Below T_s, the electron concentration is controlled by the ionization of the donors. (b) Between T_s and T_i, the electron concentration is equal to the concentration of donors since they would all have ionized. (c) At high temperatures, thermally generated electrons from the VB exceed the number of electrons from ionized donors and the semiconductor behaves as if intrinsic.

similar to the intrinsic case, that is,

$$n = (N_c N_v)^{1/2} \exp\left(-\frac{E_g}{2kT}\right) \quad [5.20]$$

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 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

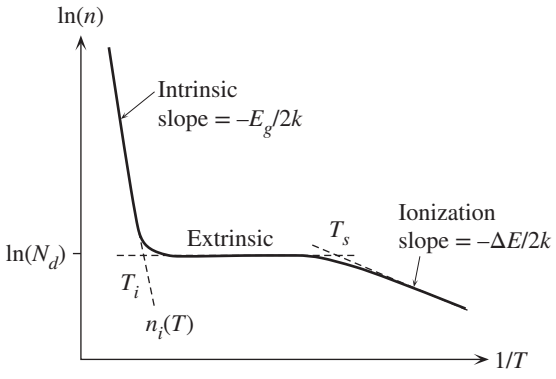


Figure 5.15 The temperature dependence of the electron concentration in an *n*-type semiconductor.

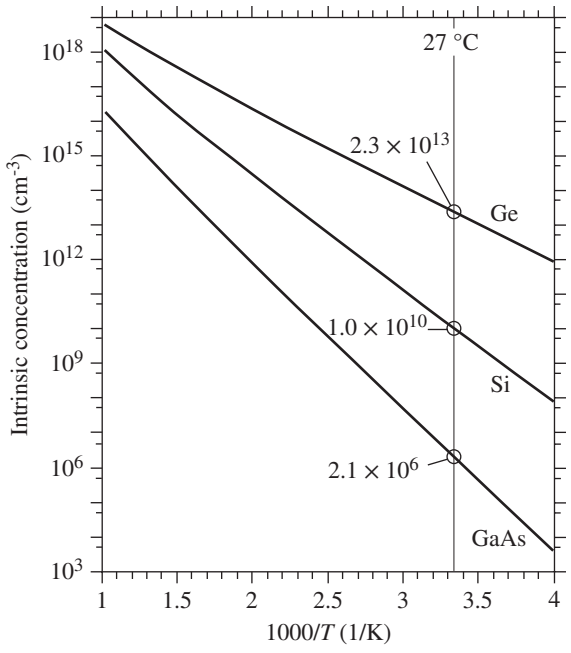


Figure 5.16 The temperature dependence of the intrinsic concentration.

$-(\Delta E/2k)$, 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 5.16 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 5.8

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:

$$n = \left(\frac{1}{2}N_cN_d\right)^{1/2} \exp\left(-\frac{\Delta E}{2kT_s}\right) \approx N_d$$

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,

$$T_s = \frac{\Delta E}{k \ln\left(\dfrac{N_c}{2N_d}\right)} = \frac{0.045\ \text{eV}}{(8.62 \times 10^{-5}\ \text{eV K}^{-1}) \ln\left[\dfrac{2.8 \times 10^{19}\ \text{cm}^{-3}}{2(1.0 \times 10^{15}\ \text{cm}^{-3})}\right]} = 54.7\ \text{K}$$

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 K to 541 K or −198 °C to about 268 °C.

EXAMPLE 5.9

Electron concentration in the ionization region

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

$$n = \left(\frac{1}{2}N_cN_d\right)^{1/2} \exp\left(-\frac{\Delta E}{2kT}\right)$$

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.**

$$\begin{aligned} N_d^+ &= N_d \times (\text{probability of not finding an electron at } E_d) = N_d[1 - f_d(E_d)] \\ &= \frac{N_d}{1 + 2\exp\left[\dfrac{(E_F - E_d)}{kT}\right]} \end{aligned} \tag{4}$$

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

Ionization region

$$n = \left(\frac{1}{2}N_c N_d\right)^{1/2} \exp\left[-\frac{(E_c - E_d)}{2kT}\right]$$

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

$$\mu = \frac{e\tau}{m_e^*} \qquad [5.21]$$

in which

$$\tau = \frac{1}{Sv_{\text{th}}N_s} \qquad [5.22]$$

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,

$$\frac{1}{2}m_e^* v_{\text{th}}^2 = \frac{3}{2}kT$$

so that $v_{\text{th}} \propto T^{1/2}$. Thus the mean time τ_L between scattering events from lattice vibrations is[5]

$$\tau_L = \frac{1}{(\pi a^2)v_{\text{th}}N_s} \propto \frac{1}{(T)(T^{1/2})} \propto T^{-3/2}$$

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

[5] The present arguments are totally classical whereas in terms of modern physics, the electrons are scattered by phonons and the phonon concentration increases with temperature. An analogy may help. The light intensity classically depends on E^2 whereas in quantum physics it is given by the photon flux density.

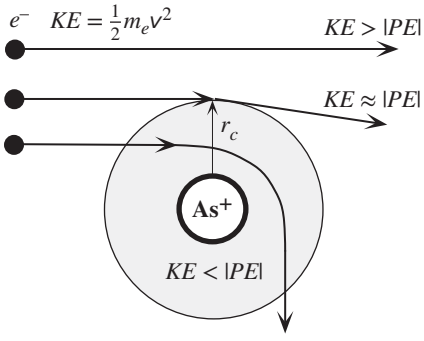


Figure 5.17 Scattering of electrons by an ionized impurity.

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 the 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

$$\tau_I = \frac{1}{Sv_{th}N_I} \propto \frac{1}{(T^{-2})(T^{1/2})N_I} \propto \frac{T^{3/2}}{N_I}$$

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

$$\mu_I \propto \frac{T^{3/2}}{N_I} \qquad [5.24]$$

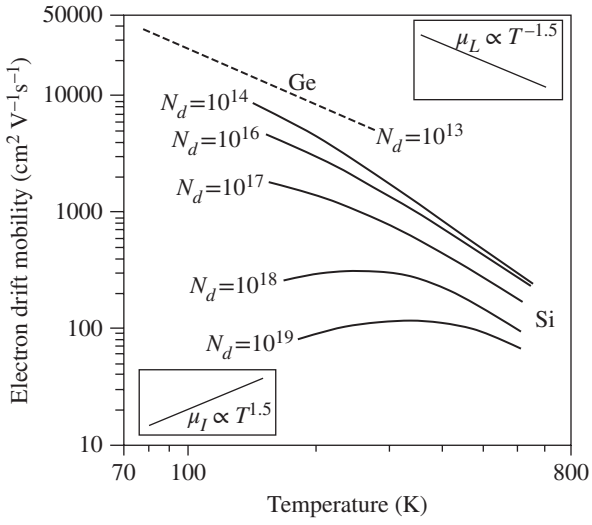


Figure 5.18 Log–log plot of drift mobility versus temperature for *n*-type Ge and *n*-type Si samples. Various donor concentrations for Si are shown. N_d are in cm^{-3}. The upper right inset is the simple theory for lattice limited mobility, whereas the lower left inset is the simple theory for impurity scattering limited mobility.

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.*, $N_d = 10^{13}$ 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) at high temperatures, however, exhibit a $\mu_e \propto T^{-2.3}$ 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.

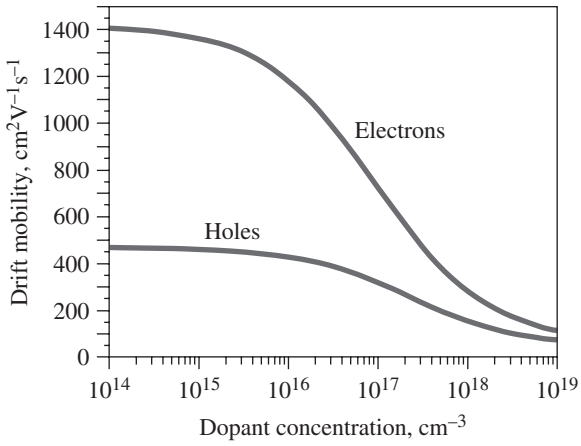


Figure 5.19 The variation of the drift mobility with dopant concentration in Si for electrons and holes at 300 K.

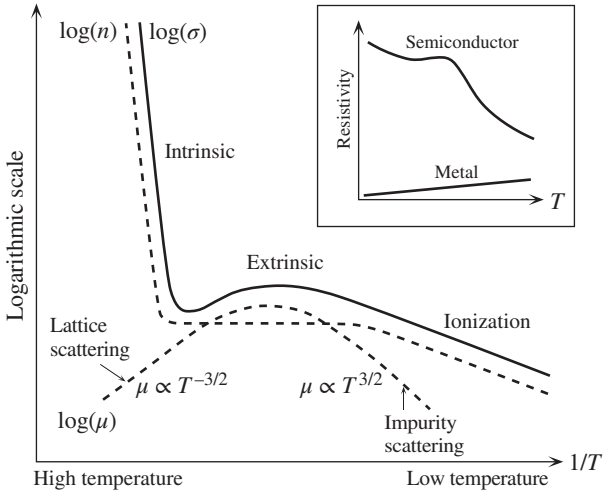


Figure 5.20 Schematic illustration of the temperature dependence of electrical conductivity for a doped (*n*-type) semiconductor.

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

$$n = \left(\frac{1}{2}N_cN_d\right)^{1/2} \exp\left[-\frac{(E_c - E_d)}{2kT}\right]$$

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.

EXAMPLE 5.10 **COMPENSATION-DOPED Si**

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 approximately 700 cm^2 V^{-1} s^{-1} from the drift mobility versus dopant concentration graph in Figure 5.19, so

$$\begin{aligned}\sigma &= en\mu_e + ep\mu_h = eN_d\mu_e \\ &= (1.6 \times 10^{-19}\ \text{C})(10^{17}\ \text{cm}^{-3})(700\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}) = 11.2\ \Omega^{-1}\ \text{cm}^{-1}\end{aligned}$$

At $T = 127$ °C = 400 K from the μ_e vs. T graph in Figure 5.18,

$$\mu_e \approx 450\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$$

Thus,

$$\sigma = eN_d\mu_e = 7.20\ \Omega^{-1}\ \text{cm}^{-1}$$

b. With further doping we have $N_a = 9 \times 10^{16}$ cm^{-3}, so from the compensation effect

$$N_d - N_a = 1 \times 10^{17} - 9 \times 10^{16} = 10^{16}\ \text{cm}^{-3}$$

Since $N_d - N_a \gg n_i$, we still have an *n*-type material with $n = N_d - N_a = 10^{16}$ cm^{-3}. But the drift mobility now is about ~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 $10^{17} + 9 \times 10^{16}$ $cm^{-3} = 1.9 \times 10^{17}$ cm^{-3}, which gives an electron drift mobility of ~600 cm^2 V^{-1} s^{-1} at 300 K (Figure 5.19) and ~400 cm^2 V^{-1} s^{-1} at 400 K (Figure 5.18). Then, neglecting the hole concentration $p = n_i^2/(N_d - N_a)$, we have

At 300 K, $$\begin{aligned}\sigma &= e(N_d - N_a)\mu_e \approx (1.6 \times 10^{-19}\ \text{C})(10^{16}\ \text{cm}^{-3})(600\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}) \\ &= 0.96\ \Omega^{-1}\ \text{cm}^{-1}\end{aligned}$$

At 400 K, $$\begin{aligned}\sigma &= e(N_d - N_a)\mu_e \approx (1.6 \times 10^{-19}\ \text{C})(10^{16}\ \text{cm}^{-3})(400\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}) \\ &= 0.64\ \Omega^{-1}\ \text{cm}^{-1}\end{aligned}$$

EXAMPLE 5.11

COMPENSATION DOPING IN Si Consider a *p*-type Si crystal that has been doped uniformly with B with a concentration of 10^{15} cm^{-3}. We wish to convert this to an *n*-Si with a conductivity 20 Ω^{-1} cm^{-1} within 10 percent. What is the donor concentration we need?

SOLUTION

The starting *p*-Si has $N_a = 10^{15}$ cm^{-3} which means that, Figure 5.19, $\mu_e \approx 1350$ cm^2 V^{-1} s^{-1}. Upon compensation doping, we would have *n*-Si in which the electron concentration $n = N_d - N_a$ and the conductivity is

$$\sigma = en\mu_e = e(N_d - N_a)\mu_e = 20\ \Omega^{-1}\ \text{cm}^{-1}$$

We know N_a, and if μ_e was independent of doping, we could readily solve this for N_d. However, as shown in Figure 5.19, μ_e depends on the ionized dopant concentration, $N_d + N_a$. We start by first using $\mu_e = 1350$ cm^2 V^{-1} s^{-1} in the starting crystal so that

$$\sigma = (1.60 \times 10^{-19})(N_d - 10^{15}\ \text{cm}^{-3})(1350\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}) = 20\ \Omega^{-1}\ \text{cm}^{-1}$$

which we can solve and find $N_d = 9.36 \times 10^{16}$ cm^{-3}. This is almost two order of magnitude larger than N_a so we may as well neglect N_a in the conductivity equation. The ionized dopant concentration, $N_d + N_a$ is also approximately N_d and at this N_d, from Figure 5.19, $\mu_e' \approx 750$ cm^2 V^{-1} s^{-1}. Therefore, the actual conductivity σ' is $(750/1350)\sigma$ or 11.1 Ω^{-1} cm^{-1}, roughly half of what we need. We can improve our calculation by using the new mobility μ_e'. So we can now write σ with this new mobility μ_e' as

$$\sigma = eN_d\mu_e' = 1.60 \times 10^{-19} \times N_d \times 750\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1} = 20\ \Omega^{-1}\ \text{cm}^{-1}$$

and solving this we find $N_d = 1.7 \times 10^{17}$ cm^{-3}. From Figure 5.19, at $N_d = 1.7 \times 10^{17}$ cm^{-3}, $\mu_e'' \approx 600$ cm^2 V^{-1} s^{-1}. The actual conductivity σ'' is $(600/750)\sigma$ or 16 Ω^{-1} cm^{-1}. Obviously, we are getting closer to 20 Ω^{-1} cm^{-1}. The next iteration will have

$$\sigma = eN_d\mu_e'' = 1.60 \times 10^{-19} \times N_d \times 600\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1} = 20\ \Omega^{-1}\ \text{cm}^{-1}$$

which upon solving gives $N_d = 2.1 \times 10^{17}$ cm^{-3}. At this donor concentration the mobility $\mu_e''' \approx 550$ cm^2 which yields a conductivity of 18.5 Ω^{-1} cm^{-1}, within 10 percent of our target 20 Ω^{-1} cm^{-1}.

Clearly in $\sigma = eN_d\mu_e$, the drift mobility μ_e depends on N_d as in Figure 5.19, so the solution for N_d above took a tedious number of iterative calculations and look-ups in Figure 5.19. We can always represent the μ_e versus N_d curve with an empirical equation $\mu_e(N_d)$ and then solve $\sigma = eN_d\mu_e(N_d)$ numerically; an approach taken in Question 5.7.

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} 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 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.

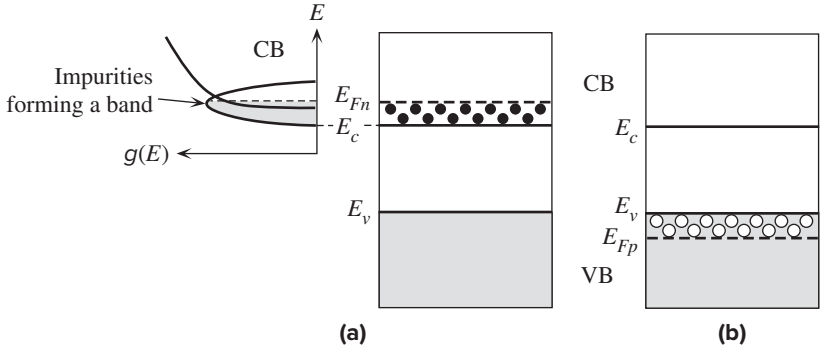


Figure 5.21 (a) Degenerate n-type semiconductor. Large number of donors form a band that overlaps the CB. (b) Degenerate p-type semiconductor.

EXAMPLE 5.12

DEGENERATE *n*-TYPE Si Consider a degenerate *n*-type Si crystal in which the donor concentration (*e.g.*, P) is 10^{20} cm^{-3} (or 10^{26} m^{-3}). Where is the Fermi level with respect to the bottom (E_c) of the CB, that is $E_{Fn} - E_c$? What is your conclusion?

SOLUTION

Clearly, $N_d > N_c$, and if we attempt to use Equation 5.6, that is we assume Boltzman statistics, then

$$\Delta E_{Fn} = E_{Fn} - E_c = kT \ln(N_d/N_c) = (0.02585 \text{ eV})\ln[(10^{20})/(2.8 \times 10^{19})] = 0.033 \text{ eV}$$

Remember that Boltzman statistics inherent in Equation 5.6 does not obey the Pauli exclusion principle; no two electrons can be in the same state (same wavefunction, including spin). When E_{Fn} is within the CB, electrons need follow the Pauli exclusion principle and look for higher energy states to avoid sharing the same state. So, we expect E_{Fn} to be greater than 0.033 eV. The electron concentration in the CB is given by the integration of the product of density of states $g_{cb}(E)$ and the Fermi–Dirac function $f(E)$,

Degenerate n-type semiconductor

$$n = N_d = \int_0^\infty \frac{g_{cb}(E)dE}{1 + \exp[(E - E_{Fn})/kT]}$$

This is the same procedure we used in the case of metals in Chapter 4 to relate the Fermi energy to the electron concentration. Recall that the Fermi energy $\Delta E_{Fn}(0)$ at absolute zero with respect to the bottom of the band is given by Equation 4.22

$$\Delta E_{Fn}(0) = \frac{h^2}{8m_e^*}\left(\frac{3n}{\pi}\right)^{2/3} = \frac{(6.626 \times 10^{-34})^2}{8(1.08 \times 9.11 \times 10^{-31})}\left[\frac{3(1 \times 10^{26})}{\pi}\right]^{2/3} = 0.0727 \text{ eV}$$

Notice that we used the effective mass m_e^* related to the density of states. While, as expected, this is larger than that from Boltzman statistics, it is still not correct because it is at 0 K. At a finite temperature T, we argued that for metals $E_{Fm}(0) \gg kT$ and the Fermi energy from the above integration approximates to Equation 4.23

$$\Delta E_{Fn} = \Delta E_{Fn}(0)\left(1 - \frac{\pi^2}{12}\left[\frac{kT}{\Delta E_{Fn}(0)}\right]^2\right) = (0.0727)\left(1 - \frac{\pi^2}{12}\left[\frac{0.02585}{0.0727}\right]^2\right) = 0.0652 \text{ eV}$$

or 65 meV above E_c. We can, of course, find E_{Fn} by trial and error until the above integration generates $n = N_d$. The final result is very close to 65 meV.[6] It is clear that the description of degenerate semiconductors follows the same concepts we used in the case of metals.

5.4 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,

[6] The Joyce–Dixon equation that is used in advanced semiconductor textbooks allows a good approximation to ΔE_{Fn} and gives 66 meV.

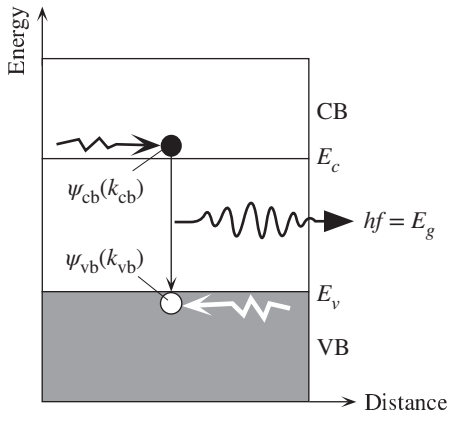


Figure 5.22 Direct recombination in GaAs. $k_{cb} = k_{vb}$ so that momentum conservation is satisfied.

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 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 $hf = E_g$. 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, ψ_{cb}, is a traveling wave with a certain wavevector k_{cb}. The actual electron wavefunctions are discussed later in this chapter, but for now we simply accept the fact if we were to solve the Schrodinger equation for an electron in a crystal in which the electron potential energy $V(x)$ is periodic (due to a periodic arrangement of atoms), we would find traveling wave solutions. For example, the electron wavefunctions ψ_{cb} in the CB will be traveling waves each with an energy E and a wavevector k_{cb}. The quantity, $\hbar k_{cb}$, just as in the case of a photon, can be used to represent the momentum of the electron in the CB. In fact, in response to an external force F_{ext}, the electron's momentum $\hbar k_{cb}$ will change according to $F_{ext} = d(\hbar k_{cb})/dt$, exactly as we expect a momentum to change in mechanics. The quantity $\hbar k_{cb}$ is called the **electron's crystal momentum** because it represents the momentum that we need in describing the behavior of the electron inside the crystal in response to an external force.[7] Similarly, the electron wavefunction, ψ_{vb} in the VB will have a momentum $\hbar k_{cb}$ associated with it.

[7] The rate of change of electron's true momentum would be due to external and internal forces summed together. However, this is not a useful approach inasmuch we would like to know the effect of external forces on the behavior of the electron. We can account for the internal forces by using a periodic potential energy in the Schrodinger equation, and once we have done this, $\hbar k$ turns out to be a useful momentum quantity that follows our usual experience that external force (F_{ext}) is $d(\hbar k)/dt$.

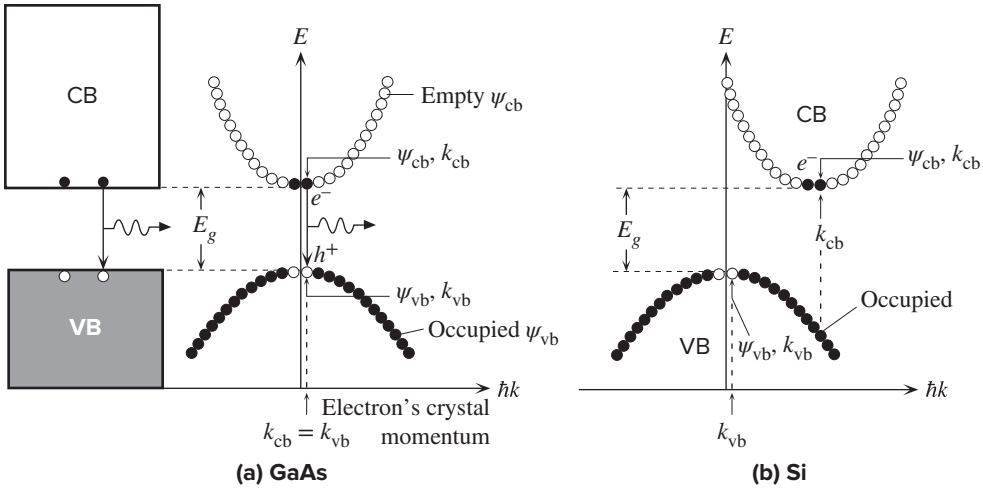


Figure 5.23 (a) The electron energy (E) versus electron's crystal momentum ($\hbar k$) in a direct bandgap semiconductor. Each circle represents a possible state, an electron wavefunction (ψ), a solution of Schrödinger's equation in a crystal, with a wavevector k. These solutions fall either into the CB or the VB; there are no solutions within the bandgap. The sketches are highly exaggerated because the circles are so close that they form a continuous energy versus momentum behavior. (b) Energy versus crystal momentum for an indirect bandgap semiconductor such as Si.

If we were to plot the energy E of each ψ_{cb} against $\hbar k_{cb}$ for the CB wavefunctions, we would find the E versus $\hbar k$ behavior shown in Figure 5.23a. Each circle is a wavefuncion ψ_{cb} with an energy E and wavevector k_{cb}. The circles represent electron states. These are normally so close to each other that they form a continuum; Figure 5.23a is highly exaggerated. Notice that E increases parabolically with $\hbar k_{cb}$ near the bottom of the CB, as we would expect classically from $E = p^2/(2m_e^*)$, where p is electron's momentum. Similar arguments, of course, apply to the VB, and we can plot E versus $\hbar k_{vb}$ as well in this case as shown in Figure 5.23a. The hole energy increases downwards (in the opposite direction to the electron energy), so that the hole energy near the top of the VB also shows a parabolic behavior with momentum, that is, $E_{hole} = p^2/(2m_h^*)$, where p the hole momentum and m_h^* is the hole effective mass.

Conservation of linear momentum during recombination requires that when the electron drops from the CB to the VB, its wavevector should remain the same, $k_{vb} = k_{cb}$, because the momentum carried away by the photon is negligibly small. This is indeed the case for GaAs whose E versus $\hbar k$ behavior follows that shown in Figure 5.23a. Such semiconductors are called **direct bandgap semiconductors.** The top of the valence band is immediately below the bottom of the CB on the E versus $\hbar k$ diagram as in Figure 5.23a. Thus, for direct bandgap semiconductors, such as GaAs, the states with $k_{vb} = k_{cb}$ are right at the top of the valence band where there are many empty states (*i.e.,* 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.

For the elemental semiconductors, Si and Ge, the electron energy versus crystal momentum (E vs. $\hbar k$) behavior is such that the bottom of the CB is displaced with respect to the top of the VB in terms of $\hbar k$ as shown in Figure 5.23b. Such semiconductors are called **indirect bandgap semiconductors.** Those states (ψ_{vb}) with $k_{vb} = k_{cb}$ are now somewhere in the middle of the VB and they are therefore fully occupied as shown in Figure 5.23b. Consequently, there are no empty states in the VB which can satisfy $k_{vb} = k_{cb}$ and so direct recombination in Si and Ge is next to impossible.

In elemental indirect bandgap semiconductors such as 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 5.24a, 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 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.

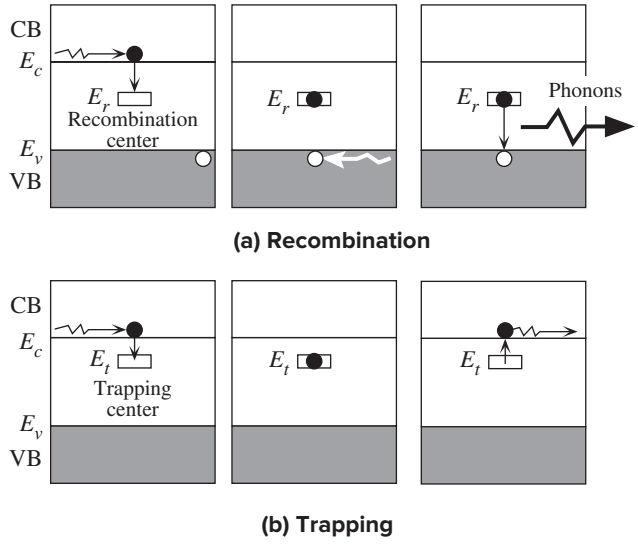


Figure 5.24 Recombination and trapping. (a) Recombination in Si via a recombination center that has a localized energy level at E_r in the bandgap, usually near the middle. (b) Trapping and detrapping of electrons by trapping centers. A trapping center has a localized energy level in the bandgap.

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 5.24b. 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 5.24b 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.

5.5 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 5.25. 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.

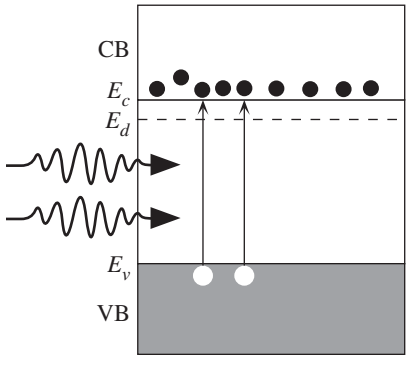


Figure 5.25 Low-level photoinjection into an n-type semiconductor in which $\Delta n_n < n_{no}$.

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

$$n_n = n_{no} + \Delta n_n \qquad \text{and} \qquad p_n = p_{no} + \Delta p_n$$

Photoexcitation creates EHPs or an equal number of electrons and holes, as shown in Figure 5.25, 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

$$\frac{dn_n}{dt} = \frac{d\Delta n_n}{dt} \qquad \text{and} \qquad \frac{dp_n}{dt} = \frac{d\Delta p_n}{dt}$$

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 5.26 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 5.27 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

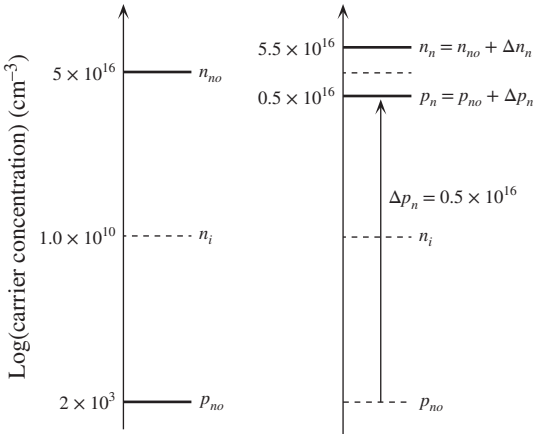


Figure 5.26 Low-level injection in an *n*-type semiconductor does not significantly affect n_n but drastically affects the minority carrier concentration p_n.

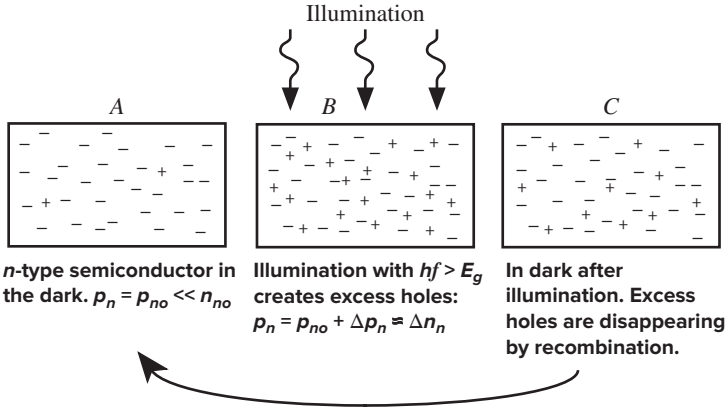


Figure 5.27 Illumination of an *n*-type semiconductor results in excess electron and hole concentrations.

After the illumination, the recombination process restores equilibrium; the excess electrons and holes simply recombine.

again. Obviously when the light is switched off, the condition $p_n = \Delta p_n$ (state B in Figure 5.27) 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 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

$$\frac{d\Delta p_n}{dt} = G_{\text{ph}} - \frac{\Delta p_n}{\tau_h} \qquad [5.27]$$

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 5.13 **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 $t = t_{\text{off}} (\gg \tau_h)$.

SOLUTION

We use Equation 5.27 with G_{ph} = constant in $0 \le t \le t_{\text{off}}$. Since Equation 5.27 is a first-order differential equation, integrating it we simply find

$$\ln\left[G_{\text{ph}} - \left(\frac{\Delta p_n}{\tau_h}\right)\right] = -\frac{t}{\tau_h} + C_1$$

where C_1 is the integration constant. At $t = 0$, $\Delta p_n = 0$, so $C_1 = \ln G_{\text{ph}}$. Therefore, the solution is

$$\Delta p_n(t) = \tau_h G_{\text{ph}}\left[1 - \exp\left(-\frac{t}{\tau_h}\right)\right] \qquad 0 \le t < t_{\text{off}} \qquad [5.28]$$

We see that as soon as the illumination is turned on, the minority carrier concentration rises exponentially toward its steady-state value $\Delta p_n(\infty) = \tau_h G_{\text{ph}}$. This is reached after a time $t > \tau_h$.

At the instant the illumination is switched off, we assume that $t_{\text{off}} \gg \tau_h$ so that from Equation 5.28,

$$\Delta p_n(t_{\text{off}}) = \tau_h G_{\text{ph}}$$

We can define t' to be the time measured from $t = t_{\text{off}}$, that is, $t' = t - t_{\text{off}}$. Then

$$\Delta p_n(t' = 0) = \tau_h G_{\text{ph}}$$

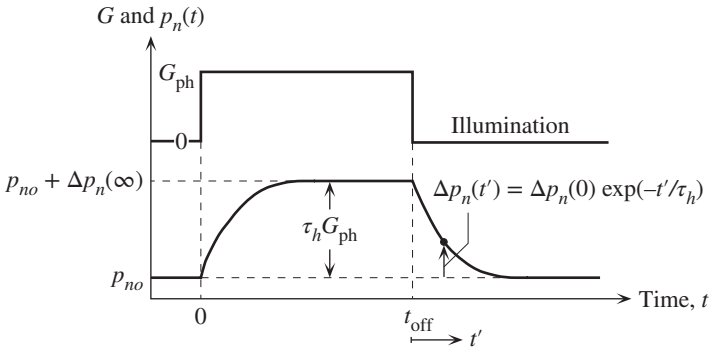


Figure 5.28 Illumination is switched on at time $t = 0$ and then off at $t = t_{off}$. The excess minority carrier concentration $\Delta p_n(t)$ rises exponentially to its steady-state value with a time constant τ_h. From t_{off}, the excess minority carrier concentration decays exponentially to its equilibrium value.

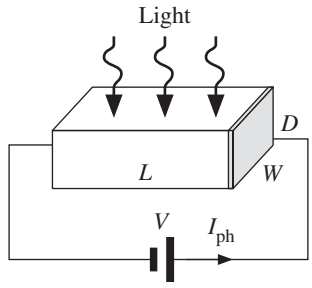


Figure 5.29 A semiconductor slab of length L, width W, and depth D is illuminated with light of wavelength λ. I_{ph} is the steady-state photocurrent.

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

$$\Delta p_n(t') = \tau_h G_{ph} \exp\left(-\frac{t'}{\tau_h}\right) \quad [5.29]$$

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 5.28.

EXAMPLE 5.14

PHOTOCONDUCTIVITY Suppose that a direct bandgap semiconductor with no traps is illuminated with light of intensity $\mathcal{I}(\lambda)$ and wavelength λ that will cause photogeneration as shown in Figure 5.29. The area of illumination is $A = (L \times W)$, and the thickness (depth) of the semiconductor is D. Assume that all incident photons are absorbed. 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

$$\Delta\sigma = \frac{e\eta\mathcal{I}\lambda\tau(\mu_e + \mu_h)}{hcD} \quad [5.30]$$

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, assuming that all the incident light is absorbed

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 density), then $\Gamma_{ph} = I/hf$ where I is the light intensity (energy flowing per unit area per second) and hf is the photon energy. 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

$$G_{ph} = \frac{\eta A \Gamma_{ph}}{AD} = \frac{\eta\left(\frac{I}{hf}\right)}{D} = \frac{\eta I \lambda}{hcD}$$

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, the steady-state photoconductivity,

$$\Delta\sigma = e\mu_e \Delta n + e\mu_h \Delta p = e\Delta n(\mu_e + \mu_h)$$

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,

$$\Delta\sigma = \frac{e\eta\mathcal{I}\lambda\tau(\mu_e + \mu_h)}{hcD}$$

That is

$$\Delta\sigma = \frac{(1.6 \times 10^{-19}\ \text{C})(1)(10^{-3} \times 10^{4}\ \text{J s}^{-1}\ \text{m}^{-2})(450 \times 10^{-9}\ \text{m})(1 \times 10^{-3}\ \text{s})(0.0358\ \text{m}^2\ \text{V}^{-1}\ \text{s}^{-1})}{(6.63 \times 10^{-34}\ \text{J s})(3 \times 10^{8}\ \text{m s}^{-1})(0.1 \times 10^{-3}\ \text{m})}$$

$$= 1.30\ \Omega^{-1}\ \text{m}^{-1}$$

c. Photocurrent density will be

$$\Delta J = E\Delta\sigma = (1.30\ \Omega^{-1}\ \text{m}^{-1})(50\ \text{V}/10^{-3}\ \text{m}) = 6.50 \times 10^{4}\ \text{A m}^{-2}$$

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. If this is not the case, the photoconductivity and hence the photocurrent will be smaller. Further we assumed that the photogeneration of carriers is uniform over the area LW and along the thickness D. Usually photogeneration along D is not uniform.

5.6 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 density** Γ just like current density, 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,

Definition of particle flux density

$$\Gamma = \frac{\Delta N}{A\ \Delta t} \qquad [5.31]$$

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 density, is related to the particle flux density 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 5.30a. 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

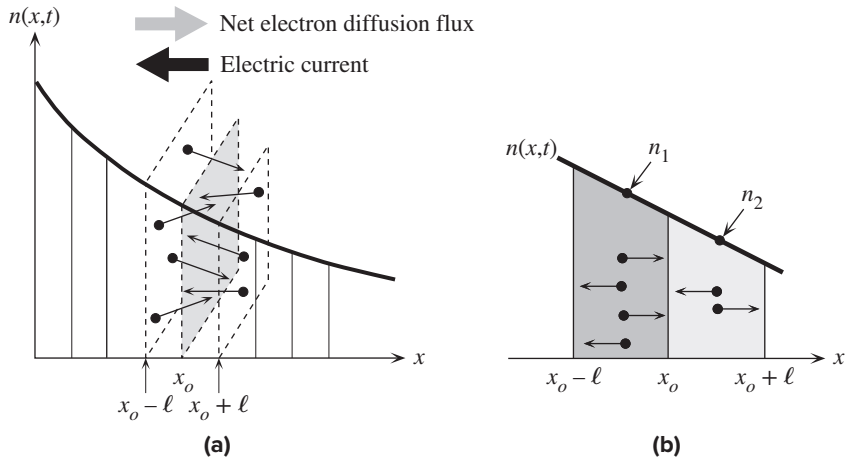


Figure 5.30 (a) Arbitrary electron concentration $n(x, t)$ profile in a semiconductor. There is a net diffusion (flux) of electrons from higher to lower concentrations. (b) Expanded view of two adjacent sections at x_o. There are more electrons crossing x_o coming from the left ($x_o - \ell$) than coming from the right ($x_o + \ell$).

so that the diffusion of electrons can be simplified to a one-dimensional problem as depicted in Figure 5.30a. 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 5.30b. 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 cross x_o is $\frac{1}{2}n_1A\ell$ 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 $\frac{1}{2}n_2A\ell$ 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 density Γ_e,

$$\Gamma_e = \frac{\frac{1}{2}n_1A\ell - \frac{1}{2}n_2A\ell}{A\tau}$$

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 density 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 density Γ_e at a position x is proportional to the concentration gradient and the diffusion coefficient. The steeper this gradient, the larger the flux density 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: $J = \sigma E = -\sigma(dV/dx)$.

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 density Γ_e at a position x where the electron concentration gradient is dn/dx. Since from Figure 5.30, 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

$$J_{D,e} = -e\Gamma_e = eD_e\frac{dn}{dx} \qquad [5.35]$$

Electron diffusion current density

In the case of a hole concentration gradient, as shown in Figure 5.31, 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 5.31, 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

$$J_{D,h} = e\Gamma_h = -eD_h\frac{dp}{dx} \qquad [5.36]$$

Hole diffusion current density

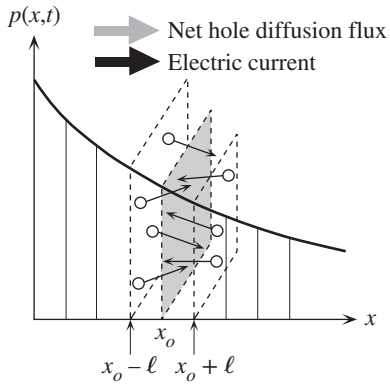


Figure 5.31 Arbitrary hole concentration $p(x, t)$ profile in a semiconductor.

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 ($x_o + \ell$).

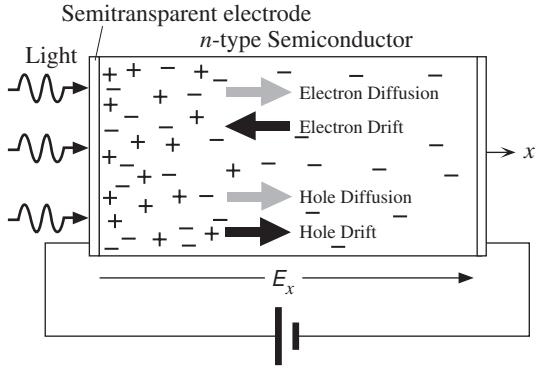


Figure 5.32 When there is an electric field and also a concentration gradient, charge carriers move both by diffusion and drift.

Suppose that there is also a positive electric field E_x acting along $+x$ in Figures 5.30 and 5.31. A practical example is shown in Figure 5.32 in which a semiconductor is sandwiched between two electrodes, the left one semitransparent. By connecting a battery to the electrodes, an applied field of 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 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

$$J_e = en\mu_e E_x + eD_e \frac{dn}{dx} \qquad [5.37]$$

We note that as 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.

Total hole current due to drift and diffusion

Similarly, the hole current due to holes drifting and diffusing, Equation 5.36, is given by

$$J_h = ep\mu_h E_x - eD_h \frac{dp}{dx} \qquad [5.38]$$

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,**

$$\frac{D_e}{\mu_e} = \frac{kT}{e} \qquad \text{and} \qquad \frac{D_h}{\mu_h} = \frac{kT}{e} \tag{5.39}$$

Einstein relation

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 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 5.15.

We equated the diffusion coefficient D to $\ell^2/2\tau$ in Equation 5.34. Our analysis, as represented in Figure 5.30, 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

$$D = \frac{\ell^2}{\tau} \tag{5.40}$$

Diffusion coefficient

EXAMPLE 5.15

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. (A more rigorous derivation can be found in Question 5.24.)

EXAMPLE 5.16 **DIFFUSION COEFFICIENT OF ELECTRONS IN Si** Calculate the diffusion coefficient of electrons at 27 °C in n-type Si doped with 10^{16} As atoms cm^{-3}.

SOLUTION

From the μ_e versus dopant concentration graph in Figure 5.19, the electron drift mobility μ_e at a donor concentration of 10^{16} cm^{-3} is about 1200 cm^2 V^{-1} s^{-1}, so

$$D_e = \frac{\mu_e kT}{e} = (1200 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1})(0.0259 \text{ V}) = 31.08 \text{ cm}^2 \text{ s}^{-1}$$

EXAMPLE 5.17 **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 semiconductor where the electron concentrations are n_1 and n_2? If the donor profile in an n-type semiconductor is $N_d(x) = N_{do} \exp(-x/b)$, where b is a characteristic of the exponential doping profile, evaluate the built-in field E_x. What is your conclusion?

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 5.33.

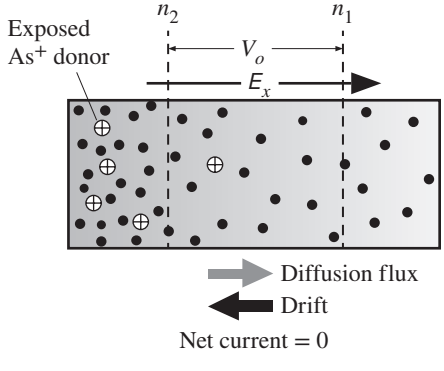


Figure 5.33 Nonuniform doping profile results in electron diffusion toward the less concentrated regions.

This exposes positively charged donors and sets up a built-in field E_x. In the steady state, the diffusion of electrons toward the right is balanced by their drift toward the left.

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),

$$J_e = en\mu_e E_x + eD_e \frac{dn}{dx} = 0$$

But the field is related to the potential difference by $E_x = -(dV/dx)$, so

$$-en\mu_e \frac{dV}{dx} + eD_e \frac{dn}{dx} = 0$$

We can now use the Einstein relation $De/\mu_e = kT/e$ 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

$$V_2 - V_1 = \frac{kT}{e} \ln\left(\frac{n_2}{n_1}\right) \quad [5.41]$$

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

$$n(x) \approx N_d(x) = N_o \exp\left(-\frac{x}{b}\right)$$

Substituting into the equation for $J_e = 0$, and again using the Einstein relation, we obtain E_x as

Built-in field

$$E_x = \frac{kT}{be} \quad [5.42]$$

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 E_x in Equation 5.42 acts to drift minority carriers faster and therefore speeds up the transistor operation as discussed in Chapter 6.

5.7 CONTINUITY EQUATION[8]

5.7.1 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 within a small volume

[8] This section may be skipped without loss of continuity. (No pun intended.)

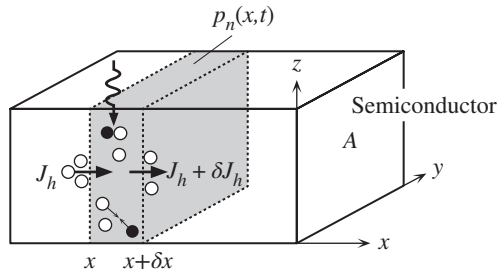


Figure 5.34 Consider an elemental volume $A\ \delta x$ in which the instantaneous hole concentration is $p(x, t)$. The electric current flow and hole drift are in the same direction.

at that location in the semiconductor. Consider an *n*-type semiconductor slab as shown in Figure 5.34 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 5.34 in which the instantaneous 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 *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,

$$\frac{1}{A\ \delta x}\left(\frac{-A\ \delta J_h}{e}\right) = \begin{array}{c}\text{Rate of increase in hole concentration} \\ \text{due to the change in } J_h\end{array} \qquad [5.43]$$

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

Continuity equation for holes

$$\frac{\partial p_n}{\partial t} = -\frac{1}{e}\left(\frac{\partial J_h}{\partial x}\right) - \frac{p_n - p_{no}}{\tau_h} + G_{\text{ph}} \qquad [5.44]$$

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 Equation 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 because the electron charge is negative. Put differently, the electron flow is in the opposite direction to the conventional current flow. (The decrease in the current density actually decreases the electron concentration in $A\delta x$.)

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 5.29. 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

Continuity equation with uniform photo-generation

$$\frac{\partial \Delta p_n}{\partial t} = -\frac{\Delta p_n}{\tau_h} + G_{\text{ph}} \qquad [5.45]$$

which is identical to the semiquantitatively derived Equation 5.27 from which photoconductivity was calculated in Example 5.14.

5.7.2 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 5.35a.[9] There is no bulk photogeneration, so $G_{\text{ph}} = 0$. 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

$$\frac{1}{e}\left(\frac{\partial J_h}{\partial x}\right) = -\frac{p_n - p_{no}}{\tau_h} \qquad [5.46]$$

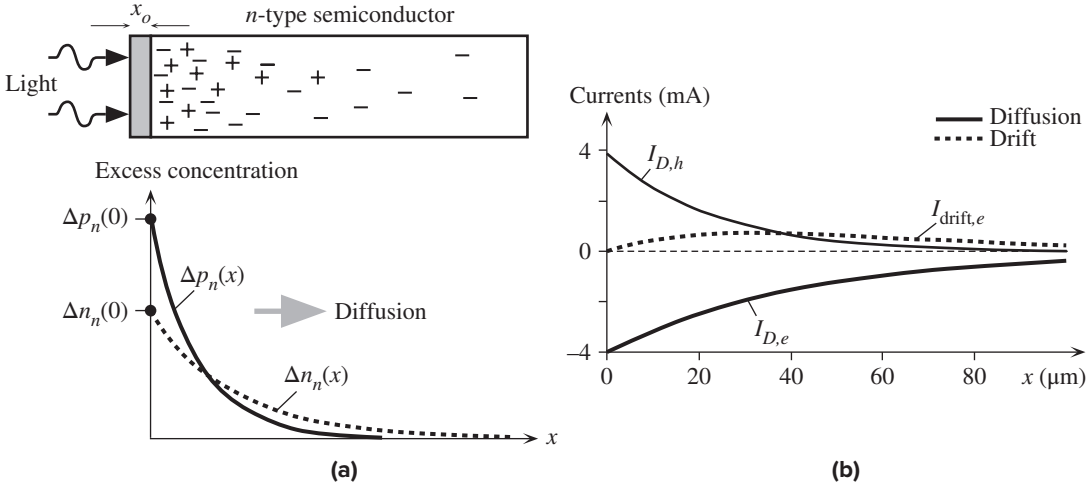


Figure 5.35 (a) Steady-state excess carrier concentration profiles in an *n*-type semiconductor that is continuously illuminated at one end. (b) Majority and minority carrier current components in open circuit. Total current is zero.

[9] One can take x_o to be very roughly the absorption depth of the incident light in the semiconductor. For simplicity, we will assume uniform photogeneration within x_o.

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 $E \approx 0$ 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 $E = 0$

$$\frac{d^2 \Delta p_n}{dx^2} = \frac{\Delta p_n}{L_h^2} \qquad [5.47]$$

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 5.35a, both excess electrons and holes are photogenerated at the surface, but the percentage increase in the concentration of holes is much more dramatic since $p_{no} \ll n_{no}$. 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,

$$Ax_o G_{\text{ph}} = \frac{1}{e} I_{D,h}(0) = \frac{AD_h}{L_h} \Delta p_n(0)$$

or

$$\Delta p_n(0) = x_o G_{\text{ph}} \left(\frac{\tau_h}{D_h}\right)^{1/2} \qquad [5.50]$$

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

$$I_{D,e} = AeD_e \frac{dn_n(x)}{dx} = -\frac{AeD_e}{L_e} \Delta n_n(0) \exp\left(-\frac{x}{L_e}\right) \quad [5.52]$$

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,

$$\Delta n_n(0) = x_o G_{\text{ph}} \left(\frac{\tau_h}{D_e}\right)^{1/2} \quad [5.53]$$

so that

$$\frac{\Delta p_n(0)}{\Delta n_n(0)} = \left(\frac{D_e}{D_h}\right)^{1/2} \quad [5.54]$$

which is greater than unity for Si.

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 $I_{\text{drift},e}$ is the electron drift current, then in an open circuit the total current $I = I_{D,h} + I_{D,e} + I_{\text{drift},e} = 0$, so

Electron drift current

$$I_{\text{drift},e} = -I_{D,h} - I_{D,e} \quad [5.55]$$

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 E in the sample, however small, to provide the necessary drift to balance the currents to zero. The field can be found from $I_{\text{drift},e} \approx Aen_{no}\mu_e E$, inasmuch as n_{no} does not change significantly (weak injection),

Electric field

$$E = \frac{I_{\text{drift},e}}{Aen_{no}\mu_e} \quad [5.56]$$

The hole drift current due to this field is

Hole drift current

$$I_{\text{drift},h} = Ae\mu_h p_n(x) E \quad [5.57]$$

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 $\Delta p_n(0) = 0.05 n_{no} = 5 \times 10^{14}$ 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

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 $I_{\text{drift},h}$ (mA)	Minority Diffusion $I_{D,e}$ (mA)	Majority Drift $I_{\text{drift},e}$ (mA)	Field 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

above as shown in Figure 5.35b. The actual values at two locations, $x = 0$ and $x = L_e = 41$ μm, are shown in Table 5.3.[10] The photoinduced charge separation and hence the generation of a potential difference as in Figure 5.35 is called the **photo-Dember effect.**

EXAMPLE 5.18 **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 5.35. 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 5.35a. To find the mean position of the photoinjected holes, we use the definition of the "mean," that is,

$$\bar{x} = \frac{\int_0^\infty x\, \Delta p_n(x)\, dx}{\int_0^\infty \Delta p_n(x)\, dx}$$

Substituting for $\Delta p_n(x)$ from Equation 5.59 and carrying out the integration gives $\bar{x} = L_h$. 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.

[10] Remember that 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.

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.

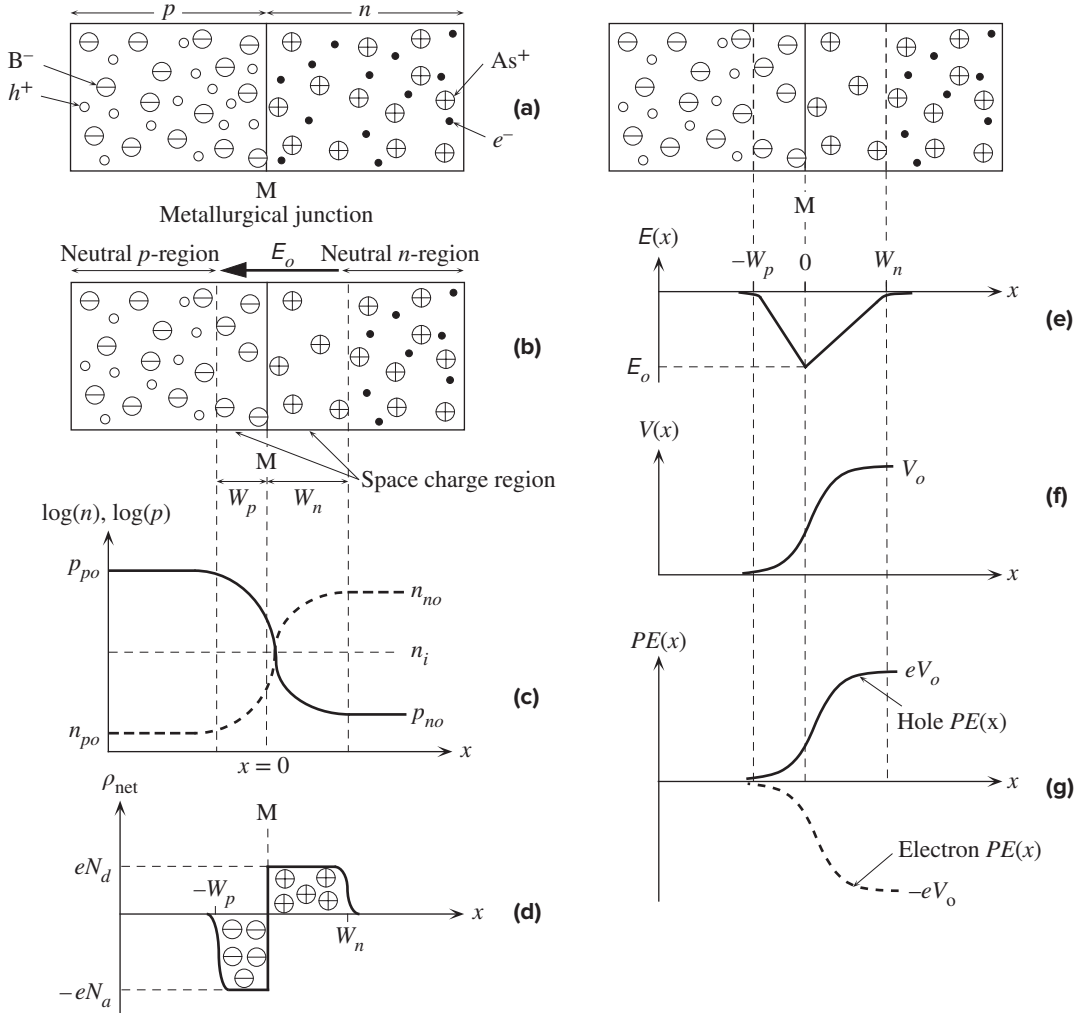


Figure 6.1 Properties of the *pn* junction.

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 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. Notice that the depletion region in Figure 6.1c has been depleted of its normal concentration of carriers, which exposes the donor and acceptor ions. The carrier concentrations are not zero. The depletion region is also called the **depletion layer** or, less commonly, the transition region.

It is clear that there is an internal electric field 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, 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 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 $\rho_{\text{net}}(x)$ 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 $E(x)$ and the net space charge density $\rho_{net}(x)$ at a point are related in electrostatics[1] by

Field and net space charge density

$$\frac{dE}{dx} = \frac{\rho_{net}(x)}{\varepsilon}$$

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 $E(x)$, that is,

Field in depletion region

$$E(x) = \frac{1}{\varepsilon}\int_{-W_p}^{x} \rho_{net}(x)\, dx \qquad [6.2]$$

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 $E(x)$ reaches a maximum value 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 $E = -dV/dx$. 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

$$E_o = -\frac{eN_d W_n}{\varepsilon} = -\frac{eN_a W_p}{\varepsilon} \qquad [6.3]$$

where $\varepsilon = \varepsilon_o \varepsilon_r$. We can integrate the expression for $E(x)$ 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

$$V_o = -\frac{1}{2}E_o W_o = \frac{eN_a N_d W_o^2}{2\varepsilon(N_a + N_d)} \qquad [6.4]$$

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,

[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.

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 $E = -eV_o$ on the n-side away from M where $n = n_{no}$. Thus

$$\frac{n_{po}}{n_{no}} = \exp\left(-\frac{eV_o}{kT}\right) \quad [6.5a]$$

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

$$\frac{p_{no}}{p_{po}} = \exp\left(-\frac{eV_o}{kT}\right) \quad [6.5b]$$

Thus, rearranging Equations 6.5a and b we obtain

$$V_o = \frac{kT}{e}\ln\left(\frac{n_{no}}{n_{po}}\right) \qquad \text{and} \qquad V_o = \frac{kT}{e}\ln\left(\frac{p_{po}}{p_{no}}\right)$$

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

$$V_o = \frac{kT}{e}\ln\left(\frac{N_a N_d}{n_i^2}\right) \quad [6.6]$$

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

$$W_o = \left[\frac{2\varepsilon(N_a + N_d)V_o}{eN_aN_d}\right]^{1/2} \quad [6.7]$$

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.

[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.

EXAMPLE 6.1 **THE BUILT-IN POTENTIALS FOR Ge, Si, InP, 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, InP, 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:

Table 6.1 Typical built-in voltages

Semiconductor	E_g(eV)	n_i(cm^{-3})	V_o(V)
Ge	0.66	2.4×10^{13}	0.37
Si	1.10	1.0×10^{10}	0.78
InP	1.34	1.3×10^{7}	1.12
GaAs	1.42	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 four semiconductors are summarized in the last column of Table 6.1 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)

$$Q = eN_aW_p = eN_dW_n$$

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 $E(x)$ 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,

$$J_{\text{hole}}(x) = ep(x)\mu_h E(x) - eD_h \frac{dp}{dx} = 0$$

The electric field is defined by $E = -dV/dx$, 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

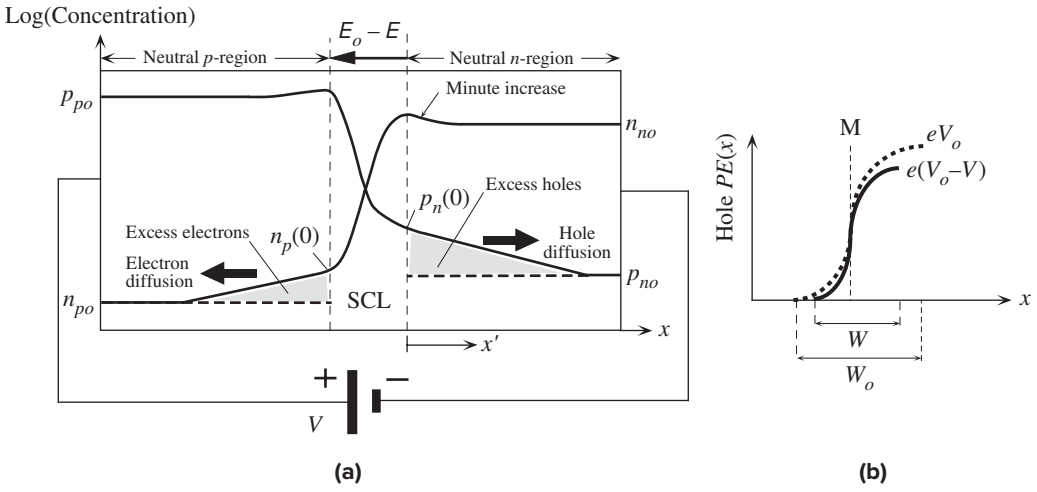


Figure 6.2 Forward-biased *pn* junction and the injection of minority carriers. (a) Carrier concentration profiles across the device under forward bias. (b) The hole potential energy with and without an applied bias. *W* is the width of the SCL with forward bias.

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.

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,

$$n_p(0) = n_{po} \exp\left(\frac{eV}{kT}\right) \tag{6.10}$$

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_{no}, that is,

$$\Delta p_n(x') = \Delta p_n(0) \exp\left(-\frac{x'}{L_h}\right) \tag{6.11}$$

Excess minority carrier profile

where

$$\Delta p_n(x') = p_n(x') - p_{no}$$

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

$$J_{D,\text{hole}} = -eD_h \frac{dp_n(x')}{dx'} = -eD_h \frac{d\Delta p_n(x')}{dx'}$$

[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.

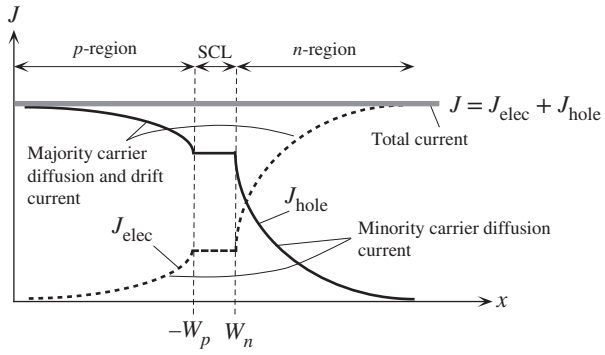


Figure 6.3 The total current anywhere in the device is constant. Just outside the depletion region, it is due to the diffusion of minority carriers.

that is,

$$J_{D,\text{hole}} = \left(\frac{eD_h}{L_h}\right) \Delta p_n(0) \exp\left(-\frac{x'}{L_h}\right)$$

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

$$J = J_{so}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \qquad [6.12a]$$

Ideal diode (Shockley) equation

This is the familiar diode equation with

$$J_{so} = \left[\left(\frac{eD_h}{L_h N_d}\right) + \left(\frac{eD_e}{L_e N_a}\right)\right] n_i^2 \qquad [6.12b]$$

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 6.12a 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

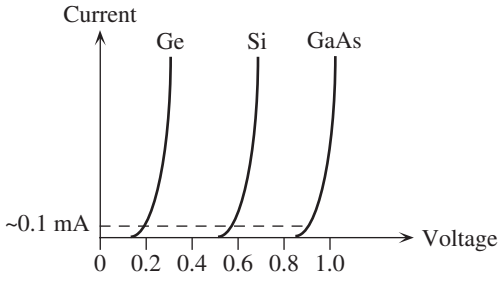


Figure 6.4 Schematic sketch of the *I–V* characteristics of Ge, Si, and GaAs *pn* junctions.

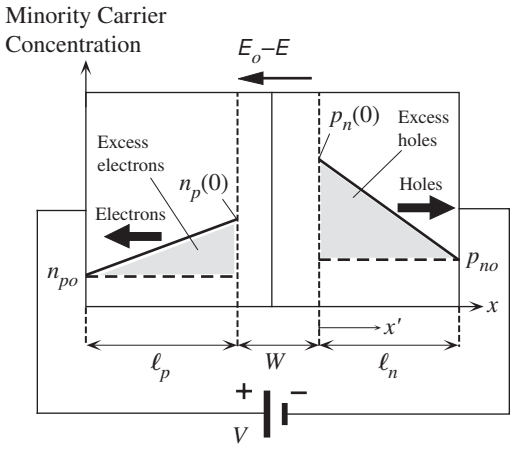


Figure 6.5 Minority carrier injection and diffusion in a short diode.

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 6.12a, 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 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,

$$J = \left(\frac{eD_h}{\ell_n N_d} + \frac{eD_e}{\ell_p N_a}\right) n_i^2 \left[\exp\left(\frac{eV}{kT}\right) - 1\right] \qquad [6.14]$$

Short diode

It is clear that this expression is identical to that of a long diode, that is, Equations 6.12a and b, 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 6.6a 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 6.6b. Suppose that we can describe the average rate of hole recombination in W_n by assigning holes a **mean hole recombination time** τ_h in this region. (Strictly we should call this an *effective* recombination time[4] as it represents an average over

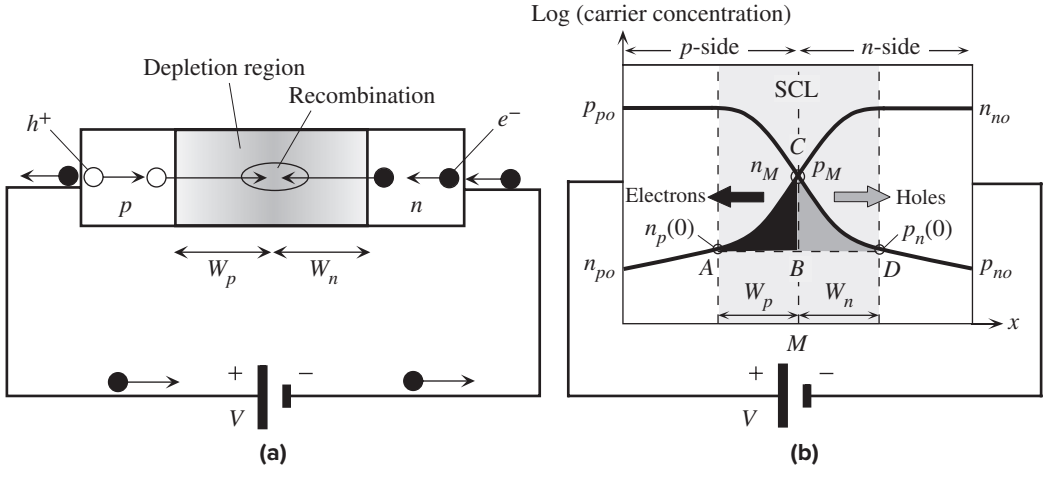


Figure 6.6 (a) Forward-biased *pn* junction and the injection of carriers and their recombination in SCL. (b) The calculation of the rate of recombination in the depletion region for a symmetric *pn* junction involves finding the two black and gray shades areas *ABC* and *BCD*.

[4] The exact analysis involves what is known as Shockley-Read-Hall indirect recombination statistics, which is discussed in more advanced textbooks. The use of effective lifetimes in the two depletion regions is equivalent averaging recombination rates in W_p and W_n. Further, the treatment here applies to indirect recombination, that is, through defects and impurities.

the rates of recombination in W_n.) Similarly, the **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 energy. At A, the potential is zero and at M it is $\frac{1}{2}e(V_o - V)$, so

$$\frac{p_M}{p_{po}} = \exp\left[-\frac{e(V_o - V)}{2kT}\right]$$

There is a similar equation for n_M/n_{no}. Further as the pn junction is symmetric $p_M = n_M$. Since V_o depends on dopant concentrations and n_i as in Equation 6.6 and further $p_{po} = N_a$ and $n_{no} = N_d$, we can simplify the above 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[5]

Recombination current

$$J_{\text{recom}} = J_{ro}\left[\exp\left(\frac{eV}{2kT}\right) - 1\right] \qquad [6.16]$$

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 6.12a and 6.15. For $V > kT/e$,

Total diode current = diffusion + recombination

$$J = J_{so}\exp\left(\frac{eV}{kT}\right) + J_{ro}\exp\left(\frac{eV}{2kT}\right)$$

[5] This is generally proved in advanced texts.

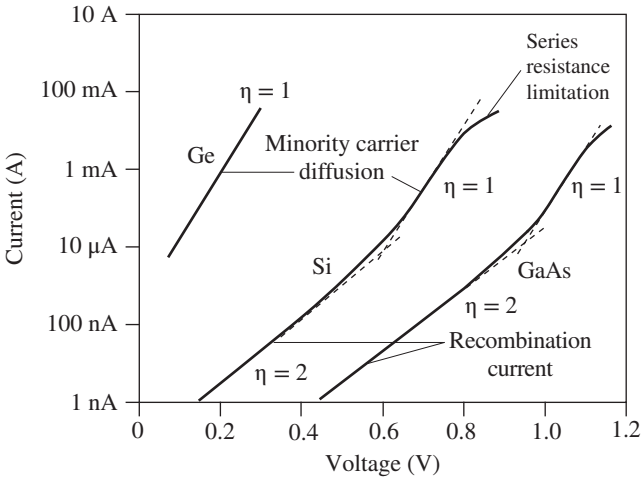


Figure 6.7 Schematic sketch of typical *I–V* characteristics of Ge, Si, and GaAs *pn* junctions as log(*I*) versus *V*.
The slope indicates $e/(\eta kT)$.

This expression is often lumped into a single exponential as

$$J = J_o \exp\left(\frac{eV}{\eta kT}\right) \qquad [6.17]$$

The diode equation for $V > kT/e$

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 both Si and GaAs, η is 2 over a wide current range but, at higher currents, it changes to 1. The current is initially controlled by recombination in the space charge layer but at high at sufficiently high voltages, it is due to by minority carrier diffusion in the neutral regions, 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.

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.

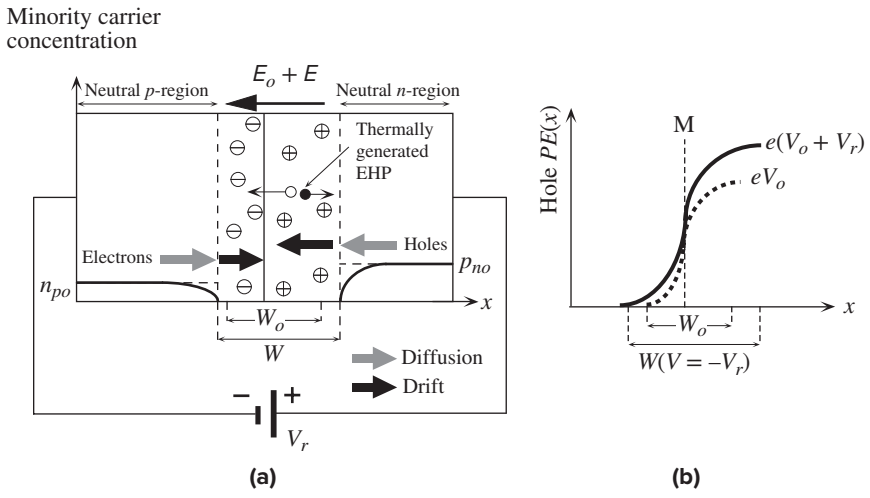


Figure 6.8 Reverse-biased *pn* junction. (a) Minority carrier profiles and the origin of the reverse current. (b) Hole *PE* across the junction under reverse bias.

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 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 6.12a 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 ($Vr > kT/e$). 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

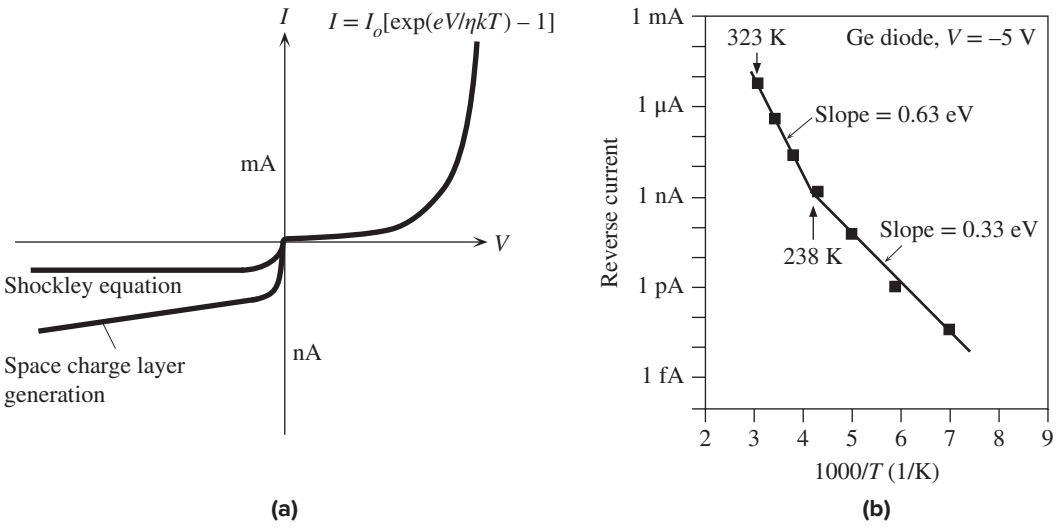


Figure 6.9 (a) Forward and reverse *I–V* characteristics of a *pn* junction (the positive and negative current axes have different scales and hence the discontinuity at the origin). (b) Reverse diode current in a Ge *pn* junction as a function of temperature in a $\ln(I_{rev})$ versus $1/T$ plot. Above 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) Data extracted from Scansen, D., and Kasap, S.O., *Canadian Journal of Physics,* 70, 1070, 1992.

processes via recombination centers, which is discussed in advanced texts. Suppose that τ_g is the **mean time to generate an EHP** 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

$$J_{\text{gen}} = \frac{eWn_i}{\tau_g} \qquad [6.18]$$

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 components,

Total reverse current

$$J_{\text{rev}} = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a}\right) n_i^2 + \frac{eWn_i}{\tau_g} \qquad [6.19]$$

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. (See Figure 6.8b.)

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 420 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}	$\mu_e \approx 150$ cm^2 V^{-1} s^{-1}
With	$N_d = 10^{16}$ cm^{-3}	$\mu_h \approx 430$ cm^2 V^{-1} s^{-1}

Thus, with $kT/e = 0.2585$ V at 300 K, we have

$$D_e = \frac{kT\mu_e}{e} \approx (0.02585 \text{ V})(150 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}) = 3.88 \text{ cm}^2 \text{ s}^{-1}$$

$$D_h = \frac{kT\mu_h}{e} \approx (0.02585 \text{ V})(430 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}) = 11.12 \text{ cm}^2 \text{ s}^{-1}$$

Diffusion lengths are

$$L_e = \sqrt{D_e\tau_e} = \sqrt{[(3.88\ \text{cm}^2\ \text{s}^{-1})(5\times10^{-9}\ \text{s})]}$$
$$= 1.39\times10^{-4}\ \text{cm} \quad \text{or} \quad 1.39\ \mu\text{m} < 5\ \mu\text{m}$$
$$L_h = \sqrt{D_h\tau_h} = \sqrt{[(11.12\ \text{cm}^2\ \text{s}^{-1})(420\times10^{-9}\ \text{s})]}$$
$$= 21.6\times10^{-4}\ \text{cm} \quad \text{or} \quad 21.6\ \mu\text{m} < 100\ \mu\text{m}$$

We therefore have a long diode. The built-in potential is

$$V_o = \left(\frac{kT}{e}\right)\ln\left(\frac{N_dN_a}{n_i^2}\right) = (0.02585\ \text{V})\ln\left[\frac{(10^{16}\times5\times10^{18})}{(1.0\times10^{10})^2}\right] = 0.875\ \text{V}$$

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,

$$I = I_{so}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \approx I_{so}\exp\left(\frac{eV}{kT}\right) \qquad \text{for } V \gg \frac{kT}{e} \qquad (= 0.02585\ \text{V})$$

where

$$I_{so} = A\,J_{so} = Aen_i^2\left[\left(\frac{D_h}{L_hN_d}\right) + \left(\frac{D_e}{L_eN_a}\right)\right] \approx \frac{Aen_i^2D_h}{L_hN_d}$$

as $N_a \gg N_d$. In other words, the current is mainly due to the diffusion of holes in the *n*-region. Thus,

$$I_{so} = \frac{(0.01\ \text{cm}^2)(1.602\times10^{-19}\ \text{C})(1.0\times10^{10}\ \text{cm}^{-3})^2\ (11.12\ \text{cm}^2\ \text{s}^{-1})}{(21.6\times10^{-4}\ \text{cm})(10^{16}\ \text{cm}^{-3})}$$
$$= 8.24\times10^{-14}\ \text{A} \quad \text{or} \quad 0.082\ \text{pA}$$

Then the diode current is

$$I \approx I_{so}\exp\left(\frac{eV}{kT}\right) = (8.24\times10^{-14}\ \text{A})\exp\left[\frac{(0.6\ \text{V})}{(0.0259\ \text{V})}\right]$$
$$= 0.99\times10^{-3}\ \text{A} \quad \text{or} \quad 1.0\ \text{mA}$$

We note that when a forward bias of 0.6 V is applied, the built-in potential is reduced from 0.875 V to 0.275 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, $n_i \approx 1.0\times10^{12}\ \text{cm}^{-3}$ (approximately from n_i versus $1/T$ graph in Figure 5.16), so

$$I_{so}(373\ \text{K}) \approx I_{so}(300\ \text{K})\left[\frac{n_i(373\ \text{K})}{n_i(300\ \text{K})}\right]^2$$
$$\approx (8.24\times10^{-14})\left(\frac{1.0\times10^{12}}{1.0\times10^{10}}\right)^2 = 8.24\times10^{-10}\ \text{A} \quad \text{or} \quad 0.824\ \text{nA}$$

At 100 °C, the forward current with 0.6 V across the diode is

$$I = I_{so}\exp\left(\frac{eV}{kT}\right) = (8.24\times10^{-10}\ \text{A})\exp\left[\frac{(0.6\ \text{V})(300\ \text{K})}{(0.02585\ \text{V})(373\ \text{K})}\right] = 0.10\ \text{A}$$

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

$$W = \left[\frac{2\varepsilon(V_o + V_r)}{eN_d}\right]^{1/2} = \left[\frac{2(11.9)(8.85 \times 10^{-12})(0.875 + 5)}{(1.6 \times 10^{-19})(10^{22})}\right]^{1/2}$$
$$= 0.88 \times 10^{-6} \text{ m} \quad \text{or} \quad 0.88 \ \mu\text{m}$$

The thermal generation current with $V_r = 5$ V is

$$I_{\text{gen}} = \frac{eAWn_i}{\tau_g} = \frac{(1.602 \times 10^{-19} \text{ C})(0.01 \text{ cm}^2)(0.88 \times 10^{-4} \text{ cm})(1.0 \times 10^{10} \text{ cm}^{-3})}{(10^{-6} \text{ s})}$$
$$= 1.41 \times 10^{-9} \text{ A} \quad \text{or} \quad 1.4 \text{ nA}$$

This thermal generation current is much greater than the reverse saturation current I_{so} (= 0.0842 pA). The reverse current is therefore dominated by I_{gen} and it is 1.4 nA.

EXAMPLE 6.5

A DIRECT BANDGAP *pn* JUNCTION In direct bandgap semiconductors, an electron and a hole can recombine directly, without needing a recombination center. Such a direct recombination leads to photon emission and is the basis of LEDs as discussed later in this chapter. Consider holes injected into the *n*-side of a *pn* junction from a direct bandgap semiconductor such as GaAs. Assume *weak injection* so that the excess hole concentration Δp_n is much less than the equilibrium majority carrier concentration n_{no}. If τ_h' is the mean lifetime due direct recombination, then the probability per unit time $1/\tau_h'$ that a hole directly recombines with an electron depends on the concentration of electrons n_{no} in the *n*-side, that is

Minority carrier lifetime in direct recombination

$$\tau_h' = \frac{1}{Bn_{no}} \qquad [6.20]$$

where B is a constant called the **direct recombination coefficient.** In addition, there will also be indirect recombination, which depends on the concentration of impurities and defects. Suppose that $1/\tau_h''$ is the probability per unit time for indirect recombination, then the overall probability of recombination per unit time $1/\tau_h$ will be

Minority carrier lifetime in direct and indirect recombination

$$\frac{1}{\tau_h} = \frac{1}{\tau_h'} + \frac{1}{\tau_h''} = Bn_n + \frac{1}{\tau_h''} \qquad [6.21]$$

where τ_h is the **effective lifetime.** The quantities τ_h' and τ_h'' are known as hole **radiative** and **non-radiative lifetimes** and are often written as τ_r and τ_{nr}. We can use the above expression for the recombination of injected carriers in the neutral regions as well as the depletion region.[6] Within the depletion region, n_n will be small and the hole lifetime will be due to indirect recombination. Similar arguments can be applied to electrons on the *p*-side with similar expressions.

Consider a symmetrical GaAs *pn* junction in which the *p*-side doping N_a is equal to the *n*-side doping N_d and both are 10^{17} cm^{-3}. The direct recombination coefficient $B \approx 2 \times 10^{-16}$ m^3 s^{-1}, cross sectional area $A = 1$ mm^2. The indirect recombination lifetime is roughly 200 ns. At these doping levels and at 300 K, the electron and hole drift mobilities are roughly $\mu_e \approx 4500$ cm^2 V^{-1} s^{-1} in the *p*-side and $\mu_h \approx 270$ cm^2 V^{-1} s^{-1} in the *n*-side. From the Einstein relation ($D = \mu kT/e$), the corresponding diffusion coefficients are $D_h = 6.98 \times 10^{-4}$ m^2 s^{-1} and $D_e = 1.16 \times 10^{-2}$ m^2 s^{-1}. Calculate the diffusion and recombination currents for this GaAs *pn* junction when the forward bias is 0.8 V and 1.0 V. What is your conclusion?

[6] There is also another recombination mechanism called Auger recombination, which occurs at high carrier concentrations, but this is ignored in this introductory treatment.

SOLUTION

We can calculate the direct recombination lifetimes τ_e' and τ_h' for electrons and holes recombining in the neutral *p*- and *n*-regions, respectively. In the *n*-side $n_n = n_{no} = N_d = 10^{17}$ cm^{-3}, and since this is a symmetric device

$$\tau_e' = \tau_h' = \frac{1}{Bn_{no}} = \frac{1}{BN_d} = \frac{1}{(2.0 \times 10^{-16}\ \text{m}^3\ \text{s}^{-1})(1 \times 10^{23}\ \text{m}^{-3})} = 50.0\ \text{ns}$$

The effective lifetime τ_h is given by Equation 6.21

$$\frac{1}{\tau_h} = \frac{1}{\tau_h'} + \frac{1}{\tau_h''} = \frac{1}{50 \times 10^{-9}} + \frac{1}{200 \times 10^{-9}}$$

which gives $\tau_h = \tau_e = 40$ ns.

To find the Shockley current in Equation 6.12a we need the diffusion lengths,

$$L_h = (D_h\tau_h)^{1/2} = [6.98 \times 10^{-4}\ \text{m}^2\ \text{s}^{-1})(40.0 \times 10^{-9}\ \text{s})]^{1/2} = 5.28 \times 10^{-6}\ \text{m},$$

and

$$L_e = (D_e\tau_e)^{1/2} = [(1.16 \times 10^{-2}\ \text{m}^2\ \text{s}^{-1})(40.0 \times 10^{-9}\ \text{s})]^{1/2} = 2.16 \times 10^{-5}\ \text{m}.$$

Notice that the electrons diffuse much further in the *p*-side due to their higher mobility. From Table 5.1, $n_i = 2.1 \times 10^{12}$ m^{-3}, so that reverse saturation current due to diffusion in the neutral regions is

$$\begin{aligned} I_{so} &= A\left(\frac{D_h}{L_hN_d} + \frac{D_e}{L_eN_a}\right)en_i^2 \\ &= (1 \times 10^{-6})\left[\frac{6.98 \times 10^{-4}}{(5.28 \times 10^{-6})(10^{23})} + \frac{1.16 \times 10^{-2}}{(2.16 \times 10^{-5})(10^{23})}\right](1.602 \times 10^{-19})(2.1 \times 10^{12})^2 \\ &\approx 4.7 \times 10^{-21}\ \text{A} \end{aligned}$$

Thus, the forward diffusion current is

$$\begin{aligned} I_{\text{diff}} &= I_{so}\exp\left(\frac{eV}{kT}\right) \\ &= (4.7 \times 10^{-21}\ \text{A})\exp\left[\frac{0.80\ \text{V}}{0.02585\ \text{V}}\right] = 1.3 \times 10^{-7}\ \text{A} \qquad \text{or} \qquad 0.13\ \mu\text{A} \end{aligned}$$

To calculate recombination component of the current, we need to know the SCL width *W* and the mean electron and hole recombination times in the depletion region.

The built-in voltage V_o is

$$V_o = \frac{kT}{e}\ln\left(\frac{N_aN_d}{n_i^2}\right) = (0.02585)\ln\left[\frac{10^{23}10^{23}}{(2.1 \times 10^{12})^2}\right] = 1.27\ \text{V}$$

Depletion region width *W* is

$$\begin{aligned} W &= \left[\frac{2\varepsilon(N_a + N_d)(V_o - V)}{eN_aN_d}\right]^{1/2} \\ &= \left[\frac{2(13)(8.85 \times 10^{-12}\ \text{F m}^{-1})(10^{23} + 10^{23}\ \text{m}^{-3})(1.27 - 0.80\ \text{V})}{(1.602 \times 10^{-19}\ \text{C})(10^{23}\ \text{m}^{-3})(10^{23}\ \text{m}^{-3})}\right]^{1/2} \\ &= 1.16 \times 10^{-7}\ \text{m}, \qquad \text{or} \qquad 0.116\ \mu\text{m}. \end{aligned}$$

In the depletion region both electron and hole concentrations are much less than n_{no} and p_{no} respectively, which means that the direct recombination rate will be small. Put differently, in the depletion region Bn_n and Bp_p are both small and, as first order, we can ignore radiative recombination. Thus, $\tau_h = \tau_e \approx 200$ ns.

As this is a symmetric diode, $W_p = W_n = (1/2)W$. The preexponential I_{ro} is

$$I_{ro} = \frac{Aen_i}{2}\left[\frac{W_p}{\tau_e} + \frac{W_n}{\tau_h}\right] = \frac{Aen_i}{2}\left(\frac{W}{\tau_h}\right)$$
$$= \frac{(10^{-6})(1.602 \times 10^{-19})(2.1 \times 10^{12})}{2}\left(\frac{1.16 \times 10^{-7}}{200 \times 10^{-9}}\right) \approx 9.8 \times 10^{-14}\ \text{A}$$

so that at $V = 0.8$ V,

$$I_{\text{recom}} \approx I_{ro} \exp\left(\frac{eV}{2kT}\right)$$
$$\approx (9.8 \times 10^{-14}\ \text{A}) \exp\left[\frac{0.8\ \text{V}}{2(0.02585\ \text{V})}\right] = 5.1 \times 10^{-7} \quad \text{A} \quad \text{or} \quad 0.51\ \mu\text{A}$$

The recombination current is more than the diffusion current. If we repeat the calculation for a voltage of 1.0 V across the device, we would find $I_{\text{diff}} = 0.30$ mA and $I_{\text{recom}} = 0.025$ mA, where I_{diff} dominates the current. Thus, as the voltage increases across a GaAs *pn* junction, the ideality factor η is initially 2 but then becomes 1 as shown in Figure 6.7. It is apparent that the *I–V* characteristics depend very much on the relative values of the radiative and nonradiative lifetimes.

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

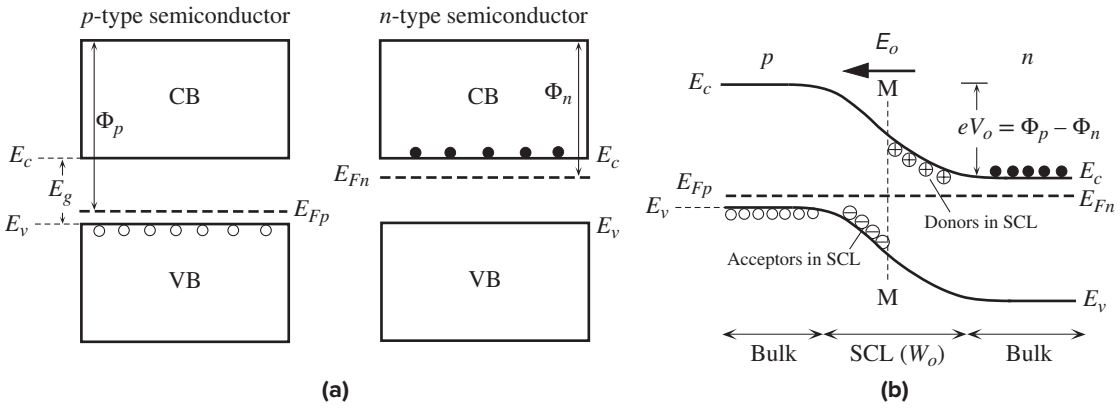


Figure 6.10 (a) Two isolated *p*- and *n*-type semiconductors (same material). (b) A *pn* junction band diagram when the two semiconductors are in contact. The Fermi level must be uniform in equilibrium. The metallurgical junction is at M. The region around M contains the space charge layer (SCL). On the *n*-side of M, SCL has the exposed positively charged donors, whereas on the *p*-side it has the exposed negatively charged acceptors.

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?

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 (majority carrier) 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 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

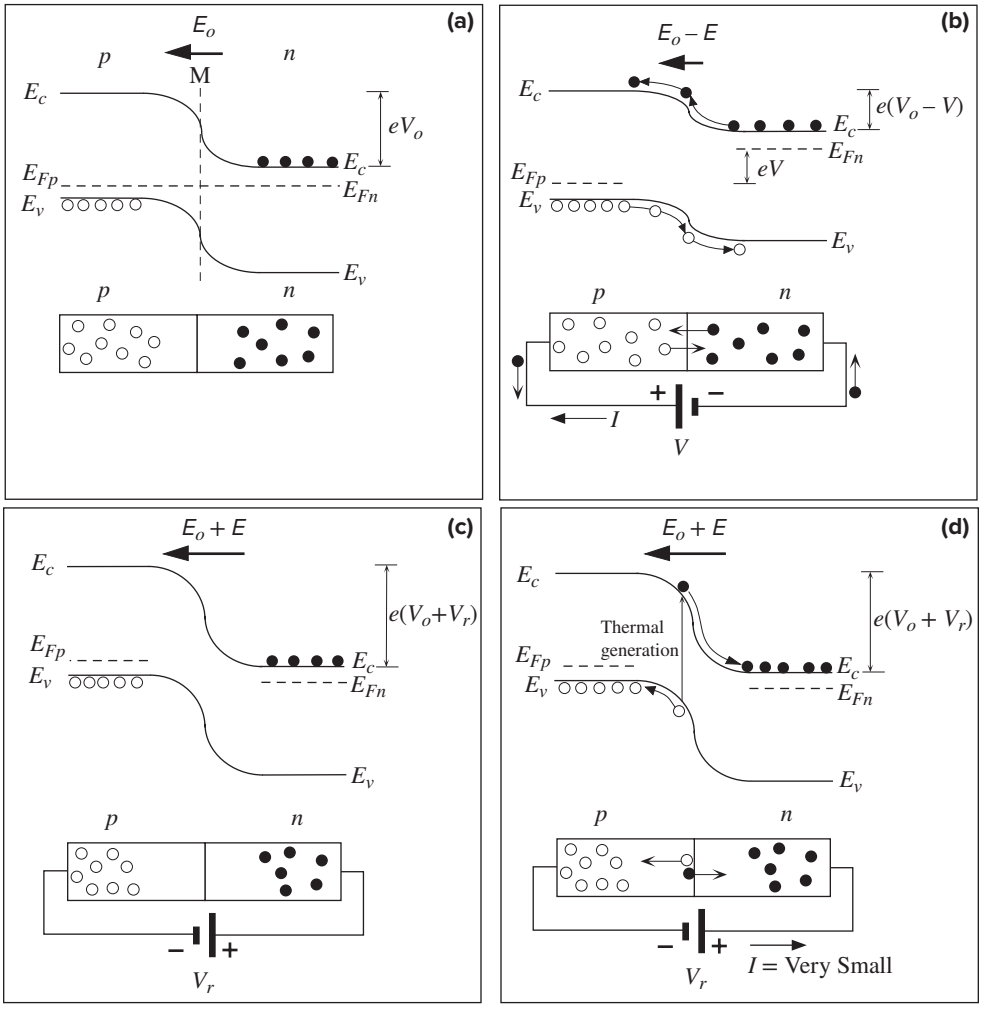


Figure 6.11 Energy band diagrams for a *pn* junction: (a) open circuit, (b) forward bias, (c) reverse bias conditions, (d) thermal generation of EHP in the depletion region results in a small reverse current.

current density J_{diff}. The probability of overcoming the *PE* barrier is proportional to $\exp(-eV_o/kT)$. Therefore, under zero bias,

$$J_{\text{diff}}(0) = B\exp\left(-\frac{eV_o}{kT}\right) \quad [6.22]$$

$$J_{\text{net}}(0) = J_{\text{diff}}(0) + J_{\text{drift}}(0) = 0 \quad [6.23]$$

where B is a proportionality constant and $J_{\text{drift}}(0)$ is the current due to the drift of electrons by 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* 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 $E_o - E$ (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 E_o has decreased to $E_o - E$, 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

pn Junction I–V characteristics

familiar diode equation,

$$J = J_o\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \qquad [6.24]$$

where J_o is a temperature-dependent constant.[7]

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 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 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 6.6 **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

$$n_{po} = N_c \exp\left[-\frac{(E_c - E_{Fp})}{kT}\right]$$

$$n_{no} = N_c \exp\left[-\frac{(E_c - E_{Fn})}{kT}\right]$$

[7] The derivation is similar to that for the Schottky diode, but there are more assumptions here.

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,

$$V_o = \left(\frac{kT}{e}\right) \ln\left[\frac{(N_a N_d)}{n_i^2}\right]$$

Built-in voltage

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.

With an applied voltage V, the width of the depletion region is given by Equation 6.7

$$W = \left[\frac{2\varepsilon(N_a + N_d)(V_o - V)}{eN_a N_d}\right]^{1/2} \quad [6.25]$$

Depletion region width

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 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 6.25, 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

$$C_{\text{dep}} = \left|\frac{dQ}{dV}\right| \quad [6.26]$$

Definition of depletion layer capacitance

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 6.25 in terms of Q and then differentiate it to obtain dQ/dV. The final result for the depletion capacitance is

$$C_{\text{dep}} = \frac{\varepsilon A}{W} = \frac{A}{(V_o - V)^{1/2}}\left[\frac{e\varepsilon(N_a N_d)}{2(N_a + N_d)}\right]^{1/2} \quad [6.27]$$

Depletion Capacitance

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 6.25. The C_{dep} versus V behavior is sketched in Figure 6.12b. Notice that C_{dep} decreases

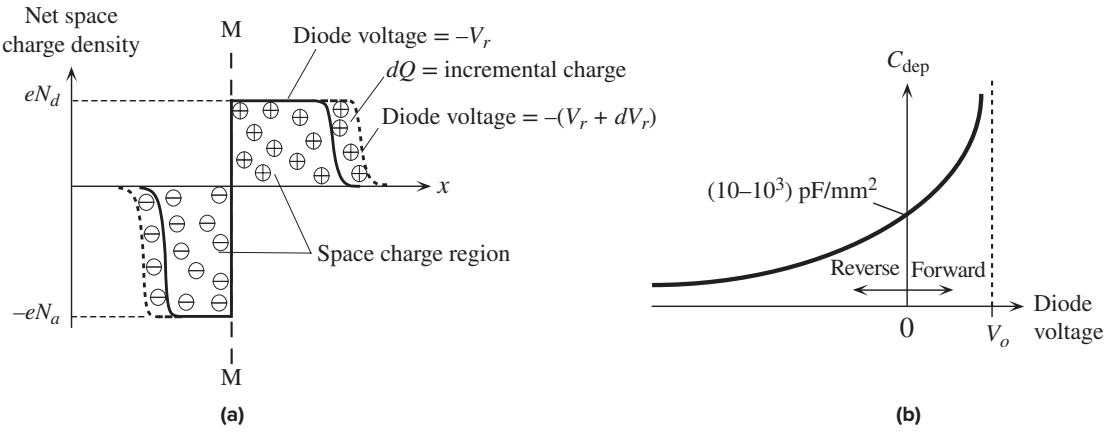


Figure 6.12 The depletion region behaves like a capacitor. (a) The charge in the depletion region depends on the applied voltage just as in a capacitor. A reverse bias example is shown. (b) The incremental capacitance of the depletion region increases with forward bias and decreases with reverse bias. Its value is typically in the range of picofarads per mm^2 of device area.

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 simple parallel plate capacitance expression $C_{\text{dep}} = \varepsilon A/W$ in Equation 6.27 was derived for an abrupt junction in which both the p and n-sides have uniform acceptor and donor concentration. It may seem unusual but it turns out that $C_{\text{dep}} = \varepsilon A/W$ is generally valid whatever the dopant concentration profiles are. Consider a one-sided pn junction in which the p-side is much more heavily doped, denoted as p^+, than the n-side as shown in Figure 6.13a. The depletion width extends almost entirely into the lightly doped n-side and we can take $W \approx W_n$. Under a reverse bias of V_r, the $+Q$ charge in the n-side is within W. When we increase V_r to $V_r + dV_r$, the charge Q increases to $Q + dQ$ as shown in Figure 6.13b. Take the net space charge density $\rho_{\text{net}} = eN_d(x)$ in the n-side depletion region. The total charge and the maximum field E_o from Equation 6.2 are given by

$$Q = A\int_0^W \rho_{\text{net}}\,dx \qquad \text{and} \qquad E_o = -\frac{1}{\varepsilon}\int_0^W \rho_{\text{net}}\,dx$$

so that $E_o = -Q/A\varepsilon$ and thus $dE_o = -dQ/A\varepsilon$. Further, the integration of $|E(x)|$ over x upto W gives V_r which is the area under the curve of $|E(x)|$ as indicated in Figure 6.13c. When we increase V_r to $V_r + dV_r$, this area increases by an amount shown as dark grey, which is dV_r. The additional dV_r drops across W and gives rise to $-dE_o$ so that[8] $dV_r/W = -dE_o$. Thus,

$$C_{\text{dep}} = dQ/dV_r = (-\varepsilon A dE_o)/(-W dE_o)$$

[8] It seems intuitively correct that $dV_r/W = |dE_o|$, but a rigorous proof is by no means trivial. The field depends on the integration of ρ_{net} and V_r depends on the double integration of ρ_{net}. We then have to differentiate the latter integral to obtain $dV_r/W = |dE_o|$.

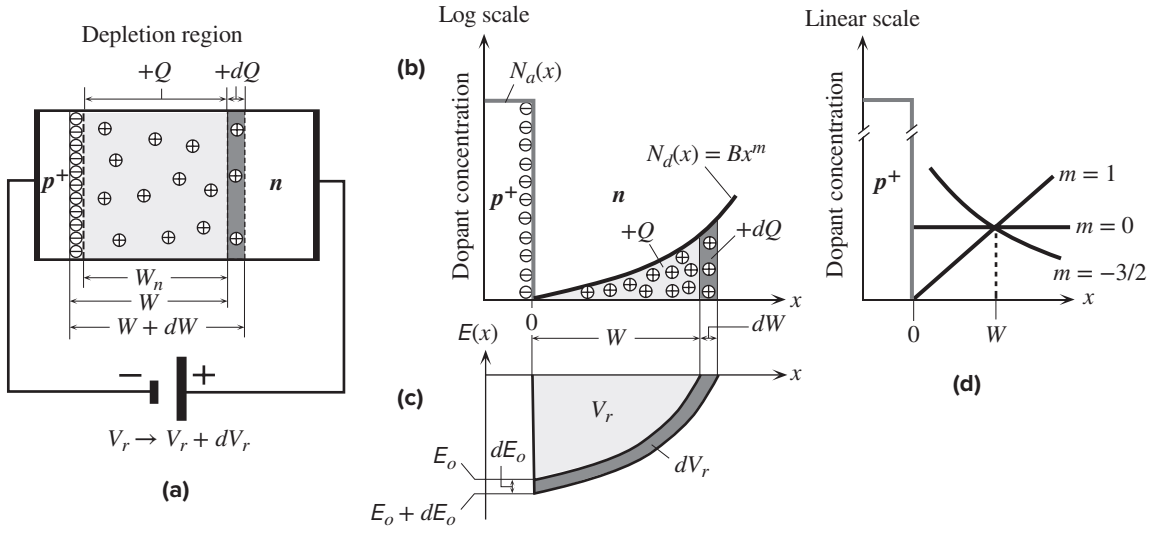


Figure 6.13 (a) A one-sided p^+n junction under reverse bias V_r in which $W_n \gg W_p$ and $W \approx W_n$. The n-side depletion region has exposed positive donors with total charge $+Q$. When V_r increased by dV_r, $+Q$ increases by $+dQ$. There is also an increase in the negative charge by the same amount in the p^+-side depletion region but this is not shown since it is very narrow. (b) An arbitrary donor concentration $N_d(x)$ on the n-side and the regions of $+Q$ and $+dQ$ corresponding to V_r and dV_r. (c) The field is almost totally on the n-side, maximum at the metallurgical junction at $x = 0$, and falls rapidly into the p^+-side. The area under the electric field $|E(x)|$ is the voltage across the depletion region. (d) Shapes of the donor concentration $N_d(x) = Bx^m$ profiles for $m = 0$ (abrupt), 1 (linear), and $-3/2$ (hyperabrupt).

that is

General depletion layer capacitance

$$C_{\text{dep}} = \frac{\varepsilon A}{W} \qquad [6.28]$$

Equation 6.28 is generally valid even if we do not have a one-sided junction, and is basically Equation 6.27 for a uniformly doped abrupt junction. Since W depends on the voltage, so does the depletion capacitance.

Suppose that we assume that the donor concentration in the p^+n junction follows $N_d(x) = Bx^m$ as shown in Figure 6.13b; and d for three m values. Obviously, $m = 0$ is the abrupt junction case. If we integrate $\rho_{\text{net}} = eBx^m$ across the depletion region W, we would get the field and if we integrate it again, we would find the total voltage across the depletion region, $V_o - V$ or $V_o + V_r$ as a function of W, that is the dependence of W on $(V_o - V)$. We can then substitute for W in Equation 6.28 and find C_{dep} as

General depletion layer capacitance

$$C_{\text{dep}} = A\left[\frac{e\varepsilon^{m+1}B}{(m+2)(V_o - V)}\right]^{1/(m+2)} \qquad [6.29]$$

in which $V = -V_r$ for reverse bias. Clearly under suitable reverse bias $V_r > V_o$, and $C_{\text{dep}} \propto V_r^{-1/(m+2)}$ which implies that we should design a pn junction whose C_{dep} dependence on the external V_r can be controlled. Notice that $m = 1$ gives $C_{\text{dep}} \propto V_r^{-1/2}$ as expected from Equation 6.27. For many pn junctions, the dopant concentration

on both or on one side can be approximated as a linear variation ($m = 1$) so that $C_{\text{dep}} \propto V_r^{-1/3}$.

The voltage dependence of the depletion capacitance is utilized in **varactor diodes (varicaps),** which are used 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. The resonant frequency of an *LC* circuit with a varactor will be

$$f_o = \frac{1}{2\pi\sqrt{LC_{\text{dep}}}} \propto (V_o - V)^{1/2(m+2)}$$

f_o will be linear in V_r if $1/(m + 2) = 1$ or $m = -3/2$, which is shown in Figure 6.13d. *pn* junctions with such or similar sharp dopant profiles are called **hyperabrupt junctions.**[9]

EXAMPLE 6.7 **DEPLETION REGION CAPACITANCE** Table 6.2 provides data on the capacitance C between the terminals of a reverse-biased Si diode at various reverse voltages V_r. The diode is a single sided p^+n junction (fabricated by ion implantation) with a circular electrode that is approximately 500 μm in diameter. The stray capacitance or the packaging capacitance between the terminals is estimated to be 0.5–0.7 pF. Find the built-in voltage V_o and the donor concentration N_d. What is your conclusion?

SOLUTION

p^+n junction depletion capacitance

Since this a single-sided p^+n type Si diode, from Equation 6.27, with $N_a \gg N_d$, we have

$$C_{\text{dep}} = A\left[\frac{e\varepsilon}{2N_d(V_o - V)}\right]^{1/2} \qquad [6.30]$$

and substituting $V = -V_r$ and rearranging the equation,

$$\frac{1}{C_{\text{dep}}^2} = \frac{2N_d}{A^2 e\varepsilon}(V_o + V_r)$$

A plot of $1/C_{\text{dep}}^2$ against V_r should be straight line and we can find V_o and N_d from the intercept and the slope. However, the measured C is not exactly C_{dep} but $C_{\text{dep}} + C_s$, where C_s is the stray capacitance 0.6 ± 0.1 pF. Table 6.2 shows a third row in which $1/C_{\text{dep}}^2$ has been calculated from the second row (C) by subtracting $C_s = 0.6$ pF. Figure 6.14 shows the plot of $1/C_{\text{dep}}^2$ against V_r, which follows the expected behavior quite well with the best line being

Table 6.2 Capacitance of a reverse-biased Si *pn* junction diode at 23 °C

V_r (V)	0.5	1.0	2.0	4.0	8.0	10	15
C (pF)	42.6	36.4	29.2	22.4	16.6	15.3	12.6
$1/C_{\text{dep}}^2 \times 10^{-4}$ (pF^{-2})	5.67	7.80	12.2	21.04	39.1	46.3	69.4

[9] See Question 6.10 on varactor diodes. The term hyperabrupt is commonly used for doping profiles in which m is negative, *i.e.*, the donor concentration decreases with x in Figure 6.13d.

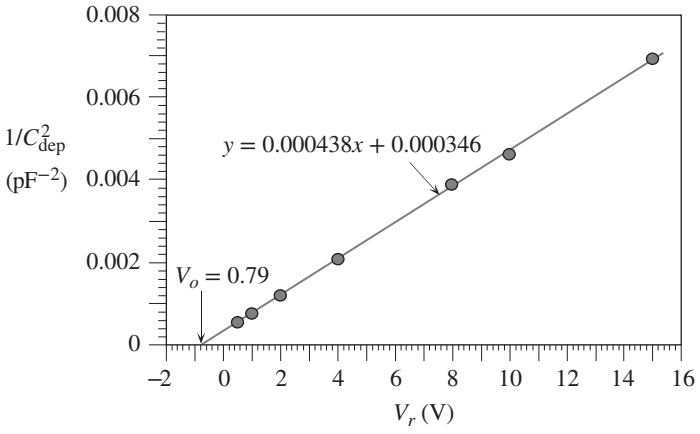


Figure 6.14 Plot of $1/C_{dep}^2$ against V_r for data in Table 6.2. The solid line is the best fit to the data.

Table 6.3 Extraction of *pn* junction characteristics from diode capacitance measurements

C_s (pF)	0	0.5	0.6	0.7	1
V_o (V)	0.96	0.82	0.79	0.75	0.67
N_d (cm^{-3})	7.8×10^{15}	7.1×10^{15}	7.0×10^{15}	6.9×10^{15}	6.5×10^{15}
N_a (cm^{-3})	3.1×10^{20}	1.2×10^{18}	4.0×10^{17}	8.1×10^{16}	4.7×10^{15}

$y = 0.000438x + 0.000346$ (easily obtained from a graphic software such as Excel). The intercept on the V_r axis gives $-V_o$ so that

$$V_o = 0.000346/0.000438 = 0.79 \text{ V}.$$

The slope is

$$\text{Slope} = \frac{2N_d}{A^2 e\varepsilon} = 0.000438 \text{ V pF}^{-2},$$

so that substituting $A = \pi(250 \times 10^{-6} \text{ m})^2 = 1.97 \times 10^{-7} \text{ m}^2$, $\varepsilon = \varepsilon_o\varepsilon_r$, $\varepsilon_r = 11.9$, we find $N_d = 7.0 \times 10^{21} \text{ m}^{-3}$ or $N_d = 7.0 \times 10^{15} \text{ cm}^{-3}$.

We can also extract N_a by using $V_o = (kT/e)\ln(N_aN_d/n_i^2)$, which gives $N_a = 4.0 \times 10^{23} \text{ m}^{-3}$ or $4.0 \times 10^{17} \text{ cm}^{-3}$; a reasonable value. While these are reasonable values, they do depend on the stray capacitance, especially N_a. If we repeat the above calculations for different C_s we would find the results in Table 6.3. Notice that while N_d values are comparable between different C_s values, N_a is extremely sensitive to stray capacitance and varies by five orders of magnitude. Clearly, stray capacitance correction is very important, assuming everything else has been accounted (including the assumption of an abrupt junction).

EXAMPLE 6.8

LINEARLY GRADED *pn* JUNCTIONS The simplest way to fabricate a *pn* junction is to diffuse dopants into a Si wafer at a high temperature in a diffusion chamber. Consider an *n*-type Si crystal and we expose one surface of the crystal to a boron gas at a high temperature in a diffusion chamber. B-atoms from the gas enter and diffuse into the Si-crystal as depicted in Figure 6.15. The boron (acceptor) concentration N_a decays with x as shown in Figure 6.15 at two times t_1 and t_2 where $t_2 > t_1$. The whole acceptor concentration profile $N_a(x)$ widens

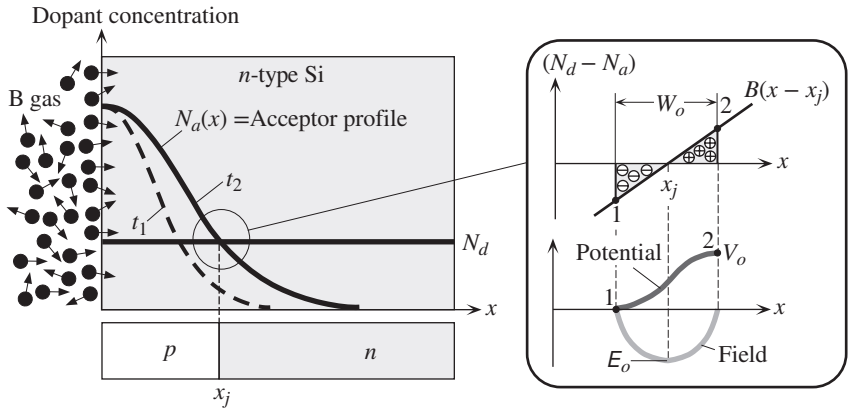


Figure 6.15 Formation of a linearly graded junction in a diffused *pn* junction. The B-atoms from B-gas on the surface of an *n*-Si wafer diffuse into the crystal. $N_a(x)$ is the acceptor concentration profile at arbitrary times t_1 and t_2 ($>t_1$). Acceptors diffuse from the surface and at time $t = t_2$, at $x = x_j$, the acceptor and donor concentrations are the same. This is the metallurgical junction. The diffusion is terminated when x_j reaches (approximately) the desired value. The net dopant concentration ($N_d - N_a$) around x_j depends linearly on x.

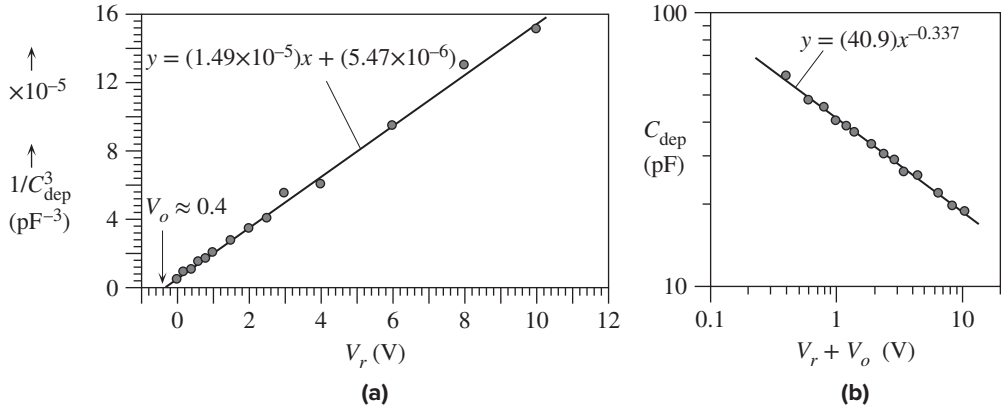


Figure 6.16 (a) Plot of $1/C^3_{dep}$ against V_r using data from diode capacitance measurements on a diffused Si power diode. The solid line is the best fit. (b) C_{dep} against $V_r + V_o$ with $V_o = 0.4$ from (a). (Measurements were carried out by Peyman Pourhaj, P. Eng.)

into the crystal as time lapses because more and more B-atoms diffuse further into the bulk. The B-gas provides a constant flux of B-atoms to the surface (an infinite source). The point $x = x_j$ where $N_a = N_d$ defines the metallurgical junction. To the left, $x < x_j$, $N_a > N_d$, and this side is *p*-type. To the right, $x > x_j$, $N_d > N_a$, and this side is *n*-type. A *pn* junction is formed with its junction at $x = x_j$ and there is a depletion region of width W_o around this junction as shown in Figure 6.15. The problem is similar to the one-sided junction and the depletion layer capacitance is given by Equation 6.29 with $m = 1$. Figure 6.16a shows a plot of $1/C^3_{dep}$ against V_r for a commercial diffused junction Si power diode and the data seem to confirm a linearly graded junction behavior and the best line is $y = (1.49 \times 10^{-5})x + 5.47 \times 10^{-6}$ which gives a built-in voltage $V_o = 0.37$ V or roughly 0.4 V on the V_r axis; the determination of the intercept for V_o is quite sensitive to stray capacitances. We can further check the linearly

graded junction assumption by plotting C_{dep} against $V_r + V_o$ on a log–log plot as in Figure 6.15b which shows a best power fit of $C_{dep} \propto (V_r + V_o)^{-0.337}$. Clearly, the assumption is well supported for this diode and the junction is linearly graded.

Suppose we take x_j as $x = 0$, then $N_d - N_a = Bx$, where B is the gradient of the doping profile. We can easily find the built-in potential V_o by noting that, as shown in Figure 6.15, the hole concentration at positions 1 and 2 in equilibrium are $p_{po}(1) = BW_o/2$ and $p_{no}(2) = n_i^2/n_{no}(2) = n_i^2/(BW_o/2)$. If we apply Boltzmann statistics (*i.e.*, assume a nondegenerate semiconductor) we can write

$$\frac{p_{no}(2)}{p_{po}(1)} = \frac{2n_i^2/BW_o}{BW_o/2} = \exp\left(-\frac{eV_o}{kT}\right)$$

so that

Built-in voltage, linear junction

$$V_o = \frac{kT}{e}\ln\left(\frac{BW_o}{2n_i}\right)^2 \qquad [6.31]$$

Further, for a lineary graded junction $m = 1$, and since $C_{dep} = \varepsilon A/W$, then from Equation 6.29, W_o at $V_r = 0$ is

Depletion layer width, linear junction

$$W_o = \left[\frac{12\varepsilon V_o}{eB}\right]^{1/3} \qquad [6.32]$$

Capacitance measurements under reverse bias in Figure 6.16a, in principle, provide V_o. We then have two equations with two unknowns, B and W_o in Equations 6.31 and 6.32, and hence we can find B and W_o. Thus, using $V_o \approx 0.4$ V in Equations 6.31 and 6.32, we find,

$$B \approx 5.5 \times 10^{16}\ \text{cm}^{-4} \qquad \text{and} \qquad W_o \approx 8.3 \times 10^{-6}\ \text{m or } 8.3\ \mu\text{m}$$

which are quite sensitive to the exact value of V_o and hence to experimental uncertainties, that is parasitic capacitances and whether the linear doping profile is linear over the whole depletion width. As we move away from the junction, the linearity will be lost. If we know the cross sectional area of the *pn* junction we can use the slope of the best line in Figure 6.16a to find B.

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 6.17, 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 $C_{diff} = dQ/dV$. 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

$$Q = \tau_h I = \tau_h I_o\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \qquad [6.33]$$

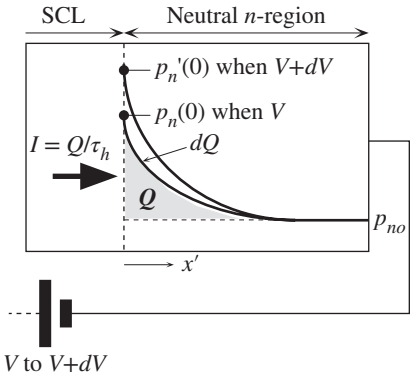


Figure 6.17 Consider the injection of holes into the *n*-side during forward bias. Storage or diffusion capacitance arises because when the diode voltage increases from V to $V + dV$, more minority carriers are injected and more minority carrier charge is stored in the *n*-region.

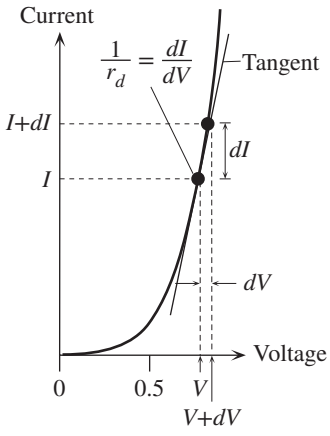


Figure 6.18 The dynamic resistance of the diode is defined as dV/dI, which is the inverse of the tangent at I.

Thus,

Diffusion capacitance

$$C_{\text{diff}} = \frac{dQ}{dV} = \frac{\tau_h e I}{kT} = \frac{\tau_h I(\text{mA})}{25} \qquad [6.34]$$

where we used $e/kT \approx 1/0.025$ at room temperature. (Note that $1/0.026$ is also commonly used.) 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 6.18. 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

$$r_d = \frac{dV}{dI} = \frac{kT}{eI} = \frac{25}{I(\text{mA})} \qquad [6.35]$$

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

Dynamic Conductance

$$g_d = \frac{dI}{dV} = \frac{I}{r_d}$$

Diffusion capacitance of a long diode

From Equations 6.34 and 6.35 we have

$$r_d\, C_{\text{diff}} = \tau_h \qquad [6.36a]$$

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}.

Equation 6.36a applies to a long diode, and cannot be used for a short diode. The reason is that the injected minority carriers simply diffuse and reach the collecting electrodes. The minority carrier profile is a straight line whose gradient determines the diffusion current as in Figure 6.5. The diode current I supplies the minority carriers that diffuse through the neutral regions and reach the electrodes. Consider a p^+n junction and the diffusion of holes on the n-side as in Figure 6.5. If τ_t is the **diffusion time** of holes across ℓ_n, then we know from Chapter 1 that $\ell_n = (2D_h\tau_t)^{1/2}$. If the total charge injected into the neutral n-side is Q (the grey area in Figure 6.17 under the $p_n(x)$ profile) then this charge takes τ_t seconds to diffuse across ℓ_n and the current I must replace Q every τ_t seconds so that $I = Q/\tau_t$. Thus $Q = I\tau_t$, and following along the lines above for the long diode, we can easily show that

Diffusion capacitance of a short diode

$$r_d\, C_{\text{diff}} = \tau_t \qquad \text{[6.36b]}$$

The short diode diffusion capacitance is always less than that of the long diode.

EXAMPLE 6.9

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 dynamic 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

$$I = I_{so}\exp\left(\frac{eV}{kT}\right) \qquad \text{or} \qquad V = \left(\frac{kT}{e}\right)\ln\left(\frac{I}{I_{so}}\right) = (0.0259)\ln\left(\frac{5\times10^{-3}}{0.0836\times10^{-12}}\right) = 0.643 \text{ V}$$

The dynamic diode resistance is given by

$$r_d = \frac{25}{I\,(\text{mA})} = \frac{25}{5} = 5\ \Omega$$

The depletion capacitance with $N_a \gg N_d$ is

$$C_{\text{dep}} = A\left[\frac{e\varepsilon(N_aN_d)}{2(N_a+N_d)(V_o-V)}\right]^{1/2} \approx A\left[\frac{e\varepsilon N_d}{2(V_o-V)}\right]^{1/2}$$

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

$$C_{\text{dep}} = 10^{-6}\left[\frac{(1.6\times10^{-19})(11.9)(8.85\times10^{-12})(10^{22})}{2(0.877-0.643)}\right]^{1/2}$$
$$= 6.0\times10^{-10}\text{ F} \qquad \text{or} \qquad 600\text{ pF}$$

The incremental diffusion capacitance C_{diff} due to holes injected and stored in the *n*-region is

$$C_{\text{diff}} = \frac{\tau_h I(\text{mA})}{25} = \frac{(417 \times 10^{-9})(5)}{25} = 8.3 \times 10^{-8}\,\text{F} \qquad \text{or} \qquad 83\,\text{nF}$$

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 6.19. In the $V = -V_{\text{br}}$ region, the reverse current increases dramatically with the reverse bias. If unlimited, the large 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 rupture a Si–Si bond. The phenomenon by which a drifting electron gains sufficient energy from the field to ionize a host crystal atom by

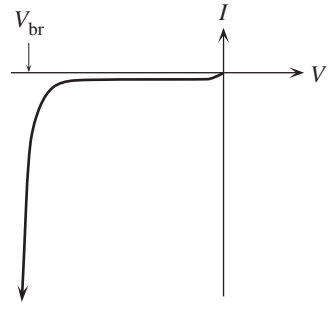


Figure 6.19 Reverse *I–V*, characteristics of a *pn* junction.

as discussed in Chapter 1. Amorphous Si:H has an E_g of about 1.8 eV. The alloying of a-Si:H with Ge to produce a-SiGe:H decreases E_g. Further, E_g of a-SiGe:H can be graded by controlling the Ge content.

6.11 BIPOLAR TRANSISTOR (BJT)

6.11.1 COMMON BASE (CB) DC CHARACTERISTICS

As an example, we will consider the *pnp* bipolar junction transistor (BJT) whose basic structure is shown in Figure 6.48a. 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 6.48a 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 6.48a.

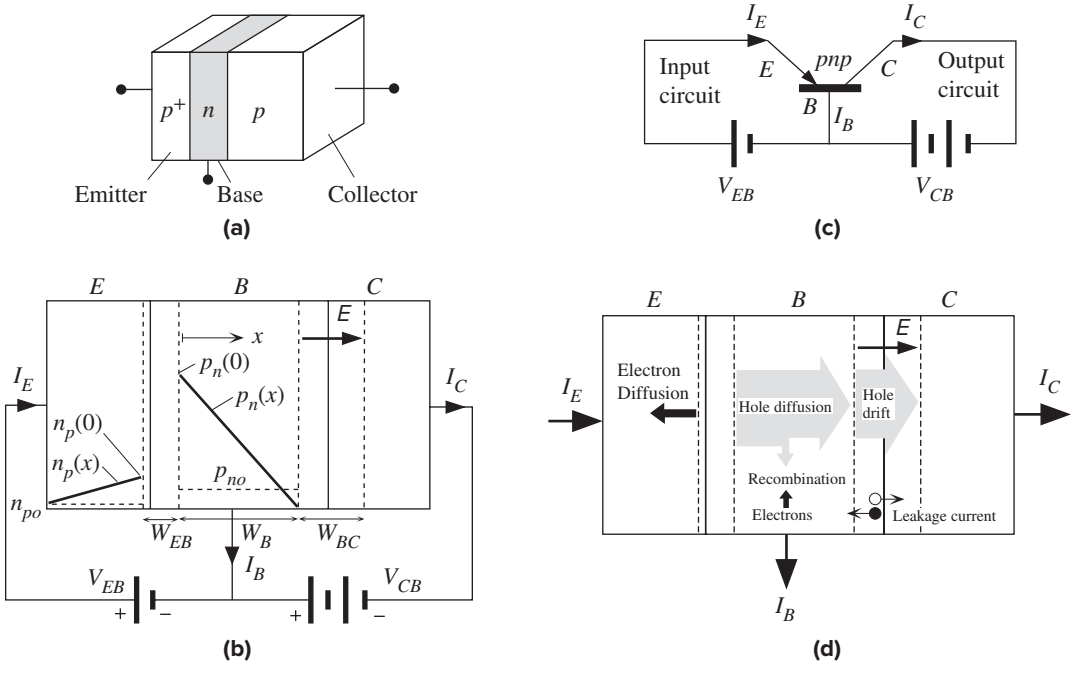
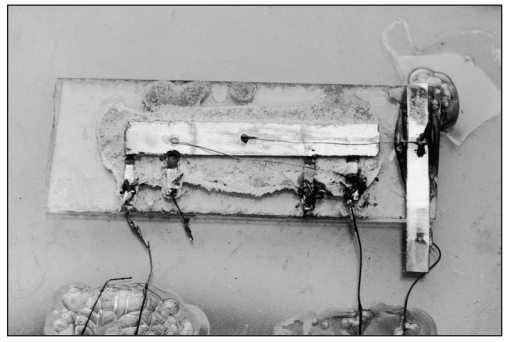


Figure 6.48 (a) A schematic illustration of the *pnp* bipolar transistor with three differently doped regions. (b) The *pnp* bipolar operated under normal and active conditions. (c) The CB configuration with input and output circuits identified. (d) The illustration of various current components under normal and active conditions.



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.

| Left: © AP Photo. Right: © Fotosearch/Getty Images.

This first commercial pocket transistor radio (Regency TR- 1) was released in 1954. It had 4 *npn* Ge transistors from Texas Instruments and was sold at $49.99, roughly $450 in today's dollars.

| © Bettmann/Getty Images.

Left to right: Andrew Grove (1936–2016), Robert Noyce (1927–1990), and Gordon Moore (born 1929), 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.

| Courtesy of Intel Corp.

The *pnp* BJT connected as shown in Figure 6.48b 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 6.48b, in which the base is common to both the collector and emitter bias voltages, is known as the common base (CB) configuration.[15] Figure 6.48c 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 6.48c 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 6.48b.

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 6.48d. 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 6.48b),

$$p_n(0) = p_{no} \exp\left(\frac{eV_{EB}}{kT}\right) \qquad [6.55]$$

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 E in W_{BC}. It is apparent that all the injected holes from the emitter become collected by the collector. The collector

[15] CB should not be confused with the conduction band abbreviation.

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 (input) circuit to the collector (output) 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.[16] 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 6.48b. This is only true in the absence of any recombination in the base as in the short diode case. The emitter current is then

$$I_E = -eAD_h\left(\frac{dp_n}{dx}\right)_{x=0} = eAD_h\frac{p_n(0)}{W_B}$$

We can substitute for $p_n(0)$ from Equation 6.55 to obtain

Emitter current

$$I_E = \frac{eAD_h p_{no}}{W_B}\exp\left(\frac{eV_{EB}}{kT}\right) \qquad [6.56]$$

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 6.48c 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,

Definition of CB current gain

$$\alpha = \frac{I_C}{I_E} \qquad [6.57]$$

Typically, α is less than unity, in the range 0.990–0.999, 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,

Total emitter current

$$I_E = I_{E(\text{hole})} + I_{E(\text{electron})}$$

Only the holes injected into the base are useful in giving a collector current because only they can reach the collector. The emitter injection efficiency is defined as

Emitter injection efficiency

$$\gamma = \frac{I_{E(\text{hole})}}{I_{E(\text{hole})} + I_{E(\text{electron})}} = \frac{1}{1 + \dfrac{I_{E(\text{electron})}}{I_{E(\text{hole})}}} \qquad [6.58]$$

[16] The actual concentration profile can be calculated by solving the steady-state continuity equation, which can be found in more advanced texts.

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 6.19.

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 6.48d. 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

Base transport factor

$$\alpha_T = \frac{I_C}{I_{E(\text{hole})}} = \frac{I_C}{\gamma I_E} \qquad [6.59]$$

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

$$\tau_t = \frac{W_B^2}{2D_h} \qquad [6.60]$$

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 6.59 gives the base transport factor α_T,

Base transport factor

$$\alpha_T = \frac{I_C}{I_{E(\text{hole})}} = 1 - \frac{\tau_t}{\tau_h} \qquad [6.61]$$

Using Equations 6.57, 6.59, and 6.61 we can find the total **CB current gain** α:

CB current gain

$$\alpha = \alpha_T \gamma = \left(1 - \frac{\tau_t}{\tau_h}\right)\gamma \qquad [6.62]$$

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 6.48d. 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 6.48d. 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

$$I_B = \left(\frac{\tau_t}{\tau_h}\right) I_{E(\text{hole})} + I_{E(\text{electron})} = \gamma \frac{\tau_t}{\tau_h} I_E + (1 - \gamma) I_E \qquad [6.63]$$

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.)

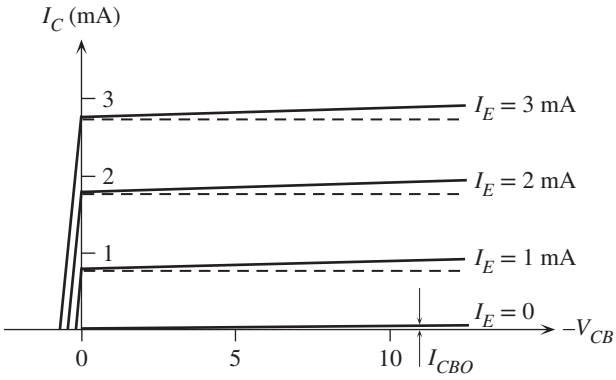


Figure 6.49 DC *I–V* characteristics of the *pnp* bipolar transistor (exaggerated to highlight various effects).

The ratio of the collector current to the base current is defined as the **current gain** β of the transistor.[17] By using Equations 6.57, 6.62, and 6.63, we can relate β to α:

Base-to-collector current gain

$$\beta = \frac{I_C}{I_B} = \frac{\alpha}{1-\alpha} \approx \frac{\gamma \tau_h}{\tau_t} \qquad [6.64]$$

The base–collector junction in Figure 6.48b 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 EHPs in the depletion region W_{BC} being drifted by the internal field, as schematically illustrated in Figure 6.48d. 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

$$I_C = \alpha I_E + I_{CBO} \qquad [6.65]$$

Active region base current

$$I_B = (1-\alpha) I_E - I_{CBO} \qquad [6.66]$$

Equations 6.65 and 6.66 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 6.48b 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 6.49. 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

[17] 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.

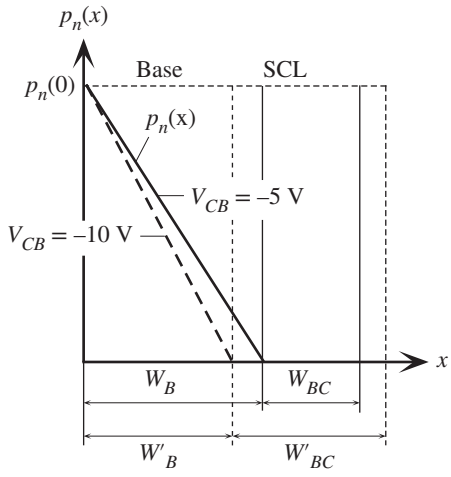


Figure 6.50 The Early effect.
When the BC reverse bias increases, the depletion width W_{BC} increases to W'_{BC}, which reduces the base width W_B to W'_B. As $p_n(0)$ is constant (constant V_{EB}), the minority carrier concentration gradient becomes steeper and the collector current, I_C increases.

$I_C = \alpha I_E + I_{CBO}$, 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 above 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 V_{CB}. 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 6.50. 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 6.61 and 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 versus V_{CB} lines at a fixed I_E are very small in Figure 6.49. The base width modulation by V_{CB} is called the **Early effect.**

EXAMPLE 6.18

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 $\mu_h = 400\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$ (minority carriers in the base). From the Einstein relationship we can easily find the diffusion coefficient of holes,

$$D_h = \left(\frac{kT}{e}\right)\mu_h = (0.02585\ \text{V})(400\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}) = 10.34\ \text{cm}^2\ \text{s}^{-1}$$

The minority carrier transit time τ_t across the base is

$$\tau_t = \frac{W_B^2}{2D_h} = \frac{(2 \times 10^{-4}\ \text{cm})^2}{2(10.34\ \text{cm}^2\ \text{s}^{-1})} = 1.93 \times 10^{-9}\ \text{s} \qquad \text{or} \qquad 1.93\ \text{ns}$$

The base transport factor and hence the CB current gain is

$$\alpha = \gamma\alpha_T = 1 - \frac{\tau_t}{\tau_h} = 1 - \frac{1.93 \times 10^{-9}\ \text{s}}{400 \times 10^{-9}\ \text{s}} = 0.99517$$

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

$$\begin{aligned} I_{EO} &= \frac{eAD_hP_{no}}{W_B} = \frac{eAD_hn_i^2}{N_dW_B} \\ &= \frac{(1.6 \times 10^{-19}\ \text{C})(0.02 \times 10^{-2}\ \text{cm}^2)(10.34\ \text{cm s}^{-1})(1.0 \times 10^{10}\ \text{cm}^{-3})^2}{(1 \times 10^{16}\ \text{cm}^{-3})(2 \times 10^{-4}\ \text{cm})} \\ &= 1.66 \times 10^{-14}\ \text{A} \end{aligned}$$

Thus,

$$V_{EB} = \frac{kT}{e}\ln\left(\frac{I_E}{I_{EO}}\right) = (0.02585\ \text{V})\ln\left(\frac{1 \times 10^{-3}\ \text{A}}{1.66 \times 10^{-14}\ \text{A}}\right) = 0.64\ \text{V}$$

The major assumption is $\gamma = 1$, which is generally not true, as shown in Example 6.19. 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.

EXAMPLE 6.19 EMITTER INJECTION EFFICIENCY γ

a. Consider a *pnp* transistor with the parameters as defined in Figure 6.48. 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 6.18, which has an acceptor doping of 2×10^{18} cm^{-3} in the emitter, donor doping of 1×10^{16} 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

$$\alpha = \gamma\left(1 - \frac{\tau_t}{\tau_h}\right)$$

In the emitter, $N_{a(\text{emitter})} = 2 \times 10^{18}$ cm^{-3} and $\mu_{e(\text{emitter})} = 200$ cm^2 V^{-1} s^{-1}, and in the base, $N_{d(\text{base})} = 1 \times 10^{16}$ cm^{-3} and $\mu_{h(\text{base})} = 400$ cm^2 V^{-1} 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 6.18 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 this 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, $I_{E(\text{hole})}$ would have to be even smaller compared with the total I_E, which would make γ and hence β even lower.

6.11.2 COMMON BASE AMPLIFIER

According to Equation 6.56 the emitter current depends exponentially on V_{EB},

$$I_E = I_{EO} \exp\left(\frac{eV_{EB}}{kT}\right) \qquad [6.67]$$

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 6.51. 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 6.55, modulates the injected hole concentration $p_n(0)$ up and down about the dc value determined by V_{EE} as depicted in Figure 6.51. 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 6.51. However, the output is commonly taken between the collector, and the base and this voltage V_{CB} is

$$V_{CB} = -V_{CC} + R_C I_C$$

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 6.51.

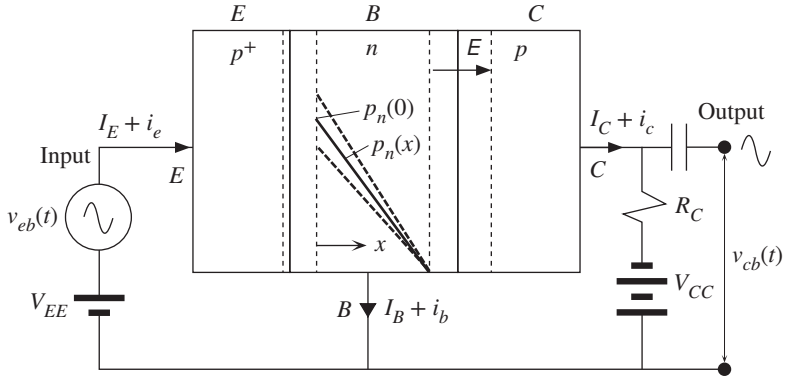


Figure 6.51 A *pnp* transistor operated in the active region in the common base amplifier configuration.

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}.

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 6.67. 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 6.67,

$$\frac{\delta I_E}{\delta V_{EB}} = \frac{e}{kT} I_E$$

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 ac input resistance r_e seen by the source v_{eb}.

Small signal input resistance

$$r_e = \frac{\delta V_{EB}}{\delta I_E} = \frac{kT}{eI_E} = \frac{25}{I_E(\text{mA})} \qquad [6.68]$$

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

$$A_V = \frac{v_{cb}}{v_{eb}} = \frac{R_C}{r_e} \qquad [6.69]$$

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 6.20

A COMMON BASE AMPLIFIER Consider a *pnp* Si BJT that has been connected as in Figure 6.51. The BJT has a $\beta = 135$ and has been biased to operate with a 10 mA collector current. What is the small-signal input resistance? What is the required R_C that will provide a voltage gain of 100? What is the base current? What should be the V_{CC} in Figure 6.51? Suppose $V_{CC} = -6$ V, what is the largest swing in the output voltage V_{CB} in Figure 6.51 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 6.68,

$$r_e = \frac{25}{I_E\,(\text{mA})} = \frac{25}{10} = 2.5\ \Omega$$

The voltage gain A_V from Equation 6.69 is

$$A_V = \frac{R_C}{r_e} \qquad \text{or} \qquad 100 = \frac{R_C}{2.5\ \Omega}$$

so a gain of 100 requires $R_C = 250\ \Omega$.

$$\text{Base current } I_B = \frac{I_C}{\beta} = \frac{10\text{ mA}}{135} = 0.074\text{ mA} \qquad \text{or} \qquad 74\ \mu\text{A}$$

There is a dc voltage across R_C given by $I_C R_C = (0.010\text{ A})(250\ \Omega) = 2.5$ 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 V + 2.5 V = -3.5 V. As we increase the signal v_{eb}, V_{EB} and hence I_C increase until the collector point C becomes nearly zero,[18] that is, $V_{CB} = 0$, which occurs when I_C is maximum at $I_{C\max} = |V_{CC}|/R_C$ or 24 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 -3.5 V to 0 V (for increasing input until $I_C = I_{C\max}$), or from -3.5 to -6 V (for decreasing input until $I_C = 0$).

6.11.3 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 6.52a. 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 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

[18] Various saturation effects are ignored in this approximate discussion.

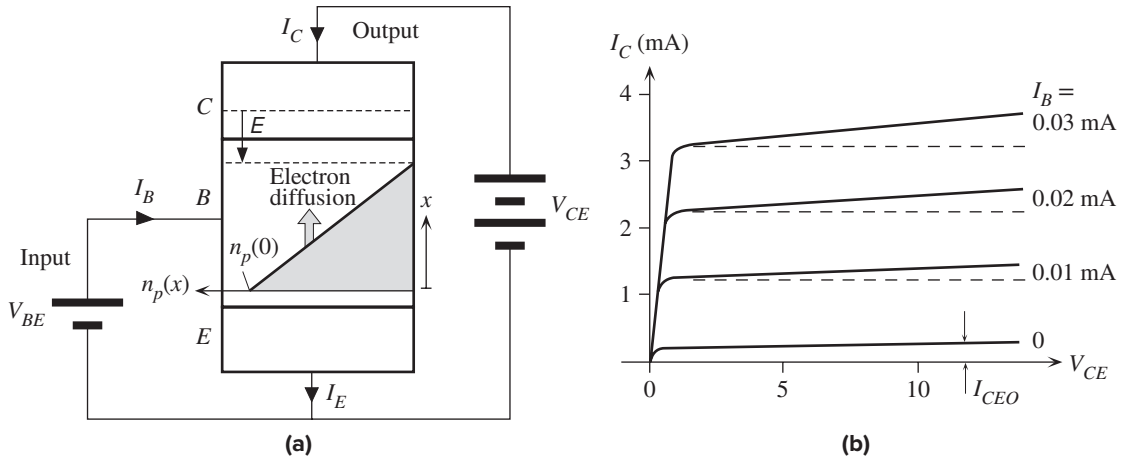


Figure 6.52 (a) An *npn* transistor operated in the active region in the common emitter configuration. The input current is the current that flows between V_{BE} and the base which is I_B. (b) DC *I–V* characteristics of the *npn* bipolar transistor in the CE configuration. (Exaggerated to highlight various effects.)

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 6.52b. The characteristics can be readily understood by Equations 6.65 and 6.66. We should note that, in practice, we are essentially adjusting V_{BE} to obtain the desired I_B because, by Equation 6.66,

$$I_B = (1 - \alpha)I_E - I_{CBO}$$

and I_E depends on V_{BE} via Equation 6.67.

Increasing I_B requires increasing V_{BE}, which increases I_C. Using Equations 6.65 and 6.66, we can obtain I_C in terms of I_B alone,

Active region collector current

$$I_C = \beta I_B + \frac{1}{(1 - \alpha)} I_{CBO}$$

or

$$I_C = \beta I_B + I_{CEO} \qquad [6.70]$$

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 6.70 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 6.50. 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 ($\approx \tau_h/\tau_t$).

When V_{CE} is less than V_{BE}, the collector junction becomes forward biased and Equation 6.70 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.**

6.11.4 LOW-FREQUENCY SMALL-SIGNAL MODEL

The *npn* bipolar transistor in the CE (common emitter) amplifier configuration is shown in Figure 6.53. 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. 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,

$$I_E = I_{EO} \exp\left(\frac{eV_{BE}}{kT}\right) \quad [6.71]$$

Emitter current and V_{BE}

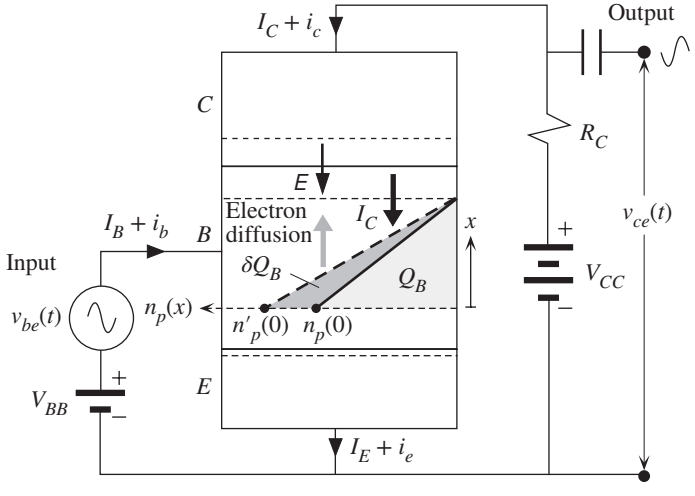


Figure 6.53 An *npn* transistor operated in the active region in the common emitter amplifier configuration.

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}.

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 $v_{be} = \delta V_{BE}$, $i_b = \delta I_B$, $i_e = \delta I_E$, $i_c = \delta I_C$. Then from Equation 6.70 we see that $\delta I_C = \beta\ \delta I_B$, so $i_c = \beta i_b$. Since $\alpha \approx 1$, $i_e \approx i_c$.

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

CE input resistance

$$r_{be} = \frac{v_{be}}{i_b} = \frac{\delta V_{BE}}{\delta I_B} \approx \beta \frac{\delta V_{BE}}{\delta I_E} = \frac{\beta kT}{eI_E} \approx \frac{\beta 25}{I_C(\text{mA})} \qquad [6.72]$$

where we differentiated Equation 6.71.

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

CE voltage gain

$$A_V = \frac{v_{ce}}{v_{be}} = \frac{-R_C i_c}{r_{be} i_b} = \frac{-R_C \beta}{r_{be}} \approx -\frac{R_C I_C(\text{mA})}{25} \qquad [6.73]$$

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

$$g_m = \frac{i_c}{v_{be}} \approx \frac{\delta I_E}{\delta V_{BE}} = \frac{I_E(\text{mA})}{25} = \frac{1}{r_e} \qquad [6.74]$$

The voltage amplification of the CE amplifier is then

Voltage gain

$$A_V = -g_m R_C \qquad [6.75]$$

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 6.54. 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

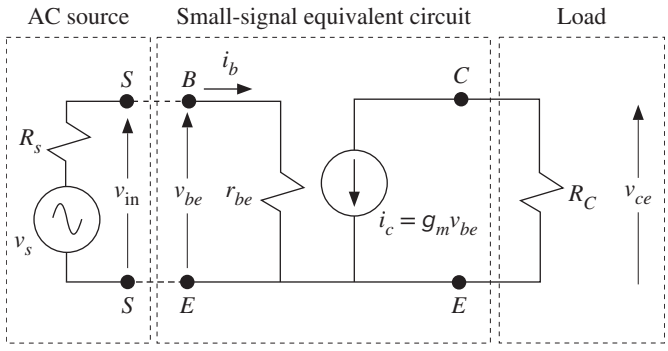


Figure 6.54 Low-frequency small-signal simplified equivalent circuit of the bipolar transistor in the CE configuration with a load resistor R_C in the collector circuit.

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.

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

$$\delta I_C = \beta \delta I_B \qquad \text{or} \qquad i_c = \beta i_b$$

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 6.72 and 6.74, we have

β at low frequencies

$$\beta = g_m r_{be} \tag{6.76}$$

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 6.21

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 6.54), 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 6.54).

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

$$P_{out} = P_{in} A_P = (9.06 \times 10^{-10})(10{,}000) = 9.06 \times 10^{-6}\text{ W} \quad \text{or} \quad 9.06\ \mu\text{W}$$

6.12 JUNCTION FIELD EFFECT TRANSISTOR (JFET)

6.12.1 GENERAL PRINCIPLES

The basic structure of the junction field effect transistor (JFET) with an n-type channel (n-channel) is depicted in Figure 6.55a. 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 6.55a. 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 6.55a. 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 6.55a is useful in explaining the principle of operation as discussed later but does not truly represent the

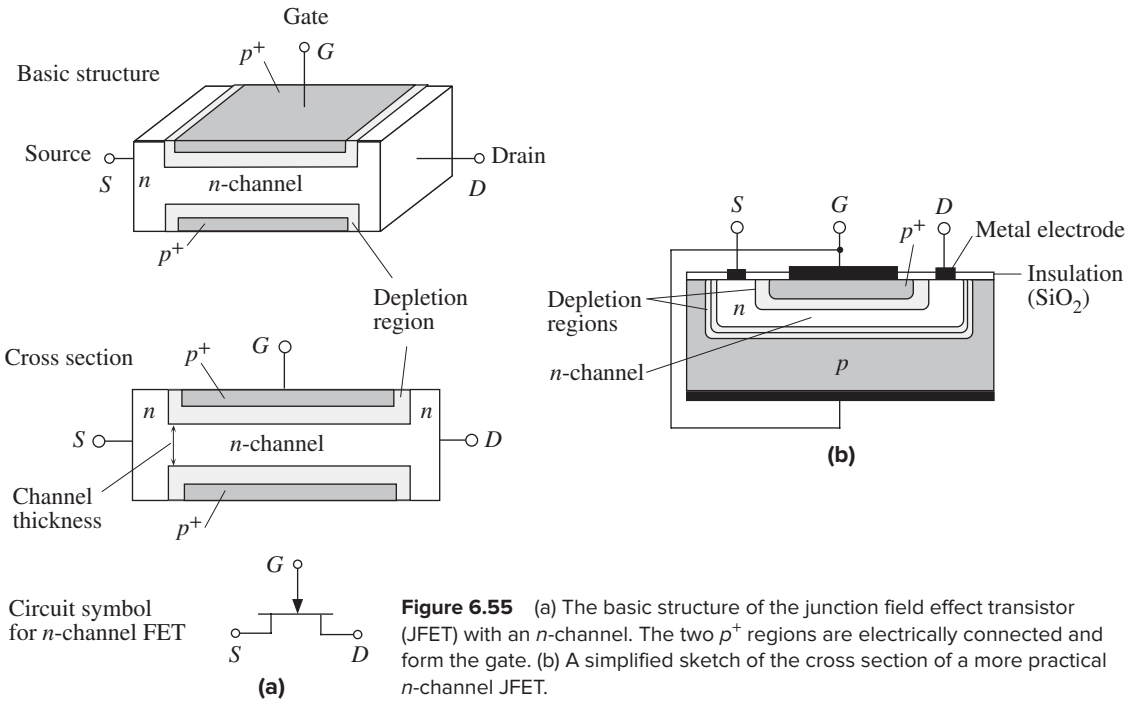


Figure 6.55 (a) The basic structure of the junction field effect transistor (JFET) with an *n*-channel. The two p^+ regions are electrically connected and form the gate. (b) A simplified sketch of the cross section of a more practical *n*-channel JFET.

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 6.55b 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 6.56a. 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 6.56a. 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

$$I_D = \frac{V_{DS}}{R_{AB}}$$

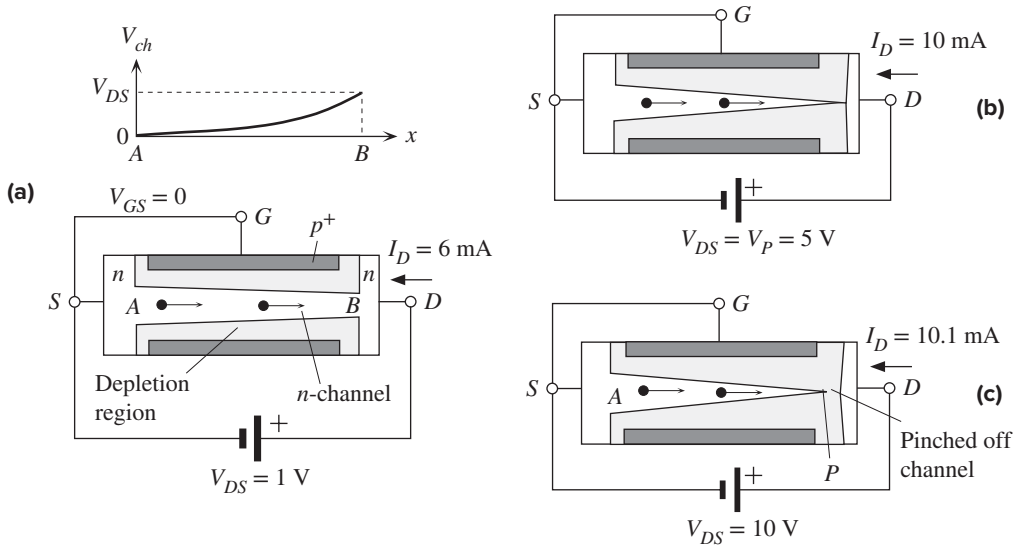


Figure 6.56 (a) The gate and source are shorted ($V_{GS} = 0$) and V_{DS} is small. (b) V_{DS} has increased to a value that allows the two depletion layers to just touch, when $V_{DS} = V_P (= 5\text{ V})$ and the p^+n junction voltage at the drain end, $V_{GD} = -V_{DS} = -V_P = -5$ V. (c) V_{DS} is large ($V_{DS} > V_P$), so a short length of the channel is pinched off.

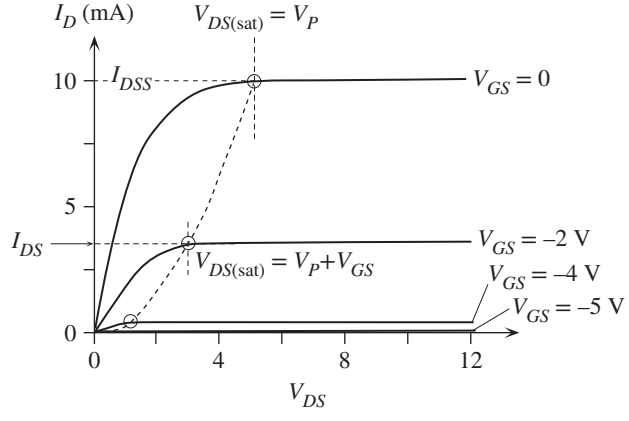


Figure 6.57 Typical I_D versus V_{DS} characteristics of a JFET for various fixed gate voltages V_{GS}.

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 6.57.

As V_{DS} increases further, the depletion layers extend more into the channel and eventually, when $V_{DS} = V_P$ (= 5 V), the two depletion layers around B meet at point P at the drain end of the channel, as depicted in Figure 6.56b. 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 make them just touch at the drain end. Since the actual bias

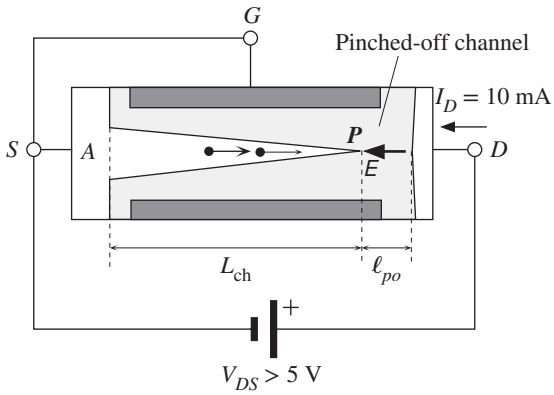


Figure 6.58 The pinched-off channel and conduction for $V_{DS} > V_P$ (= 5 V).

voltage across the p^+n junctions at the drain end (B) is V_{GD}, the pinch-off occurs whenever

$$V_{GD} = -V_P \qquad [6.77]$$

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 6.57, 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 6.58. There is a very strong electric field 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 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 6.58 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 6.57.

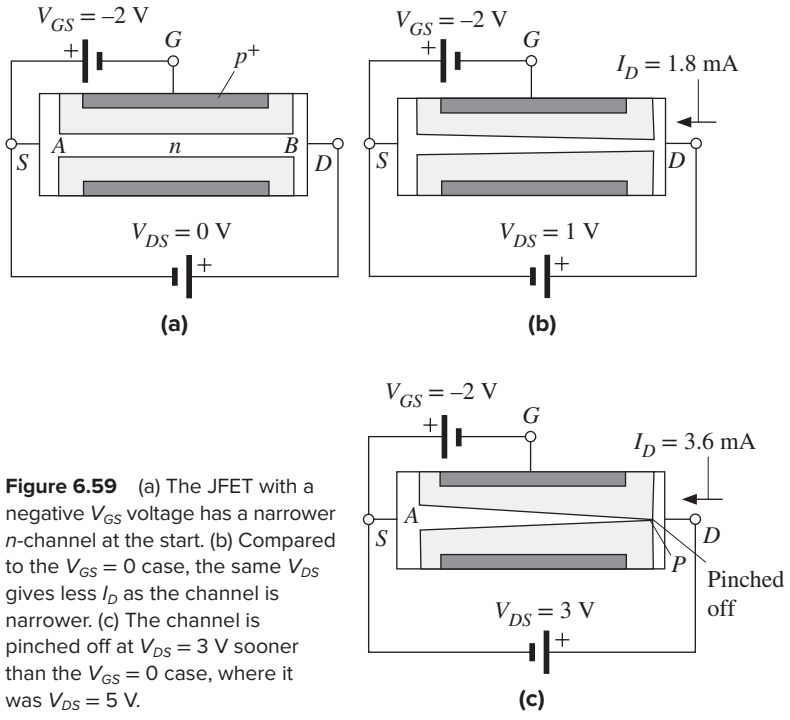


Figure 6.59 (a) The JFET with a negative V_{GS} voltage has a narrower *n*-channel at the start. (b) Compared to the $V_{GS} = 0$ case, the same V_{DS} gives less I_D as the channel is narrower. (c) The channel is pinched off at $V_{DS} = 3$ V sooner than the $V_{GS} = 0$ case, where it was $V_{DS} = 5$ 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 6.59a 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 6.59b, is now smaller than in the $V_{GS} = 0$ case as apparent in Figure 6.57. 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 6.59c. 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 6.56 and 6.59. In the presence of V_{GS}, as apparent from Figure 6.57, the pinch-off occurs at $V_{DS} = V_{DS(\text{sat})}$, and from Equation 6.77.

Pinch-off condition

$$V_{DS(\text{sat})} = V_P + V_{GS} \qquad [6.78]$$

where V_{GS} is a negative voltage (reducing V_P). Beyond pinch-off when $V_{DS} > V_{DS(\text{sat})}$, the point P where the channel is *just pinched* still remains at potential $V_{DS(\text{sat})}$, given by Equation 6.78.

For $V_{DS} > V_{DS(\text{sat})}$, I_D becomes nearly saturated at a value denoted as I_{DS}, which is indicated in Figure 6.57. 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}, the drain current I_D is

$$I_D \approx I_{DS} \approx \frac{V_{DS(\text{sat})}}{R_{AP}(V_{GS})} = \frac{V_P + V_{GS}}{R_{AP}\,(V_{GS})} \qquad V_{DS} > V_{DS(\text{sat})} \qquad [6.79]$$

where $R_{AP}(V_{GS})$ is the effective resistance of the conducting n-channel from A to P (Figure 6.59b), 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 $V_{DS(\text{sat})}$ decreasing with negative V_{GS}, as apparent in Equation 6.79. Figure 6.57 shows the effect of the gate voltage on the I_D versus V_{DS} behavior. The two effects, that from $V_{DS(\text{sat})}$ and that from $R_{AP}(V_{GS})$ in Equation 6.79, 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 6.60. 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 6.57 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 6.79. This is analogous to the BJT in which the collector current I_C is controlled by the base–emitter bias voltage V_{BE}. Figure 6.61a 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 6.61b. This control is only possible if $V_{DS} > V_{DS(\text{sat})}$. 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 $V_{GS(\text{off})}$ 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 6.55a, has an arrow to identify the gate and the pn junction direction.

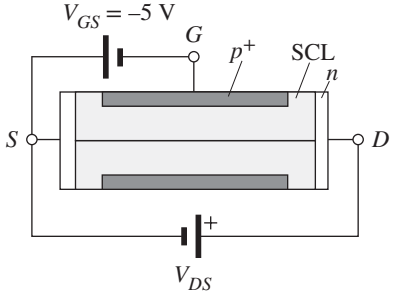


Figure 6.60 When $V_{GS} = -5$ V, the depletion layers close the whole channel from the start, at $V_{DS} = 0$. As V_{DS} is increased, there is a very small drain current, which is the small reverse leakage current due to thermal generation of carriers in the depletion layers.

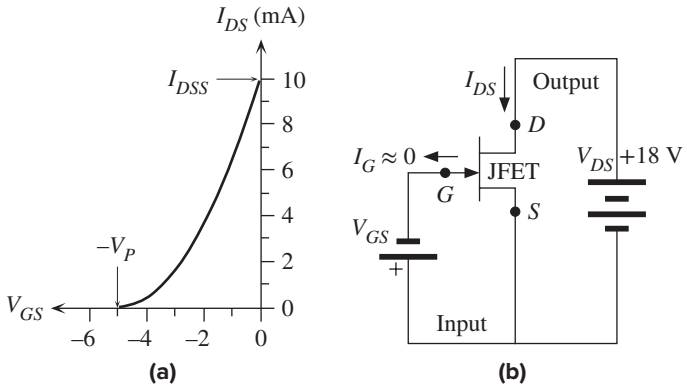


Figure 6.61 (a) Typical I_{DS} versus V_{GS} characteristics of a JFET. (b) The dc circuit where V_{GS} in the gate–source circuit (input) controls the drain current I_{DS} in the drain–source (output) circuit in which V_{DS} is kept constant and large ($V_{DS} > V_P$).

JFET equation beyond pinch-off

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 6.79. It turns out that a simple parabolic dependence seems to represent the data reasonably well,

$$I_{DS} = I_{DSS}\left[1 - \left(\frac{V_{GS}}{V_{GS(\text{off})}}\right)\right]^2 \qquad [6.80]$$

where I_{DSS} is the drain current when $V_{GS} = 0$ (Figure 6.61) and $V_{GS(\text{off})}$ 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 $V_{DS(\text{sat})}$. $V_{GS(\text{off})}$ 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 6.79 and 6.80 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.

6.12.2 JFET AMPLIFIER

The transistor action in the JFET is the control of I_{DS} by V_{GS}, as shown in Figure 6.61. 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

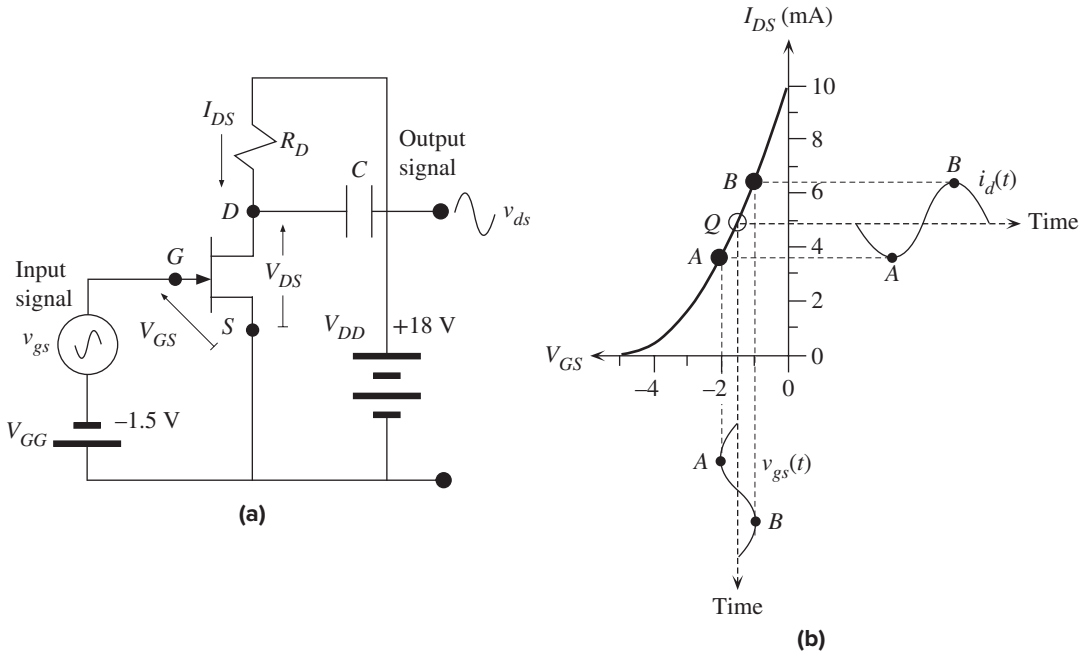


Figure 6.62 (a) Common source (CS) ac amplifier using a JFET. (b) Explanation of how I_D is modulated by the signal v_{gs} in series with the dc bias voltage V_{GG}.

gate to source voltage. With V_{GS} limited to negative voltages, the maximum current in the output circuit can only be I_{DSS}, as shown in Figure 6.61a. The maximum input voltage V_{GS} should therefore give an I_{DS} less than I_{DSS}.

Figure 6.62a 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 6.62b 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 6.62b. Suppose that v_{gs} varies sinusoidally between −0.5 V and +0.5 V. Then, as shown in Figure 6.62b 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 converted to a drain current variation from 3.6 mA to 6.4 mA as indicated in Figure 6.62b. We could have just as easily calculated the drain current from Equation 6.80.

Table 6.6 Voltage and current in the common source amplifier of Figure 6.62a

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 Ω.

Table 6.6 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 6.80, 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

$$V_{DS} = V_{DD} - I_{DS}\, R_D \qquad [6.81]$$

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 6.62a. The capacitor *C* simply blocks the dc. Suppose that $R_D = 2000\ \Omega$ and $V_{DD} = 18$ V, then using Equation 6.81 we can calculate the dc value of V_{DS} and also the minimum and maximum values of V_{DS}, as shown in Table 6.6.

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 6.81, the change in V_{DS} is related to the change in I_{DS} by

$$v_{ds} = -R_D i_d \qquad [6.82]$$

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 6.6 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 $V_{DS(\text{sat})}$ (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,

Definition of JFET transconductance

$$g_m = \frac{dI_{DS}}{dV_{GS}} \approx \frac{\delta I_{DS}}{\delta V_{GS}} = \frac{i_d}{v_{gs}}$$

This transconductance can be found by differentiating Equation 6.80,

JFET transconductance

$$g_m = \frac{dI_{DS}}{dV_{GS}} = -\frac{2I_{DSS}}{V_{GS(\text{off})}}\left[1 - \left(\frac{V_{GS}}{V_{GS(\text{off})}}\right)\right] = -\frac{2[I_{DSS}I_{DS}]^{1/2}}{V_{GS(\text{off})}} \qquad [6.83]$$

The output signal current is

$$i_d = g_m v_{gs}$$

so using Equation 6.82, the small-signal voltage amplification is

Small-signal voltage gain

$$A_V = \frac{v_{ds}}{v_{gs}} = \frac{-R_D(g_m v_{gs})}{v_{gs}} = -g_m R_D \qquad [6.84]$$

Equation 6.84 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 6.22

THE JFET AMPLIFIER Consider the *n*-channel JFET common source amplifier shown in Figure 6.62a. The JFET has an I_{DSS} of 10 mA and a pinch-off voltage V_P of 5 V as in Figure 6.62b. 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 6.62b. 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 6.80 with $V_{GS(\text{off})} = -V_P = -5$ V:

$$I_{DS} = I_{DSS}\left[1 - \left(\frac{V_{GS}}{V_{GS(\text{off})}}\right)\right]^2 = (10\text{ mA})\left[1 - \left(\frac{-1.5}{-5}\right)\right]^2 = 4.9\text{ mA}$$

The transconductance at this dc current (at Q) is given by Equation 6.83,

$$g_m = -\frac{2(I_{DSS}I_{DS})^{1/2}}{V_{GS(\text{off})}} = -\frac{2[(10\times10^{-3})(4.9\times10^{-3})]^{1/2}}{-5} = 2.8\times10^{-3}\ \text{A/V}$$

The voltage amplification of small signals about point Q is

$$A_V = -g_m R_D = -(2.8\times10^{-3})(2000) = -5.6$$

This turns out to be the same as the peak-to-peak voltage amplification we calculated in Table 6.6. When the input ac signal v_{gs} varies between −0.5 and +0.5 V, as in Table 6.6, 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.**

6.13 METAL-OXIDE-SEMICONDUCTOR FIELD EFFECT TRANSISTOR (MOSFET)

6.13.1 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 6.63a. 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 $E = V/L$. 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 E displaces some of the conduction electrons to the surface to form $-Q$. In the top plate E displaces some of the 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 2×10^{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.

The same is not true when one of the electrodes is a semiconductor, as shown in Figure 6.63b where the "capacitor" now is of a **metal-insulator-semiconductor** (MOS) device. Suppose that we replace the lower metal in Figure 6.63a with a p-type semiconductor with an acceptor concentration of 10^{15} cm^{-3}. The number of acceptor atoms on the surface[19] is 1×10^{10} cm^{-2}. We may assume that at room temperature

[19] Surface concentration of atoms (atoms per unit area) can be found from $n_{\text{surf}} \approx (n_{\text{bulk}})^{2/3}$.

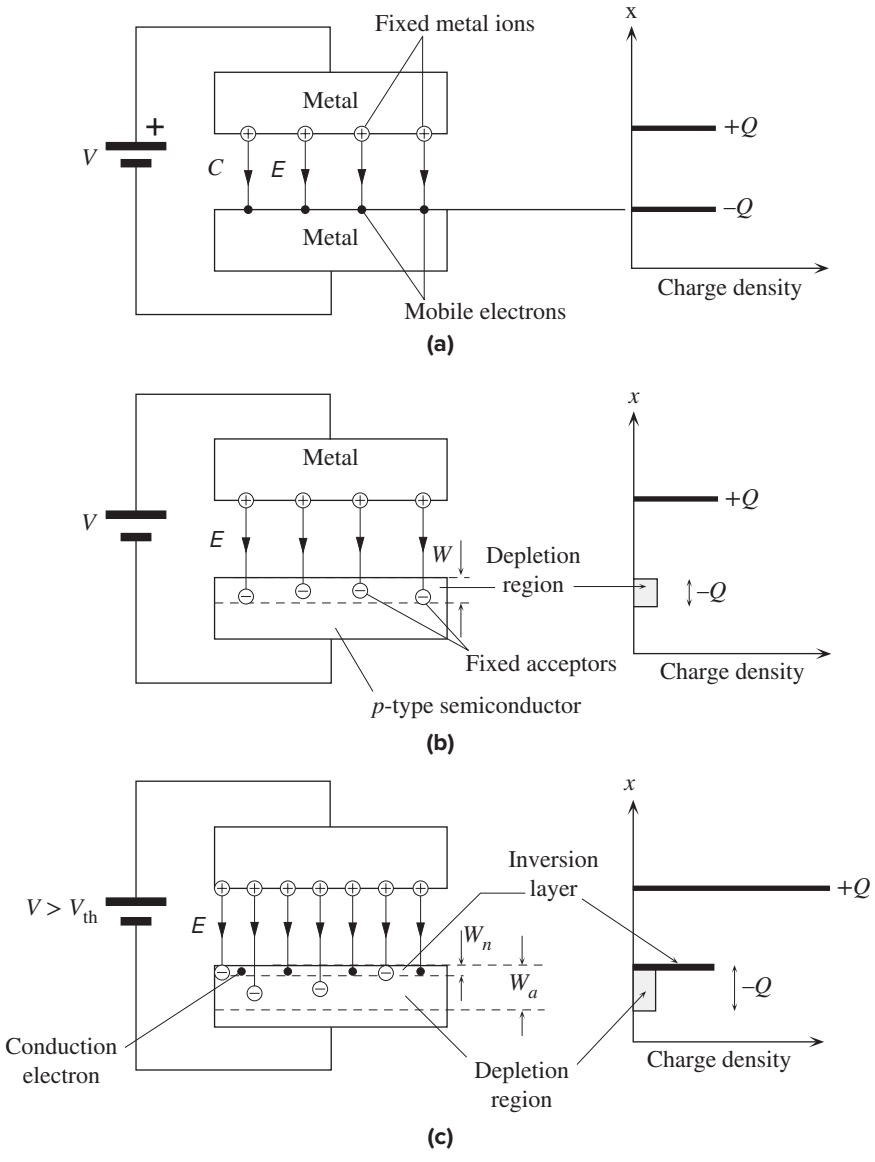


Figure 6.63 The field effect. (a) In a metal-air-metal capacitor, all the charges reside on the surface. (b) Illustration of field penetration into a *p*-type semiconductor. (c) As the field increases, eventually when $V > V_{th}$, an inversion layer is created near the surface in which there are conduction electrons.

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 region still has p-type characteristics as holes are in the majority.

If the voltage increases further, $-Q$ also increases in magnitude, 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 6.63c. 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 EHPs 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 $V > V_{\text{th}}$, 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 of fixed negative charge remain constant. The insulator between the metal and the semiconductor, that is, a vacuum in Figure 6.63, is typically SiO_2 in many devices.

6.13.2 ENHANCEMENT MOSFET

Figure 6.64 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 a metal 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

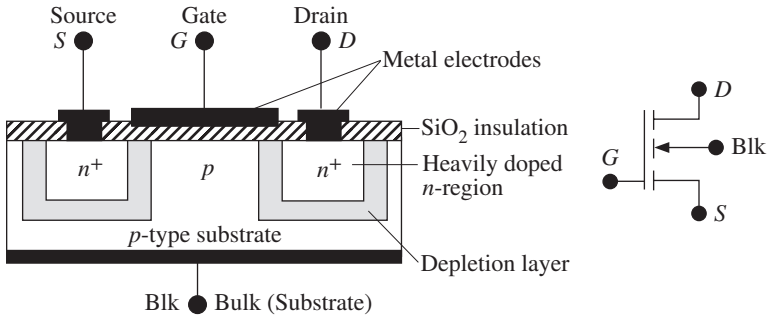


Figure 6.64 The basic structure of the enhancement MOSFET and its circuit symbol.

to the p-type Si substrate (or the bulk), which in many devices is connected to the source terminal as shown in Figure 6.64. 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.

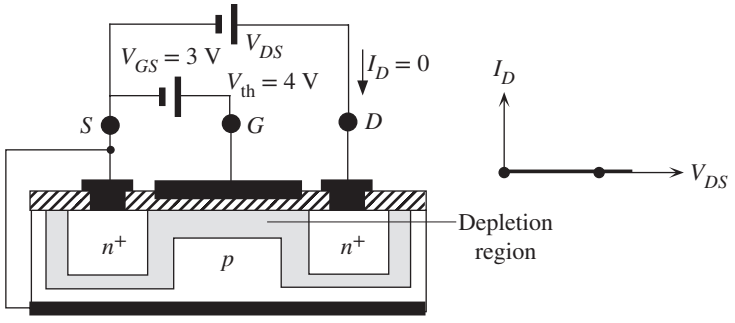
When a positive voltage less than V_{th} is applied to the gate, $V_{GS} < V_{\text{th}}$, as shown in Figure 6.65a, the p-type semiconductor under the gate develops a depletion layer as a result of the expulsion of holes into the bulk, just as in Figure 6.63b. Since S and D are isolated by a low-conductivity 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 6.65b. 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 $R_{n\text{-ch}}$:

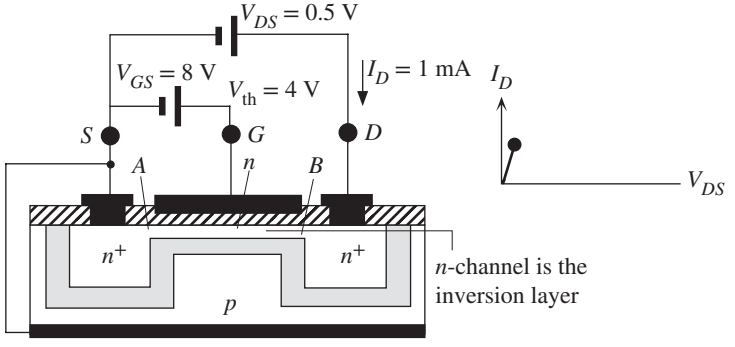
$$I_D = \frac{V_{DS}}{R_{n\text{-ch}}} \qquad [6.85]$$

Thus, I_D initially increases with V_{DS} almost linearly, as shown in Figure 6.65b.

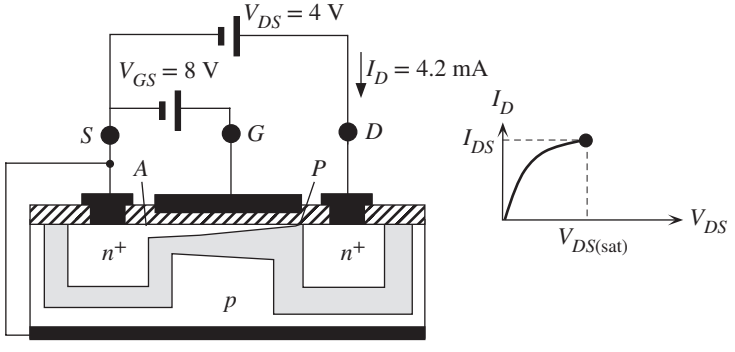
The voltage variation along the channel is from zero at A (source end) to V_{DS} at B (drain end). The gate to the n-channel voltage is then V_{GS} at A and $V_{GD} = V_{GS} - V_{DS}$ at B. Thus point A depends only on V_{GS} and remains undisturbed by V_{DS}. As V_{DS} increases, the voltage at B (V_{GD}) 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



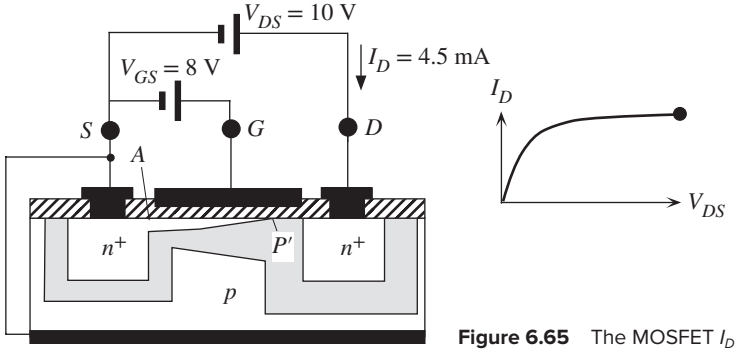
(a) Below threshold $V_{GS} < V_{th}$ and $V_{DS} > 0$



(b) Above threshold $V_{GS} > V_{th}$ and $V_{DS} < V_{DS(sat)}$



(c) Above threshold $V_{GS} > V_{th}$ and saturation, $V_{DS} = V_{DS(sat)}$



(d) Above threshold $V_{GS} > V_{th}$ and saturation region, $V_{DS} > V_{DS(sat)}$

Figure 6.65 The MOSFET I_D versus V_{DS} characteristics.

Figure 6.65c. The n-channel becomes pinched off at this point P. This occurs when $V_{DS} = V_{DS(\text{sat})}$, satisfying

$$V_{GD} = V_{GS} - V_{DS(\text{sat})} = V_{\text{th}} \qquad [6.86]$$

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 $V_{DS(\text{sat})}$, 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 6.65d. 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 I_D is then nearly saturated at I_{DS},

$$I_D \approx I_{DS} \approx \frac{V_{DS(\text{sat})}}{R_{AP'n\text{-ch}}} \qquad V_{DS} > V_{DS(\text{sat})} \qquad [6.87]$$

As $V_{DS(\text{sat})}$ 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 6.66a. It can be seen that there is only a slight increase in I_{DS} with V_{DS} beyond $V_{DS(\text{sat})}$. The I_{DS} versus V_{GS} when $V_{DS} > V_{DS(\text{sat})}$ characteristics are shown in Figure 6.66b. It is apparent that as long as $V_{DS} > V_{DS(\text{sat})}$, 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

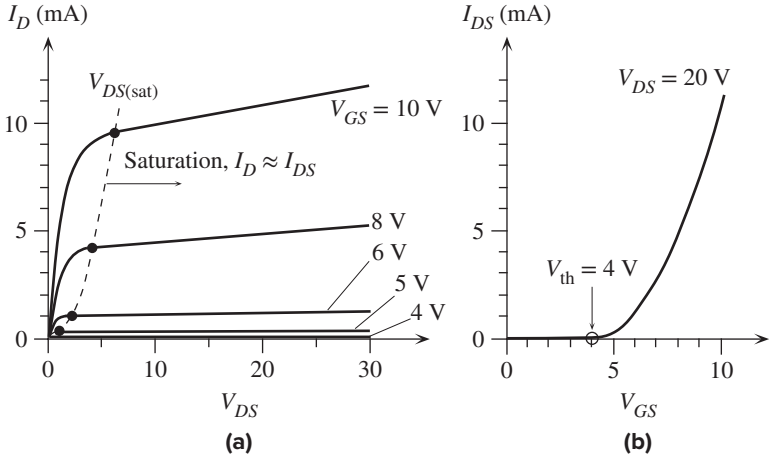


Figure 6.66 (a) Typical I_D versus V_{DS} characteristics of an enhancement MOSFET ($V_{\text{th}} = 4$ V) for various fixed gate voltages V_{GS}. (b) Dependence of I_{DS} on V_{GS} at a given V_{DS} ($>V_{DS(\text{sat})}$).

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 $V_{DS} > V_{DS(\text{sat})}$) 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 $V_{GS} > V_{\text{th}}$ so I_{DS} exists only when $V_{GS} > V_{\text{th}}$,

Enhancement NMOSFET

$$I_{DS} = K(V_{GS} - V_{\text{th}})^2 \qquad [6.88]$$

where K is a constant. For an ideal MOSFET, it can be expressed as

Enhancement NMOSFET constant

$$K = \frac{Z\mu_e\varepsilon}{2Lt_{\text{ox}}} \qquad [6.89]$$

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 ($\varepsilon_r\varepsilon_o$) and thickness of the oxide insulation under the gate. According to Equation 6.88, I_{DS} is independent of V_{DS}. The shallow slopes of the I_D versus V_{DS} lines beyond $V_{DS(\text{sat})}$ in Figure 6.66a can be accounted for by writing Equation 6.88 as

Enhancement NMOSFET

$$I_{DS} = K(V_{GS} - V_{\text{th}})^2(1 + \lambda V_{DS}) \qquad [6.90]$$

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.

The drift mobility μ_e in Equation 6.89 represents the drift of electrons in the channel near the surface of the semiconductor. This region also has the field from the gate penetrating into it as well as a longitudinal field along the channel. μ_e is not the same as the drift mobility in the bulk of p-Si but depends on the field penetrating into the channel, and defects and dopants in this region, especially near the semiconductor–oxide interface. μ_e is therefore a field effect mobility and should be viewed as an *effective mobility in the channel*.

EXAMPLE 6.23 **THE ENHANCEMENT NMOSFET** A particular discrete 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. Its threshold voltage is 4 V. Estimate the drain current when $V_{GS} = 8$ V and $V_{DS} = 20$ V, given $\lambda = 0.01$. The effective electron drift mobility μ_e is roughly 700 $\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$.

SOLUTION

Since $V_{DS} > V_{\text{th}}$, we can assume that the drain current is saturated and we can use the I_{DS} versus V_{GS} relationship in Equation 6.90,

$$I_{DS} = K(V_{GS} - V_{\text{th}})^2(1 + \lambda V_{DS})$$

where the constant K is given by Equation 6.89

$$K = \frac{Z\mu_e\varepsilon_r\varepsilon_o}{2Lt_{ox}} = \frac{(50 \times 10^{-6})(700 \times 10^{-4})(3.9 \times 8.85 \times 10^{-12})}{2(10 \times 10^{-6})(450 \times 10^{-10})} = 0.000134 \text{ AV}^{-1}$$

When $V_{GS} = 8$ V and $V_{DS} = 20$ V, with $\lambda = 0.01$, from Equation 6.90,

$$I_{DS} = 0.000134(8 - 4)^2[1 + (0.01)(20)]$$
$$= 0.0026 \text{ A} \quad \text{or} \quad 2.6 \text{ mA}$$

6.13.3 THRESHOLD VOLTAGE

The threshold voltage is an important parameter in MOSFET devices. Its control in device fabrication is therefore essential. Figure 6.67a 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 6.67a. The voltage drop across the MOS structure, however, 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 $E = -dV/dx$, as shown in Figure 6.67a. The field terminates on both electrons in the

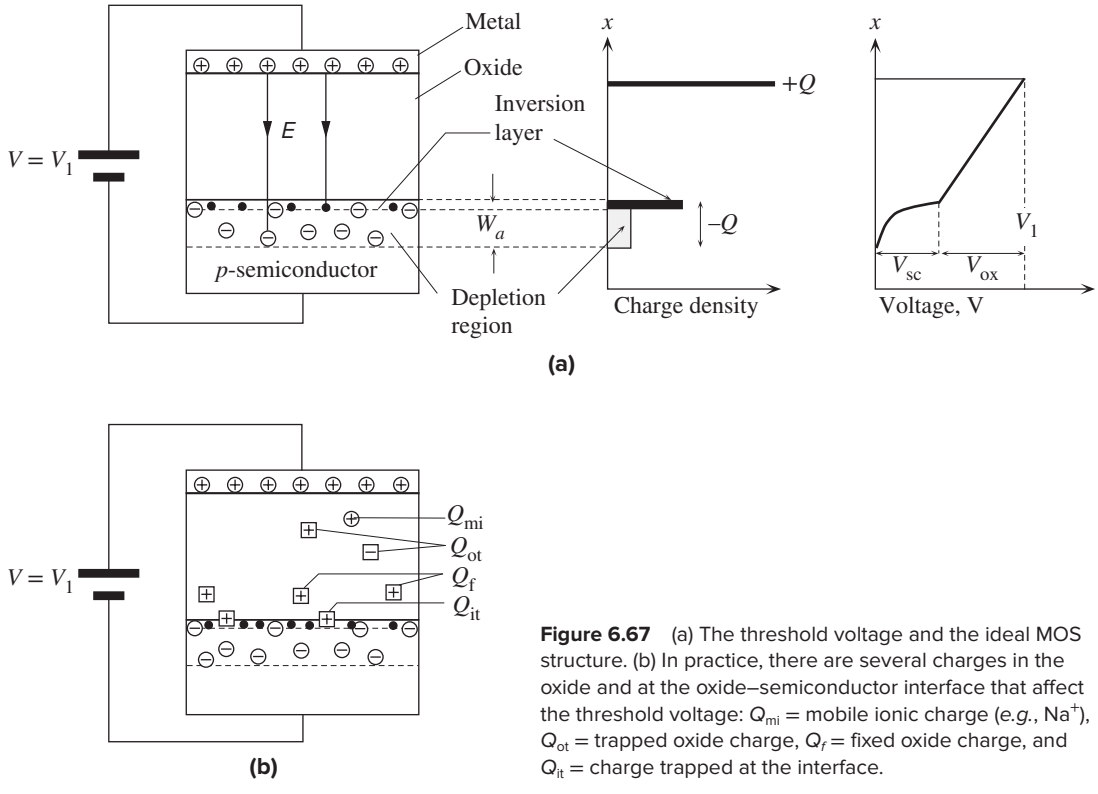


Figure 6.67 (a) The threshold voltage and the ideal MOS structure. (b) In practice, there are several charges in the oxide and at the oxide–semiconductor interface that affect the threshold voltage: Q_{mi} = mobile ionic charge (*e.g.*, Na^+), Q_{ot} = trapped oxide charge, Q_f = fixed oxide charge, and Q_{it} = charge trapped at the interface.

inversion layer and acceptors in W_a, so within the semiconductor 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 6.67a. If 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 6.67b 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 6.67b. 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-trapped charges are represented as Q_{it} in Figure 6.67b. 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 6.67b, 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 6.64, 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.

6.13.4 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 6.64). 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 6.67b (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 chamber 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 6.68. 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

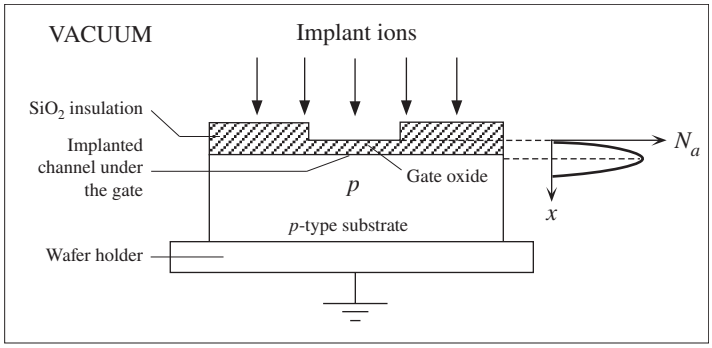


Figure 6.68 Schematic illustration of ion implantation for the control of V_{th}.

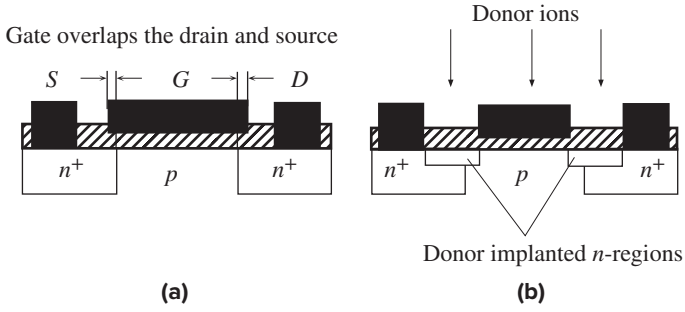


Figure 6.69 (a) There is an overlap of the gate electrode with the source and drain regions and hence additional capacitance between the gate and drain. (b) n^+-type ion implantation extends the drain and source to line up with the gate.

within the implanted region. The defects are almost totally eliminated by annealing the device at an 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 6.69a, 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 6.69b. 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 a metal for the gate electrode. This technique is called **silicon gate technology.** Typically, the silicon for the gate is vacuum deposited (*e.g.*, by chemical vapor deposition using silane

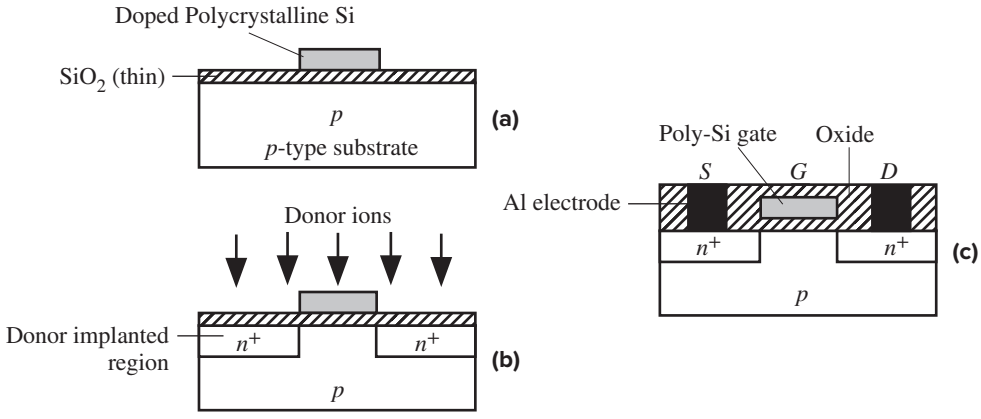


Figure 6.70 The poly-Si gate technology. (a) Poly-Si is deposited onto the oxide, and the areas outside the gate dimensions are etched away. (b) The poly-Si gate acts as a mask during ion implantation of donors to form the n^+ source and drain regions. (c) A simplified schematic sketch of the final poly-Si MOS transistor.

gas) onto the oxide, as shown in Figure 6.70. 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 during fabrication whereas a metal such as Al would melt 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 6.70.

ADDITIONAL TOPICS

6.14 *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 6.71a. 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