CHAPTER

6

Semiconductor Devices

Most diodes are essentially *pn* junctions fabricated by forming a contact between a *p*-type and an *n*-type semiconductor. The junction possesses rectifying properties in that a current in one direction can flow quite easily whereas in the other direction it is limited by a leakage current that is generally very small. A transistor is a three-terminal solid-state device in which a current flowing between two electrodes is controlled by the voltage between the third and one of the other terminals. Transistors are capable of providing current and voltage gains thereby enabling weak signals to be amplified. Transistors can also be used as switches just like electromagnetic relays. Indeed, the whole microcomputer industry is based on transistor switches. The majority of the transistors in microelectronics are of essentially two types: **bipolar junction transistors** (BJTs) and **field effect transistors** (FETs). The appreciation of the underlying principles of the *pn* junction is essential to understanding the operation of not only the bipolar transistor but also a variety of related devices. The central fundamental concept is the **minority carrier injection** as purported by William Shockley in his explanations of the transistor operation. Field effect transistors operate on a totally different principle than BJTs. Their characteristics arise from the effect of the applied field on a conducting channel between two terminals. The last two decades have seen enormous advances and developments in optoelectronic and photonic devices which we now take for granted, the best examples being **light emitting diodes** (LEDs), **semiconductor lasers, photodetectors,** and **solar cells.** Nearly all these devices are based on *pn* junction principles. The present chapter takes the semiconductor concepts developed in Chapter 5 to device level applications, from the basic *pn* junction to heterojunction laser diodes.

6.1 IDEAL *pn* JUNCTION

6.1.1 NO APPLIED BIAS: OPEN CIRCUIT

Consider what happens when one side of a sample of Si is doped *n*-type and the other *p*-type, as shown in Figure 6.1a. We assume that there is an abrupt discontinuity between the *p*- and *n*-regions, which we call the **metallurgical junction** and label as M in Figure 6.1a, where the fixed (immobile) ionized donors and the free electrons (in the conduction band, CB) in the *n*-region and fixed ionized acceptors and holes (in the valence band, VB) in the *p*-region are also shown.

Due to the hole concentration gradient from the *p*-side, where $p = p_{po}$, to the *n*-side, where $p = p_{no}$, holes diffuse toward the right. Similarly the electron concentration

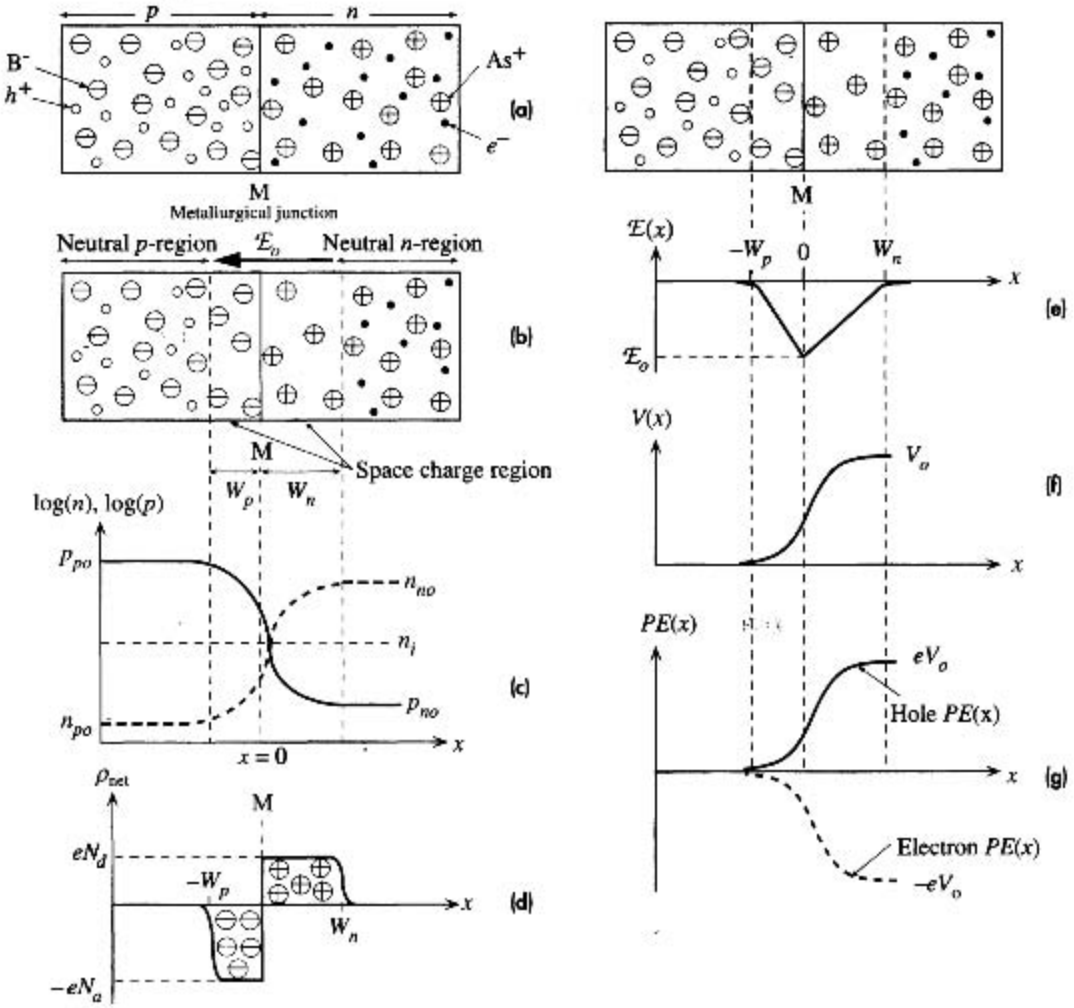


Figure 6.1 Properties of the *pn* junction.

gradient drives the electrons by diffusion toward the left. Holes diffusing and entering the n-side recombine with the electrons in the n-side near the junction. Similarly, electrons diffusing and entering the p-side recombine with holes in the p-side near the junction. The junction region consequently becomes depleted of free carriers in comparison with the bulk p- and n-regions far away from the junction. Note that we must, under equilibrium conditions (*e.g.*, no applied bias or photoexcitation), have $pn = n_i^2$ everywhere. Electrons leaving the n-side near the junction M leave behind exposed positively charged donor ions, say As^+, of concentration N_d. Similarly, holes leaving the p-region near M expose negatively charged acceptor ions, say B^-, of concentration N_a. There is therefore a **space charge layer** (SCL) around M. Figure 6.1b shows the **depletion region,** or the space charge layer, around M, whereas Figure 6.1c illustrates the hole and electron concentration profiles in which the vertical concentration scale is logarithmic. The depletion region is also called the transition region.

It is clear that there is an internal electric field $\mathcal{E}_o$ from positive ions to negative ions, that is, in the $-x$ direction, that tries to drift the holes back into the p-region and electrons back into the n-region. This field drives the holes in the opposite direction to their diffusion. As shown in Figure 6.1b, $\mathcal{E}_o$ imposes a drift force on holes in the $-x$ direction, whereas the hole diffusion flux is in the $+x$ direction. A similar situation also applies for electrons with the electric field attempting to drift the electrons against diffusion from the n-region to the p-region. It is apparent that as more and more holes diffuse toward the right, and electrons toward the left, the internal field around M will increase until eventually an "equilibrium" is reached when the rate of holes diffusing toward the right is just balanced by holes drifting back to the left, driven by the field $\mathcal{E}_o$. The electron diffusion and drift fluxes will also be balanced in equilibrium.

For uniformly doped p- and n-regions, the net space charge density $\rho_{\text{net}}(x)$ 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

$$N_a W_p = N_d W_n \tag{6.1}$$

In Figure 6.1, we arbitrarily assumed that the donor concentration is less than the acceptor concentration, $N_d < N_a$. From Equation 6.1 this implies that $W_n > W_p$; that is, the depletion region penetrates the n-side, the lightly doped side, more than the p-side, the heavily doped side. Indeed, if $N_a \gg N_d$, then the depletion region is almost entirely on the n-side. We generally indicate heavily doped regions with the plus sign as a superscript, that is, p^+.

The electric field $\mathcal{E}(x)$ and the net space charge density $\rho_{\text{net}}(x)$ at a point are related in electrostatics[1] by

Field and net space charge density

$$\frac{d\mathcal{E}}{dx} = \frac{\rho_{\text{net}}(x)}{\varepsilon}$$

[1] This is called **Gauss's law in point** form and comes from Gauss's law in electrostatics. Gauss's law is discussed in Section 7.5.

where $\varepsilon = \varepsilon_o \varepsilon_r$ is the permittivity of the medium and ε_o and ε_r are the absolute permittivity and relative permittivity of the semiconductor material. We can thus integrate $\rho_{net}(x)$ across the diode and thus determine the electric field $\mathcal{E}(x)$, that is,

Field in depletion region

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

The variation of the electric field across the *pn* junction is shown in Figure 6.1e. The negative field means that it is in the $-x$ direction. Note that $\mathcal{E}(x)$ reaches a maximum value $\mathcal{E}_o$ at the metallurgical junction M.

The potential $V(x)$ at any point x can be found by integrating the electric field since by definition $\mathcal{E} = -dV/dx$. Taking the potential on the *p*-side far away from M as zero (we have no applied voltage), which is an arbitrary reference level, then $V(x)$ increases in the depletion region toward the *n*-side, as indicated in Figure 6.1f. Its functional form can be determined by integrating Equation 6.2, which is, of course, a parabola. Notice that on the *n*-side the potential reaches V_o, which is called the **built-in potential.**

The fact that we are considering an abrupt *pn* junction means that $\rho_{net}(x)$ can simply be described by step functions, as displayed in Figure 6.1d. Using the step form of $\rho_{net}(x)$ in Figure 6.1d in the integration of Equation 6.2 gives the electric field at M as

Built-in field

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

where $\varepsilon = \varepsilon_o \varepsilon_r$. We can integrate the expression for $\mathcal{E}(x)$ in Figure 6.1e to evaluate the potential $V(x)$ and thus find V_o by putting in $x = W_n$. The graphical representation of this integration is the step from Figure 6.1e to f. The result is

Built-in voltage

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

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, in equilibrium, Boltzmann statistics[2] demands that the concentrations n_1 and n_2 of carriers at potential energies E_1 and E_2 are related by

$$\frac{n_2}{n_1} = \exp\left[-\frac{(E_2 - E_1)}{kT}\right]$$

where $E = qV$, where q is the charge of the carrier. Considering electrons ($q = -e$), we see from Figure 6.1g that $E = 0$ on the *p*-side far away from M where $n = n_{po}$, and

[2] We use Boltzmann statistics, that is, $n|E| \propto \exp|-E/kT|$, because the concentration of electrons in the conduction band, whether on the *n*-side or *p*-side, is never so large that the Pauli exclusion principle becomes important. As long as the carrier concentration in the conduction band is much smaller than N_c, we can use Boltzmann statistics.

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

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

Boltzmann statistics for electrons

This shows that V_o depends on n_{no} and n_{po} and hence on N_d and N_a. The corresponding equation for hole concentrations is clearly

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

Thus, rearranging Equations 6.5a and b we obtain

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

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

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

Built-in voltage

Clearly V_o has been conveniently related to the dopant and material properties via N_a, N_d, and n_i^2. The built-in voltage (V_o) is the voltage across a *pn* junction, going from *p*- to *n*-type semiconductor, in an open circuit. It is *not* the voltage across the diode, which is made up of V_o as well as the contact potentials at the metal-to-semiconductor junctions at the electrodes. If we add V_o and the contact potentials at the electroded ends, we will find zero.

Once we know the built-in potential from Equation 6.6, we can then calculate the width of the depletion region from Equation 6.4, namely

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

Depletion region width

Notice that the depletion width $W_o \propto V_o^{1/2}$. This results in the capacitance of the depletion region being voltage dependent, as we will see in Section 6.3.

EXAMPLE 6.1

THE BUILT-IN POTENTIALS FOR Ge, Si, AND GaAs *pn* JUNCTIONS A *pn* junction diode has a concentration of 10^{16} acceptor atoms cm^{-3} on the p-side and a concentration of 10^{17} donor atoms cm^{-3} on the n-side. What will be the built-in potential for the semiconductor materials Ge, Si, and GaAs?

SOLUTION

The built-in potential is given by Equation 6.6, which requires the knowledge of the intrinsic concentration for each semiconductor. From Chapter 5 we can tabulate the following

at 300 K:

Semiconductor	E_g (eV)	n_i (cm^{-3})	V_o (V)
Ge	0.7	2.40×10^{13}	0.37
Si	1.1	1.0×10^{10}	0.78
GaAs	1.4	2.1×10^{6}	1.21

Using

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

for Si with $N_d = 10^{17}$ cm^{-3} and $N_a = 10^{16}$ cm^{-3}, $kT/e = 0.0259$ V at 300 K, and $n_i = 1.0 \times 10^{10}$ cm^{-3}, we obtain

$$V_o = (0.0259 \text{ V}) \ln\left[\frac{(10^{17})(10^{16})}{(1.0 \times 10^{10})^2}\right] = 0.775 \text{ V}$$

The results for all three semiconductors are summarized in the last column of the table in this example.

EXAMPLE 6.2

THE p^+n JUNCTION Consider a p^+n junction, which has a heavily doped p-side relative to the n-side, that is, $N_a \gg N_d$. Since the amount of charge Q on both sides of the metallurgical junction must be the same (so that the junction is overall neutral)

$$Q = eN_a W_p = eN_d W_n$$

it is clear that the depletion region essentially extends into the n-side. According to Equation 6.7, when $N_d \ll N_a$, the width is

$$W_o = \left[\frac{2\varepsilon V_o}{eN_d}\right]^{1/2}$$

What is the depletion width for a pn junction Si diode that has been doped with 10^{18} acceptor atoms cm^{-3} on the p-side and 10^{16} donor atoms cm^{-3} on the n-side?

SOLUTION

To apply the above equation for W_o, we need the built-in potential, which is

$$V_o = \left(\frac{kT}{e}\right) \ln\left(\frac{N_d N_a}{n_i^2}\right) = (0.0259 \text{ V}) \ln\left[\frac{(10^{16})(10^{18})}{(1.0 \times 10^{10})^2}\right] = 0.835 \text{ V}$$

Then with $N_d = 10^{16}$ cm^{-3}, that is, 10^{22} m^{-3}, $V_o = 0.835$ V, and $\varepsilon_r = 11.9$ in the equation for W_o

$$W_o = \left[\frac{2\varepsilon V_o}{eN_d}\right]^{1/2} = \left[\frac{2(11.9)(8.85 \times 10^{-12})(0.835)}{(1.6 \times 10^{-19})(10^{22})}\right]^{1/2}$$

$$= 3.32 \times 10^{-7} \text{ m} \quad \text{or} \quad 0.33 \ \mu\text{m}$$

Nearly all of this region (99 percent of it) is on the n-side.

EXAMPLE 6.3

BUILT-IN VOLTAGE There is a rigorous derivation of the built-in voltage across a *pn* junction. Inasmuch as in equilibrium there is no net current through the *pn* junction, drift of holes due to the built-in field $\mathcal{E}(x)$ must be just balanced by their diffusion due to the concentration gradient dp/dx. We can thus set the total electron and hole current densities (drift + diffusion) through the depletion region to zero. Considering holes alone, from Equation 5.38,

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

The electric field is defined by $\mathcal{E} = -dV/dx$, so substituting we find,

$$-ep\mu_h\,dV - eD_h\,dp = 0$$

We can now use the *Einstein relation* $D_h/\mu_h = kT/e$ to get

$$-ep\,dV - kT\,dp = 0$$

We can integrate this equation. According to Figure 6.1, in the *p*-side, $p = p_{po}$, $V = 0$, and in the *n*-side, $p = p_{no}$, $V = V_o$, thus,

$$\int_0^{V_o} dV + \frac{kT}{e}\int_{p_{po}}^{p_{no}} \frac{dp}{p} = 0$$

that is,

$$V_o + \frac{kT}{e}[\ln(p_{no}) - \ln(p_{po})] = 0$$

giving

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

which is the same as Equation 6.5b and hence leads to Equation 6.6.

6.1.2 FORWARD BIAS: DIFFUSION CURRENT

Consider what happens when a battery is connected across a *pn* junction so that the positive terminal of the battery is attached to the *p*-side and the negative terminal to the *n*-side. Suppose that the applied voltage is V. It is apparent that the negative polarity of the supply will reduce the potential barrier V_o by V, as shown in Figure 6.2a. The reason for this is that the bulk regions outside the depletion width have high conductivities due to plenty of majority carriers in the bulk, in comparison with the depletion region in which there are mainly immobile ions. Thus, the applied voltage drops mostly across the depletion width W. Consequently, V directly opposes V_o and the potential barrier against diffusion is reduced to $(V_o - V)$, as depicted in Figure 6.2b. This has drastic consequences because the probability that a hole will surmount this potential barrier and diffuse to the right now becomes proportional to $\exp[-e(V_o - V)/kT]$. In other words, the applied voltage effectively reduces the built-in potential and hence the built-in field, which acts against diffusion. Consequently many holes can now diffuse across the depletion region and enter the *n*-side. This results in the **injection of excess minority carriers,** holes, into the *n*-region. Similarly, excess electrons can now diffuse toward the *p*-side and enter this region and thereby become injected minority carriers.

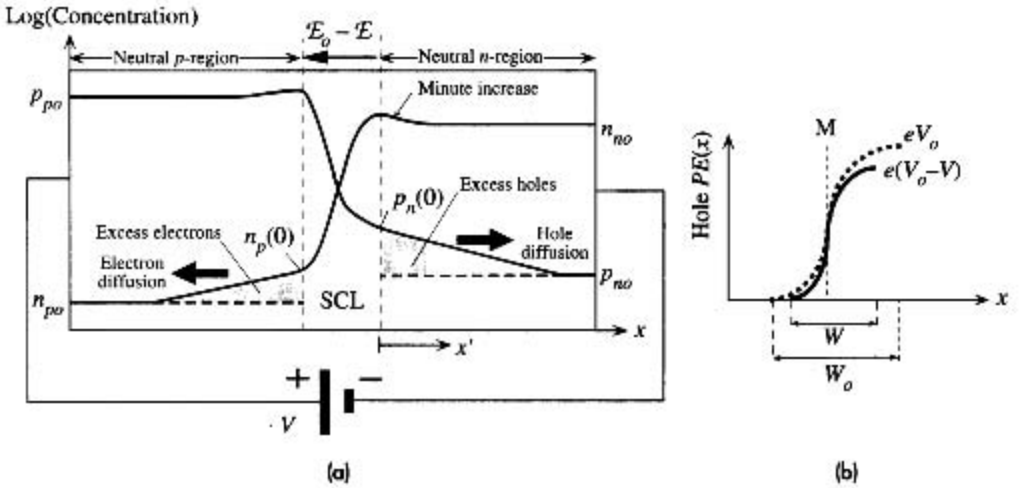


Figure 6.2 Forward-biased *pn* junction and the injection of minority carriers.
(a) Carrier concentration profiles across the device under forward bias.
(b) The hole potential energy with and without an applied bias. *W* is the width of the SCL with forward bias.

The hole concentration

$$p_n(0) = p_n(x' = 0)$$

just outside the depletion region at $x' = 0$ (x' is measured from W_n) is due to the excess of holes diffusing as a result of the reduction in the built-in potential barrier. This concentration $p_n(0)$ is determined by the probability of surmounting the new potential energy barrier $e(V_o - V)$,

$$p_n(0) = p_{po} \exp\left[-\frac{e(V_o - V)}{kT}\right] \qquad [6.8]$$

This follows directly from the Boltzmann equation, by virtue of the hole potential energy rising by $e(V_o - V)$ from $x = -W_p$ to $x = W_n$, as indicated in Figure 6.2b, and at the same time the hole concentration falling from p_{po} to $p_n(0)$. By dividing Equation 6.8 by Equation 6.5b, we obtain the effect of the applied voltage directly, which shows how the voltage V determines the amount of excess holes diffusing and arriving at the *n*-region. Equation 6.8 divided by Equation 6.5b is

Law of the junction

$$p_n(0) = p_{no} \exp\left(\frac{eV}{kT}\right) \qquad [6.9]$$

which is called the **law of the junction.** Equation 6.9 is an important equation that we will use again in dealing with *pn* junction devices. It describes the effect of the applied voltage V on the injected minority carrier concentration just outside the depletion region $p_n(0)$. Obviously, with no applied voltage, $V = 0$ and $p_n(0) = p_{no}$, which is exactly what we expect.

Injected holes diffuse in the n-region and eventually recombine with electrons in this region as there are many electrons in the n-side. Those electrons lost by recombination are readily replenished by the negative terminal of the battery connected to this side. The current due to holes diffusing in the n-region can be sustained because more holes can be supplied by the p-region, which itself can be replenished by the positive terminal of the battery.

Electrons are similarly injected from the n-side to the p-side. The electron concentration $n_p(0)$ just outside the depletion region at $x = -W_p$ is given by the equivalent of Equation 6.9 for electrons, that is,

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

Law of the junction

In the p-region, the injected electrons diffuse toward the positive terminal looking to be collected. As they diffuse they recombine with some of the many holes in this region. Those holes lost by recombination can be readily replenished by the positive terminal of the battery connected to this side. The current due to the diffusion of electrons in the p-side can be maintained by the supply of electrons from the n-side, which itself can be replenished by the negative terminal of the battery. It is apparent that an electric current can be maintained through a *pn* junction under forward bias, and that the current flow, surprisingly, seems to be due to the **diffusion of minority carriers.** There is, however, some drift of majority carriers as well.

If the lengths of the p- and n-regions are longer than the minority carrier diffusion lengths, then we will be justified to expect the hole concentration $p_n(x')$ on the n-side to fall exponentially toward the thermal equilibrium value p_{no}, that is,

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

Excess minority carrier profile

where

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

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

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

that is,

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

[3] This is simply the solution of the continuity equation in the absence of an electric field, which is discussed in Chapter 5. Equation 6.11 is identical to Equation 5.48.

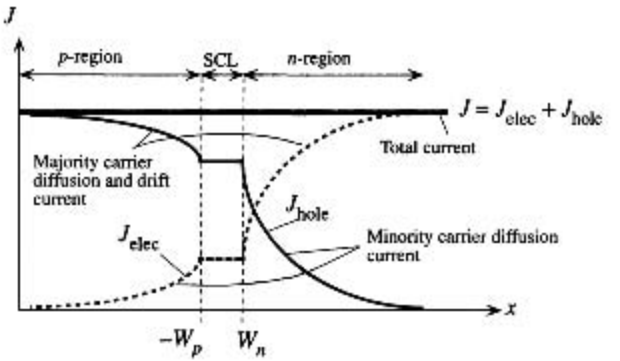


Figure 6.3 The total current anywhere in the device is constant. Just outside the depletion region, it is due to the diffusion of minority carriers.

Although this equation shows that the hole diffusion current depends on location, the total current at any location is the sum of hole and electron contributions, which is independent of x, as indicated in Figure 6.3. The decrease in the minority carrier diffusion current with x' is made up by the increase in the current due to the drift of the majority carriers, as schematically shown in Figure 6.3. The field in the neutral region is not totally zero but a small value, just sufficient to drift the huge number of majority carriers there.

At $x' = 0$, just outside the depletion region, the hole diffusion current is

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

We can now use the law of the junction to substitute for $\Delta p_n(0)$ in terms of the applied voltage V. Writing

$$\Delta p_n(0) = p_n(0) - p_{no} = p_{no}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

and substituting in $J_{D,\text{hole}}$, we get

Hole diffusion current in n-side

$$J_{D,\text{hole}} = \left(\frac{eD_h p_{no}}{L_h}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

Thermal equilibrium hole concentration p_{no} is related to the donor concentration by

$$p_{no} = \frac{n_i^2}{n_{no}} = \frac{n_i^2}{N_d}$$

Thus,

Hole diffusion current in n-side

$$J_{D,\text{hole}} = \left(\frac{eD_h n_i^2}{L_h N_d}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

There is a similar expression for the electron diffusion current density $J_{D,\text{elec}}$ in the p-region. We will assume (quite reasonably) that the electron and hole currents do not change across the depletion region because, in general, the width of this region is narrow (reality is not quite like the schematic sketches in Figures 6.2 and 6.3). The electron

current at $x = -W_p$ is the same as that at $x = W_n$. The total current density is then simply given by $J_{D,\text{hole}} + J_{D,\text{elec}}$, that is,

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

or

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

Ideal diode (Shockley) equation

This is the familiar diode equation with

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

Reverse saturation current

It is frequently called the **Shockley equation.** The constant J_{so} depends not only on the doping, N_d and N_a, but also on the material via n_i, D_h, D_e, L_h, and L_e. It is known as the **reverse saturation current density,** as explained below. Writing

$$n_i^2 = (N_c N_v) \exp\left(-\frac{eV_g}{kT}\right)$$

Intrinsic concentration

where $V_g = E_g/e$ is the bandgap energy expressed in volts, we can write Equation 6.12 as

$$J = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a}\right)\left[(N_c N_v) \exp\left(-\frac{eV_g}{kT}\right)\right]\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

that is,

$$J = J_1 \exp\left(-\frac{eV_g}{kT}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

or

$$J = J_1 \exp\left[\frac{e(V - V_g)}{kT}\right] \quad \text{for} \quad \frac{eV}{kT} \gg 1 \qquad [6.13]$$

Diode current and bandgap energy

where

$$J_1 = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a}\right)(N_c N_v)$$

is a new constant.

The significance of Equation 6.13 is that it reflects the dependence of *I–V* characteristics on the bandgap (via V_g), as displayed in Figure 6.4 for the three important semiconductors, Ge, Si, and GaAs. Notice that the voltage across the *pn* junction for an appreciable current of say ~ 0.1 mA is about 0.2 V for Ge, 0.6 V for Si, and 0.9 V for GaAs.

The diode equation, Equation 6.12, was derived by assuming that the lengths of the *p* and *n* regions outside the depletion region are long in comparison with the diffusion lengths L_h and L_e. Suppose that ℓ_p is the length of the *p*-side outside the depletion region

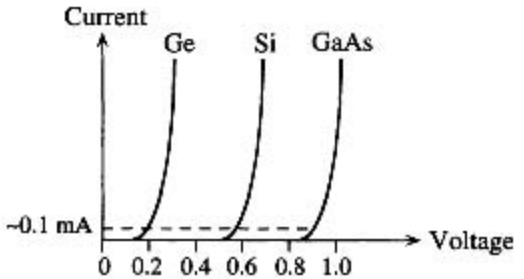


Figure 6.4 Schematic sketch of the *I–V* characteristics of Ge, Si, and GaAs *pn* junctions.

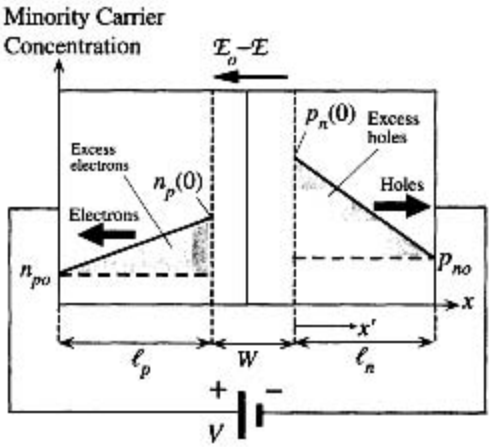


Figure 6.5 Minority carrier injection and diffusion in a short diode.

and ℓ_n is that of the n-side outside the depletion region. If ℓ_p and ℓ_n are shorter than the diffusion lengths L_e and L_h, respectively, then we have what is called a **short diode** and consequently the minority carrier distribution profiles fall almost linearly with distance from the depletion region, as depicted in Figure 6.5. This can be readily proved by solving the continuity equation, but an intuitive explanation makes it clear. At $x' = 0$, the minority carrier concentration is determined by the law of the junction, whereas at the battery terminal there can be no excess carriers as the battery will simply collect these. Since the length of the neutral region is shorter than the diffusion length, there are practically no holes lost by recombination, and therefore the hole flow is expected to be uniform across ℓ_n. This can be so only if the driving force for diffusion, the concentration gradient, is linear.

The excess minority carrier gradient is

$$\frac{d\Delta p_n(x')}{dx'} = -\frac{[p_n(0) - p_{no}]}{\ell_n}$$

The current density $J_{D,\text{hole}}$ due to the injection and diffusion of holes in the n-region as a result of forward bias is

$$J_{D,\text{hole}} = -eD_h \frac{d\Delta p_n(x')}{dx'} = eD_h \frac{[p_n(0) - p_{no}]}{\ell_n}$$

We can now use the law of the junction

$$p_n(0) = p_{no} \exp\left(\frac{eV}{kT}\right)$$

for $p_n(0)$ in the above equation and also obtain a similar equation for electrons diffusing in the p-region and then sum the two for the total current J,

Short diode

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

It is clear that this expression is identical to that of a long diode, that is, Equation 6.12, if in the latter we replace the diffusion lengths L_h and L_e by the lengths ℓ_n and ℓ_p of the n- and p-regions outside the SCL.

6.1.3 FORWARD BIAS: RECOMBINATION AND TOTAL CURRENT

So far we have assumed that, under a forward bias, the minority carriers diffusing and recombining in the neutral regions are supplied by the external current. However, some of the minority carriers will recombine in the depletion region. The external current must therefore also supply the carriers lost in the recombination process in the SCL. Consider for simplicity a symmetrical *pn* junction as in Figure 6.6 under forward bias. At the metallurgical junction at the center C, the hole and electron concentrations are p_M and n_M and are equal. We can find the SCL recombination current by considering electrons recombining in the p-side in W_p and holes recombining in the n-side in W_n as shown by the shaded areas ABC and BCD, respectively, in Figure 6.6. Suppose that the **mean hole recombination time** in W_n is τ_h and **mean electron recombination time** in W_p is τ_e. The rate at which the electrons in ABC are recombining is the area ABC (nearly all injected electrons) divided by τ_e. The electrons are replenished by the diode current. Similarly, the rate at which holes in BCD are recombining is the area BCD divided by τ_h. Thus, the recombination current density is

$$J_{\text{recom}} = \frac{eABC}{\tau_e} + \frac{eBCD}{\tau_h}$$

We can evaluate the areas ABC and BCD by taking them as triangles, $ABC \approx \frac{1}{2}W_p n_M$, etc., so that

$$J_{\text{recom}} \approx \frac{e\frac{1}{2}W_p n_M}{\tau_e} + \frac{e\frac{1}{2}W_n p_M}{\tau_h}$$

Under steady-state and equilibrium conditions, assuming a nondegenerate semiconductor, we can use Boltzmann statistics to relate these concentrations to the potential

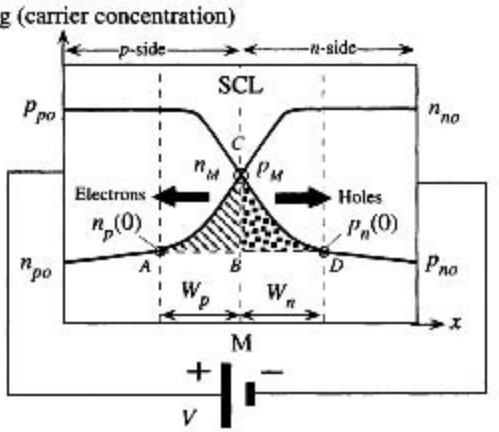


Figure 6.6 Forward-biased *pn* junction and the injection of carriers and their recombination in SCL.

energy. At A, the potential is zero and at M it is $\frac{1}{2}e(V_o - V)$, so

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

Since V_o depends on dopant concentrations and n_i as in Equation 6.6 and further $p_{po} = N_a$, we can simplify this equation to

$$p_M = n_i \exp\left(\frac{eV}{2kT}\right)$$

This means that the recombination current for $V > kT/e$ is given by

Recombination current

$$J_{\text{recom}} = \frac{en_i}{2}\left(\frac{W_p}{\tau_e} + \frac{W_n}{\tau_h}\right)\exp\left(\frac{eV}{2kT}\right) \qquad [6.15]$$

From a better quantitative analysis, the expression for the recombination current can be shown to be[4]

Recombination current

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

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 6.12 and 6.15.

Total diode current = diffusion + recombination

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

This expression is often lumped into a single exponential as

The diode equation

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

where J_o is a new constant and η is an **ideality factor,** which is 1 when the current is due to minority carrier diffusion in the neutral regions and 2 when it is due to recombination in the space charge layer. Figure 6.7 shows typical expected I–V characteristics of pn junction Ge, Si, and GaAs diodes. At the highest currents, invariably, the bulk resistances of the neutral regions limit the current (why?). For Ge diodes, typically $\eta = 1$ and the overall I–V characteristics are due to minority carrier diffusion. In the case of GaAs, $\eta \approx 2$ and the current is limited by recombination in the space charge layer. For Si, typically, η changes from 2 to 1 as the current increases, indicating that both processes play an important role. In the case of heavily doped Si diodes, heavy doping leads to short minority carrier recombination times and the current is controlled by recombination in the space charge layer so that the $\eta = 2$ region extends all the way to the onset of bulk resistance limitation.

[4] This is generally proved in advanced texts.

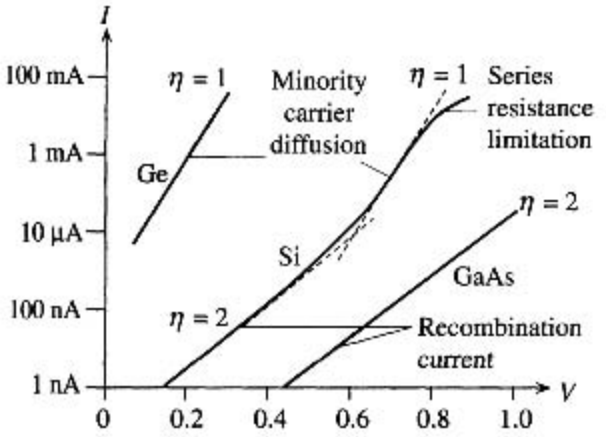


Figure 6.7 Schematic sketch of typical *I–V* characteristics of Ge, Si, and GaAs *pn* junctions as log(*I*) versus *V*.
The slope indicates $e/(\eta kT)$.

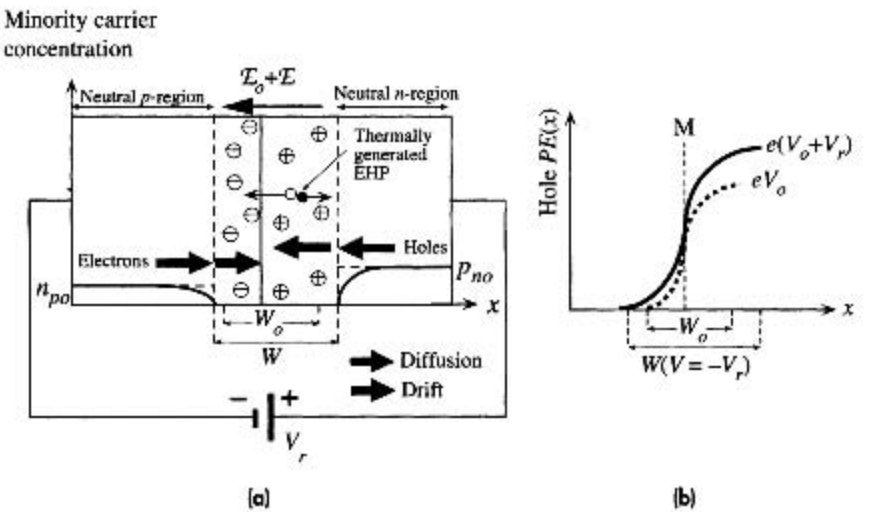


Figure 6.8 Reverse-biased *pn* junction.
(a) Minority carrier profiles and the origin of the reverse current.
(b) Hole *PE* across the junction under reverse bias.

6.1.4 REVERSE BIAS

When a *pn* junction is reverse-biased, as shown in Figure 6.8a, the applied voltage, as before, drops mainly across the depletion region, that is, the space charge layer (SCL), which becomes wider. The negative terminal will attract the holes in the *p*-side to move away from the SCL, which results in more exposed negative acceptor ions and thus a wider SCL. Similarly, the positive terminal will attract electrons away from the SCL, which exposes more positively charged donors. The depletion width on the *n*-side also widens. The movement of electrons in the *n*-region toward the positive battery

terminal cannot be sustained because there is no electron supply to this n-side. The p-side cannot supply electrons to the n-side because it has almost none. However, there is a small reverse current due to two causes.

The applied voltage increases the built-in potential barrier, as depicted in Figure 6.8b. The electric field in the SCL is larger than the built-in internal field $\mathcal{E}_o$. The small number of holes on the n-side near the SCL become extracted and swept by the field across the SCL over to the p-side. This small current can be maintained by the diffusion of holes from the n-side bulk to the SCL boundary.

Assume that the reverse bias $V_r > kT/e = 25$ mV. The hole concentration $p_n(0)$ just outside the SCL is nearly zero by the law of the junction, Equation 6.9, whereas the hole concentration in the bulk (or near the negative terminal) is the equilibrium concentration p_{no}, which is small. There is therefore a small concentration gradient and hence a small hole diffusion current toward the SCL as shown in Figure 6.8a. Similarly, there is a small electron diffusion current from bulk p-side to the SCL. Within the SCL, these carriers are drifted by the field. This minority carrier diffusion current is essentially the Shockley model. The reverse current is given by Equation 6.12 with a negative voltage which leads to a diode current density of $-J_{so}$ called the **reverse saturation current density.** The value of J_{so} depends only on the material via n_i, μ_h, μ_e, dopant concentrations, but not on the voltage ($V_r > kT/e$). Furthermore, as J_{so} depends on n_i^2, it is strongly temperature dependent. In some books it is stated that the causes of reverse current are the thermal generation of minority carriers in the neutral region within a diffusion length to the SCL, the diffusion of these carriers to the SCL, and their subsequent drift through the SCL. This description, in essence, is identical to the Shockley model we just described.

The thermal generation of electron–hole pairs (EHPs) in the SCL, as shown in Figure 6.8a, can also contribute to the observed reverse current since the internal field in this layer will separate the electron and hole and drift them toward the neutral regions. This drift will result in an external current in addition to the reverse current due to the diffusion of minority carriers. The theoretical evaluation of SCL generation current involves an in-depth knowledge of the charge carrier generation processes via recombination centers, which is discussed in advanced texts. Suppose that τ_g is the **mean time to generate an electron–hole pair** by virtue of the thermal vibrations of the lattice; τ_g is also called the **mean thermal generation time.** Given τ_g, the rate of thermal generation per unit volume must be n_i/τ_g because it takes on average τ_g seconds to create n_i number of EHPs per unit volume. Furthermore, since WA, where A is the cross-sectional area, is the volume of the depletion region, the rate of EHP, or charge carrier, generation is $(AWn_i)/\tau_g$. Both holes and electrons drift in the SCL each contributing equally to the current. The observed current density must be $e(Wn_i)/\tau_g$. Therefore the reverse current density component due to thermal generation of EHPs within the SCL should be given by

EHP thermal generation in SCL

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

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

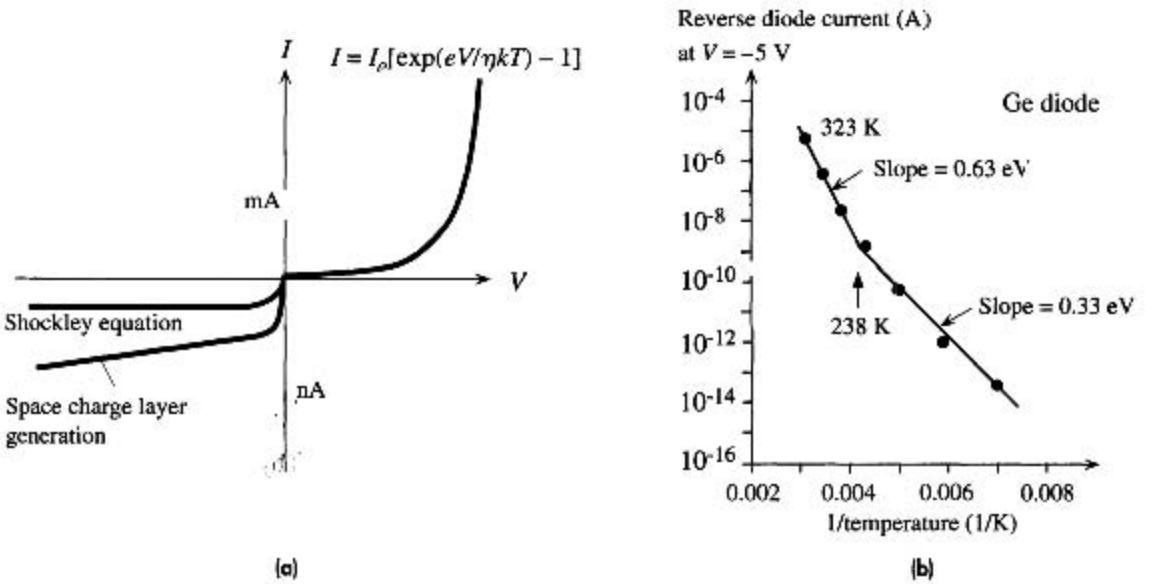


Figure 6.9

(a) Forward and reverse *I–V* characteristics of a *pn* junction (the positive and negative current axes have different scales and hence the discontinuity at the origin).

(b) Reverse diode current in a Ge *pn* junction as a function of temperature in a $\ln(I_{rev})$ versus $1/T$ plot. Above 238 K, I_{rev} is controlled by n_i^2, and below 238 K, it is controlled by n_i. The vertical axis is a logarithmic scale with actual current values.

SOURCE: (b) From D. Scansen and S. O. Kasap, *Cnd. J. Physics*, **70**, 1070, 1992.

components,

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

Total reverse current

which is shown schematically in Figure 6.9a. The thermal generation component J_{gen} in Equation 6.18 increases with reverse bias V_r because the SCL width W increases with V_r.

The terms in the reverse current in Equation 6.19 are predominantly controlled by n_i^2 and n_i. Their relative importance depends not only on the semiconductor properties but also on the temperature since $n_i \propto \exp(-E_g/2kT)$. Figure 6.9b shows the reverse current I_{rev} in dark in a Ge *pn* junction (a photodiode) plotted as $\ln(I_{rev})$ versus $1/T$ to highlight the two different processes in Equation 6.19. The measurements in Figure 6.9b show that above 238 K, I_{rev} is controlled by n_i^2 because the slope of $\ln(I_{rev})$ versus $1/T$ yields an E_g of approximately 0.63 eV, close to the expected E_g of about 0.66 eV in Ge. Below 238 K, I_{rev} is controlled by n_i because the slope of $\ln(I_{rev})$ versus $1/T$ is equivalent to $E_g/2$ of approximately 0.33 eV. In this range, the reverse current is due to EHP generation in the SCL via defects and impurities (recombination centers).

EXAMPLE 6.4 **FORWARD- AND REVERSE-BIASED Si DIODE** An abrupt Si p^+n junction diode has a cross-sectional area of 1 mm^2, an acceptor concentration of 5×10^{18} boron atoms cm^{-3} on the p-side, and a donor concentration of 10^{16} arsenic atoms cm^{-3} on the n-side. The lifetime of holes in the n-region is 417 ns, whereas that of electrons in the p-region is 5 ns due to a greater concentration of impurities (recombination centers) on that side. Mean thermal generation lifetime (τ_g) is about 1 μs. The lengths of the p- and n-regions are 5 and 100 microns, respectively.

a. Calculate the minority diffusion lengths and determine what type of a diode this is.

b. What is the built-in potential across the junction?

c. What is the current when there is a forward bias of 0.6 V across the diode at 27 °C? Assume that the current is by minority carrier diffusion.

d. Estimate the forward current at 100 °C when the voltage across the diode remains at 0.6 V. Assume that the temperature dependence of n_i dominates over those of D, L, and μ.

e. What is the reverse current when the diode is reverse-biased by a voltage $V_r = 5$ V?

SOLUTION

The general expression for the diffusion length is $L = \sqrt{D\tau}$ where D is the diffusion coefficient and τ is the carrier lifetime. D is related to the carrier mobility μ via the Einstein relationship $D/\mu = kT/e$. We therefore need to know μ to calculate D and hence L. Electrons diffuse in the p-region and holes in the n-region, so we need μ_e in the presence of N_a acceptors and μ_h in the presence of N_d donors. From the drift mobility, μ versus dopant concentration in Figure 5.19, we have the following:

With $N_a = 5 \times 10^{18}$ cm^{-3} $\quad \mu_e \approx 120$ cm^2 V^{-1} s^{-1}

With $N_d = 10^{16}$ cm^{-3} $\quad \mu_h \approx 440$ cm^2 V^{-1} s^{-1}

Thus

$$D_e = \frac{kT\mu_e}{e} \approx (0.0259 \text{ V})(120 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}) = 3.10 \text{ cm}^2 \text{ s}^{-1}$$

$$D_h = \frac{kT\mu_h}{e} \approx (0.0259 \text{ V})(440 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}) = 11.39 \text{ cm}^2 \text{ s}^{-1}$$

Diffusion lengths are

$$L_e = \sqrt{D_e\tau_e} = \sqrt{[(3.10 \text{ cm}^2 \text{ s}^{-1})(5 \times 10^{-9} \text{ s})]}$$
$$= 1.2 \times 10^{-4} \text{ cm} \quad \text{or} \quad 1.2 \text{ μm} < 5 \text{ μm}$$
$$L_h = \sqrt{D_h\tau_h} = \sqrt{[(11.39 \text{ cm}^2 \text{ s}^{-1})(417 \times 10^{-9} \text{ s})]}$$
$$= 21.8 \times 10^{-4} \text{ cm} \quad \text{or} \quad 21.8 \text{ μm} < 100 \text{ μm}$$

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

$$V_o = \left(\frac{kT}{e}\right) \ln\left(\frac{N_d N_a}{n_i^2}\right) = (0.0259 \text{ V}) \ln\left[\frac{(5 \times 10^{18} \times 10^{16})}{(1.0 \times 10^{10})^2}\right] = 0.877 \text{ V}$$

To calculate the forward current when $V = 0.6$ V, we need to evaluate both the diffusion and recombination components to the current. It is likely that the diffusion component will exceed the recombination component at this forward bias (this can be easily verified). Assuming

that the forward current is due to minority carrier diffusion in neutral regions,

$$I = I_{so}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \approx I_{so}\exp\left(\frac{eV}{kT}\right) \qquad \text{for } V \gg \frac{kT}{e} \qquad (= 0.0259 \text{ V})$$

where

$$I_{so} = AJ_{so} = Aen_i^2\left[\left(\frac{D_h}{L_h N_d}\right) + \left(\frac{D_e}{L_e N_a}\right)\right] \approx \frac{Aen_i^2 D_h}{L_h N_d}$$

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

$$I_{so} = \frac{(0.01 \text{ cm}^2)(1.6 \times 10^{-19} \text{ C})(1.0 \times 10^{10} \text{ cm}^{-3})^2(11.39 \text{ cm}^2 \text{ s}^{-1})}{(21.8 \times 10^{-4} \text{ cm})(10^{16} \text{ cm}^{-3})}$$

$$= 8.36 \times 10^{-14} \text{ A} \qquad \text{or} \qquad 0.084 \text{ pA}$$

Then the diode current is

$$I \approx I_{so}\exp\left(\frac{eV}{kT}\right) = (8.36 \times 10^{-14} \text{ A})\exp\left[\frac{(0.6 \text{ V})}{(0.0259 \text{ V})}\right]$$

$$= 0.96 \times 10^{-3} \text{ A} \qquad \text{or} \qquad 0.96 \text{ mA}$$

We note that when a forward bias of 0.6 V is applied, the built-in potential is reduced from 0.877 V to 0.256 V, which encourages minority carrier injection, that is, diffusion of holes from p- to n-side and electrons from n- to p-side. To find the current at 100 °C, first we assume that $I_{so} \propto n_i^2$. Then at $T = 273 + 100 = 373$ K, $n_i \approx 1.0 \times 10^{12}$ cm^{-3} (approximately from n_i versus $1/T$ graph in Figure 5.16), so

$$I_{so}(373 \text{ K}) \approx I_{so}(300 \text{ K})\left[\frac{n_i(373 \text{ K})}{n_i(300 \text{ K})}\right]^2$$

$$\approx (8.36 \times 10^{-14})\left(\frac{1.0 \times 10^{12}}{1.0 \times 10^{10}}\right)^2 = 8.36 \times 10^{-10} \text{A} \qquad \text{or} \qquad 0.836 \text{ nA}$$

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

$$I = I_{so}\exp\left(\frac{eV}{kT}\right) = (8.36 \times 10^{-10} \text{ A})\exp\left[\frac{(0.6 \text{ V})(300 \text{ K})}{(0.0259 \text{ V})(373 \text{ K})}\right] = 0.10 \text{ A}$$

When a reverse bias of V_r is applied, the potential difference across the depletion region becomes $V_o + V_r$ and the width W of the depletion region is

$$W = \left[\frac{2\varepsilon(V_o + V_r)}{eN_d}\right]^{1/2} = \left[\frac{2(11.9)(8.85 \times 10^{-12})(0.877 + 5)}{(1.6 \times 10^{-19})(10^{22})}\right]^{1/2}$$

$$= 0.88 \times 10^{-6} \text{ m} \qquad \text{or} \qquad 0.88 \text{ μm}$$

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

$$I_{gen} = \frac{eAWn_i}{\tau_g} = \frac{(1.6 \times 10^{-19} \text{ C})(0.01 \text{ cm}^2)(0.88 \times 10^{-4} \text{ cm})(1.0 \times 10^{10} \text{ cm}^{-3})}{(10^{-6} \text{ s})}$$

$$= 1.41 \times 10^{-9} \text{ A} \qquad \text{or} \qquad 1.4 \text{ nA}$$

This thermal generation current is much greater than the reverse saturation current I_{so}(= 0.084 pA). The reverse current is therefore dominated by I_{gen} and it is 1.4 nA.

6.2 *pn* JUNCTION BAND DIAGRAM

6.2.1 OPEN CIRCUIT

Figure 6.10a shows the energy band diagrams for a p-type and an n-type semiconductor of the same material (same E_g) when the semiconductors are isolated from each other. In the p-type material the Fermi level E_{Fp} is Φ_p below the vacuum level and is close to E_v. In the n-type material the Fermi level E_{Fn} is Φ_n below the vacuum level and is close to E_c. The separation $E_c - E_{Fn}$ determines the electron concentration n_{no} in the n-type and $E_{Fp} - E_v$ determines the hole concentration p_{po}, in the p-type semiconductor under thermal equilibrium conditions.

An important property of the Fermi energy E_F is that in a system in equilibrium, the Fermi level must be spatially continuous. A difference in Fermi levels ΔE_F is equivalent to electrical work eV, which is either done on the system or extracted from the system. When the two semiconductors are brought together, as in Figure 6.10b, the Fermi level must be uniform through the two materials and the junction at M, which marks the position of the metallurgical junction. Far away from M, in the bulk of the n-type semiconductor, we should still have an n-type semiconductor and $E_c - E_{Fn}$ should be the same as before. Similarly, $E_{Fp} - E_v$ far away from M inside the p-type material should also be the same as before. These features are sketched in Figure 6.10b keeping E_{Fp} and E_{Fn} the same through the whole system and, of course, keeping the bandgap $E_c - E_v$ the same. Clearly, to draw the energy band diagram, we have to bend the bands E_c and E_v around the junction at M because E_c on the n-side is close to E_{Fn} whereas on the p-side it is far away from E_{Fp}. How do bands bend and what does it mean?

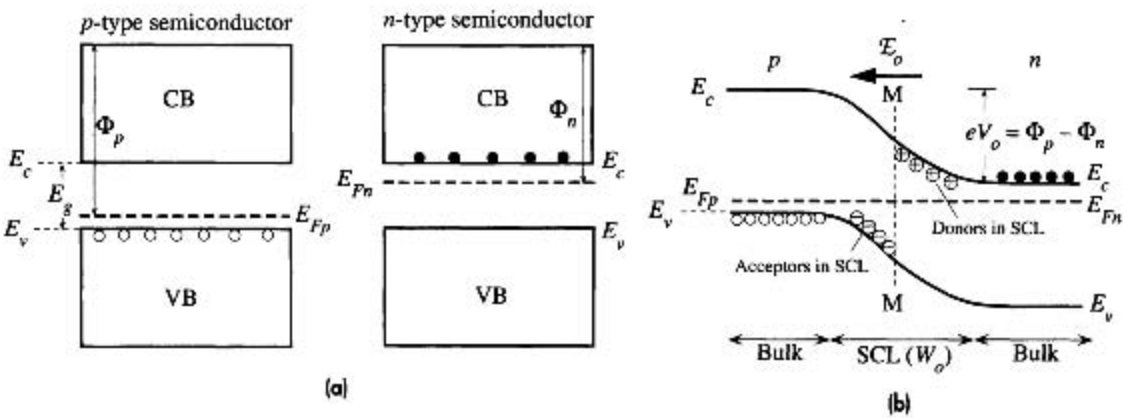


Figure 6.10

(a) Two isolated p- and n-type semiconductors (same material).

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

As soon as the two semiconductors are brought together to form the junction, electrons diffuse from the n-side to the p-side and as they do so they deplete the n-side near the junction. Thus E_c must move away from E_{Fn} toward M, which is exactly what is sketched in Figure 6.10b. Holes diffuse from the p-side to the n-side and the loss of holes in the p-type material near the junction means that E_v moves away from E_{Fp} toward M, which is also in the figure.

Furthermore, as electrons and holes diffuse toward each other, most of them recombine and disappear around M, which leads to the formation of a depletion region or the space charge layer, as we saw in Figure 6.1. The electrostatic potential energy (PE) of the electron decreases from 0 inside the p-region to $-eV_o$ inside the n-region, as shown in Figure 6.1g. The total energy of the electron must therefore decrease going from the p- to the n-region by an amount eV_o. In other words, the electron in the n-side at E_c must overcome a PE barrier to go over to E_c in the p-side. This PE barrier is eV_o, where V_o is the built-in potential that we evaluated in Section 6.1. Band bending around M therefore accounts not only for the variation of electron and hole concentrations in this region but also for the effect of the built-in potential (and hence the built-in field as the two are related).

In Figure 6.10b we have also schematically sketched in the positive donor (at E_d) and the negative acceptor (at E_a) charges in the SCL around M to emphasize that there are exposed charges near M. These charges are, of course, immobile and, generally, they are not shown in band diagrams. It should be noted that in the SCL region, marked as W_o, the Fermi level is close to neither E_c nor E_v, compared with the bulk semiconductor regions. This means that both n and p in this zone are much less than their bulk values n_{no} and p_{po}. The metallurgical junction zone has been depleted of carriers compared with the bulk. Any applied voltage must therefore drop across the SCL.

6.2.2 FORWARD AND REVERSE BIAS

The energy band diagram of the *pn* junction under open circuit conditions is shown in Figure 6.11a. There is no net current, so the diffusion current of electrons from the n- to p-side is balanced by the electron drift current from the p- to n-side driven by the built-in field $\mathcal{E}_o$. Similar arguments apply to holes. The probability that an electron diffuses from E_c in the n-side to E_c in the p-side determines the diffusion current density J_{diff}. The probability of overcoming the PE barrier is proportional to $\exp(-eV_o/kT)$. Therefore, under zero bias,

$$J_{\text{diff}}(0) = B\exp\left(-\frac{eV_o}{kT}\right) \qquad [6.20]$$

$$J_{\text{net}}(0) = J_{\text{diff}}(0) + J_{\text{drift}}(0) = 0 \qquad [6.21]$$

where B is a proportionality constant and $J_{\text{drift}}(0)$ is the current due to the drift of electrons by $\mathcal{E}_o$. Clearly $J_{\text{drift}}(0) = -J_{\text{diff}}(0)$; that is, drift is in the opposite direction to diffusion.

When the *pn* junction is forward-biased, the majority of the applied voltage drops across the depletion region, so the applied voltage is in opposition to the built-in potential V_o. Figure 6.11b shows the effect of forward bias, which is to reduce the PE

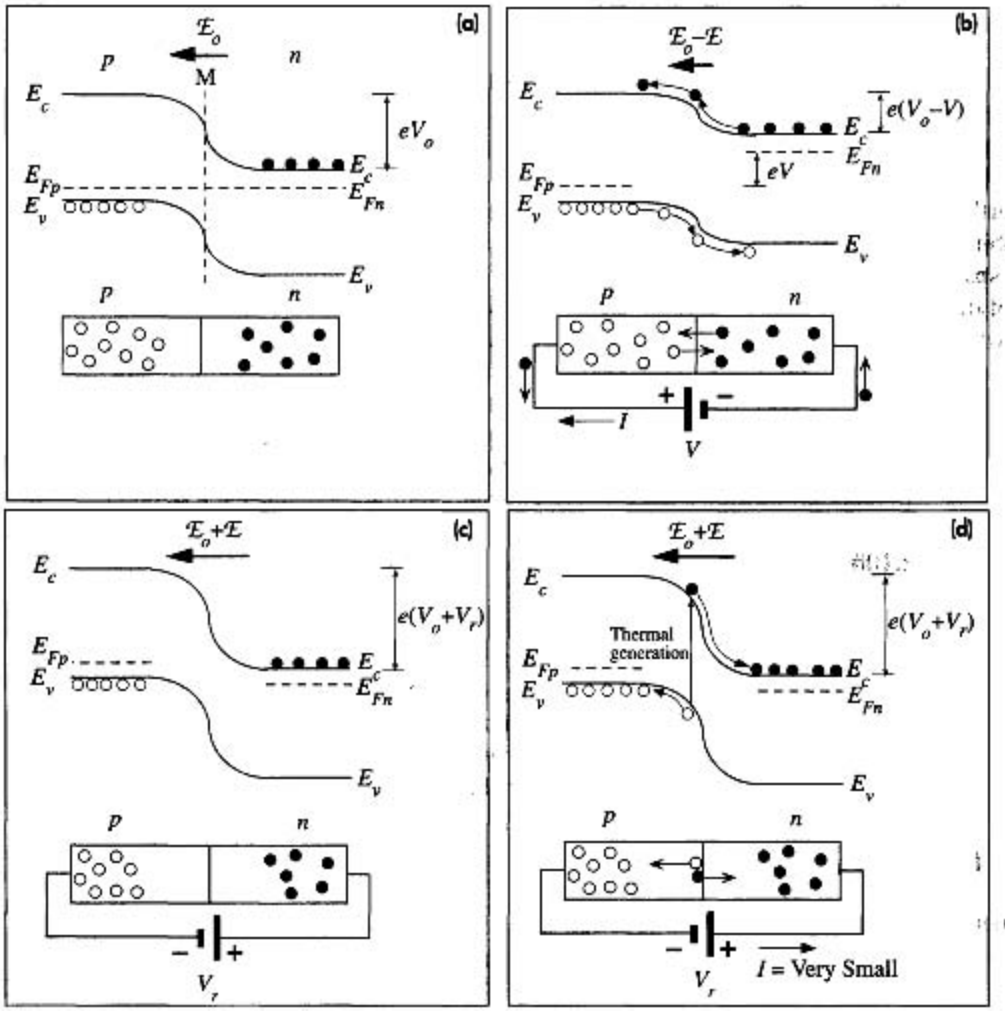


Figure 6.11 Energy band diagrams for a *pn* junction: (a) open circuit, (b) forward bias, (c) reverse bias conditions, (d) thermal generation of electron–hole pairs in the depletion region results in a small reverse current.

barrier from eV_o to $e(V_o - V)$. The electrons at E_c in the n-side can now readily overcome the PE barrier and diffuse to the p-side. The diffusing electrons from the n-side can be replenished easily by the negative terminal of the battery connected to this side. Similarly holes can now diffuse from the p- to n-side. The positive terminal of the battery can replenish those holes diffusing away from the p-side. There is therefore a current flow through the junction and around the circuit.

The probability that an electron at E_c in the n-side overcomes the new PE barrier and diffuses to E_c in the p-side is now proportional to $\exp[-e(V_o - V)/kT]$. The latter increases enormously even for small forward voltages. The new diffusion current due

to electrons diffusing from the n- to p-side is

$$J_{\text{diff}}(V) = B \exp\left[-\frac{e(V_o - V)}{kT}\right]$$

There is still a drift current due to electrons being drifted by the new field $\mathcal{E}_o - \mathcal{E}$ ($\mathcal{E}$ is the applied field) in the SCL. This drift current now has the value $J_{\text{drift}}(V)$. The net current is the diode current under forward bias

$$J = J_{\text{diff}}(V) + J_{\text{drift}}(V)$$

$J_{\text{drift}}(V)$ is difficult to evaluate. As a first approximation we can assume that although $\mathcal{E}_o$ has decreased to $\mathcal{E}_o - \mathcal{E}$, there is, however, an increase in the electron concentration in the SCL due to diffusion so that we can approximately take $J_{\text{drift}}(V)$ to remain the same as $J_{\text{drift}}(0)$. Thus

$$J \approx J_{\text{diff}}(V) + J_{\text{drift}}(0) = B \exp\left[-\frac{e(V_o - V)}{kT}\right] - B \exp\left(-\frac{eV_o}{kT}\right)$$

Factoring leads to

$$J \approx B \exp\left(-\frac{eV_o}{kT}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

We should also add to this the hole contribution, which has a similar form with a different constant B. The diode current–voltage relationship then becomes the familiar diode equation,

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

pn Junction I–V characteristics

where J_o is a temperature-dependent constant.[5]

When a reverse bias, $V = -V_r$, is applied to the *pn* junction, the voltage again drops across the SCL. In this case, however, V_r adds to the built-in potential V_o, so the *PE* barrier becomes $e(V_o + V_r)$, as shown in Figure 6.11c. The field in the SCL at M increases to $\mathcal{E}_o + \mathcal{E}$, where $\mathcal{E}$ is the applied field.

The diffusion current due to electrons diffusing from E_c in the n-side to E_c in the p-side is now almost negligible because it is proportional to $\exp[-e(V_o + V_r)/kT]$, which rapidly becomes very small with V_r. There is, however, a small reverse current arising from the drift component. When an electron–hole pair (EHP) is thermally generated in the SCL, as shown in Figure 6.11d, the field here separates the pair. The electron falls down the *PE* hill, down to E_c, in the n-side to be collected by the battery. Similarly the hole falls down its own *PE* hill (energy increases downward for holes) to make it to the p-side. The process of falling down a *PE* hill is the same process as being driven by a field, in this case by $\mathcal{E}_o + \mathcal{E}$. Under reverse bias conditions, there is therefore a small reverse current that depends on the rate of thermal generation of EHPs in the SCL. An electron in the p-side that is thermally generated within a diffusion length

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

L_e to the SCL can diffuse to the SCL and consequently can become drifted by the field, that is, roll down the *PE* hill in Figure 6.11d. Such minority carrier thermal generation in neutral regions can also give rise to a small reverse current.

EXAMPLE 6.5 **THE BUILT-IN VOLTAGE V_o FROM THE ENERGY BAND DIAGRAM** The energy band treatment allows a simple way to calculate V_o. When the junction is formed in Figure 6.10 from a to b, E_{Fp} and E_{Fn} must shift and line up. Using the energy band diagrams in this figure and semiconductor equations for n and p, derive an expression for the built-in voltage V_o in terms of the material and doping properties N_d, N_a, and n_i.

SOLUTION

The shift in E_{Fp} and E_{Fn} to line up is clearly $\Phi_p - \Phi_n$, the work function difference. Thus the *PE* barrier eV_o is $\Phi_p - \Phi_n$. From Figure 6.10, we have

$$eV_o = \Phi_p - \Phi_n = (E_c - E_{Fp}) - (E_c - E_{Fn})$$

But on the p- and n-sides, the electron concentrations in thermal equilibrium are given by

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

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

From these equations, we can now substitute for $(E_c - E_{Fp})$ and $(E_c - E_{Fn})$ in the expression for eV_o. The N_c cancel and we obtain

$$eV_o = kT \ln\left(\frac{n_{no}}{n_{po}}\right)$$

Since $n_{po} = n_i^2/N_a$ and $n_{no} = N_d$, we readily obtain the built-in potential V_o,

Built-in voltage

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

6.3 DEPLETION LAYER CAPACITANCE OF THE *pn* JUNCTION

It is apparent that the depletion region of a *pn* junction has positive and negative charges separated over a distance W similar to a parallel plate capacitor. The stored charge in the depletion region, however, unlike the case of a parallel plate capacitor, does not depend linearly on the voltage. It is useful to define an incremental capacitance that relates the incremental charge stored to an incremental voltage change across the *pn* junction.

The width of the depletion region is given by

Depletion region width

$$W = \left[\frac{2\varepsilon(N_a + N_d)(V_o - V)}{eN_a N_d}\right]^{1/2} \quad [6.22]$$

where, for forward bias, V is positive, which reduces V_o, and, for reverse bias, V is negative, so V_o is increased. We are interested in obtaining the capacitance of the

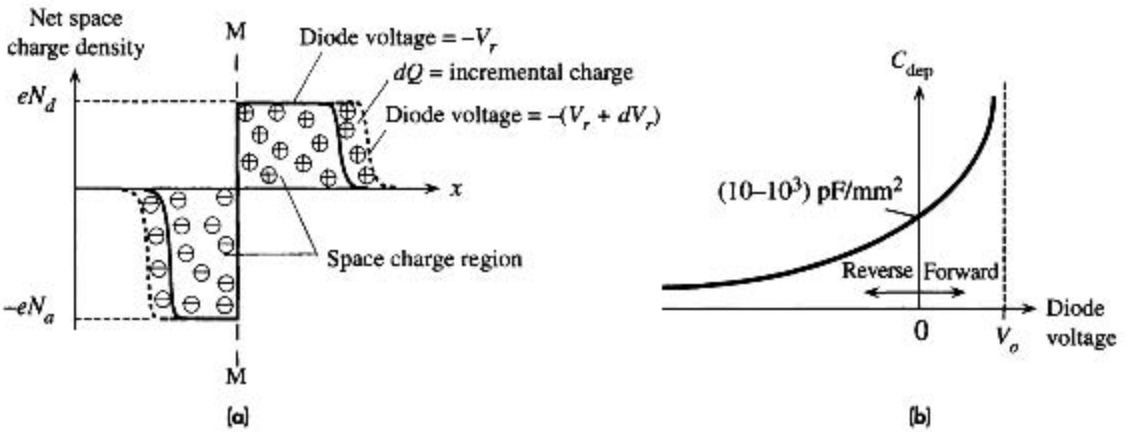


Figure 6.12 The depletion region behaves like a capacitor.

(a) The charge in the depletion region depends on the applied voltage just as in a capacitor. A reverse bias example is shown.

(b) The incremental capacitance of the depletion region increases with forward bias and decreases with reverse bias. Its value is typically in the range of picofarads per mm^2 of device area.

depletion region under dynamic conditions, that is, when V is a function of time. When the applied voltage V changes by dV, to $V + dV$, then W also changes via Equation 6.22, and as a result, the amount of charge in the depletion region becomes $Q + dQ$, as shown in Figure 6.12a for the reverse bias case, that is, $V = -V_r$ and $dV = -dV_r$. The **depletion layer capacitance** C_{dep} is defined by

Definition of depletion layer capacitance

$$C_{dep} = \left| \frac{dQ}{dV} \right| \qquad [6.23]$$

where the amount of charge (on any one side of the depletion layer) is

$$|Q| = eN_d W_n A = eN_a W_p A$$

and $W = W_n + W_p$. We can therefore substitute for W in Equation 6.22 in terms of Q and then differentiate it to obtain dQ/dV. The final result for the depletion capacitance is

Depletion capacitance

$$C_{dep} = \frac{\varepsilon A}{W} = \frac{A}{(V_o - V)^{1/2}} \left[\frac{e\varepsilon(N_a N_d)}{2(N_a + N_d)} \right]^{1/2} \qquad [6.24]$$

We should note that C_{dep} is given by the same expression as that for the parallel plate capacitor, $\varepsilon A / W$, but with W being voltage dependent by virtue of Equation 6.22. The $C_{dep} - V$ behavior is sketched in Figure 6.12b. Notice that C_{dep} decreases with increasing reverse bias, which is expected since the separation of the charges increases via $W \propto (V_o + V_r)^{1/2}$. The capacitance C_{dep} is present under both forward and reverse bias conditions.

The voltage dependence of the depletion capacitance is utilized in **varactor diodes** (varicaps), which are employed as voltage-dependent capacitors in tuning circuits. A varactor diode is reverse biased to prevent conduction, and its depletion capacitance is varied by the magnitude of the reverse bias.

6.4 DIFFUSION (STORAGE) CAPACITANCE AND DYNAMIC RESISTANCE

The diffusion or storage capacitance arises under forward bias only. As shown in Figure 6.2a, when the p^+n junction is forward biased, we have stored a positive charge on the n-side by the continuous injection and diffusion of minority carriers. Similarly, a negative charge has been stored on the p^+-side by electron injection, but the magnitude of this negative charge is small for the p^+n junction. When the applied voltage is increased from V to $V + dV$, as shown in Figure 6.13, then $p_n(0)$ changes from $p_n(0)$ to $p_n'(0)$. If dQ is the additional minority carrier charge injected into the n-side, as a result of a small increase dV in V, then the incremental **storage** or **diffusion capacitance** C_{diff} is defined as $C_{\text{diff}} = dQ/dV$. At voltage V, the injected positive charge Q on the n-side is disappearing by recombination at a rate Q/τ_h, where τ_h is the minority carrier lifetime. The diode current I is therefore Q/τ_h, from which

Injected minority carrier charge

$$Q = \tau_h I = \tau_h I_o \left[\exp\left(\frac{eV}{kT}\right) - 1\right] \qquad [6.25]$$

Thus,

Diffusion capacitance

$$C_{\text{diff}} = \frac{dQ}{dV} = \frac{\tau_h e I}{kT} = \frac{\tau_h I\,(\text{mA})}{25} \qquad [6.26]$$

where we used $e/kT \approx 1/0.025$ at room temperature. Generally the value of the diffusion capacitance, typically in the nanofarads range, far exceeds that of the depletion layer capacitance.

Suppose that the voltage V across the diode is increased by an infinitesimally small amount dV, as shown in an exaggerated way in Figure 6.14. This gives rise to a small increase dI in the diode current. We define the **dynamic** or **incremental resistance** r_d of the diode as dV/dI, so

Dynamic/ incremental resistance

$$r_d = \frac{dV}{dI} = \frac{kT}{eI} = \frac{25}{I\,(\text{mA})} \qquad [6.27]$$

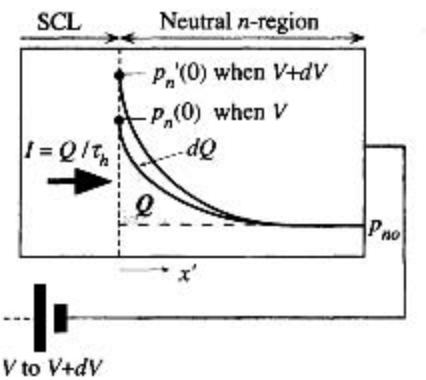


Figure 6.13 Consider the injection of holes into the n-side during forward bias.

Storage or diffusion capacitance arises because when the diode voltage increases from V to $V + dV$, more minority carriers are injected and more minority carrier charge is stored in the n-region.

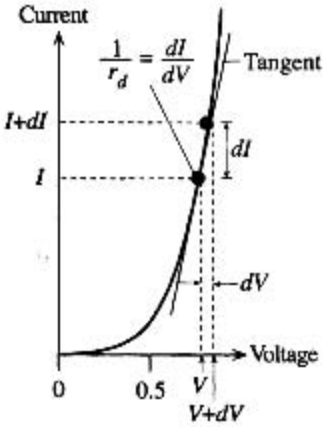


Figure 6.14 The dynamic resistance of the diode is defined as dV/dI, which is the inverse of the tangent at I.

The dynamic resistance is therefore the inverse of the slope of the I–V characteristics at a point and hence depends on the current I. It relates the changes in the diode current and voltage arising from the **diode action** alone, by which we mean the modulation of the rate of minority carrier diffusion by the diode voltage. We could have equivalently defined a dynamic conductance by

$$g_d = \frac{dI}{dV} = \frac{1}{r_d}$$

Dynamic conductance

From Equations 6.26 and 6.27 we have

$$r_d C_{\text{diff}} = \tau_h \qquad [6.28]$$

The dynamic resistance r_d and diffusion capacitance C_{diff} of a diode determine its response to small ac signals under forward bias conditions. By *small* we usually mean voltages smaller than the thermal voltage kT/e or 25 mV at room temperature. For small ac signals we can simply represent a forward-biased diode as a resistance r_d in parallel with a capacitance C_{diff}.

EXAMPLE 6.6

INCREMENTAL RESISTANCE AND CAPACITANCE An abrupt Si p^+n junction diode of cross-sectional area (A) 1 mm^2 with an acceptor concentration of 5×10^{18} boron atoms cm^{-3} on the p-side and a donor concentration of 10^{16} arsenic atoms cm^{-3} on the n-side is forward-biased to carry a current of 5 mA. The lifetime of holes in the n-region is 417 ns, whereas that of electrons in the p-region is 5 ns. What are the small-signal ac resistance, incremental storage, and depletion capacitances of the diode?

SOLUTION

This is the same diode we considered in Example 6.4 for which the built-in potential was 0.877 V and $I_{so} = 0.0836$ pA. The current through the diode is 5 mA. Thus

$$I = I_{so}\exp\left(\frac{eV}{kT}\right) \quad \text{or} \quad V = \left(\frac{kT}{e}\right)\ln\left(\frac{I}{I_{so}}\right) = (0.0259)\ln\left(\frac{5 \times 10^{-3}}{0.0836 \times 10^{-12}}\right) = 0.643 \text{ V}$$

The dynamic diode resistance is given by

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

The depletion capacitance per unit area with $N_a \gg N_d$ is

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

At $V = 0.643$ V, with $V_o = 0.877$ V, $N_d = 10^{22}\ \text{m}^{-3}$, $\varepsilon_r = 11.9$, and $A = 10^{-6}\ \text{m}^2$, the above equation gives

$$C_{dep} = 10^{-6}\left[\frac{(1.6 \times 10^{-19})(11.9)(8.85 \times 10^{-12})(10^{22})}{2(0.877 - 0.643)}\right]^{1/2}$$

$$= 6.0 \times 10^{-10}\ \text{F} \quad \text{or} \quad 600\ \text{pF}$$

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

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

Clearly the diffusion capacitance (83 nF) that arises during forward bias completely overwhelms the depletion capacitance (600 pF).

We note that there is also a diffusion capacitance due to electrons injected and stored in the p-region. However, electron lifetime in the p-region is very short (here 5 ns), so the value of this capacitance is much smaller than that due to holes in the n-region. In calculating the diffusion capacitance, we normally consider the minority carriers that have the longest recombination lifetime, here τ_h. These are the carriers that take a long time to disappear by recombination when the bias is suddenly switched off.

6.5 REVERSE BREAKDOWN: AVALANCHE AND ZENER BREAKDOWN

The reverse voltage across a *pn* junction cannot be increased without limit. Eventually the *pn* junction breaks down either by the Avalanche or Zener breakdown mechanisms, which lead to large reverse currents, as shown in Figure 6.15. In the $V = -V_{br}$ region, the reverse current increases dramatically with the reverse bias. If unlimited, the large

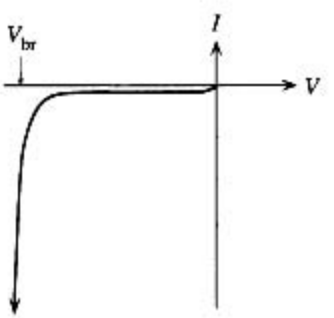


Figure 6.15 Reverse I–V, characteristics of a *pn* junction.

reverse current will increase the power dissipated, which in turn raises the temperature of the device, which leads to a further increase in the reverse current and so on. If the temperature does not burn out the device, for example, by melting the contacts, then the breakdown is recoverable. If the current is limited by an external resistance to a value within the power dissipation specifications, then there is no reason why the device cannot operate under breakdown conditions.

6.5.1 AVALANCHE BREAKDOWN

As the reverse bias increases, the field in the SCL can become so large that an electron drifting in this region can gain sufficient kinetic energy to impact on a Si atom and ionize it, or break a Si–Si bond. The phenomenon by which a drifting electron gains sufficient energy from the field to ionize a host crystal atom by bombardment is termed **impact ionization.** The accelerated electron must gain at least an energy equal to E_g as impact ionization breaks a Si–Si bond, which is tantamount to exciting an electron from the valence band to the conduction band. Thus an additional electron–hole pair is created by this process.

Consider what happens when a thermally generated electron just inside the SCL in the p-side is accelerated by the field. The electron accelerates and gains sufficient energy to collide with a host Si atom and release an EHP by impact ionization, as depicted in Figure 6.16. It will lose at least E_g amount of energy, but it can accelerate and head for another ionizing collision further along the depletion region until it reaches the neutral n-region. The EHPs generated by impact ionization themselves can now be accelerated by the field and will themselves give rise to further EHPs by ionizing collisions and so on, leading to an **avalanche effect.** One initial carrier can thus create many carriers in the SCL through an avalanche of impact ionizations.

If the reverse current in the SCL in the absence of impact ionization is I_o, then due to the avalanche of ionizing collisions in the SCL, the reverse current becomes MI_o where M is the multiplication. It is the net number of carriers generated by the avalanche effect per carrier in the SCL. Impact ionization depends strongly on the electric field. Small increases in the reverse bias can lead to dramatic increases in the

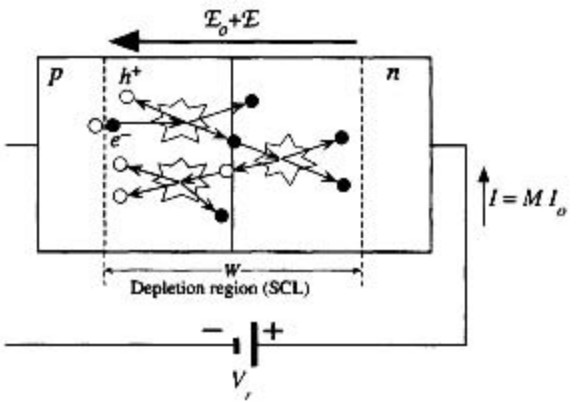


Figure 6.16 Avalanche breakdown by impact ionization.

multiplication process. Typically

$$M = \frac{1}{1 - \left(\frac{V_r}{V_{br}}\right)^n} \quad [6.29]$$

where V_r is the reverse bias, V_{br} is the breakdown voltage, and n is an index in the range 3 to 5. It is clear that the reverse current MI_o increases sharply with V_r near V_{br}, as depicted in Figure 6.15. Indeed, the voltage across a diode under reverse breakdown remains around V_{br} for very large current variations (several orders of magnitude). If the reverse current under breakdown is limited by an appropriate external resistor R, as shown in Figure 6.17, to prevent destructive power dissipation in the diode, then the voltage across the diode remains approximately at V_{br}. Thus, as long as $V_r > V_{br}$, the diode clamps the voltage between A and B to approximately V_{br}. The reverse current in the circuit is then $(V_r - V_{br})/R$.

Since the electric field in the SCL depends on the width of the depletion region W, which in turn depends on the doping parameters, V_{br} also depends on the doping, as discussed in Example 6.7.

6.5.2 ZENER BREAKDOWN

Heavily doped *pn* junctions have narrow depletion widths, which lead to large electric fields within this region. When a reverse bias is applied to a *pn* junction, the energy band diagram of the *n*-side can be viewed as being lowered with respect to the *p*-side, as depicted in Figure 6.18. For a sufficient reverse bias (typically less than 10 V), E_c

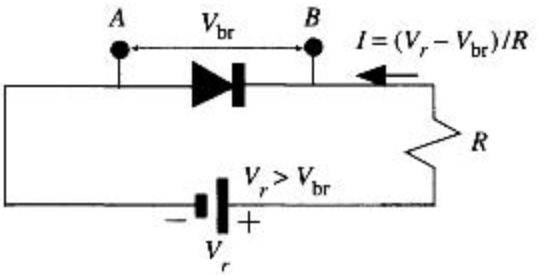


Figure 6.17 If the reverse breakdown current when $V_r > V_{br}$ is limited by an external resistance R to prevent destructive power dissipation, then the diode can be used to clamp the voltage between A and B to remain approximately V_{br}.

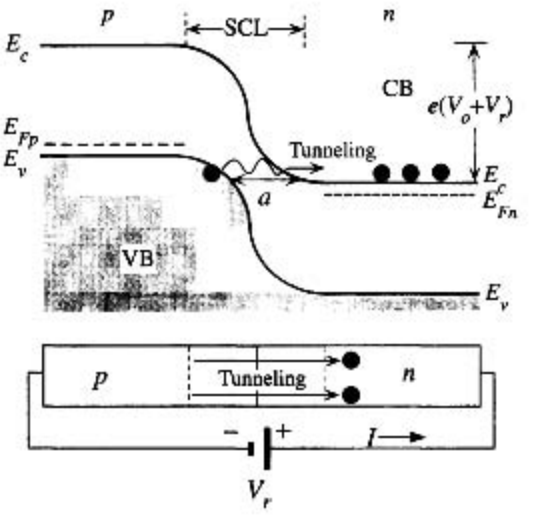


Figure 6.18 Zener breakdown involves electrons tunneling from the VB of *p*-side to the CB of *n*-side when the reverse bias reduces E_c to line up with E_v.

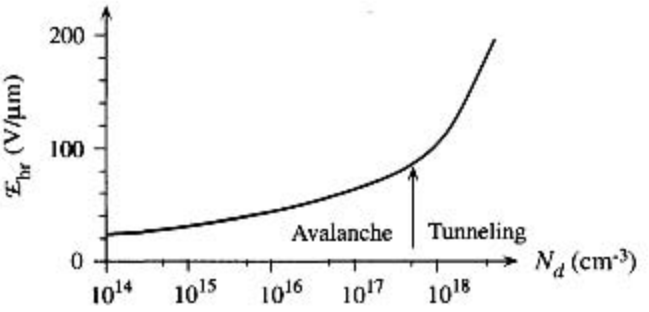


Figure 6.19 The breakdown field $\mathcal{E}_{br}$ in the depletion layer for the onset of reverse breakdown versus doping concentration N_d in the lightly doped region in a one-sided (p^+n or pn^+) abrupt pn junction.

Avalanche and tunneling mechanisms are separated by the arrow.

SOURCE: Data extracted from M. Sze and G. Gibbons, *Solid State Electronics*, 9, no. 831, 1966.

on the n-side may be lowered to be below E_v on the p-side. This means that electrons at the top of the VB in the p-side are now at the same energy level as the empty states in the CB in the n-side. As the separation between the VB and CB narrows, shown as a ($< W$), the electrons easily tunnel from the VB in the p-side to the CB in the n-side, which leads to a current. This process is called the **Zener effect.** As there are many electrons in the VB and many empty states in the CB, the tunneling current can be substantial. The reverse voltage V_r, which starts the tunneling current and hence the Zener breakdown, is clearly that which lowers E_c on the n-side to be below E_v on the p-side and thereby gives a separation that encourages tunneling. In nonquantum mechanical terms, one may intuitively view the Zener effect as the strong electric field in the depletion region ripping out some of those electrons in the Si–Si bonds and thereby releasing them for conduction.

Figure 6.19 shows the dependence of the breakdown field $\mathcal{E}_{br}$ in the depletion region for the onset of avalanche or Zener breakdown in a one-sided (p^+n or pn^+) abrupt junction on the dopant concentration N_d in the lightly doped side. At high fields, the tunneling becomes the dominant reverse breakdown mechanism.

EXAMPLE 6.7

AVALANCHE BREAKDOWN Consider a uniformly doped abrupt p^+n junction ($N_a \gg N_d$) reverse biased by $V = -V_r$.

a. What is the relationship between the depletion width W and the potential difference $(V_o + V_r)$ across W?

b. If avalanche breakdown occurs when the maximum field in the depletion region $\mathcal{E}_o$ reaches the breakdown field $\mathcal{E}_{br}$, show that the breakdown voltage V_{br} ($\gg V_o$) is then given by

$$V_{br} = \frac{\varepsilon \mathcal{E}_{br}^2}{2eN_d}$$

c. An abrupt Si p^+n junction has boron doping of 10^{19} cm^{-3} on the p-side and phosphorus doping of 10^{16} cm^{-3} on the n-side. The dependence of the avalanche breakdown field on the impurity concentration is shown in Figure 6.19.

1. What is the reverse breakdown voltage of this Si diode?
2. Calculate the reverse breakdown voltage when the phosphorus doping is increased to 10^{17} cm^{-3}.

SOLUTION

One can assume that all the applied reverse bias drops across the depletion layer so that the new voltage across W is now $V_o + V_r$. We have to integrate $d\mathcal{E}/dx = \rho_{net}/\varepsilon$ as before across W to find the maximum field. The most important fact to remember here is that the *pn* junction equations relating W, $\mathcal{E}_o$, V_o, N_a, N_d, and so on remain the same but with V_o replaced with $V_o + V_r$ since the applied reverse bias of V_r increases V_o to $V_o + V_r$. Then from Equation 6.4,

$$W^2 = \frac{2\varepsilon(V_o + V_r)(N_a^{-1} + N_d^{-1})}{e} \approx \frac{2\varepsilon(V_o + V_r)}{eN_d}$$

since $N_a \gg N_d$. The maximum field that corresponds to the breakdown field $\mathcal{E}_{br}$ is given by

Maximum field and reverse bias

$$\mathcal{E}_o = -\frac{2(V_o + V_r)}{W}$$

Thus, from these two equations we can eliminate W and obtain $V_{br} = V_r$ as

Breakdown voltage and doping

$$V_{br} = \frac{\varepsilon \mathcal{E}_{br}^2}{2eN_d}$$

Given $N_a \gg N_d$ we have a p^+n junction with $N_d = 10^{16}$ cm^{-3}. The depletion region extends into the n-region, so the maximum field actually occurs in the n-region. Here the breakdown field $\mathcal{E}_{br}$ depends on the doping level as given in the graph of the critical field at breakdown $\mathcal{E}_{br}$ versus doping concentration N_d in Figure 6.19. Taking $\mathcal{E}_{br} \approx 40$ V/μm or 4.0×10^5 V cm^{-1} at $N_d = 10^{16}$ cm^{-3} and using the above equation for V_{br}, we get $V_{br} = 53$ V.

When $N_d = 10^{17}$ cm^{-3}, $\mathcal{E}_{br}$ from the graph is about 6×10^5 V cm^{-1}, which leads to $V_{br} = 11.8$ V.

6.6 BIPOLAR TRANSISTOR (BJT)

6.6.1 COMMON BASE (CB) DC CHARACTERISTICS

As an example, we will consider the *pnp* bipolar junction transistor (BJT) whose basic structure is shown in Figure 6.20a. The *pnp* transistor has three differently doped semiconductor regions. These regions of different doping occur within the same single crystal by the variation of acceptor and donor concentrations resulting from the fabrication process. The most heavily doped p-region (p^+) is called the **emitter.** In contact with this region is the lightly doped n-region, which is called the **base.** The next region is the p-type doped **collector.** The base region has the most narrow width for reasons discussed below. Although the three regions in Figure 6.20a have identical cross-sectional areas, in practice, due to the fabrication process, the cross-sectional area increases from the emitter to the collector and the collector region has an extended width. For simplicity, we will assume that the cross-sectional area is uniform, as in Figure 6.20a.

The *pnp* BJT connected as shown in Figure 6.20b is said to be operating under normal and active conditions, which means that the base–emitter (BE) junction is forward biased and the base–collector (BC) junction is reverse biased. The circuit in

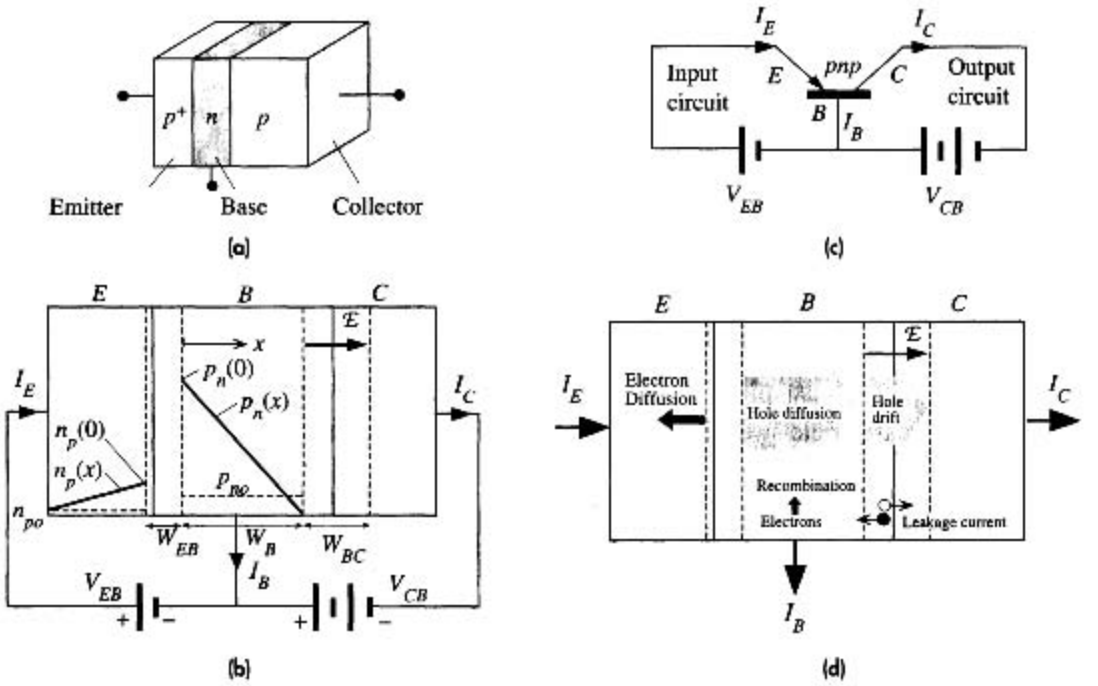


Figure 6.20

(a) A schematic illustration of the *pnp* bipolar transistor with three differently doped regions.

(b) The *pnp* bipolar operated under normal and active conditions.

(c) The CB configuration with input and output circuits identified.

(d) The illustration of various current components under normal and active conditions.

Figure 6.20b, in which the base is common to both the collector and emitter bias voltages, is known as the common base (CB) configuration.[6] Figure 6.20c shows the CB transistor circuit with the BJT represented by its circuit symbol. The arrow identifies the emitter junction and points in the direction of current flow when the EB junction is forward biased. Figure 6.20c also identifies the emitter circuit, where V_{EB} is connected, as the input circuit. The collector circuit, where V_{CB} is connected, is the output circuit.

The base–emitter junction is simply called the **emitter junction** and the base-collector junction is called the **collector junction.** As the emitter is heavily doped, the base–emitter depletion region W_{EB} extends almost entirely into the base. Generally, the base and collector regions have comparable doping, so the base–collector depletion region W_{BC} extends to both sides. The width of the neutral base region outside the depletion regions is labeled as W_B. All these parameters are shown and defined in Figure 6.20b.

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

We should note that all the applied voltages drop across the depletion widths. The applied collector–base voltage V_{CB} reverse biases the BC junction and hence increases the field in the depletion region at the collector junction.

Since the EB junction is forward-biased, minority carriers are then injected into the emitter and base exactly as they are in the forward-biased diode. Holes are injected into the base and electrons into the emitter, as depicted in Figure 6.20d. Hole injection into the base, however, far exceeds the electron injection into the emitter because the emitter is heavily doped. We can then assume that the emitter current is almost entirely due to holes injected from the emitter into the base. Thus, when forward biased, the emitter "emits," that is, injects holes into the base.

Injected holes into the base must diffuse toward the collector junction because there is a hole concentration gradient in the base. Hole concentration $p_n(W_B)$ just outside the depletion region at the collector junction is negligibly small because the increased field sweeps nearly all the holes here across the junction into the collector (the collector junction is reverse biased).

The hole concentration $p_n(0)$ in the base just outside the emitter junction depletion region is given by the law of the junction. Measuring x from this point (Figure 6.20b),

$$p_n(0) = p_{no} \exp\left(\frac{eV_{EB}}{kT}\right) \qquad [6.30]$$

whereas at the collector end, $x = W_B$, $p_n(W_B) \approx 0$.

If no holes are lost by recombination in the base, then all the injected holes diffuse to the collector junction. There is no field in the base to drift the holes. Their motion is by diffusion. When they reach the collector junction, they are quickly swept across into the collector by the internal field $\mathcal{E}$ in W_{BC}. It is apparent that all the injected holes from the emitter become collected by the collector. The collector current is then the same as the emitter current. The only difference is that the emitter current flows across a smaller voltage difference V_{EB}, whereas the collector current flows through a larger voltage difference V_{CB}. This means a *net gain in power* from the emitter circuit to the collector circuit.

Since the current in the base is by diffusion, to evaluate the emitter and collector currents we must know the hole concentration gradient at $x = 0$ and $x = W_B$ and therefore we must know the hole concentration profile $p_n(x)$ across the base.[7] In the first instance, we can approximate the $p_n(x)$ profile in the base as a straight line from $p_n(0)$ to $p_n(W_B) = 0$, as shown in Figure 6.20b. This is only true in the absence of any recombination in the base as in the short diode case. The emitter current is then

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

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

We can substitute for $p_n(0)$ from Equation 6.30 to obtain

$$I_E = \frac{eAD_h p_{no}}{W_B} \exp\left(\frac{eV_{EB}}{kT}\right) \quad [6.31]$$

Emitter current

It is apparent that I_E is determined by V_{EB}, the forward bias applied across the EB junction, and the base width W_B. In the absence of recombination, the collector current is the same as the emitter current, $I_C = I_E$. The control of the collector current I_C in the output (collector) circuit by V_{EB} in the input (emitter) circuit is what constitutes the **transistor action.** The common base circuit has a **power gain** because I_C in the output in Figure 6.20c flows around a larger voltage difference V_{CB} compared with I_E in the input, which flows across V_{EB} (about 0.6 V).

The ratio of the collector current I_C to the emitter current I_E is defined as the **CB current gain** or **current transfer ratio** α of the transistor,

$$\alpha = \frac{I_C}{I_E} \quad [6.32]$$

Definition of CB current gain

Typically, α is less than unity, in the range 0.99–0.999, due to two reasons. First is the limitation due to the emitter injection efficiency. When the *BE* junction is forward-biased, holes are injected from the emitter into the base, giving an emitter current $I_{E(\text{hole})}$, and electrons are injected from the base into the emitter, giving an emitter current $I_{E(\text{electron})}$. The total emitter current is, therefore,

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

Total emitter current

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

$$\gamma = \frac{I_{E(\text{hole})}}{I_{E(\text{hole})} + I_{E(\text{electron})}} = \frac{1}{1 + \dfrac{I_{E(\text{electron})}}{I_{E(\text{hole})}}} \quad [6.33]$$

Emitter injection efficiency

Consequently, the collector current, which depends on $I_{E(\text{hole})}$ only, is less than the emitter current. We would like γ to be as close to unity as possible; $I_{E(\text{hole})} \gg I_{E(\text{electron})}$. γ can be readily calculated for the forward-biased pn junction current equations as shown in Example 6.9.

Secondly, a small number of the diffusing holes in the narrow base inevitably become lost by recombination with the large number of electrons present in this region as depicted in Figure 6.20d. Thus, a fraction of $I_{E(\text{hole})}$ is lost in the base due to recombination, which further reduces the collector current. We define the **base transport factor** α_T as

$$\alpha_T = \frac{I_C}{I_{E(\text{hole})}} = \frac{I_C}{\gamma I_E} \quad [6.34]$$

Base transport factor

If the emitter were a perfect injector, $I_E = I_{E(\text{hole})}$, then the current gain α would be α_T. If τ_h is the hole (minority carrier) lifetime in the base, then $1/\tau_h$ is the probability per unit time that a hole will recombine and disappear. We also know that in

time t, a particle diffuses a distance x, given by $x = \sqrt{2Dt}$ where D is the diffusion coefficient. The time τ_t it takes for a hole to diffuse across W_B is then given by

Base minority carrier transit time

$$\tau_t = \frac{W_B^2}{2D_h} \quad [6.35]$$

This diffusion time is called the **transit time** of the minority carriers across the base.

The probability of recombination in time τ_t is then τ_t/τ_h. The probability of not recombining and therefore diffusing across is $(1 - \tau_t/\tau_h)$. Since $I_{E(\text{hole})}$ represents the holes entering the base per unit time, $I_{E(\text{hole})}(1 - \tau_t/\tau_h)$ represents the number of holes leaving the base per unit time (without recombining) which is the collector current I_C. Substituting for I_C and $I_{E(\text{hole})}$ in Equation 6.34 gives the base transport factor α_T,

Base transport factor

$$\alpha_T = \frac{I_C}{I_{E(\text{hole})}} = 1 - \frac{\tau_t}{\tau_h} \quad [6.36]$$

Using Equations 6.32, 6.34, and 6.36 we can find the total **CB current gain** α:

CB current gain

$$\alpha = \alpha_T \gamma = \left(1 - \frac{\tau_t}{\tau_h}\right)\gamma \quad [6.37]$$

The recombination of holes with electrons in the base means that the base must be replenished with electrons, which are supplied by the external battery in the form of a small base current I_B, as shown in Figure 6.20d. In addition, the base current also has to supply the electrons injected from the base into the emitter, that is, $I_{E(\text{electron})}$, and shown as electron diffusion in the emitter in Figure 6.20d. The number of holes entering the base per unit time is represented by $I_{E(\text{hole})}$, and the number recombining per unit time is then $I_{E(\text{hole})}(\tau_t/\tau_h)$. Thus, I_B is

Base current

$$I_B = \left(\frac{\tau_t}{\tau_h}\right) I_{E(\text{hole})} + I_{E(\text{electron})} = \gamma \frac{\tau_t}{\tau_h} I_E + (1 - \gamma) I_E \quad [6.38]$$

which further simplifies to $I_E - I_C$; the difference between the emitter current and the collector current is the base current. (This is exactly what we expect from Kirchoff's current law.)

The ratio of the collector current to the base current is defined as the **current gain** β of the transistor.[8] By using Equations 6.32, 6.37, and 6.38, we can relate β to α:

Base-to-collector current gain

$$\beta = \frac{I_C}{I_B} = \frac{\alpha}{1 - \alpha} \approx \frac{\gamma \tau_h}{\tau_t} \quad [6.39]$$

The base–collector junction in Figure 6.20b is reverse biased, which leads to a leakage current into the collector terminal even in the absence of an emitter current. This leakage current is due to thermally generated electron–hole pairs in the depletion region W_{BC} being drifted by the internal field, as schematically illustrated in Figure 6.20d.

[8] β is a useful parameter when the transistor is used in what is called the common emitter (CE) configuration, in which the input current is made to flow into the base of the transistor, and the collector current is made to flow in the output circuit.

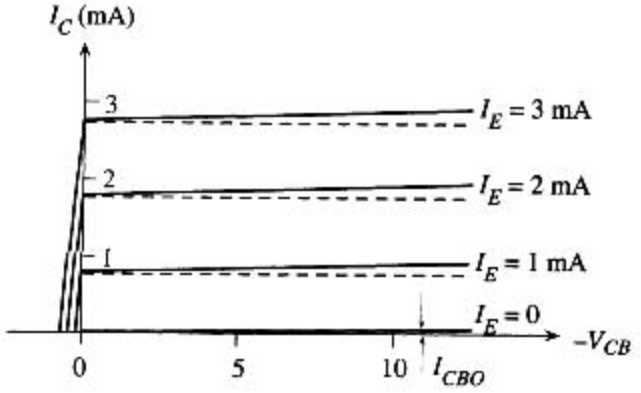


Figure 6.21 DC I–V characteristics of the *pnp* bipolar transistor (exaggerated to highlight various effects).

Suppose that we open circuit the emitter ($I_E = 0$). Then the collector current is simply the leakage current, denoted by I_{CBO}. The base current is then $-I_{CBO}$ (flowing out from the base terminal). In the presence of an emitter current I_E, we have

Active region collector current

$$I_C = \alpha I_E + I_{CBO} \quad [6.40]$$

Active region base current

$$I_B = (1 - \alpha) I_E - I_{CBO} \quad [6.41]$$

Equations 6.40 and 6.41 give the collector and base currents in terms of the input current I_E, which in turn depends on V_{EB}. They only hold when the collector junction is reverse biased and the emitter junction is forward biased, which is defined as the **active region** of the BJT. It should be emphasized that what constitutes the transistor action is the control of I_E, and hence I_C, by V_{EB}.

The dc characteristics of the CB-connected BJT as in Figure 6.20b are normally represented by plotting the collector current I_C as a function of V_{CB} for various fixed values of the emitter current. A typical example of such dc characteristics for a *pnp* transistor is illustrated in Figure 6.21. The following characteristics are apparent. The collector current when $I_E = 0$ is the CB junction leakage current I_{CBO}, typically a fraction of a microampere. As long as the collector is negatively biased with respect to the base, the CB junction is reverse biased and the collector current is given by $I_C = \alpha I_E + I_{CBO}$, which is close to the emitter current when $I_E \gg I_{CBO}$. When the polarity of V_{CB} is changed, the CB junction becomes forward biased. The collector junction is then like a forward biased diode and the collector current is the difference between the forward biased CB junction current and the forward biased EB junction current. As they are in opposite directions, they subtract.

We note that I_C increases slightly with the magnitude of V_{CB} even when I_E is constant. In our treatment I_C did not directly depend on V_{CB}, which simply reverse biased the collector junction to collect the diffusing holes. In our discussions we assumed that the base width W_B does not depend on the applied voltages. This is only approximately true. Suppose that we increase the reverse bias V_{CB} (for example, from -5 to -10 V). Then the base–collector depletion width W_{BC} also increases, as schematically depicted in Figure 6.22. Consequently the base width W_B gets slightly narrower, which leads to a slightly shorter base transit time τ_t. The base transport factor α_T in Equation 6.36 and

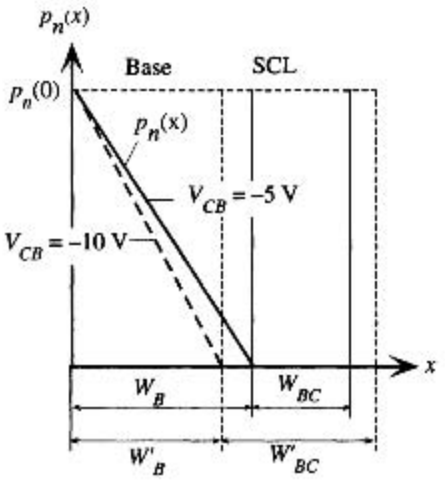


Figure 6.22 The Early effect. When the BC reverse bias increases, the depletion width W_{BC} increases to W'_{BC}, which reduces the base width W_B to W'_B. As $p_n(0)$ is constant (constant V_{EB}), the minority carrier concentration gradient becomes steeper and the collector current, I_C increases.

hence α are then slightly larger, which leads to a small increase in I_C. The modulation of the base width W_B by V_{CB} is not very strong, which means that the slopes of the $I_C - V_{CB}$ lines at a fixed I_E are very small in Figure 6.21. The base width modulation by V_{CB} is called the **Early effect.**

EXAMPLE 6.8 **A *pnp* TRANSISTOR** Consider a *pnp* Si BJT that has the following properties. The emitter region mean acceptor doping is 2×10^{18} cm^{-3}, the base region mean donor doping is 1×10^{16} cm^{-3}, and the collector region mean acceptor doping is 1×10^{16} cm^{-3}. The hole drift mobility in the base is 400 cm^2 V^{-1} s^{-1}, and the electron drift mobility in the emitter is 200 cm^2 V^{-1} s^{-1}. The transistor emitter and base neutral region widths are about 2 μm each when the transistor is under normal operating conditions, that is, when the EB junction is forward-biased and the BC junction is reverse-biased. The effective cross-sectional area of the device is 0.02 mm^2. The hole lifetime in the base is approximately 400 ns. Assume that the emitter has 100 percent injection efficiency, $\gamma = 1$. Calculate the CB current transfer ratio α and the current gain β. What is the emitter–base voltage if the emitter current is 1 mA?

SOLUTION

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

$$D_h = \left(\frac{kT}{e}\right)\mu_h = (0.0259 \text{ V})(400 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}) = 10.36 \text{ cm}^2 \text{ s}^{-1}$$

The minority carrier transit time τ_t across the base is

$$\tau_t = \frac{W_B^2}{2D_h} = \frac{(2 \times 10^{-4} \text{ cm})^2}{2(10.36 \text{ cm}^2 \text{ s}^{-1})} = 1.93 \times 10^{-9} \text{ s} \quad \text{or} \quad 1.93 \text{ ns}$$

The base transport factor and hence the CB current gain is

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

The current gain β of the transistor is

$$\beta = \frac{\alpha}{1-\alpha} = \frac{0.99517}{1-0.99517} = 206.2$$

The emitter current is due to holes diffusing in the base ($\gamma = 1$),

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

where

$$I_{EO} = \frac{eAD_h p_{no}}{W_B} = \frac{eAD_h n_i^2}{N_d W_B}$$

$$= \frac{(1.6 \times 10^{-19}\ \text{C})(0.02 \times 10^{-2}\ \text{cm}^2)(10.36\ \text{cm s}^{-1})(1.0 \times 10^{10}\ \text{cm}^{-3})^2}{(1 \times 10^{16}\ \text{cm}^{-3})(2 \times 10^{-4}\ \text{cm})}$$

$$= 1.66 \times 10^{-14}\ \text{A}$$

Thus,

$$V_{EB} = \frac{kT}{e} \ln\left(\frac{I_E}{I_{EO}}\right) = (0.0259\ \text{V}) \ln\left(\frac{1 \times 10^{-3}\ \text{A}}{1.66 \times 10^{-14}\ \text{A}}\right) = 0.64\ \text{V}$$

The major assumption is $\gamma = 1$, which is generally not true, as shown in Example 6.9. The actual α and hence β will be smaller due to less than 100 percent emitter injection. Note also that W_B is the *neutral region width*, that is, the region of base outside the depletion regions. It is not difficult to calculate the depletion layer widths within the base, which are about 0.2 μm on the emitter side and roughly about 0.7 μm on the collector side, so that the total base width junction to junction is $2 + 0.2 + 0.7 = 2.9$ μm.

The transit time of minority carriers across the base is τ_t. If the input signal changes before the minority carriers have diffused across the base, then the collector current cannot respond to the changes in the input. Thus, if the frequency of the input signal is greater than $1/\tau_t$, the minority carriers will not have time to transit the base and the collector current will remain unmodulated by the input signal. One can set the upper frequency limit at $\sim 1/\tau_t$ which is 518 MHz.

EMITTER INJECTION EFFICIENCY γ **EXAMPLE 6.9**

a. Consider a *pnp* transistor with the parameters as defined in Figure 6.20. Show that the **injection efficiency of the emitter**, defined as

$$\gamma = \frac{\text{Emitter current due to minority carriers injected into the base}}{\text{Total emitter current}}$$

is given by

$$\gamma = \frac{1}{1 + \dfrac{N_d W_B \mu_{e(\text{emitter})}}{N_a W_E \mu_{h(\text{base})}}}$$

b. How would you modify the CB current gain α to include the emitter injection efficiency?

c. Calculate the emitter injection efficiency for the *pnp* transistor in Example 6.8, which has an acceptor doping of 2×10^{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

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

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 6.8 had a β of 206. It is clear that the emitter injection efficiency γ and the base transport factor α_T have comparable impacts in controlling the overall gain in the example. We neglected the recombination of

electrons and holes in the EB depletion region. In fact, if we were to also consider this recombination component of the emitter current, $I_{E(\text{hole})}$ would have to be even smaller compared with the total I_E, which would make γ and hence β even lower.

6.6.2 COMMON BASE AMPLIFIER

According to Equation 6.31 the emitter current depends exponentially on V_{EB},

$$I_E = I_{EO} \exp\left(\frac{eV_{EB}}{kT}\right) \qquad [6.42]$$

It is therefore apparent that small changes in V_{EB} lead to large changes in I_E. Since $I_C \approx I_E$, we see that small variations in V_{EB} cause large changes in I_C in the collector circuit. This can be fruitfully used to obtain voltage amplification as shown in Figure 6.23. The battery V_{CC}, through R_C, provides a reverse bias for the base–collector junction. The dc voltage V_{EE} forward biases the EB junction, which means that it provides a dc current I_E. The input signal is the ac voltage v_{eb} applied in series with the dc bias voltage V_{EE} to the EB junction. The applied signal v_{eb} modulates the total voltage V_{EB} across the EB junction and hence, by virtue of Equation 6.30, modulates the injected hole concentration $p_n(0)$ up and down about the dc value determined by V_{EE} as depicted in Figure 6.23. This variation in $p_n(0)$ alters the concentration gradient and therefore gives rise to a change in I_E, and hence a nearly identical change in I_C. The change in the collector current can be converted to a voltage change by using a resistor R_C in the collector circuit as shown in Figure 6.23. However, the output is commonly taken between the collector, and the base and this voltage V_{CB} is

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

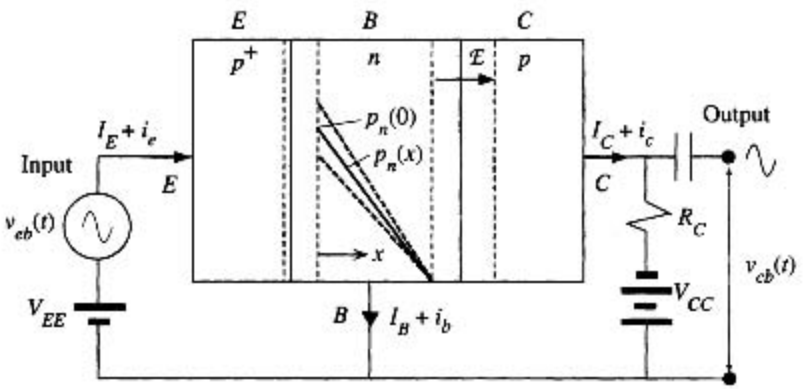


Figure 6.23 A *pnp* transistor operated in the active region in the common base amplifier configuration.

The applied (input) signal v_{eb} modulates the dc voltage across the EB junction and hence modulates the injected hole concentration up and down about the dc value $p_n(0)$. The solid line shows $p_n(x)$ when only the dc bias V_{EE} is present. The dashed lines show how $p_n(x)$ is modulated up and down by the signal v_{eb} superimposed on V_{EE}.

Increasing the emitter–base voltage V_{EB} (by increasing v_{eb}) increases I_C, which increases V_{CB}. Since we are interested in ac signals, that voltage variation across CB is tapped out through a dc blocking capacitor in Figure 6.23.

For simplicity we will assume that changes δV_{EB} and δI_E in the dc values of V_{EB} and I_E are small, which means that δV_{EB} and δI_E can be related by differentiating Equation 6.42. We are hence tacitly assuming an operation under small signals. Further, we will take the changes to represent the ac signal magnitudes, $v_{eb} = \delta V_{EB}$, $i_e = \delta I_E$, $i_c = \delta I_C \approx \delta I_E \approx i_e$, $v_{cb} = \delta V_{CB}$.

The output signal voltage v_{cb} corresponds to the change in V_{CB},

$$v_{cb} = \delta V_{CB} = R_C\,\delta I_C = R_C\,\delta I_E$$

The variation in the emitter current δI_E depends on the variation δV_{EB} in V_{EB}, which can be determined by differentiating Equation 6.42,

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

By definition, δV_{EB} is the input signal v_{eb}. The change δI_E in I_E is the input signal current (i_e) flowing into the emitter as a result of δV_{EB}. Therefore the quantity $\delta V_{EB}/\delta I_E$ represents an input resistance r_e seen by the source v_{eb}.

Input resistance

$$r_e = \frac{\delta V_{EB}}{\delta I_E} = \frac{kT}{eI_E} = \frac{25}{I_E(\text{mA})} \qquad [6.43]$$

The output signal is then

$$v_{cb} = R_C\,\delta I_E = R_C \frac{v_{eb}}{r_e}$$

so the voltage amplification is

CB voltage gain

$$A_V = \frac{v_{cb}}{v_{eb}} = \frac{R_C}{r_e} \qquad [6.44]$$

To obtain a voltage gain we obviously need $R_C > r_e$, which is invariably the case by the appropriate choice of I_E, hence r_e, and R_C. For example, when the BJT is biased so that I_E is 10 mA and r_e is 2.5 Ω, and if R_C is chosen to be 50 Ω, then the gain is 20.

EXAMPLE 6.10 **A COMMON BASE AMPLIFIER** Consider a *pnp* Si BJT that has been connected as in Figure 6.23. The BJT has a $\beta = 135$ and has been biased to operate with a 5 mA collector current. What is the small-signal input resistance? What is the required R_C that will provide a voltage gain of 20? What is the base current? What should be the V_{CC} in Figure 6.23? Suppose $V_{CC} = -6$ V, what is the largest swing in the output voltage V_{CB} in Figure 6.23 as the input signal is increased and decreased about the bias point V_{EE}, taken as 0.65 V?

SOLUTION

The emitter and collector currents are approximately the same. From Equation 6.43,

$$r_e = \frac{25}{I_E\,(\text{mA})} = \frac{25}{5} = 5\ \Omega$$

The voltage gain A_V from Equation 6.44 is

$$A_V = \frac{R_C}{r_e} \quad \text{or} \quad 20 = \frac{R_C}{5\ \Omega}$$

so a gain of 20 requires $R_C = 100\ \Omega$.

$$\text{Base current } I_B = \frac{I_C}{\beta} = \frac{5\text{ mA}}{135} = 0.037\text{ mA} \quad \text{or} \quad 37\ \mu\text{A}$$

There is a dc voltage across R_C given by $I_C R_C = (0.005\text{ A})(100\ \Omega) = 0.5$ V. V_{CC} has to provide the latter voltage across R_C and also a sufficient voltage to keep the BC junction reverse biased at all times under normal operation. Let us set $V_{CC} = -6$ V. Thus, in the absence of any input signal v_{eb}, V_{CB} is set to $-6\text{ V} + 0.5\text{ V} = -5.5$ V. As we increase the signal v_{eb}, V_{EB} and hence I_C increase until the point C becomes nearly zero,[9] that is, $V_{CB} = 0$, which occurs when I_C is maximum at $I_{C\max} = |V_{CC}|/R_C$ or 60 mA. As v_{eb} decreases, so does V_{EB} and hence I_C. Eventually I_C will simply become zero, and point C will be at -6 V, so $V_{CB} = V_{CC}$. Thus, V_{CB} can only swing from -5.5 V to 0 V (for increasing input until $I_C = I_{C\max}$), or from -5.5 to -6 V (for decreasing input until $I_C = 0$).

6.6.3 COMMON EMITTER (CE) DC CHARACTERISTICS

An *npn* bipolar transistor when connected in the common emitter (CE) configuration has the emitter common to both the input and output circuits, as shown in Figure 6.24a. The dc voltage V_{BE} forward biases the BE junction and thereby injects electrons as minority carriers into the base. These electrons diffuse to the collector junction where the field $\mathcal{E}$ sweeps them into the collector to constitute the collector current I_C. V_{BE} controls the current I_E and hence I_B and I_C. The advantage of the CE configuration is that the **input current** is the current flowing between the ac source and the base, which is the base current I_B. This current is much smaller than the emitter current by about a factor of β. The output current is the current flowing between V_{CE} and the collector, which is I_C. In the CE configuration, the dc voltage V_{CE} must be greater than V_{BE} to reverse bias the collector junction and collect the diffusing electrons in the base.

The dc characteristics of the BJT in the CE configuration are normally given as I_C versus V_{CE} for various values of fixed base currents I_B, as shown in Figure 6.24b. The characteristics can be readily understood by Equations 6.40 and 6.41. We should note that, in practice, we are essentially adjusting V_{BE} to obtain the desired I_B because, by Equation 6.41,

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

and I_E depends on V_{BE} via Equation 6.42.

Increasing I_B requires increasing V_{BE}, which increases I_C. Using Equations 6.40 and 6.41, we can obtain I_C in terms of I_B alone,

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

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

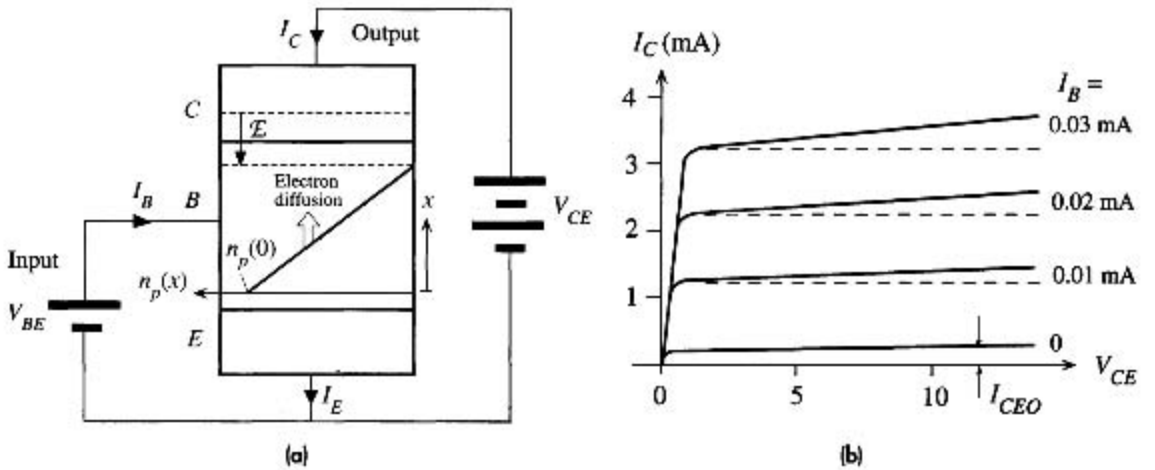


Figure 6.24

(a) An *npn* transistor operated in the active region in the common emitter configuration. The input current is the current that flows between V_{BE} and the base which is I_B.

(b) DC I–V characteristics of the *npn* bipolar transistor in the CE configuration. (Exaggerated to highlight various effects.)

Active region collector current

or

$$I_C = \beta I_B + I_{CEO} \qquad [6.45]$$

where

$$I_{CEO} = \frac{I_{CBO}}{(1-\alpha)} \approx \beta I_{CBO}$$

is the leakage current into the collector when the base is open circuited. This is much larger in the CE circuit than in the CB configuration.

Even when I_B is kept constant, I_C still exhibits a small increase with V_{CE}, which, according to Equation 6.45, indicates an increase in the current gain β with V_{CE}. This is due to the Early effect or modulation of the base width by V_{CB}, shown in Figure 6.22. Increasing V_{CE} increases V_{CB}, which increases W_{BC}, reduces W_B, and hence shortens τ_t. The resulting effect is a larger β ($\approx \tau_h/\tau_t$).

When V_{CE} is less than V_{BE}, the collector junction becomes forward biased and Equation 6.45 is not valid. The collector current is then the difference between forward currents of emitter and collector junctions. The transistor operating in this region is said to be **saturated.**

6.6.4 LOW-FREQUENCY SMALL-SIGNAL MODEL

The *npn* bipolar transistor in the CE (common emitter) amplifier configuration is shown in Figure 6.25. The input circuit has a dc bias V_{BB} to forward bias the base–emitter (BE) junction and the output circuit has a dc voltage V_{CC} (larger than V_{BB}) to reverse bias the base–collector (BC) junction through a collector resistor R_C.

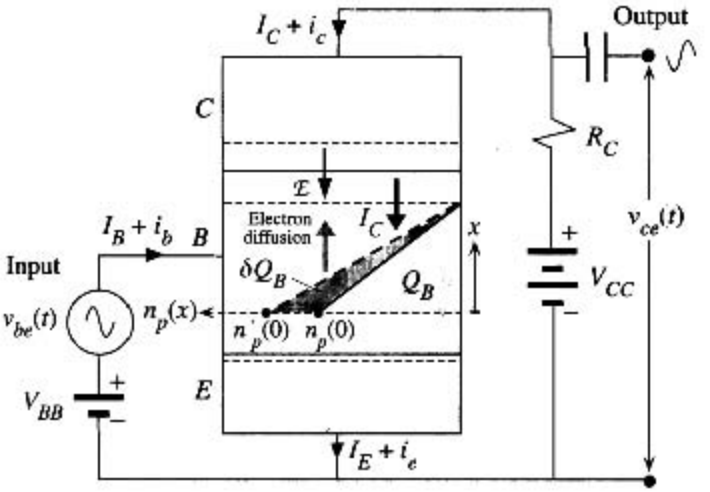


Figure 6.25 An *npn* transistor operated in the active region in the common emitter amplifier configuration.
The applied signal v_{be} modulates the dc voltage across the BE junction and hence modulates the injected electron concentration up and down about the dc value $n_p(0)$. The solid line shows $n_p(x)$ when only the dc bias V_{BB} is present. The dashed line shows how $n_p(x)$ is modulated up by a positive small signal v_{be} superimposed on V_{BB}.

The actual reverse bias voltage across the BC junction is $V_{CE} - V_{BE}$, where V_{CE} is

$$V_{CE} = V_{CC} - I_C R_C$$

An input signal in the form of a small ac signal v_{be} is applied in series with the bias voltage V_{BB} and modulates the voltage V_{BE} across the BE junction about its dc value V_{BB}. The varying voltage across the BE modulates $n_p(0)$ up and down about its dc value, which leads to a varying emitter current and hence to an almost identically varying collector current in the output circuit. The variation in the collector current is converted to an output voltage signal by the collector resistance R_C. Note that increasing V_{BE} increases I_C, which leads to a decrease in V_{CE}. Thus, the output voltage is 180° out of phase with the input voltage.

Since the BE junction is forward-biased, the relationship between I_E and V_{BE} is exponential,

$$I_E = I_{EO} \exp\left(\frac{eV_{BE}}{kT}\right) \qquad [6.46]$$

Emitter current and V_{BE}

where I_{EO} is a constant. We can differentiate this expression to relate small variations in I_E and V_{BE} as in the presence of small signals superimposed on dc values. For small signals, we have $v_{be} = \delta V_{BE}$, $i_b = \delta I_B$, $i_e = \delta I_E$, $i_c = \delta I_C$. Then from Equation 6.45 we see that $\delta I_C = \beta\, \delta I_B$, so $i_c = \beta i_b$. Since $\alpha \approx 1$, $i_e \approx i_c$.

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

$$r_{be} = \frac{v_{be}}{i_b} = \frac{\delta V_{BE}}{\delta I_B} \approx \beta \frac{\delta V_{BE}}{\delta I_E} = \frac{\beta kT}{eI_E} \approx \frac{\beta 25}{I_C(\text{mA})} \qquad [6.47]$$

Input resistance

where we differentiated Equation 6.46.

The output ac signal v_{ce} develops across the CE and is tapped out through a capacitor. Since $V_{CE} = V_{CC} - I_C R_C$, as I_C increases, V_{CE} decreases. Thus,

$$v_{ce} = \delta V_{CE} = -R_C\, \delta I_C = -R_C i_c$$

The voltage amplification is

Voltage gain

$$A_V = \frac{v_{ce}}{v_{be}} = \frac{-R_C i_c}{r_{be} i_b} = \frac{-R_C \beta}{r_{be}} \approx -\frac{R_C I_C(\text{mA})}{25} \qquad [6.48]$$

which is the same as that in the CB configuration. However, in the CE configuration the output to input current ratio $i_c/i_b = \beta$, whereas this is almost unity in the CB configuration. Consequently, the CE configuration provides a greater power amplification, which is the second advantage of the CE circuit.

The input signal v_{be} gives rise to an output current i_c. This input voltage to output current conversion is defined in a parameter called the **mutual conductance,** or **transconductance,** g_m.

Transconductance

$$g_m = \frac{i_c}{v_{be}} \approx \frac{\delta I_E}{\delta V_{BE}} = \frac{I_E(\text{mA})}{25} = \frac{1}{r_e} \qquad [6.49]$$

The voltage amplification of the CE amplifier is then

Voltage gain

$$A_V = -g_m R_C \qquad [6.50]$$

We generally find it convenient to use a small-signal equivalent circuit for the low-frequency behavior of a BJT in the CE configuration. Between the base and emitter, the applied ac source voltage v_s sees only an input resistance of r_{be}, as shown in Figure 6.26. To underline the importance of the transistor input resistance, the output (or the internal) resistance R_s of the ac source is also shown. In the output circuit there is a voltage-controlled current source i_c which generates a current of $g_m v_{be}$. The current i_c passes through the load (or collector) resistance R_C across which the voltage signal develops. As we are only interested in ac signals, the batteries are taken as a short-circuit path for the ac current, which means that the internal resistances of the batteries are taken as zero. This model, of course, is valid only under normal and active operating conditions and small signals about dc values, and at low frequencies.

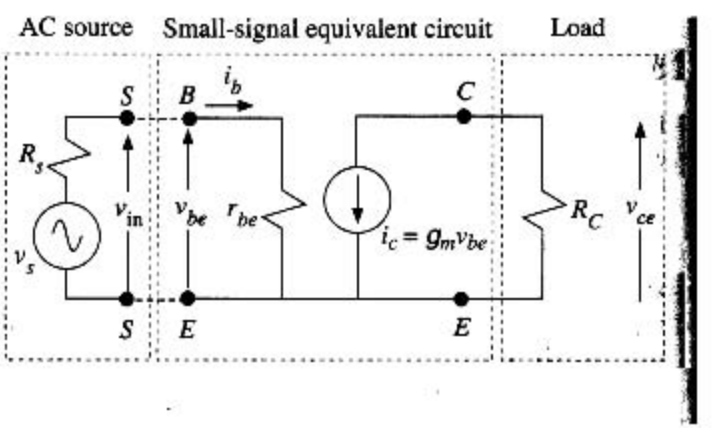
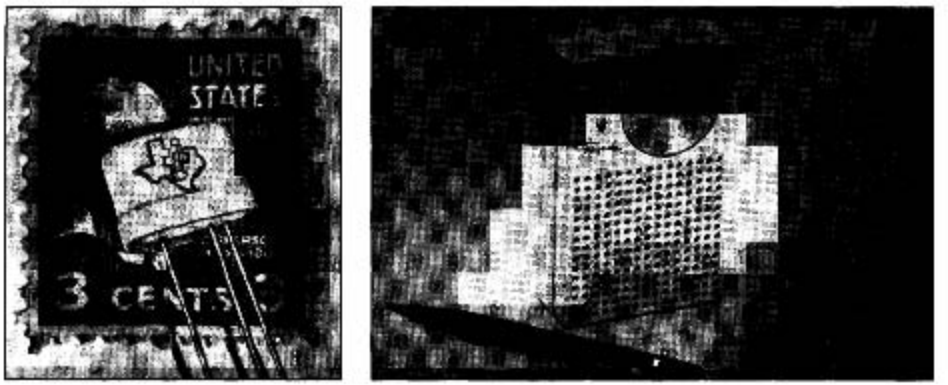


Figure 6.26 Low-frequency small-signal simplified equivalent circuit of the bipolar transistor in the CE configuration with a load resistor R_C in the collector circuit.



Left: The first commercial Si transistor from Texas Instruments (1954). Right: The first transistor pocket radio (1954). It had four Ge *npn* transistors.

SOURCE: Courtesy of Texas Instruments.

The bipolar transistor general dc current equation $I_C = \beta I_B$, where $\beta \approx \tau_h/\tau_t$ is a material-dependent constant, implies that the ac small-signal collector current is

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

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 6.47 and 6.49, we have

β at low frequencies

$$\beta = g_m r_{be} \qquad \textbf{[6.51]}$$

In transistor data books, the dc current gain I_C/I_B is denoted as h_{FE} whereas the small-signal ac current gain i_c/i_b is denoted as h_{fe}. Except at high frequencies, $h_{fe} \approx h_{FE}$.

EXAMPLE 6.11

CE LOW-FREQUENCY SMALL-SIGNAL EQUIVALENT CIRCUIT Consider a BJT with a β of 100, used in a CE amplifier in which the collector current is 2.5 mA and R_C is 1 kΩ. If the ac source has an rms voltage of 1 mV and an output resistance R_s of 50 Ω, what is the rms output voltage? What is the input and output power and the overall power amplification?

SOLUTION

As the collector current is 2.5 mA, the input resistance and the transconductance are

$$r_{be} = \frac{\beta 25}{I_C(\text{mA})} = \frac{(100)(25)}{2.5} = 1000\ \Omega$$

and

$$g_m = \frac{I_C(\text{mA})}{25} = \frac{2.5}{25} = 0.1\ \text{A/V}$$

The *magnitude* of the voltage gain of the BJT small-signal equivalent circuit is

$$A_V = \frac{v_{ce}}{v_{be}} = g_m R_C = (0.1)(1000) = 100$$

When the ac source is connected to the B and E terminals (Figure 6.26), the input resistance r_{be} of the BJT loads the ac source, so v_{be} across BE is

$$v_{be} = v_s \frac{r_{be}}{(r_{be} + R_s)} = (1\text{ mV})\frac{1000\ \Omega}{(1000\ \Omega + 50\ \Omega)} = 0.952\text{ mV}$$

The output voltage (rms) is, therefore,

$$v_{ce} = A_V v_{be} = 100(0.952\text{ mV}) = 95.2\text{ mV}$$

The loading effect makes the output less than 100 mV. To reduce the loading of the ac source, we need to increase r_{be}, *i.e.*, reduce the collector current, but that also reduces the gain. So to keep the gain the same, we need to reduce I_C and increase R_C. However, R_C cannot be increased indefinitely because R_C itself is loaded by the input of the next stage and, in addition, there is an incremental resistance between the collector and emitter terminals (typically $\sim$100 kΩ) that shunts R_C (not shown in Figure 6.26).

The power amplification of the CE BJT itself is

$$A_P = \frac{i_c v_{ce}}{i_b v_{be}} = \beta A_V = (100)(100) = 10{,}000$$

The input power into the BE terminals is

$$P_{in} = v_{be} i_b = \frac{v_{be}^2}{r_{be}} = \frac{(0.952 \times 10^{-3}\text{ V})^2}{1000\ \Omega} = 9.06 \times 10^{-10}\text{ W} \quad \text{or} \quad 0.906\text{ nW}$$

The output power is

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

6.7 JUNCTION FIELD EFFECT TRANSISTOR (JFET)

6.7.1 GENERAL PRINCIPLES

The basic structure of the junction field effect transistor (JFET) with an n-type channel (n-channel) is depicted in Figure 6.27a. An n-type semiconductor slab is provided with contacts at its ends to pass current through it. These terminals are called **source** (S) and **drain** (D). Two of the opposite faces of the n-type semiconductor are heavily p-type doped to some small depth so that an n-type channel is formed between the source and drain terminals, as shown in Figure 6.27a. The two p^+ regions are normally electrically connected and are called the **gate** (G). As the gate is heavily doped, the depletion layers extend almost entirely into the n-channel, as shown in Figure 6.27. For simplicity we will assume that the two gate regions are identical (both p^+ type) and that the doping in the n-type semiconductor is uniform. We will define the n-channel to be the region of conducting n-type material contained between the two depletion layers.

The basic and idealized symmetric structure in Figure 6.27a is useful in explaining the principle of operation as discussed later but does not truly represent

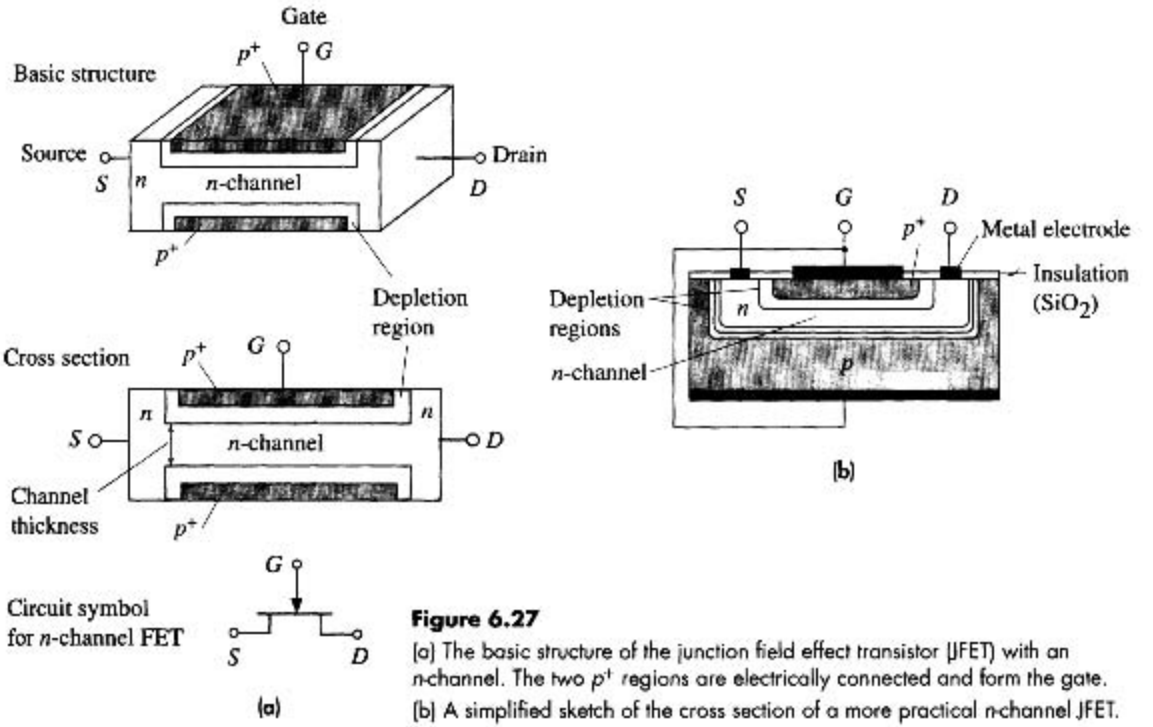


Figure 6.27
(a) The basic structure of the junction field effect transistor (JFET) with an *n*-channel. The two p^+ regions are electrically connected and form the gate. (b) A simplified sketch of the cross section of a more practical *n*-channel JFET.

the structure of a typical practical device. A simplified schematic sketch of the cross section of a more practical device (as, for example, fabricated by the planar technology) is shown in Figure 6.27b where it is apparent that the two gate regions do not have identical doping and that, except for one of the gates, all contacts are on one surface.

We first consider the behavior of the JFET with the gate and source shorted ($V_{GS} = 0$), as shown in Figure 6.28a. The resistance between S and D is essentially the resistance of the conducting n-channel between A and B, R_{AB}. When a positive voltage is applied to D with respect to S ($V_{DS} > 0$), then a current flows from D to S, which is called the **drain current** I_D. There is a voltage drop along the channel, between A and B, as indicated in Figure 6.28a. The voltage in the n-channel is zero at A and V_{DS} at B. As the voltage along the n-channel is positive, the p^-n junctions between the gates and the n-channel become progressively more reverse-biased from A to B. Consequently the depletion layers extend more into the channel and thereby decrease the thickness of the conducting channel from A to B.

Increasing V_{DS} increases the widths of the depletion layers, which penetrate more into the channel and hence result in more channel narrowing toward the drain. The resistance of the n-channel R_{AB} therefore increases with V_{DS}. The drain current therefore does not increase linearly with V_{DS} but falls below it because

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

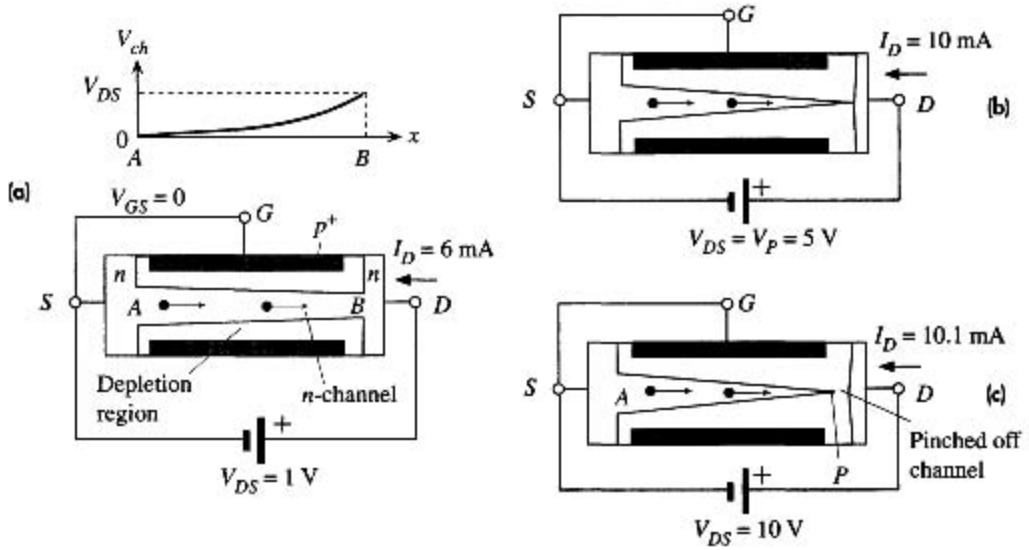


Figure 6.28

(a) The gate and source are shorted ($V_{GS} = 0$) and V_{DS} is small.

(b) V_{DS} has increased to a value that allows the two depletion layers to just touch, when $V_{DS} = V_P$ (= 5 V) and the p^+n junction voltage at the drain end, $V_{GD} = -V_{DS} = -V_P = -5$ V.

(c) V_{DS} is large ($V_{DS} > V_P$), so a short length of the channel is pinched off.

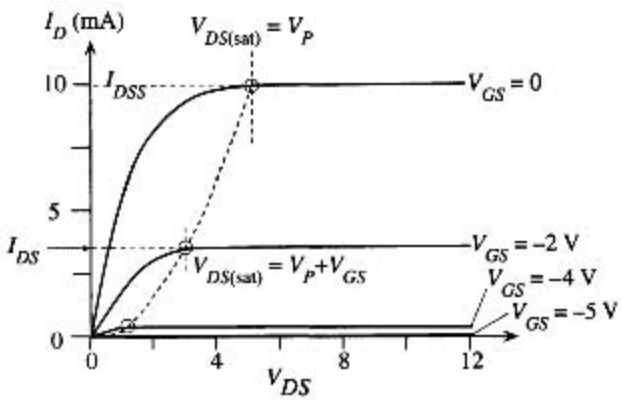


Figure 6.29 Typical I_D versus V_{DS} characteristics of a JFET for various fixed gate voltages V_{GS}.

and R_{AB} increases with V_{DS}. Thus I_D versus V_{DS} exhibits a sublinear behavior, as shown in the $V_{DS} < 5$ V region in Figure 6.29.

As V_{DS} increases further, the depletion layers extend more into the channel and eventually, when $V_{DS} = V_P$ (= 5 V), the two depletion layers around B meet at point P at the drain end of the channel, as depicted in Figure 6.28b. The channel is then said to be "pinched off" by the two depletion layers. The voltage V_P is called the **pinch-off voltage.** It is equal to the magnitude of reverse bias needed across the p^+n junctions to

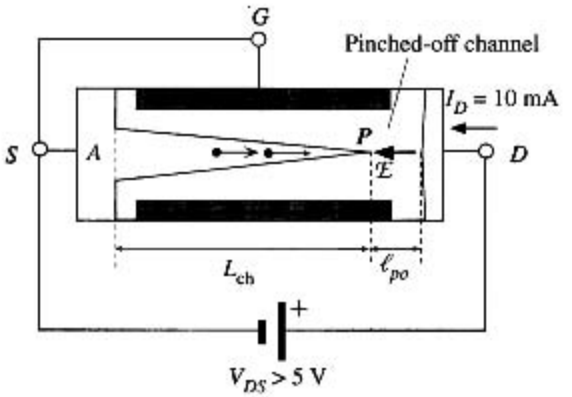


Figure 6.30 The pinched-off channel and conduction for $V_{DS} > V_P$ (= 5 V).

make them just touch at the drain end. Since the actual bias voltage across the p^+n junctions at the drain end (B) is V_{GD}, the pinch-off occurs whenever

$$V_{GD} = -V_P \qquad [6.52]$$

Pinch-off condition

In the present case, gate to source is shorted, $V_{GS} = 0$, so $V_{GD} = -V_{DS}$ and pinch-off occurs when $V_{DS} = V_P$ (5 V). The drain current from pinch-off onwards, as shown in Figure 6.29, does not increase significantly with V_{DS} for reasons given below. Beyond $V_{DS} = V_P$, there is a short pinched-off channel of length ℓ_{po}.

The pinched-off channel is a reverse-biased depletion region that separates the drain from the n-channel, as depicted in Figure 6.30. There is a very strong electric field $\mathcal{E}$ in this pinched-off region in the D to S direction. This field is the vector sum of the fields from positive donors to negative acceptors in the depletion regions of the channel and the gate on the drain side. Electrons in the n-channel drift toward P, and when they arrive at P, they are swept across the pinched-off channel by $\mathcal{E}$. This process is similar to minority carriers in the base of a BJT reaching the collector junction depletion region, where the internal field there sweeps them across the depletion layer into the collector. Consequently the drain current is actually determined by the resistance of the conducting n-channel over L_{ch} from A to P in Figure 6.30 and not by the pinched-off channel.

As V_{DS} increases, most of the additional voltage simply drops across ℓ_{po} as this region is depleted of carriers and hence highly resistive. Point P, where the depletion layers first meet, moves slightly toward A, thereby slightly reducing the channel length L_{ch}. Point P must still be at a potential V_P because it is this potential that just makes the depletion layers touch. Thus the voltage drop across L_{ch} remains as V_P. Beyond pinch-off then

$$I_D = \frac{V_P}{R_{AP}} \qquad (V_{DS} > V_P)$$

Since R_{AP} is determined by L_{ch}, which decreases slightly with V_{DS}, I_D increases slightly with V_{DS}. In many cases, I_D is conveniently taken to be saturated at a value I_{DSS} for $V_{DS} > V_P$. Typical I_D versus V_{DS} behavior is shown in Figure 6.29.

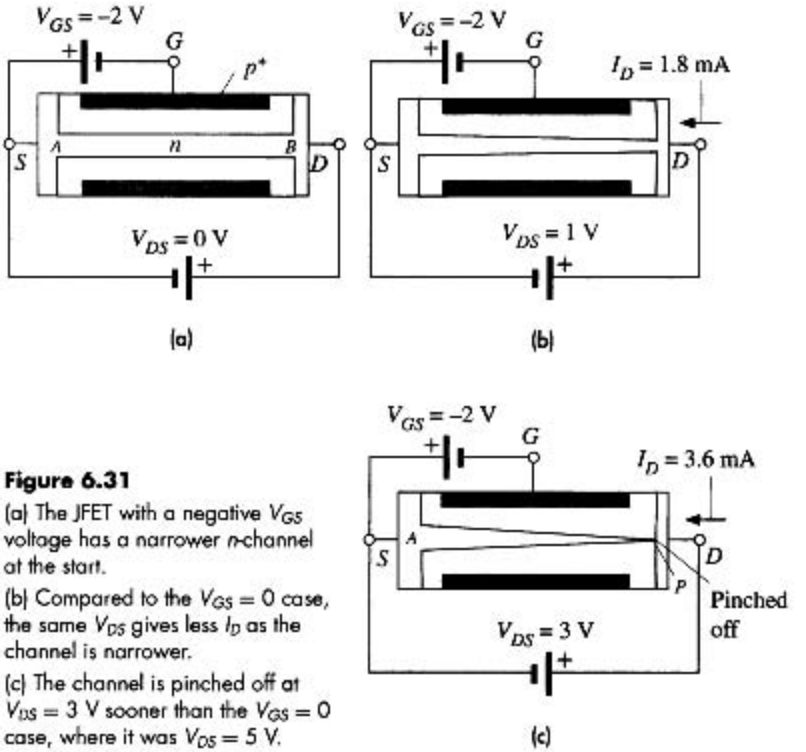


Figure 6.31

(a) The JFET with a negative V_{GS} voltage has a narrower *n*-channel at the start.

(b) Compared to the $V_{GS} = 0$ case, the same V_{DS} gives less I_D as the channel is narrower.

(c) The channel is pinched off at $V_{DS} = 3$ V sooner than the $V_{GS} = 0$ case, where it was $V_{DS} = 5$ V.

We now consider what happens when a negative voltage, say $V_{GS} = -2$ V, is applied to the gate with respect to the source, as shown in Figure 6.31a with $V_{DS} = 0$. The p^+n junctions are now reverse-biased from the start, the channel is narrower, and the channel resistance is now larger than in the $V_{GS} = 0$ case. The drain current that flows when a small V_{DS} is applied, as in Figure 6.31b, is now smaller than in the $V_{GS} = 0$ case as apparent in Figure 6.29. The p^+n junctions are now progressively more reverse-biased from V_{GS} at the source end to $V_{GD} = V_{GS} - V_{DS}$ at the drain end. We therefore need a smaller V_{DS} (= 3 V) to pinch off the channel, as shown in Figure 6.31c. When $V_{DS} = 3$ V, the G to D voltage V_{GD} across the p^+n junctions at the drain end is -5 V, which is $-V_P$, so the channel becomes pinched off. Beyond pinch-off, I_D is nearly saturated just as in the $V_{GS} = 0$ case, but its magnitude is obviously smaller as the thickness of the channel at A is smaller; compare Figures 6.28 and 6.31. In the presence of V_{GS}, the pinch-off occurs at $V_{DS} = V_{DS(sat)}$, and from Equation 6.52.

Pinch-off condition

$$V_{DS(sat)} = V_P + V_{GS} \qquad [6.53]$$

where V_{GS} is a negative voltage (reducing V_P). Beyond pinch-off when $V_{DS} > V_{DS(sat)}$, the point P where the channel is *just pinched* still remains at potential $V_{DS(sat)}$, given by Equation 6.53.

For $V_{DS} > V_{DS(sat)}$, I_D becomes nearly saturated at a value denoted as I_{DS}, which is indicated in Figure 6.29. When G and S are shorted ($V_{GS} = 0$), I_{DS} is called I_{DSS} (which

stands for I_{DS} with shorted gate to source). Beyond pinch-off, with negative V_{GS}, I_{DS} is

$$I_D \approx I_{DS} \approx \frac{V_{DS(sat)}}{R_{AP}(V_{GS})} = \frac{V_P + V_{GS}}{R_{AP}(V_{GS})} \qquad V_{DS} > V_{DS(sat)} \qquad (6.54)$$

where R_{AP} (V_{GS}) is the effective resistance of the conducting n-channel from A to P (Figure 6.31b), which depends on the channel thickness and hence on V_{GS}. The resistance increases with more negative gate voltage as this increases the reverse bias across the p^+n junctions, which leads to the narrowing of the channel. For example, when $V_{GS} = -4$ V, the channel thickness at A becomes narrower than in the case with $V_{GS} = -2$ V, thereby increasing the resistance, R_{AP}, of the conducting channel and therefore decreasing I_{DS}. Further, there is also a reduction in the drain current by virtue of $V_{DS(sat)}$ decreasing with negative V_{GS}, as apparent in Equation 6.54. Figure 6.29 shows the effect of the gate voltage on the I_D versus V_{DS} behavior. The two effects, that from $V_{DS(sat)}$ and that from R_{AP} (V_{GS}) in Equation 6.54, lead to I_{DS} almost decreasing parabolically with $-V_{GS}$.

When the gate voltage is such that $V_{GS} = -V_P$ $(= -5$ V) with the source and drain shorted ($V_{DS} = 0$), then the two depletion layers touch over the entire channel length and the whole channel is closed, as illustrated in Figure 6.32. The channel is said to be off. The only drain current that flows when a V_{DS} is applied is due to the thermally generated carriers in the depletion layers. This current is very small.

Figure 6.29 summarizes the full I_D versus V_{DS} characteristics of the n-channel JFET at various gate voltages V_{GS}. It is apparent that I_{DS} is relatively independent of V_{DS} and that it is controlled by the gate voltage V_{GS}, as expected by Equation 6.54. This is analogous to the BJT in which the collector current I_C is controlled by the base–emitter bias voltage V_{BE}. Figure 6.33a shows the dependence of I_{DS} on the gate voltage V_{GS}. The transistor action is the control of the drain current I_{DS}, in the drain–source (output) circuit by the voltage V_{GS} in the gate–source (input circuit), as shown in Figure 6.33b. This control is only possible if $V_{DS} > V_{DS(sat)}$. When $V_{GS} = -V_P$, the drain current is nearly zero because the channel has been totally pinched off. This gate–source voltage is denoted by $V_{GS(off)}$ as the drain current has been switched off. Furthermore, we should note that as V_{GS} reverse biases the p^+n junction, the current into the gate I_G is the reverse leakage current of these junctions. It is usually very small. In some JFETs, I_G is as low as a fraction of a nanoampere. We should also note that the circuit symbol for the JFET, as shown in Figure 6.27a, has an arrow to identify the gate and the pn junction direction.

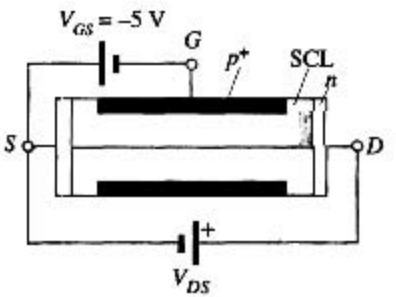


Figure 6.32 When $V_{GS} = -5$ V, the depletion layers close the whole channel from the start, at $V_{DS} = 0$. As V_{DS} is increased, there is a very small drain current, which is the small reverse leakage current due to thermal generation of carriers in the depletion layers.

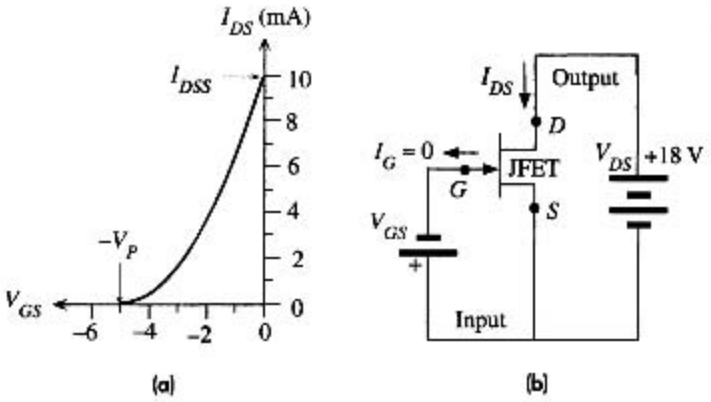


Figure 6.33
(a) Typical I_{DS} versus V_{GS} characteristics of a JFET.
(b) The dc circuit where V_{GS} in the gate–source circuit (input) controls the drain current I_{DS} in the drain–source (output) circuit in which V_{DS} is kept constant and large ($V_{DS} > V_P$).

Is there a convenient relationship between I_{DS} and V_{GS}? If we calculate the effective resistance R_{AP} of the n-channel between A and P, we can obtain its dependence on the channel thickness, and thus on the widths of the depletion layers and hence on V_{GS}. We can then find I_{DS} from Equation 6.54. It turns out that a simple parabolic dependence seems to represent the data reasonably well,

Beyond pinch-off

$$I_{DS} = I_{DSS}\left[1 - \left(\frac{V_{GS}}{V_{GS(\text{off})}}\right)\right]^2 \qquad [6.55]$$

where I_{DSS} is the drain current when $V_{GS} = 0$ (Figure 6.33) 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 6.54 and 6.55 hold, is commonly called the **current saturation region,** as well as **constant current region** and **pentode region.** The term **saturation** should not be confused with similar terms used for saturation effects in bipolar transistors. A saturated BJT cannot be used as an amplifier, but JFETs are invariably used as amplifiers in the saturated current region.

6.7.2 JFET AMPLIFIER

The transistor action in the JFET is the control of I_{DS} by V_{GS}, as shown in Figure 6.33. The input circuit is therefore the gate–source circuit containing V_{GS} and the output circuit is the drain–source circuit in which the drain current I_{DS} flows. The JFET is almost never used with the pn junction between the gate and channel forward-biased ($V_{GS} > 0$) as this would lead to a very large gate current and near shorting of the gate to source

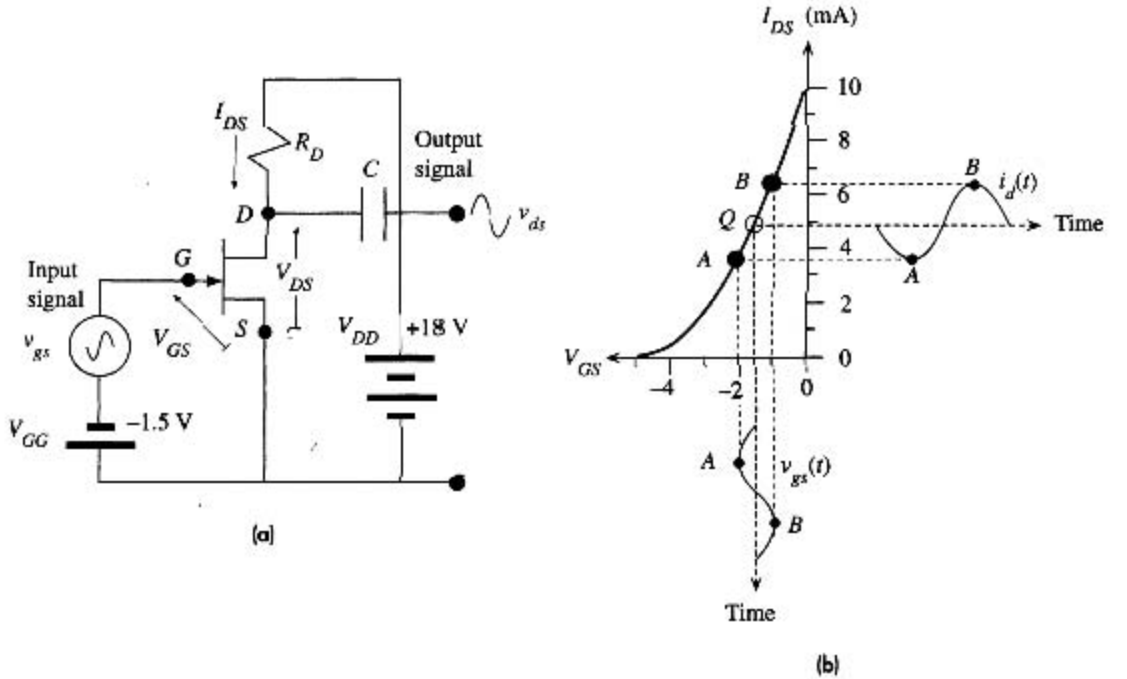


Figure 6.34

(a) Common source (CS) ac amplifier using a JFET.

(b) Explanation of how I_D is modulated by the signal v_{gs} in series with the dc bias voltage V_{GG}.

voltage. With V_{GS} limited to negative voltages, the maximum current in the output circuit can only be I_{DSS}, as shown in Figure 6.33a. The maximum input voltage V_{GS} should therefore give an I_{DS} less than I_{DSS}.

Figure 6.34a shows a simplified illustration of a typical JFET voltage amplifier. As the source is common to both the input and output circuits, this is called a **common source** (CS) **amplifier.** The input signal is the ac source v_{gs} connected in series with a negative dc bias voltage V_{GG} of -1.5 V in the GS circuit. First we will find out what happens when there is no ac signal in the circuit ($v_{gs} = 0$). The dc supply (-1.5 V) in the input provides a negative dc voltage to the gate and therefore gives a dc current I_{DS} in the output circuit (less than I_{DSS}). Figure 6.34b shows that when $V_{GS} = -1.5$ V, point Q on the I_{DS} versus V_{GS} characteristics gives $I_{DS} = 4.9$ mA. Point Q, which determines the dc operation, is called the **quiescent point.**

The ac source v_{gs} is connected in series with the negative dc bias voltage V_{GS}. It therefore modulates V_{GS} up and down about -1.5 V with time, as shown in Figure 6.34b. Suppose that v_{gs} varies sinusoidally between -0.5 V and $+0.5$ V. Then, as shown in Figure 6.34b when v_{gs} is -0.5 V (point A), $V_{GS} = -2.0$ V and the drain current is given by point A on the I_{DS}–V_{GS} curve and is about 3.6 mA. When v_{gs} is $+0.5$ V (point B), then $V_{GS} = -1.0$ V and the drain current is given by point B on the I_{DS}–V_{GS} curve and is about 6.4 mA. The input variation from -0.5 V to $+0.5$ V has thus been

Table 6.1 Voltage and current in the common source amplifier of Figure 6.34a

v_{gs} (V)	V_{GS} (V)	I_{DS} (mA)	i_d (mA)	$V_{DS} = V_{DD} - I_{DS}R_D$	v_{ds} (V)	Voltage Gain	Comment
0	−1.5	4.9	0	8.2	0		dc conditions, point Q
−0.5	−2.0	3.6	−1.3	10.8	+2.6	−5.2	Point A
+0.5	−1.0	6.4	+1.5	5.2	−3.0	−6	Point B

NOTE: $V_{DD} = 18$ V and $R_D = 2000\ \Omega$.

converted to a drain current variation from 3.6 mA to 6.4 mA as indicated in Figure 6.34b. We could have just as easily calculated the drain current from Equation 6.55. Table 6.1 summarizes what happens to the drain current as the ac input voltage is varied about zero.

The change in the drain current with respect to its dc value is the output signal current denoted as i_d. Thus at A,

$$i_d = 3.6 - 4.9 = -1.3 \text{ mA}$$

and at B,

$$i_d = 6.4 - 4.9 = 1.5 \text{ mA}$$

The variation in the output current is not quite symmetric as that in the input signal v_{gs} because the I_{DS}–V_{GS} relationship, Equation 6.55, is not linear.

The drain current variations in the DS circuit are converted to voltage variations by the resistance R_D. The voltage across DS is

$$V_{DS} = V_{DD} - I_{DS} R_D \qquad [6.56]$$

where V_{DD} is the bias battery voltage in the DS circuit. Thus, variations in I_{DS} result in variations in V_{DS} that are in the opposite direction or 180° out of phase. The ac output voltage between D and S is tapped out through a capacitor C, as shown in Figure 6.34a. The capacitor C simply blocks the dc. Suppose that $R_D = 2000\ \Omega$ and $V_{DD} = 18$ V, then using Equation 6.56 we can calculate the dc value of V_{DS} and also the minimum and maximum values of V_{DS}, as shown in Table 6.1.

It is apparent that as v_{gs} varies from −0.5 V, at A, to +0.5 V, at B, V_{DS} varies from 10.8 V to 5.2 V, respectively. The change in V_{DS} with respect to dc is what constitutes the output signal v_{ds}, as only the ac is tapped out. From Equation 6.56, the change in V_{DS} is related to the change in I_{DS} by

$$v_{ds} = -R_D i_d \qquad [6.57]$$

Thus the output, v_{ds}, changes from −3.0 V to 2.6 V. The peak-to-peak voltage amplification is

$$A_{V(\text{pk-pk})} = \frac{\Delta V_{DS}}{\Delta V_{GS}} = \frac{v_{ds(\text{pk-pk})}}{v_{gs(\text{pk-pk})}} = \frac{-3 \text{ V} - (2.6 \text{ V})}{0.5 \text{ V} - (-0.5 \text{ V})} = -5.6$$

The negative sign represents the fact that the output and input voltages are out of phase by 180°. This can also be seen from Table 6.1 where a negative v_{gs} results in a positive v_{ds}. Even though the ac input signal v_{gs} is symmetric about zero, ±0.5 V, the ac output signal v_{ds} is not symmetric, which is due to the I_{DS} versus V_{GS} curve being nonlinear, and thus varies between −3.0 V and 2.6 V. If we were to calculate the voltage amplification for the most negative input signal, we would find −5.2, whereas for the most positive input signal, it would be −6. The peak-to-peak voltage amplification, which was −5.6, represents a mean gain taking both negative and positive input signals into account.

The amplification can of course be increased by increasing R_D, but we must maintain V_{DS} at all times above $V_{DS(\text{sat})}$ (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,

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

Definition of JFET transconductance

This transconductance can be found by differentiating Equation 6.55,

$$g_m = \frac{dI_{DS}}{dV_{GS}} = -\frac{2I_{DSS}}{V_{GS(\text{off})}}\left[1 - \left(\frac{V_{GS}}{V_{GS(\text{off})}}\right)\right] = -\frac{2[I_{DSS}I_{DS}]^{1/2}}{V_{GS(\text{off})}} \qquad [6.58]$$

JFET transconductance

The output signal current is

$$i_d = g_m v_{gs}$$

so using Equation 6.57, the small-signal voltage amplification is

$$A_V = \frac{v_{ds}}{v_{gs}} = \frac{-R_D(g_m v_{gs})}{v_{gs}} = -g_m R_D \qquad [6.59]$$

Small-signal voltage gain

Equation 6.59 is only valid under small-signal conditions in which the variations about the dc values are small compared with the dc values themselves. The negative sign indicates that v_{ds} and v_{gs} are 180° out of phase.

EXAMPLE 6.12

THE JFET AMPLIFIER Consider the *n*-channel JFET common source amplifier shown in Figure 6.34a. The JFET has an I_{DSS} of 10 mA and a pinch-off voltage V_P of 5 V as in Figure 6.34b. Suppose that the gate dc bias voltage supply $V_{GG} = -1.5$ V, the drain circuit supply $V_{DD} = 18$ V, and $R_D = 2000\ \Omega$. What is the voltage amplification for small signals? How does this compare with the peak-to-peak amplification of −5.6 found for an input signal that had a peak-to-peak value of 1 V ?

SOLUTION

We first calculate the operating conditions at the bias point with no ac signals. This corresponds to point Q in Figure 6.34b. The dc bias voltage V_{GS} across the gate to source is −1.5 V. The

resulting dc drain current I_{DS} can be calculated from Equation 6.55 with $V_{GS(off)} = -V_P = -5$ V:

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

The transconductance at this dc current (at Q) is given by Equation 6.58,

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

The voltage amplification of small signals about point Q is

$$A_V = -g_m R_D = -(2.8 \times 10^{-3})(2000) = -5.6$$

This turns out to be the same as the peak-to-peak voltage amplification we calculated in Table 6.1. When the input ac signal v_{gs} varies between -0.5 and $+0.5$ V, as in Table 6.1, the output signal is not symmetric. It varies between -3 V and 2.8 V, so the voltage gain depends on the input signal. The amplifier is then said to exhibit **nonlinearity.**

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

6.8.1 FIELD EFFECT AND INVERSION

The metal-oxide-semiconductor field effect transistor is based on the effect of a field penetrating into a semiconductor. Its operation can be understood by first considering a parallel plate capacitor with metal electrodes and a vacuum as insulation in between, as shown in Figure 6.35a. When a voltage V is applied between the plates, charges $+Q$ and $-Q$ (where $Q = CV$) appear on the plates and there is an electric field given by $\mathcal{E} = V/L$. The origins of these charges are the conduction electrons for $-Q$ and exposed positively charged metal ions for $+Q$. Metallic bonding is based on all the valence electrons forming a sea of conduction electrons and permeating the space between metal ions that are fixed at crystal lattice sites. Since the electrons are mobile, they are readily displaced by the field. Thus in the lower plate $\mathcal{E}$ displaces some of the conduction electrons to the surface to form $-Q$. In the top plate $\mathcal{E}$ displaces some electrons from the surface into the bulk to expose positively charged metal ions to form $+Q$.

Suppose that the plate area is 1 cm^2 and spacing is 0.1 μm and that we apply 2 V across it. The capacitance C is 8.85 nF and the magnitude of charge Q on each plate is 1.77×10^{-8} C, which corresponds to 1.1×10^{11} electrons. A typical metal such as copper has something like 1.9×10^{15} atoms per cm^2 on the surface. Thus, there will be that number of positive metal ions and electrons on the surface (assuming one conduction electron per atom). The charges $+Q$ and $-Q$ can therefore be generated by the electrons and metal ions at the surface alone. For example, if one in every 1.7×10^4 electrons on the surface moves one atomic spacing ($\sim$0.3 nm) into the bulk, then the surface will have a charge of $+Q$ due to exposed positive metal ions. It is clear that, for all practical purposes, the electric field does not penetrate into the metal and terminates at the metal surface.

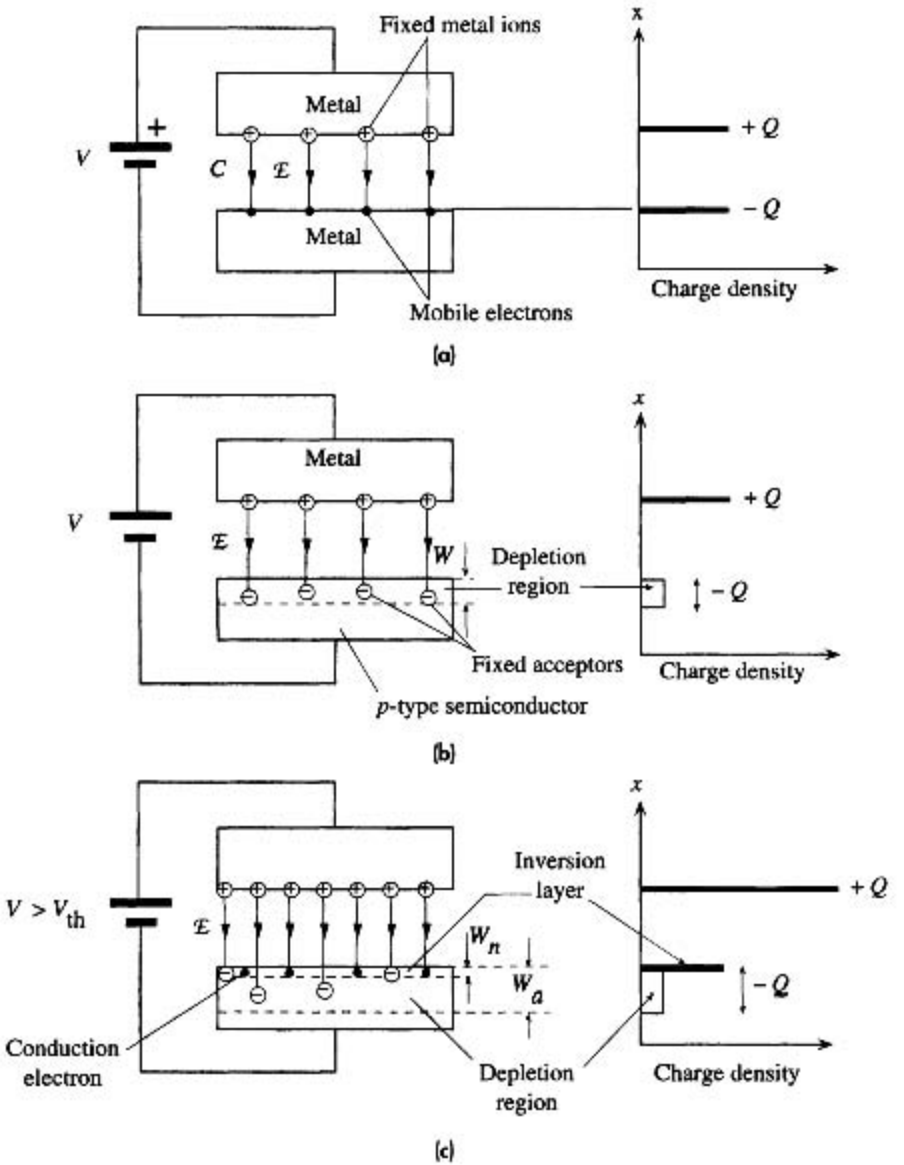


Figure 6.35 The field effect.

(a) In a metal-air-metal capacitor, all the charges reside on the surface.

(b) Illustration of field penetration into a *p*-type semiconductor.

(c) As the field increases, eventually when $V > V_{th}$, an inversion layer is created near the surface in which there are conduction electrons.

The same is not true when one of the electrodes is a semiconductor, as shown in Figure 6.35b where the structure now is of the metal-insulator-semiconductor type. Suppose that we replace the lower metal in Figure 6.35a with a p-type semiconductor with an acceptor concentration of 10^{15} cm^{-3}. The number of acceptor atoms on the surface[10] is 1×10^{10} cm^{-2}. We may assume that at room temperature all the acceptors are ionized and thus negatively charged. It is immediately apparent that we do not have a sufficient number of negative acceptors at the surface to generate the charge $-Q$. We must therefore also expose negative acceptors in the bulk, which means that the field must penetrate into the semiconductor. Holes in the surface region of the semiconductor become repelled toward the bulk and thereby expose more negative acceptors. We can estimate the width W into which the field penetrates since the total negative charge exposed $eAWN_a$ must be Q. We find that W is of the order of 1 μm, which is something like 4000 atomic layers. Our conclusion is that the field penetrates into a semiconductor by an amount that depends on the doping concentration.

The penetrating field into the semiconductor drifts away most of the holes in this region and thereby exposes negatively charged acceptors to make up the charge $-Q$. The region into which the field penetrates has lost holes and is therefore depleted of its equilibrium concentration of holes. We refer to this region as a **depletion layer.** As long as $p > n$ even though $p \ll N_a$, this still has p-type characteristics as holes are in the majority.

If the voltage increases further, $-Q$ also increases, as the field becomes stronger and penetrates more into the semiconductor but eventually it becomes more difficult to make up the charge $-Q$ by simply extending the depletion layer width W into the bulk. It becomes possible (and more favorable) to attract conduction electrons into the depletion layer and form a thin electron layer of width W_n near the surface. The charge $-Q$ is now made up of the fixed negative charge of acceptors in W_a and of conduction electrons in W_n, as shown in Figure 6.35c. Further increases in the voltage do not change the width W_a of the depletion layer but simply increase the electron concentration in W_n. Where do these electrons come from as the semiconductor is doped p-type? Some are attracted into the depletion layer from the bulk, where they were minority carriers. But most are thermally generated by the breaking of Si–Si bonds (*i.e.*, across the bandgap) in the depleted layer. Thermal generation in the depletion layer generates electron–hole pairs that become separated by the field. The holes are then drifted by the field into the bulk and the electrons toward the surface. Recombination of the thermally generated electrons and holes with other carriers is greatly reduced because the depletion layer has so few carriers. Since the electron concentration in the electron layer exceeds the hole concentration and this layer is within a normally p-type semiconductor, we call this an **inversion layer.**

It is now apparent that increasing the field in the metal-insulator-semiconductor device first creates a depletion layer and then an inversion layer at the surface when the voltage exceeds some threshold value V_{th}. This is the basic principle of the field effect device. As long as $V > V_{th}$, any increase in the field and hence $-Q$ leads to more electrons in the inversion layer, whereas the width of the depletion layer W_a and hence the quantity

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

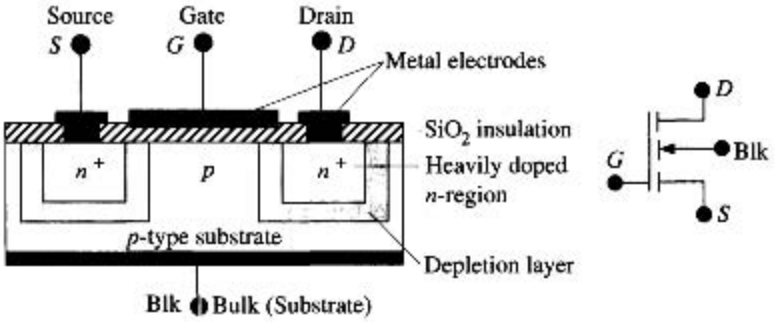


Figure 6.36 The basic structure of the enhancement MOSFET and its circuit symbol.

of fixed negative charge remain constant. The insulator between the metal and the semiconductor, that is, a vacuum in Figure 6.35, is typically SiO_2 in many devices.

6.8.2 ENHANCEMENT MOSFET

Figure 6.36 shows the basic structure of an enhancement n-channel MOSFET device (NMOSFET). A metal-insulator-semiconductor structure is formed between a p-type Si substrate and an aluminum electrode, which is called the gate (G). The insulator is the SiO_2 oxide grown during fabrication. There are two n^+ doped regions at the ends of the MOS device that form the source (S) and drain (D). A metal contact is also made to the p-type Si substrate (or the bulk), which in many devices is connected to the source terminal as shown in Figure 6.36. Further, many MOSFETs have a degenerately doped polycrystalline Si material as the gate that serves the same function as the metal electrode.

With no voltage applied to the gate, S to D is an n^+pn^+ structure that is always reverse-biased whatever the polarity of the source to drain voltage. However, if the substrate (bulk) is connected to the source, a negative V_{DS} will forward bias the n^+p junction between the drain and the substrate. As the n-channel MOSFET device is not normally used with a negative V_{DS}, we will not consider this polarity.

When a positive voltage less than V_{th} is applied to the gate, $V_{GS} < V_{th}$, as shown in Figure 6.37a, the p-type semiconductor under the gate develops a depletion layer as a result of the expulsion of holes into the bulk, just as in Figure 6.35b. Since S and D are isolated by a low-conductivity p-doped region that has a depletion layer from S to D, no current can flow for any positive V_{DS}.

With $V_{DS} = 0$, as soon as V_{GS} is increased beyond the threshold voltage V_{th}, an n-channel inversion layer is formed within the depletion layer under the gate and immediately below the surface, as shown in Figure 6.37b. This n-channel links the two n^+ regions of source and drain. We then have a continuous n-type material with electrons as mobile carriers between the source and drain. When a small V_{DS} is applied, a drain current I_D flows that is limited by the resistance of the n-channel $R_{n\text{-ch}}$:

$$I_D = \frac{V_{DS}}{R_{n\text{-ch}}} \qquad [6.60]$$

Thus, I_D initially increases with V_{DS} almost linearly, as shown in Figure 6.37b.

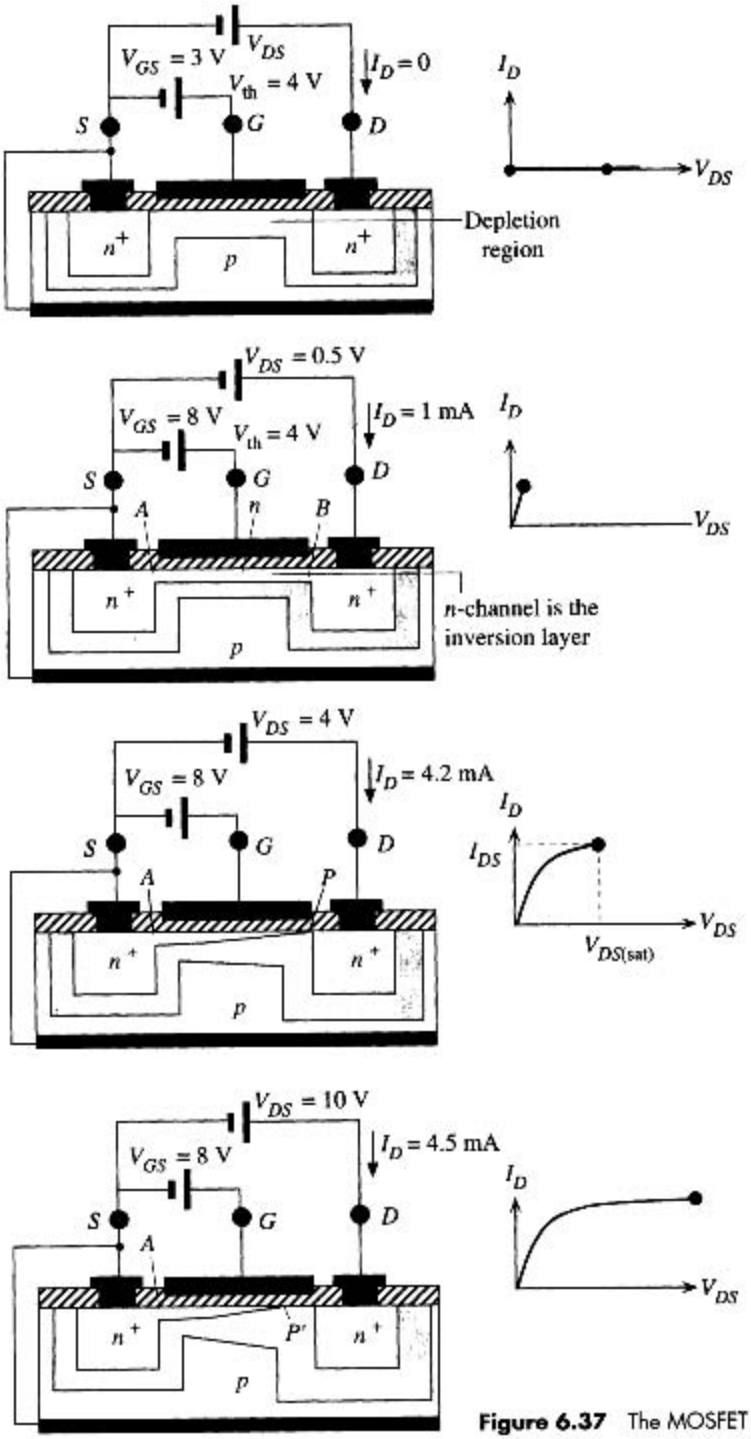


Figure 6.37 The MOSFET I_D versus V_{DS} characteristics.

The voltage variation along the channel is from zero at A (source end) to V_{DS} at B (drain end). The gate to the n-channel voltage is then V_{GS} at A and $V_{GD} = V_{GS} - V_{DS}$ at B. Thus point A depends only on V_{GS} and remains undisturbed by V_{DS}. As V_{DS} increases, the voltage at B (V_{GD}) decreases and thereby causes less inversion. This means that the channel gets narrower from A to B and its resistance $R_{n\text{-ch}}$ increases with V_{DS}. I_D versus V_{DS} then falls increasingly below the $I_D \propto V_{DS}$ line. Eventually when the gate to n-channel voltage at B decreases to just below V_{th}, the inversion layer at B disappears and a depletion layer is exposed, as illustrated in Figure 6.37c. The n-channel becomes pinched off at this point P. This occurs when $V_{DS} = V_{DS(\text{sat})}$, satisfying

$$V_{GD} = V_{GS} - V_{DS(\text{sat})} = V_{\text{th}} \qquad [6.61]$$

It is apparent that the whole process of the narrowing of the n-channel and its eventual pinch-off is similar to the operation of the n-channel JFET. When the drifting electrons in the n-channel reach P, the large electric field within the very narrow depletion layer at P sweeps the electrons across into the n^+ drain. The current is limited by the supply of electrons from the n-channel to the depletion layer at P, which means that it is limited by the effective resistance of the n-channel between A and P.

When V_{DS} exceeds $V_{DS(\text{sat})}$, 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 6.37d. At P', the gate to channel voltage must still be just V_{th} as this is the voltage required to just pinch off the channel and just eliminate inversion. The widening of the depletion layer (from B to P') at the drain end with V_{DS}, however, is small compared with the channel length AB. The resistance of the channel from A to P' does not change significantly with increasing V_{DS}, which means that the drain current is then nearly saturated at I_{DS},

$$I_D \approx I_{DS} \approx \frac{V_{DS(\text{sat})}}{R_{AP'n\text{-ch}}} \qquad V_{DS} > V_{DS(\text{sat})} \qquad [6.62]$$

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 6.38a. 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 6.38b. 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 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

$$I_{DS} = K(V_{GS} - V_{\text{th}})^2 \qquad [6.63]$$

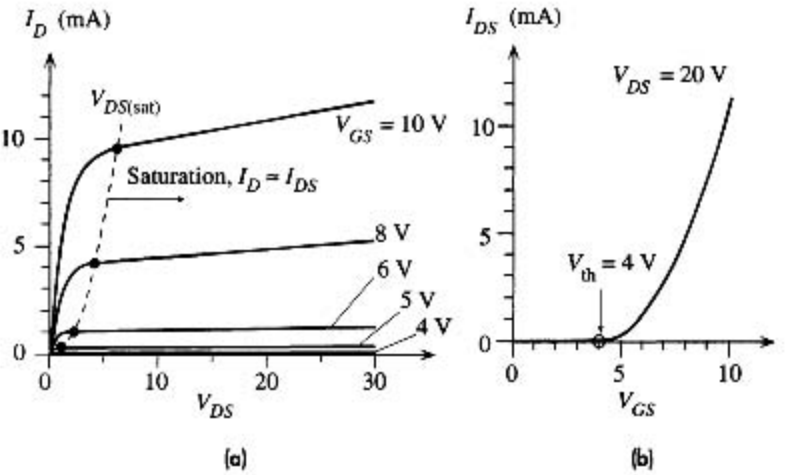


Figure 6.38
(a) Typical I_D versus V_{DS} characteristics of an enhancement MOSFET ($V_{th} = 4$ V) for various fixed gate voltages V_{GS}. (b) Dependence of I_{DS} on V_{GS} at a given V_{DS} (> $V_{DS\,(sat)}$).

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

Enhancement NMOSFET constant

$$K = \frac{Z\mu_e\varepsilon}{2Lt_{ox}}$$

where μ_e is the electron drift mobility in the channel, L and Z are the length and width of the gate controlling the channel, and ε and t_{ox} are the permittivity ($\varepsilon_r\varepsilon_o$) and thickness of the oxide insulation under the gate. According to Equation 6.63, I_{DS} is independent of V_{DS}. The shallow slopes of the I_D versus V_{DS} lines beyond $V_{DS(sat)}$ in Figure 6.38a can be accounted for by writing Equation 6.63 as

Enhancement NMOSFET

$$I_{DS} = K(V_{GS} - V_{th})^2(1 + \lambda V_{DS}) \qquad [6.64]$$

where λ is a constant that is typically 0.01 V^{-1}. If we extend the I_{DS} versus V_{DS} lines, they intersect the $-V_{DS}$ axis at $1/\lambda$, which is called the **Early voltage.** It should be apparent that I_{DSS}, which is I_{DS} with the gate and source shorted ($V_{GS} = 0$), is zero and is not a useful quantity in describing the behavior of the enhancement MOSFET.

EXAMPLE 6.13 **THE ENHANCEMENT NMOSFET** A particular enhancement NMOS transistor has a gate with a width (Z) of 50 μm, length (L) of 10 μm, and SiO_2 thickness of 450 Å. The relative permittivity of SiO_2 is 3.9. The p-type bulk is doped with 10^{16} acceptors cm^{-3}. Its threshold voltage is 4 V. Estimate the drain current when $V_{GS} = 8$ V and $V_{DS} = 20$ V, given $\lambda = 0.01$. Due to the strong scattering of electrons near the crystal surface assume that the electron drift mobility μ_e in the channel is half the drift mobility in the bulk.

SOLUTION

Since $V_{DS} > V_{th}$, we can assume that the drain current is saturated and we can use the I_{DS} versus V_{GS} relationship in Equation 6.64,

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

where

$$K = \frac{Z\mu_e\varepsilon}{2Lt_{ox}}$$

The electron mobility in the bulk when $N_a = 10^{16}$ cm^{-3} is 1300 cm^2 V^{-1} s^{-1} (Chapter 5). Thus

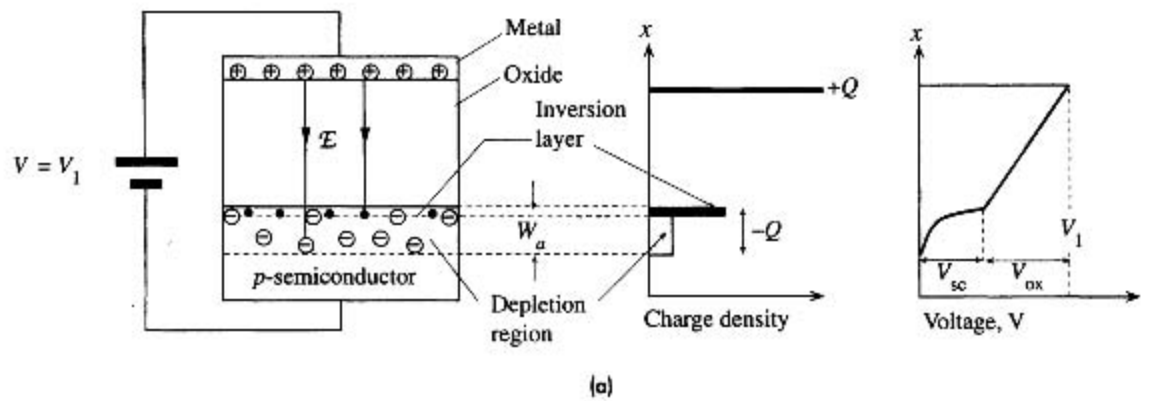
$$K = \frac{Z\mu_e\varepsilon_r\varepsilon_0}{2Lt_{ox}} = \frac{(50 \times 10^{-6})\left(\frac{1}{2} \times 1300 \times 10^{-4}\right)(3.9 \times 8.85 \times 10^{-12})}{2(10 \times 10^{-6})(450 \times 10^{-10})} = 0.000125$$

When $V_{GS} = 8$ V and $V_{DS} = 20$ V, with $\lambda = 0.01$, we have

$$I_{DS} = 0.000125(8 - 4)^2[1 + (0.01)(20)] = 0.0024 \text{ A} \quad \text{or} \quad 2.4 \text{ mA}$$

6.8.3 THRESHOLD VOLTAGE

The threshold voltage is an important parameter in MOSFET devices. Its control in device fabrication is therefore essential. Figure 6.39a shows an idealized MOS structure where all the electric field lines from the metal pass through the oxide and penetrate the p-type semiconductor. The charge $-Q$ is made up of fixed negative acceptors in a surface region of W_a and of conduction electrons in the inversion layer at the surface, as shown in Figure 6.39a. The voltage drop across the MOS structure, however,



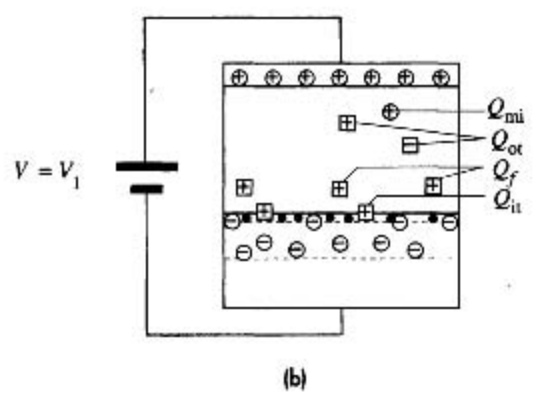


Figure 6.39

(a) The threshold voltage and the ideal MOS structure.

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

is not uniform. As the field penetrates the semiconductor, there is a voltage drop V_{sc} across the field penetration region of the semiconductor by virtue of $\mathcal{E} = -dV/dx$, as shown in Figure 6.39a. The field terminates on both electrons in the inversion layer and acceptors in W_a, so within the semiconductor $\mathcal{E}$ is not uniform and therefore the voltage drop is not constant. But the field in the oxide is uniform, as we assumed there were no charges inside the oxide. The voltage drop across the oxide is constant and is V_{ox}, as shown in Figure 6.39a. As the applied voltage is V_1, we must have $V_{sc} + V_{ox} = V_1$. The actual voltage drop V_{sc} across the semiconductor determines the condition for inversion. We can show this as follows. If the acceptor doping concentration is 10^{16} cm^{-3}, then the Fermi level E_F in the bulk of the p-type semiconductor must be 0.347 eV below E_{Fi} in intrinsic Si. To make the surface n-type we need to shift E_F at the surface to go just above E_{Fi}. Thus we need to shift E_F from bulk to surface by at least 0.347 eV. We have to bend the energy band by 0.347 eV at the surface. Since the voltage drop across the semiconductor is V_{sc} and the corresponding electrostatic PE change is eV_{sc}, this must be 0.347 eV or $V_{sc} = 0.347$ V. The gate voltage for the start of inversion will then be $V_{ox} + 0.347$ V. By inversion, however, we generally infer that the electron concentration at the surface is comparable to the hole concentration in the bulk. This means that we actually have to shift E_F above E_{Fi} by another 0.347 eV, so the gate threshold voltage V_{th} must be $V_{ox} + 0.694$ V.

In practice there are a number of other important effects that must be considered in evaluating the threshold voltage. Invariably there are charges both within the oxide and at the oxide–semiconductor interface that alter the field penetration into the semiconductor and hence the threshold voltage needed at the gate to cause inversion. Some of these are depicted in Figure 6.39b and can be qualitatively summarized as follows.

There may be some mobile ions within the SiO_2, such as alkaline ions (Na^+, K^+), which are denoted as Q_{mi} in Figure 6.39b. These may be introduced unintentionally, for example, during cleaning and etching processes in the fabrication. In addition there may be various trapped (immobile) charges within the oxide Q_{ot} due to structural defects, for example, an interstitial Si^+. Frequently these oxide trapped charges are created as a result of radiation damage (irradiation by X-rays or other high-energy beams). They can be reduced by annealing the device.

A significant number of fixed positive charges (Q_f) exist in the oxide region close to the interface. They are believed to originate from the nonstoichiometry of the oxide near the oxide–semiconductor interface. They are generally attributed to positively charged Si^+ ions. During the oxidation process, a Si atom is removed from the Si surface to react with the oxygen diffusing in through the oxide. When the oxidation process is stopped suddenly, there are unfulfilled Si ions in this region. Q_f depends on the crystal orientation and on the oxidation and annealing processes. The semiconductor to oxide interface itself is a sudden change in the structure from crystalline Si to amorphous oxide. The semiconductor surface itself will have various defects, as discussed in Chapter 1. There is some inevitable mismatch between the two structures at the interface, and consequently there are broken bonds, dangling bonds, point defects such as vacancies and Si^+, and other defects at this interface that trap charges (*e.g.*, holes). All these interface charges are represented as Q_{it} in Figure 6.39b. Q_{it} depends not only on the crystal orientation but also on the chemical composition of the interface. Both Q_f and Q_{it} overall represent a positive charge that effectively reduces the

gate voltage needed for inversion. They are smaller for the (100) surface than the (111) surface, so (100) is the preferred surface for the Si MOS device.

In addition to various charges in the oxide and at the interface shown in Figure 6.39b, there will also be a voltage difference, denoted as V_{FB}, between the semiconductor surface and the metal surface, even in the absence of an applied voltage. V_{FB} arises from the work function difference between the metal and the p-type semiconductor, as discussed in Chapter 4. The metal work function is generally smaller than the semiconductor work function, which means that the semiconductor surface will have an accumulation of electrons and the metal surface will have positive charges (exposed metal ions). The gate voltage needed for inversion will therefore also depend on V_{FB}. Since V_{FB} is normally positive and Q_f and Q_{it} are also positive, there may already be an inversion layer formed at the semiconductor surface even without a positive gate voltage. The fabrication of an enhancement MOSFET then requires special fabrication procedures, such as ion implantation, to obtain a positive and predictable V_{th}.

The simplest way to control the threshold gate voltage is to provide a separate electrode to the bulk of an enhancement MOSFET, as shown in Figure 6.36, and to apply a bias voltage to the bulk with respect to the source to obtain the desired V_{th} between the gate and source. This technique has the disadvantage of requiring an additional bias supply for the bulk and also adjusting the bulk to source voltage almost individually for each MOSFET.

6.8.4 ION IMPLANTED MOS TRANSISTORS AND POLY-SI GATES

The most accurate method of controlling the threshold voltage is by ion implantation, as the number of ions that are implanted into a device and their location can be closely controlled. Furthermore, ion implantation can also provide a self-alignment of the edges of the gate electrode with the source and drain regions. In the case of an n-channel enhancement MOSFET, it is generally desirable to keep the p-type doping in the bulk low to avoid small V_{DS} for reverse breakdown between the drain and the bulk (see Figure 6.36). Consequently, the surface, in practice, already has an inversion layer (without any gate voltage) due to various fixed positive charges residing in the oxide and at the interface, as shown in Figure 6.39b (positive Q_f and Q_{it} and V_{FB}). It then becomes necessary to implant the surface region under the gate with boron acceptors to remove the electrons and restore this region to a p-type behavior.

The ion implantation process is carried out in a vacuum where the required impurity ions are generated and then accelerated toward the device. The energy of the arriving ions and hence their penetration into the device can be readily controlled. Typically, the device is implanted with B acceptors under the gate oxide, as shown in Figure 6.40. The distribution of implanted acceptors as a function of distance into the device from the surface of the oxide is also shown in the figure. The position of the peak depends on the energy of the ions and hence on the accelerating voltage. The peak of the concentration of implanted acceptors is made to occur just below the surface of the semiconductor. Since ion implantation involves the impact of energetic ions with the crystal structure, it results in the inevitable generation of various defects within the implanted region. The defects are almost totally eliminated by annealing the device at an

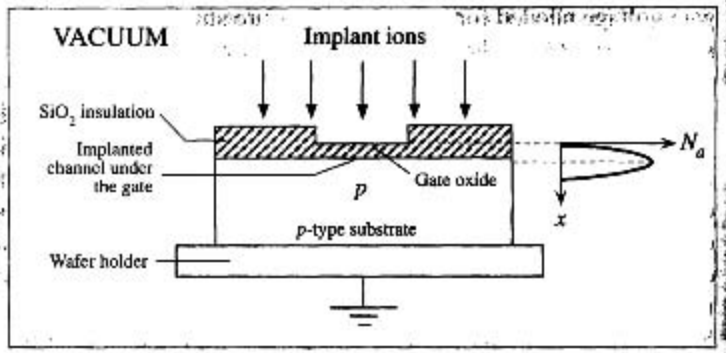


Figure 6.40 Schematic illustration of ion implantation for the control of V_{th}.

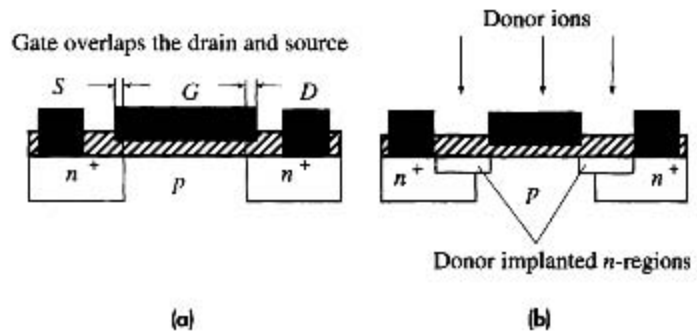


Figure 6.41
(a) There is an overlap of the gate electrode with the source and drain regions and hence additional capacitance between the gate and drain.
(b) n^+-type ion implantation extends the drain and source to line up with the gate.

elevated temperature. Annealing also broadens the acceptor implanted region as a result of increased diffusion of implanted acceptors.

Ion implantation also has the advantage of providing self-alignment of the drain and source with the edges of the gate electrode. In a MOS transistor, it is important that the gate electrode extends all the way from the source to the drain regions so that the channel formed under the gate can link the two regions; otherwise, an incomplete channel will be formed. To avoid the possibility of forming an incomplete channel, it is necessary to allow for some overlap, as shown in Figure 6.41a, between the gate and source and drain regions because of various tolerances and variations involved in the fabrication of a MOSFET by conventional masking and diffusional techniques. The overlap, however, results in additional capacitances between the gate and source and the gate and drain and adversely affects the high-frequency (or transient) response of the device. It is therefore desirable to align the edges of the gate electrode with the source and drain regions. Suppose that the gate electrode is made narrower so that it does not extend all the way between the source and drain regions, as shown in Figure 6.41b. If the device is now ion implanted with donors, then donor ions passing through the thin oxide will extend the n^+ regions up to the edges of the gate and thereby align the drain and source with the edges of the gate. The thick metal gate is practically impervious to the arriving donor ions.

Another method of controlling V_{th} is to use silicon instead of Al for the gate electrode. This technique is called **silicon gate technology.** Typically, the silicon for the

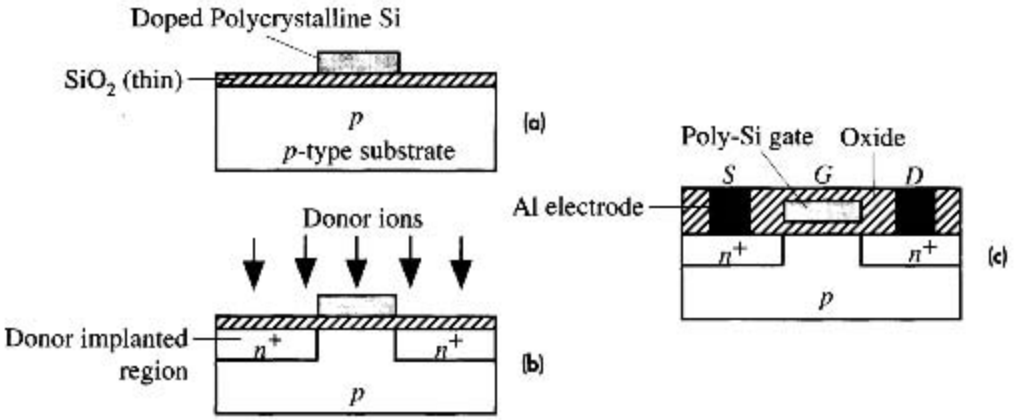


Figure 6.42 The poly-Si gate technology.
(a) Poly-Si is deposited onto the oxide, and the areas outside the gate dimensions are etched away.
(b) The poly-Si gate acts as a mask during ion implantation of donors to form the n^+ source and drain regions.
(c) A simplified schematic sketch of the final poly-Si MOS transistor.

gate is vacuum deposited (*e.g.*, by chemical vapor deposition using silane gas) onto the oxide, as shown in Figure 6.42. As the oxide is noncrystalline, the Si gate is polycrystalline (rather than a single crystal) and is therefore called a **poly-Si gate.** Normally it is heavily doped to ensure that it has sufficiently low resistivity to avoid *RC* time constant limitations in charging and discharging the gate capacitance during transient or ac operations. The advantage of the poly-Si gate is that its work function depends on the doping (type and concentration) and can be controlled so that V_{FB} and hence V_{th} can also be controlled. There are also additional advantages in using the poly-Si gate. For example, it can be raised to high temperatures (Al melts at 660 °C). It can be used as a mask over the gate region of the semiconductor during the formation of the source and drain regions. If ion implantation is used to deposit donors into the semiconductor, then the n^+ source and drain regions are self-aligned with the poly-Si gate, as shown in Figure 6.42.

6.9 LIGHT EMITTING DIODES (LED)

6.9.1 LED PRINCIPLES

A light emitting diode (LED) is essentially a *pn* junction diode typically made from a direct bandgap semiconductor, for example, GaAs, in which the electron–hole pair (EHP) recombination results in the emission of a photon. The emitted photon energy $h\nu$ is approximately equal to the bandgap energy E_g. Figure 6.43a shows the energy band diagram of an unbiased pn^+ junction device in which the *n*-side is more heavily doped than the *p*-side. The Fermi level E_F is uniform through the device, which is a requirement of equilibrium with no applied bias. The depletion region extends mainly into the *p*-side. There is a *PE* barrier eV_o from E_c on the *n*-side to E_c on the *p*-side

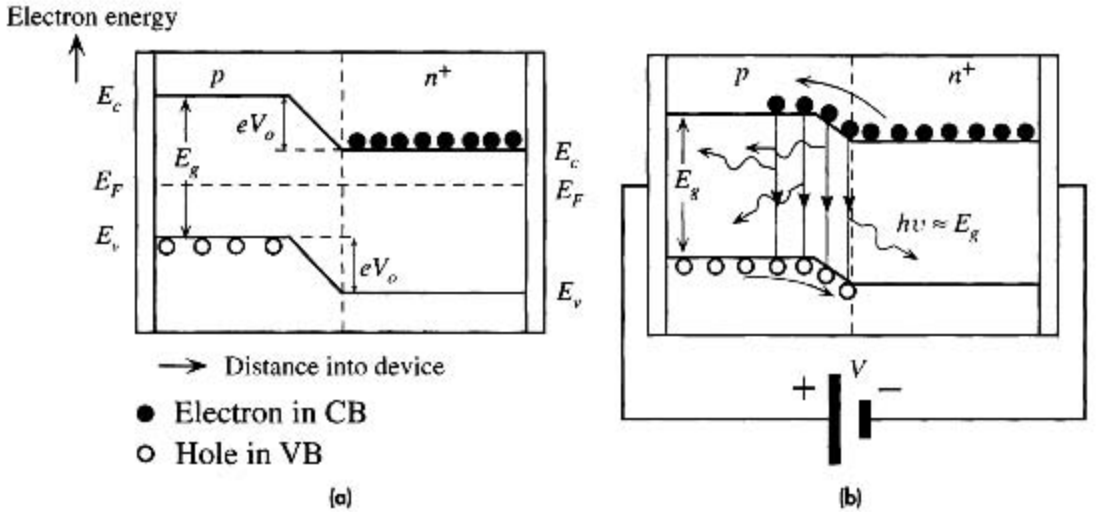


Figure 6.43 Energy band diagram of a *pn* (heavily n-type doped) junction.
(a) No bias voltage.
(b) With forward bias V. Recombination around the junction and within the diffusion length of the electrons in the *p*-side leads to photon emission.

where V_o is the *built-in voltage*. The *PE* barrier eV_o prevents the diffusion of electrons from the n-side to the p-side.

When a forward bias V is applied, the built-in potential V_o is reduced to $V_o - V$, which then allows the electrons from the n^+-side to diffuse, that is, become injected, into the p-side as depicted in Figure 6.43b. The hole injection component from p into the n^+-side is much smaller than the electron injection component from the n^+-side to the p-side. The recombination of injected electrons in the depletion region and within a volume extending over the electron diffusion length L_e in the p-side leads to photon emission. The phenomenon of light emission from the EHP recombination as a result of minority carrier injection is called **injection electroluminescence.** Due to the statistical nature of the recombination process between electrons and holes, the emitted photons are in random directions; they result from spontaneous emission processes. The LED structure has to be such that the emitted photons can escape the device without being reabsorbed by the semiconductor material. This means the p-side has to be sufficiently narrow or we have to use *heterostructure* devices as discussed below.

One very simple LED structure is shown in Figure 6.44. First a doped semiconductor layer is grown on a suitable substrate (GaAs or GaP). The growth is done **epitaxially;** that is, the crystal of the new layer is grown to follow the structure of the substrate crystal. The **substrate** is essentially a sufficiently thick crystal that serves as a mechanical support for the *pn* junction device (the doped layers) and can be of different crystal. The pn^+ junction is formed by growing another epitaxial layer but doped p-type. Those photons that are emitted toward the n-side become either absorbed or reflected back at the substrate interface depending on the substrate thickness and the exact structure of the LED. If the epitaxial layer and the substrate crystals have different

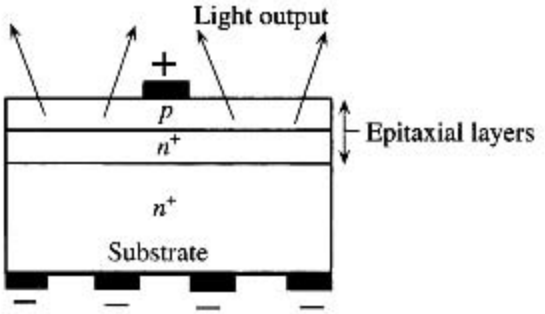


Figure 6.44 A schematic illustration of one possible LED device structure. First an n^+ layer is epitaxially grown on a substrate. A thin p layer is then epitaxially grown on the first layer.

crystal lattice parameters, then there is a lattice mismatch between the two crystal structures. This causes lattice strain in the LED layer and hence leads to crystal defects. Such crystal defects encourage radiationless EHP recombinations. That is, a defect acts as a recombination center. Such defects are reduced by lattice matching the LED epitaxial layer to the substrate crystal. It is therefore important to lattice match the LED layer to the substrate crystal. For example, one of the AlGaAs alloys is a direct bandgap semiconductor that has a bandgap in the red-emission region. It can be grown on GaAs substrates with excellent lattice match which results in high-efficiency LED devices.

There are various direct bandgap semiconductor materials that can be readily doped to make commercial pn junction LEDs which emit radiation in the red and infrared range of wavelengths. An important class of commercial semiconductor materials that covers the visible spectrum is the **III–V ternary alloys** based on alloying GaAs and GaP and denoted as $GaAs_{1-y}P_y$. In this compound, As and P atoms from Group V are distributed randomly at normal As sites in the GaAs crystal structure. When $y < 0.45$, the alloy $GaAs_{1-y}P_y$ is a direct bandgap semiconductor and hence the EHP recombination process is direct as depicted in Figure 6.45a. The rate of recombination is directly proportional to the product of electron and hole concentrations. The emitted wavelengths range from about 630 nm, red, for $y = 0.45$ ($GaAs_{0.55}P_{0.45}$) to 870 nm for $y = 0$ (GaAs).

$GaAs_{1-y}P_y$ alloys (which include GaP) with $y > 0.45$ are indirect bandgap semiconductors. The EHP recombination processes occur through recombination centers and involve lattice vibrations rather than photon emission. However, if we add

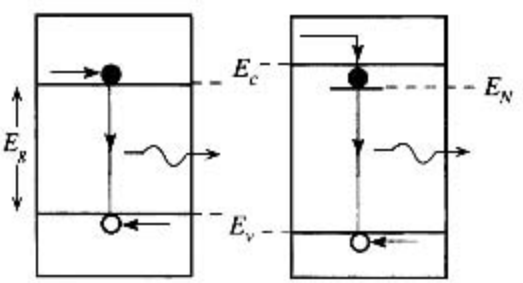


Figure 6.45
(a) Photon emission in a direct bandgap semiconductor.
(b) GaP is an indirect bandgap semiconductor. When it is doped with nitrogen, there is an electron recombination center at E_N. Direct recombination between a captured electron at E_N and a hole emits a photon.

Table 6.2 Selected LED semiconductor materials

Semiconductor Active Layer	Structure	D or I	λ (nm)	$\eta_{external}$ (%)	Comments
GaAs	DH	D	870–900	10	Infrared (IR)
$Al_xGa_{1-x}As$ $(0 < x < 0.4)$	DH	D	640–870	3–20	Red to IR
$In_{1-x}Ga_xAs_yP_{1-y}$ $(y \approx 2.20x, 0 < x < 0.47)$	DH	D	1–1.6 μm	>10	LEDs in communications
$In_{0.49}Al_xGa_{0.51-x}P$	DH	D	590–630	>10	Amber, green, red; high luminous intensity
InGaN/GaN quantum well	QW	D	450–530	5–20	Blue to green
$GaAs_{1-y}P_y$ $(y < 0.45)$	HJ	D	630–870	< 1	Red to IR
$GaAs_{1-y}P_y$ $(y > 0.45)$ (N or Zn, O doping)	HJ	I	560–700	< 1	Red, orange, yellow
SiC	HJ	I	460–470	0.02	Blue, low efficiency
GaP (Zn)	HJ	I	700	2–3	Red
GaP (N)	HJ	I	565	< 1	Green

NOTE: Optical communication channels are at 850 nm (local network) and at 1.3 and 1.55 μm (long distance). D = direct bandgap, I = indirect bandgap. $\eta_{external}$ is typical and may vary substantially depending on the device structure. DH = double heterostructure, HJ = homojunction, QW = quantum well.

isoelectronic impurities such as nitrogen (in the same Group V as P) into the semiconductor crystal, then some of these N atoms substitute for P atoms. Since N and P have the same valency, N atoms substituting for P atoms form the same number of bonds and do not act as donors or acceptors. The electronic cores of N and P, however, are different. The positive nucleus of N is less shielded by electrons compared with that of the P atom. This means that a conduction electron in the neighborhood of a N atom will be attracted and may become captured at this site. N atoms therefore introduce localized energy levels, or electron traps, E_N near the conduction band (CB) edge as depicted in Figure 6.45b. When a conduction electron is captured at E_N, it can attract a hole (in the valence band) in its vicinity by Coulombic attraction and eventually recombine with it directly and emit a photon. The emitted photon energy is only slightly less than E_g as E_N is typically close to E_c. As the recombination process depends on N doping, it is not as efficient as direct recombination. Thus, the efficiency of LEDs from N doped indirect bandgap $GaAs_{1-y}P_y$ semiconductors is less than those from direct bandgap semiconductors. Nitrogen doped indirect bandgap $GaAs_{1-y}P_y$ alloys are widely used in inexpensive green, yellow, and orange LEDs.

The **external efficiency** $\eta_{external}$ of an LED quantifies the efficiency of conversion of electric energy into an emitted external optical energy. It incorporates the internal efficiency of the radiative recombination process and the subsequent efficiency of photon extraction from the device. The input of electric power into an LED is simply the diode current and diode voltage product (IV). If P_{out} is the optical power emitted by the device, then

External efficiency

$$\eta_{external} = \frac{P_{out}(optical)}{IV} \times 100\% \qquad [6.65]$$

and some typical values are listed in Table 6.2. For indirect bandgap semiconductors, $\eta_{external}$ are generally less than 1 percent, whereas for direct bandgap semiconductors with the right device structure, $\eta_{external}$ can be substantial.

6.9.2 HETEROJUNCTION HIGH-INTENSITY LEDS

A *pn* junction between two differently doped semiconductors that are of the same material, that is, the same bandgap E_g, is called a **homojunction.** A junction between two different bandgap semiconductors is called a **heterojunction.** A semiconductor device structure that has junctions between different bandgap materials is called a **heterostructure device.**

LED constructions for increasing the intensity of the output light make use of the double heterostructure. Figure 6.46a shows a **double-heterostructure** (DH) device based on two junctions between different semiconductor materials with different bandgaps. In this case the semiconductors are AlGaAs with $E_g \approx 2$ eV and GaAs with $E_g \approx 1.4$ eV. The double heterostructure in Figure 6.46a has an n^+p heterojunction between n^+-AlGaAs and p-GaAs. There is another heterojunction between p-GaAs and p-AlGaAs. The p-GaAs region is a thin layer, typically a fraction of a micron, and it is lightly doped.

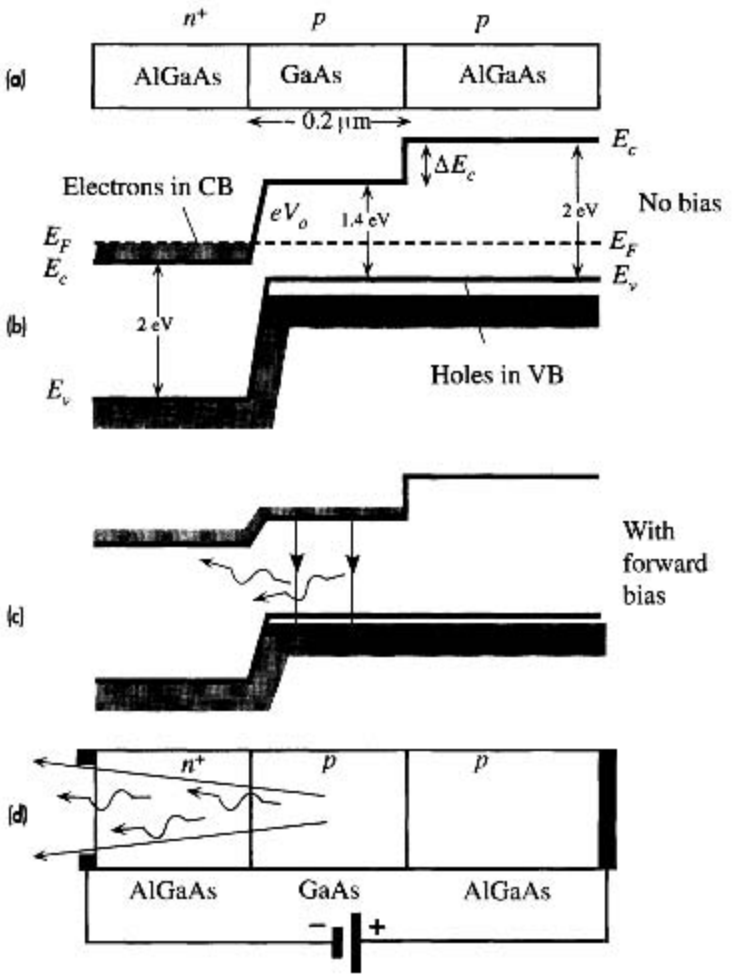


Figure 6.46

(a) A double heterostructure diode has two junctions which are between two different bandgap semiconductors (GaAs and AlGaAs).

(b) A simplified energy band diagram with exaggerated features. E_F must be uniform.

(c) Forward-biased simplified energy band diagram.

(d) Forward-biased LED. Schematic illustration of photons escaping reabsorption in the AlGaAs layer and being emitted from the device.

The simplified energy band diagram for the whole device in the absence of an applied voltage is shown in Figure 6.46b. The Fermi level E_F is continuous throughout the whole structure. There is a potential energy barrier eV_o for electrons in the CB of n^+-AlGaAs against diffusion into p-GaAs. There is a bandgap change at the junction between p-GaAs and p-AlGaAs which results in a step change ΔE_c in E_c between the two conduction bands of p-GaAs and p-AlGaAs. This ΔE_c is effectively a *potential energy barrier* that prevents any electrons in the CB in p-GaAs passing to the CB of p-AlGaAs. (There is also a step change ΔE_v in E_v, but this is small and is not shown.)

When a forward bias is applied, most of this voltage drops between the n^+-AlGaAs and p-GaAs and reduces the potential energy barrier eV_o, just as in the normal pn junction. This allows electrons in the CB of n^+-AlGaAs to be injected into p-GaAs as shown in Figure 6.46c. These electrons, however, are *confined* to the CB of p-GaAs since there is a barrier ΔE_c between p-GaAs and p-AlGaAs. The wide bandgap AlGaAs layers therefore act as **confining layers** that restrict injected electrons to the p-GaAs layer. The recombination of injected electrons and the holes already present in this p-GaAs layer results in spontaneous photon emission. Since the bandgap E_g of AlGaAs is greater than GaAs, the emitted photons do not get reabsorbed as they escape the active region and can reach the surface of the device as depicted in Figure 6.46d. Since light is also not absorbed in p-AlGaAs, it can be reflected to increase the light output.

6.9.3 LED CHARACTERISTICS

The energy of an emitted photon from an LED is not simply equal to the bandgap energy E_g because electrons in the conduction band are distributed in energy and so are the holes in the valence band (VB). Figure 6.47a and b illustrate the energy band diagram and the energy distributions of electrons and holes in the CB and VB, respectively. The electron concentration as a function of energy in the CB is given by $g(E)f(E)$ where $g(E)$ is the density of states and $f(E)$ is the Fermi–Dirac function (probability of finding an electron in a state with energy E). The product $g(E)f(E)$ represents the electron concentration per unit energy or the concentration in energy and is plotted along the horizontal axis in Figure 6.47b. There is a similar energy distribution for holes in the VB.

The electron concentration in the CB as a function of energy is asymmetrical and has a peak at $\frac{1}{2}kT$ above E_c. The energy spread of these electrons is typically $\sim 2kT$ from E_c as shown in Figure 6.47b. The hole concentration is similarly spread from E_v in the valence band. Recall the rate of direct recombination is proportional to both the electron and hole concentrations at the energies involved. The transition which is identified as 1 in Figure 6.47a involves the direct recombination of an electron at E_c and a hole at E_v. But the carrier concentrations near the band edges are very small and hence this type of recombination does not occur frequently. The relative intensity of light at this photon energy $h\nu_1$ is small as shown in Figure 6.47c. The transitions that involve the largest electron and hole concentrations occur most frequently. For example, the transition 2 in Figure 6.47a has the maximum probability as both electron and hole concentrations are largest at these energies as shown in Figure 6.47b.

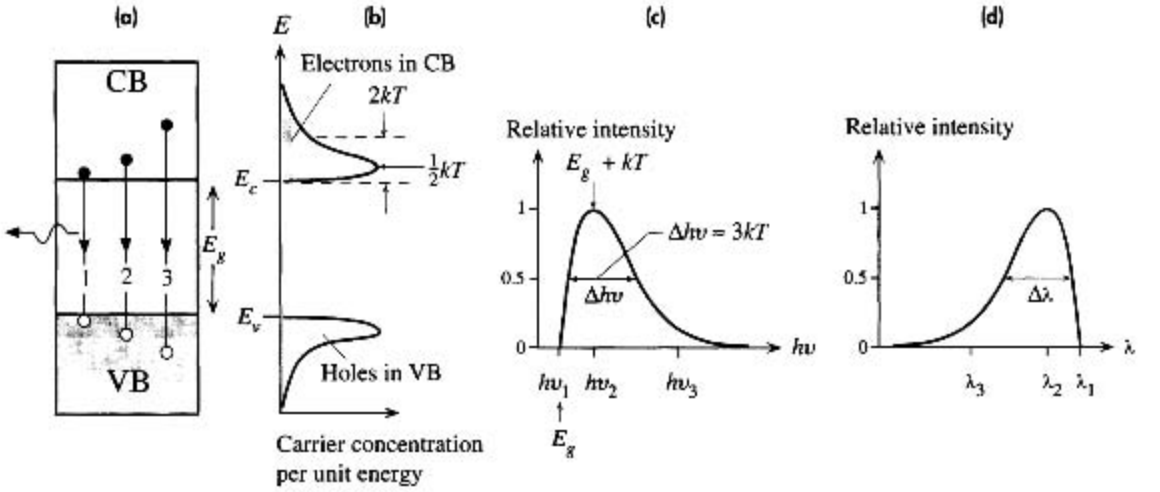


Figure 6.47

(a) Energy band diagram with possible recombination paths.

(b) Energy distribution of electrons in the CB and holes in the VB. The highest electron concentration is $\frac{1}{2}kT$ above E_c.

(c) The relative light intensity as a function of photon energy based on (b).

(d) Relative intensity as a function of wavelength in the output spectrum based on (b) and (c).

The relative intensity of light corresponding to this transition energy $h\nu_2$ is then maximum, or close to maximum, as indicated in Figure 6.47c.[11] The transitions marked as 3 in Figure 6.47a that emit relatively high energy photons $h\nu_3$ involve energetic electrons and holes whose concentrations are small as apparent in Figure 6.47b. Thus, the light intensity at these relatively high photon energies is small. The fall in light intensity with photon energy is shown in Figure 6.47c. The relative light intensity versus photon energy characteristic of the output spectrum is shown in Figure 6.47c and represents an important LED characteristic. Given the spectrum in Figure 6.47c we can also obtain the relative light intensity versus wavelength characteristic as shown in Figure 6.47d since $\lambda = c/\nu$. The **linewidth** of the output spectrum, $\Delta\nu$ or $\Delta\lambda$, is defined as the width between half-intensity points as shown in Figure 6.47c and d.

The wavelength for the peak intensity and the linewidth $\Delta\lambda$ of the emitted spectrum are obviously related to the energy distributions of the electrons and holes in the conduction and valence bands and therefore to the density of states in these bands. The photon energy for the peak emission is roughly $E_g + kT$ inasmuch as it corresponds to peak-to-peak transitions in the energy distributions of the electrons and holes in Figure 6.47b. The linewidth $\Delta(h\nu)$ of the output radiation between the half intensity points is approximately $3kT$ as shown in Figure 6.47c. It is relatively straightforward to calculate the corresponding spectral linewidth $\Delta\lambda$ in terms of wavelength as explained in Example 6.14.

[11] The intensity is not necessarily maximum when both the electron and hole concentrations are maximum, but it will be close.

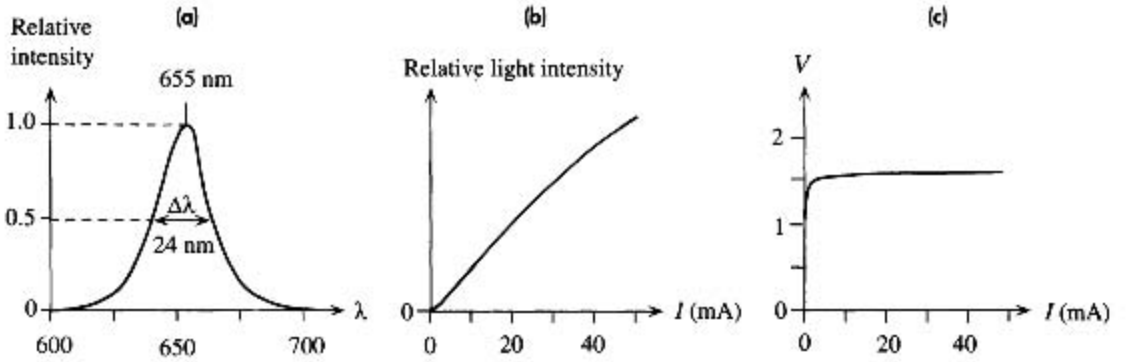


Figure 6.48

(a) A typical output spectrum from a red GaAsP LED.

(b) Typical output light power versus forward current.

(c) Typical *I–V* characteristics of a red LED. The turn-on voltage is around 1.5 V.

The output spectrum, or the relative intensity versus wavelength characteristics, from an LED depends not only on the semiconductor material but also on the structure of the *pn* junction diode, including the dopant concentration levels. The spectrum in Figure 6.47d represents an idealized spectrum without including the effects of heavy doping on the energy bands and the reabsorption of some of the photons.

Typical characteristics of a red LED (655 nm), as an example, are shown in Figure 6.48a to c. The output spectrum in Figure 6.48a exhibits less asymmetry than the idealized spectrum in Figure 6.47d. The width of the spectrum is about 24 nm, which corresponds to a width of about $2.7kT$ in the energy distribution of the emitted photons. As the LED current increases so does the injected minority carrier concentration, and thus the rate of recombination and hence the output light intensity. The increase in the output light power is not however linear with the LED current as apparent in Figure 6.48b. At high current levels, a strong injection of minority carriers leads to the recombination time depending on the injected carrier concentration and hence on the current itself; this leads to a nonlinear recombination rate with current. Typical current–voltage characteristics are shown in Figure 6.48c where it can be seen that the **turn-on,** or **cut-in, voltage** is about 1.5 V from which point the current increases very steeply with voltage. The turn-on voltage depends on the semiconductor and generally increases with the energy bandgap E_g. For example, typically, for a blue LED it is about 3.5–4.5 V, for a yellow LED it is about 2 V, and for a GaAs infrared LED it is around 1 V.

EXAMPLE 6.14 **SPECTRAL LINEWIDTH OF LEDS** We know that a spread in the output wavelengths is related to a spread in the emitted photon energies as depicted in Figure 6.47. The emitted photon energy $E_{ph} = hc/\lambda$ and the spread in the photon energies $\Delta E_{ph} = \Delta(h\nu) \approx 3kT$ between the half-intensity points as shown in Figure 6.47c. Show that the corresponding **linewidth** $\Delta\lambda$ between the *half-intensity points* in the output spectrum is

LED spectral linewidth

$$\Delta\lambda = \lambda^2 \frac{3kT}{hc} \qquad [6.66]$$

What is the spectral linewidth of an optical communications LED operating at 1550 nm and at 300 K?

SOLUTION

First consider the relationship between the photon frequency ν and λ,

$$\lambda = \frac{c}{\nu} = \frac{hc}{h\nu}$$

in which $h\nu$ is the photon energy. We can differentiate this,

$$\frac{d\lambda}{d(h\nu)} = -\frac{hc}{(h\nu)^2} = -\frac{\lambda^2}{hc}$$

The negative sign implies that increasing the photon energy decreases the wavelength. We are only interested in changes or spreads; thus $\Delta\lambda/\Delta(h\nu) \approx |d\lambda/d(h\nu)|$,

$$\Delta\lambda = \frac{\lambda^2}{hc}\Delta(h\nu) = \frac{\lambda^2}{hc}3kT$$

where we used $\Delta(h\nu) = 3kT$, and obtained Equation 6.66. We can substitute $\lambda = 1550\,\text{nm}$ and $T = 300$ K to calculate the linewidth of the 1550 nm LED:

$$\Delta\lambda = \lambda^2\frac{3kT}{hc} = (1550 \times 10^{-9})^2\frac{3(1.38 \times 10^{-23})(300)}{(6.626 \times 10^{-34})(3 \times 10^8)}$$

$$= 1.50 \times 10^{-7}\,\text{m} \quad \text{or} \quad 150\,\text{nm}$$

The spectral linewidth of an LED output is due to the spread in the photon energies, which is fundamentally about $3kT$. The only option for decreasing $\Delta\lambda$ at a given wavelength is to reduce the temperature. The output spectrum of a laser, on the other hand, has a much narrower linewidth. A single-mode laser can have an output linewidth less than 1 nm.

6.10 SOLAR CELLS

6.10.1 PHOTOVOLTAIC DEVICE PRINCIPLES

A simplified schematic diagram of a typical solar cell is shown in Figure 6.49. Consider a *pn* junction with a very narrow and more heavily doped n-region. The illumination is through the thin n-side. The depletion region (W) or the space charge layer (SCL) extends primarily into the p-side. There is a built-in field $\mathcal{E}_o$ in this depletion layer. The electrodes attached to the n-side must allow illumination to enter the device and at the same time result in a small series resistance. They are deposited on the n-side to form an array of **finger electrodes** on the surface as depicted in Figure 6.50. A thin **antireflection coating** on the surface (not shown in the figure) reduces reflections and allows more light to enter the device.

As the n-side is very narrow, most of the photons are absorbed within the depletion region (W) and within the neutral p-side (ℓ_p) and photogenerate EHPs in these regions. EHPs photogenerated in the depletion region are immediately separated by the built-in field $\mathcal{E}_o$ which drifts them apart. The electron drifts and reaches the neutral n^+-side whereupon it makes this region negative by an amount of charge $-e$. Similarly,

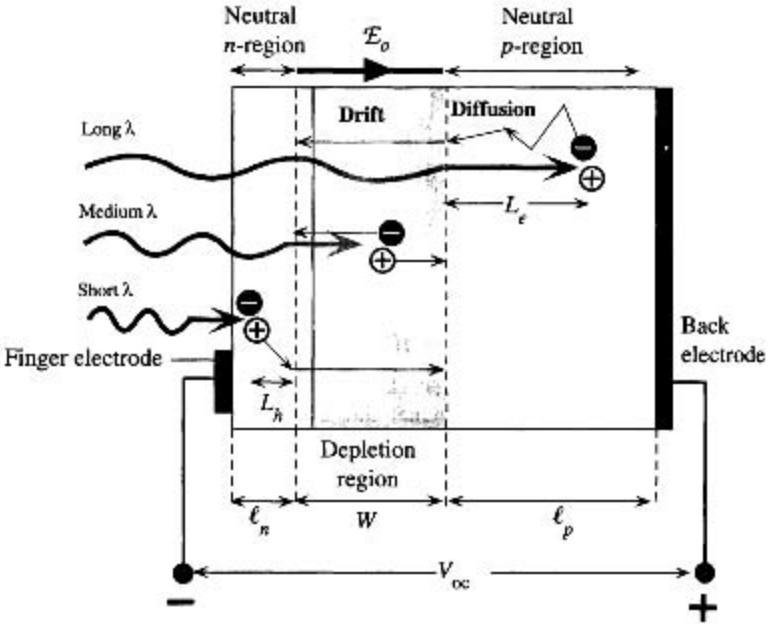


Figure 6.49 The principle of operation of the solar cell (exaggerated features to highlight principles).

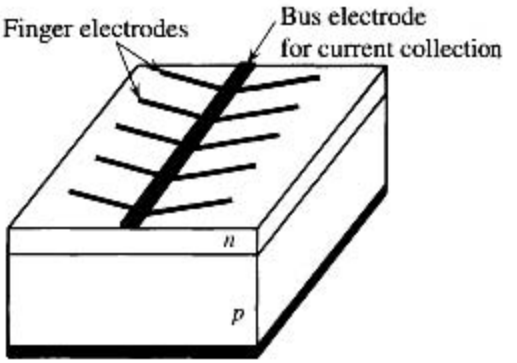


Figure 6.50 Finger electrodes on the surface of a solar cell reduce the series resistance.

the hole drifts and reaches the neutral p-side and thereby makes this side positive. Consequently an **open circuit voltage** develops between the terminals of the device with the p-side positive with respect to the n-side. If an external load is connected, then the excess electron in the n-side can travel around the external circuit, do work, and reach the p-side to recombine with the excess hole there. It is important to realize that without the internal field $\mathcal{E}_o$ it is not possible to drift apart the photogenerated EHPs and accumulate excess electrons on the n-side and excess holes on the p-side.

The EHPs photogenerated by long-wavelength photons that are absorbed in the neutral p-side diffuse around in this region as there is no electric field. If the recombination lifetime of the electron is τ_e, it diffuses a mean distance $L_e = \sqrt{2D_e\tau_e}$ where D_e is its diffusion coefficient in the p-side. Those electrons within a distance L_e to the depletion region can readily diffuse and reach this region whereupon they become drifted

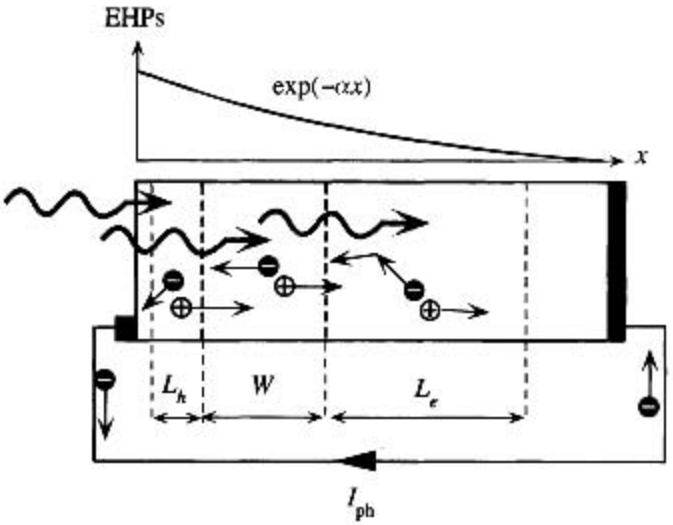


Figure 6.51 Photogenerated carriers within the volume $L_h + W + L_e$ give rise to a photocurrent I_{ph}.
The variation in the photogenerated EHP concentration with distance is also shown where α is the absorption coefficient at the wavelength of interest.

by $\mathcal{E}_o$ to the n-side as shown in Figure 6.49. Consequently only those EHPs photogenerated within the minority carrier diffusion length L_e to the depletion layer can contribute to the photovoltaic effect. Again the importance of the built-in field $\mathcal{E}_o$ is apparent. Once an electron diffuses to the depletion region, it is swept over to the n-side by $\mathcal{E}_o$ to give an additional negative charge there. Holes left behind in the p-side contribute a net positive charge to this region. Those photogenerated EHPs further away from the depletion region than L_e are lost by recombination. It is therefore important to have the minority carrier diffusion length L_e be as long as possible. This is the reason for choosing this side of a Si pn junction to be p-type which makes electrons the minority carriers; the electron diffusion length in Si is longer than the hole diffusion length. The same ideas also apply to EHPs photogenerated by short-wavelength photons absorbed in the n-side. Those holes photogenerated within a diffusion length L_h can reach the depletion layer and become swept across to the p-side. The photogeneration of EHPs that contributes to the photovoltaic effect therefore occurs in a volume covering $L_h + W + L_e$. If the terminals of the device are shorted, as in Figure 6.51, then the excess electron in the n-side can flow through the external circuit to neutralize the excess hole in the p-side. This current due to the flow of the photogenerated carriers is called the **photocurrent.**

Under a steady-state operation, there can be no net current through an *open circuit* solar cell. This means the photocurrent inside the device due to the flow of photogenerated carriers must be exactly balanced by a flow of carriers in the opposite direction. The latter carriers are minority carriers that become injected by the appearance of the photovoltaic voltage across the pn junction as in a normal diode. This is not shown in Figure 6.49.

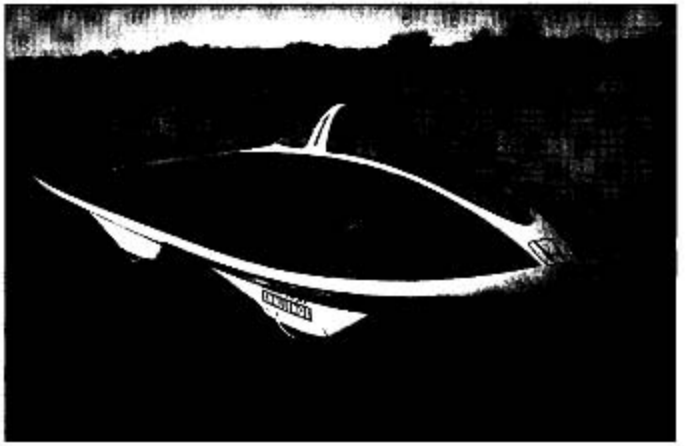
EHPs photogenerated by energetic photons absorbed in the n-side near the surface region or outside the diffusion length L_h to the depletion layer are lost by recombination as the lifetime in the n-side is generally very short (due to heavy doping). The n-side is therefore made very thin, typically less than 0.2 μm. Indeed, the length ℓ_n of

Solar cell inventors at Bell Labs (left to right): Gerald Pearson, Daryl Chapin, and Calvin Fuller. They are checking a Si solar cell sample for the amount of voltage produced (1954).

SOURCE: Courtesy of Bell Labs, Lucent Technologies.

Helios is a solar cell–powered airplane that is remotely piloted. It has been able to fly as high as about 30 km during the day. Its wingspan is 9 m. It has fuel cells to fly at night.

SOURCE: Courtesy of NASA, Dryden Flight Center.



pn Junction Si solar cells at work. Honda's two-seated Dream car is powered by photovoltaics. The Honda Dream was first to finish 3,010 km in four days in the 1996 World Solar Challenge.

SOURCE: Courtesy of Centre for Photovoltaic Engineering, University of New South Wales, Sydney, Australia.

the n-side may be shorter than the hole diffusion length L_h. The EHPs photogenerated very near the surface of the n-side, however, disappear by recombination due to various surface defects acting as recombination centers as discussed below.

At long wavelengths, around 1–1.2 μm, the absorption coefficient α of Si is small and the *absorption depth* $(1/\alpha)$ is typically greater than 100 μm. To capture these long-wavelength photons, we therefore need a thick p-side and at the same time a long minority carrier diffusion length L_e. Typically the p-side is 200–500 μm and L_e tends to be shorter than this.

Crystalline silicon has a bandgap of 1.1 eV which corresponds to a threshold wavelength of 1.1 μm. The incident energy in the wavelength region greater than 1.1 μm is then wasted; this is not a negligible amount (~25 percent). The worst part of the efficiency limitation however comes from the high-energy photons becoming absorbed near the crystal surface and being lost by recombination in the surface region. Crystal surfaces and interfaces contain a high concentration of recombination centers which facilitate the recombination of photogenerated EHPs near the surface. Losses due to EHP recombinations near or at the surface can be as high as 40 percent. These combined effects bring the efficiency down to about 45 percent. In addition, the antireflection coating is not perfect, which reduces the total collected photons by a factor of about 0.8–0.9. When we also include the limitations of the photovoltaic action itself (discussed below), the upper limit to a photovoltaic device that uses a single crystal of Si is about 24–26 percent at room temperature.

Consider an ideal pn junction photovoltaic device connected to a resistive load R as shown in Figure 6.52a. Note that I and V in the figure define the convention for the direction of positive current and positive voltage. If the load is a short circuit, then the only current in the circuit is that generated by the incident light. This is the photocurrent I_{ph} shown in Figure 6.52b which depends on the number of EHPs photogenerated within the volume enclosing the depletion region (W) and the diffusion lengths to the depletion region (Figure 6.51). The greater is the light intensity, the

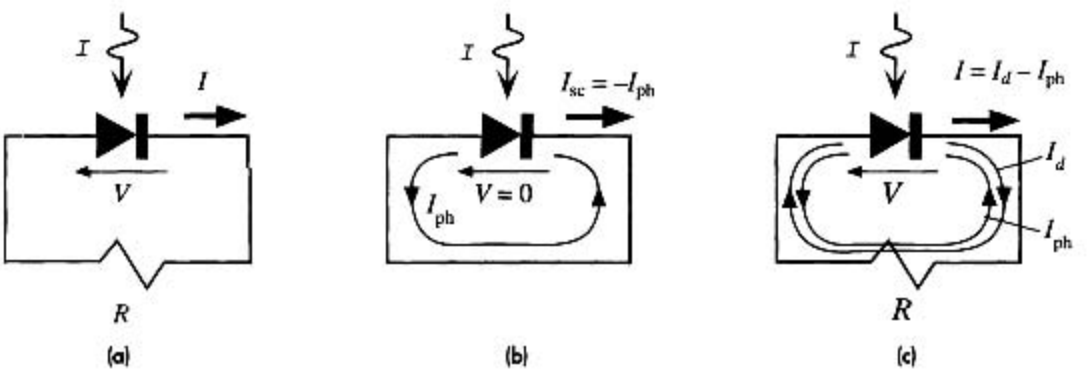


Figure 6.52

(a) The solar cell connected to an external load R and the convention for the definitions of positive voltage and positive current.

(b) The solar cell in short circuit. The current is the photocurrent I_{ph}.

(c) The solar cell driving an external load R. There is a voltage V and current I in the circuit.

higher is the photogeneration rate and the larger is I_{ph}. If I is the light intensity, then the **short circuit current** is

Short circuit solar cell current in light

$$I_{sc} = -I_{ph} = -KI \quad [6.67]$$

where K is a constant that depends on the particular device. The photocurrent does not depend on the voltage across the pn junction because there is always some internal field to drift the photogenerated EHP. We exclude the secondary effect of the voltage modulating the width of the depletion region. The photocurrent I_{ph} therefore flows even when there is not a voltage across the device.

If R is not a short circuit, then a positive voltage V appears across the pn junction as a result of the current passing through it as shown in Figure 6.52c. This voltage reduces the built-in potential of the pn junction and hence leads to minority carrier injection and diffusion just as it would in a normal diode. Thus, in addition to I_{ph} there is also a forward diode current I_d in the circuit as shown in Figure 6.52c which arises from the voltage developed across R. Since I_d is due to the normal pn junction behavior, it is given by the diode characteristics,

$$I_d = I_o\left[\exp\left(\frac{eV}{\eta kT}\right) - 1\right]$$

where I_o is the "reverse saturation current" and η is the ideality factor ($\eta = 1 - 2$). In an open circuit, the net current is zero. This means that the photocurrent I_{ph} develops just enough photovoltaic voltage V_{oc} to generate a diode current $I_d = I_{ph}$.

Thus the **total current** through the solar cell, as shown in Figure 6.52c, is

Solar cell I–V

$$I = -I_{ph} + I_o\left[\exp\left(\frac{eV}{\eta kT}\right) - 1\right] \quad [6.68]$$

The overall I–V characteristics of a typical Si solar cell are shown in Figure 6.53. It can be seen that it corresponds to the normal dark characteristics being shifted down

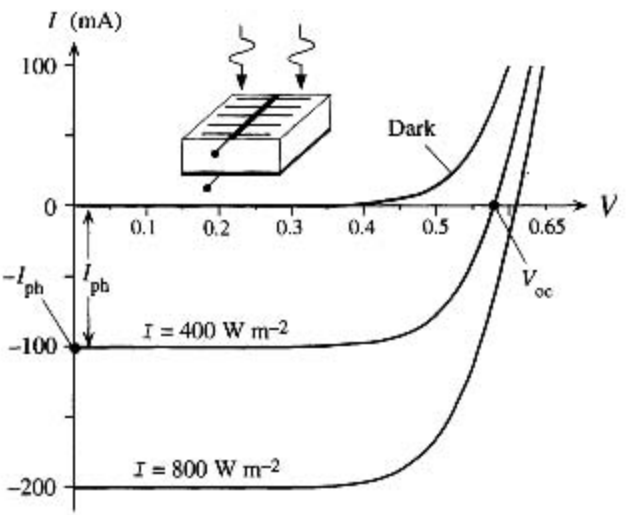


Figure 6.53 Typical I–V characteristics of a Si solar cell.
The short circuit current is I_{ph} and the open circuit voltage is V_{oc}. The I–V curves for positive current require an external bias voltage. Photovoltaic operation is always in the negative current region.

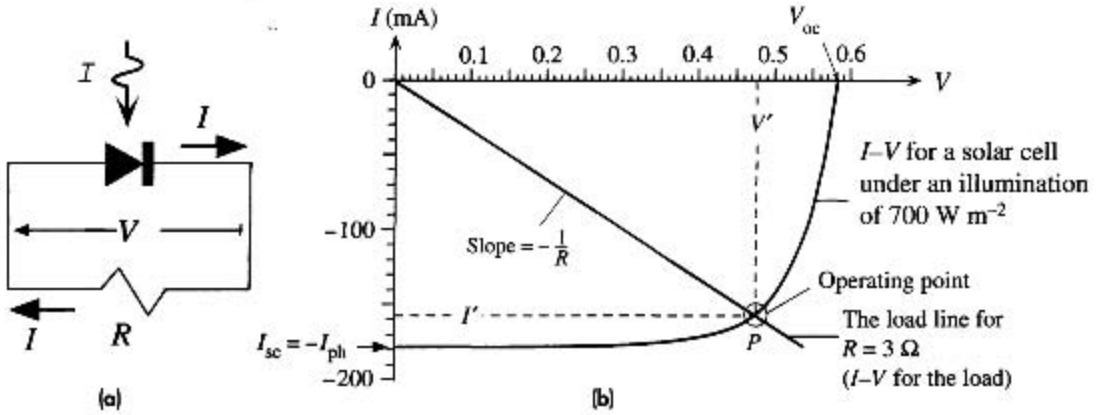


Figure 6.54

(a) When a solar cell drives a load R, R has the same voltage as the solar cell but the current through it is in the opposite direction to the convention that current flows from high to low potential.

(b) The current I' and voltage V' in the circuit of (a) can be found from a load line construction. Point P is the operating point (I', V'). The load line is for $R = 3\ \Omega$.

by the photocurrent I_{ph}, which depends on the light intensity $\mathcal{I}$. The open circuit output voltage V_{oc}, of the solar cell is given by the point where the I–V curve cuts the V axis ($I = 0$). It is apparent that although it depends on the light intensity, its value typically lies in the range 0.5–0.7 V.

Equation 6.68 gives the I–V characteristics of the solar cell. When the solar cell is connected to a load as in Figure 6.54a, the load has the same voltage as the solar cell and carries the same current. But the current I through R is now in the opposite direction to the convention that current flows from high to low potential. Thus, as shown in Figure 6.54a,

$$I = -\frac{V}{R} \qquad [6.69]$$ *The load line*

The actual current I' and voltage V' in the circuit must satisfy both the I–V characteristics of the solar cell, Equation 6.68, and that of the load, Equation 6.69. We can find I' and V' by solving these two equations simultaneously or using a graphical solution. I' and V' in the solar cell circuit are most easily found by using a **load line construction.** The I–V characteristics of the load in Equation 6.69 is a straight line with a negative slope $-1/R$. This is called the **load line** and is shown in Figure 6.54b along with the I–V characteristics of the solar cell under a given intensity of illumination. The load line cuts the solar cell characteristic at P where the load and the solar cell have the same current and voltage I' and V'. Point P therefore satisfies both Equations 6.68 and 6.69 and thus represents the **operating point of the circuit.**

The **power delivered** to the load is $P_{out} = I'V'$, which is the area of the rectangle bound by the I and V axes and the dashed lines shown in Figure 6.54b. Maximum power is delivered to the load when this rectangular area is maximized (by changing R or the intensity of illumination), when $I' = I_m$ and $V' = V_m$. Since the maximum

possible current is I_{sc} and the maximum possible voltage is V_{oc}. $I_{sc}V_{oc}$ represents the desirable goal in power delivery for a given solar cell. Therefore it makes sense to compare the maximum power output I_mV_m with $I_{sc}V_{oc}$. The **fill factor** FF, which is a figure of merit for the solar cell, is defined as

Definition of fill factor

$$\text{FF} = \frac{I_mV_m}{I_{sc}V_{oc}} \qquad [6.70]$$

The FF is a measure of the closeness of the solar cell I–V curve to the rectangular shape (the ideal shape). It is clearly advantageous to have the FF as close to unity as possible, but the exponential pn junction properties prevent this. Typically FF values are in the range 70–85 percent and depend on the device material and structure.

EXAMPLE 6.15 **A SOLAR CELL DRIVING A RESISTIVE LOAD** Consider the solar cell in Figure 6.54 that is driving a load of 3 Ω. This cell has an area of 3 cm × 3 cm and is illuminated with light of intensity 700 W m^{-2}. Find the current and voltage in the circuit. Find the power delivered to the load, the efficiency of the solar cell in this circuit, and the fill factor of the solar cell.

SOLUTION

The I–V characteristic of the load in Figure 6.54a, is the load line in Equation 6.69; that is, $I = -V/(3\ \Omega)$. The line is drawn in Figure 6.54b with a slope $1/(3\ \Omega)$. It cuts the I–V characteristics of the solar cell at $I' = 157$ mA and $V' = 0.475$ V as apparent in Figure 6.54b, which are the current and voltage, respectively, in the photovoltaic circuit of Figure 6.54a. The power delivered to the load is

$$P_{out} = I'V' = (157 \times 10^{-3})(0.475\ \text{V}) = 0.0746\ \text{W} \qquad \text{or} \qquad 74.6\ \text{mW}$$

The input of sunlight power is

$$P_{in} = (\text{Light intensity})(\text{Surface area}) = (700\ \text{W m}^{-2})(0.03\ \text{m})^2 = 0.63\ \text{W}$$

The efficiency is

$$\eta_{\text{photovoltaic}} = (100\%)\frac{P_{out}}{P_{in}} = (100\%)\frac{(0.0746\ \text{W})}{(0.63\ \text{W})} = 11.8\%$$

This will increase if the load is adjusted to extract the maximum power from the solar cell, but the increase will be small as the rectangular area $I'V'$ in Figure 6.54b is already quite close to the maximum.

The fill factor can also be calculated since point P in Figure 6.54b is close to the optimum operation, maximum output power, in which the rectangular area $I'V'$ is maximum:

$$\text{FF} = \frac{I_mV_m}{I_{sc}V_{oc}} \approx \frac{I'V'}{I_{sc}V_{oc}} = \frac{(157\ \text{mA})(0.475\ \text{V})}{(178\ \text{mA})(0.58\ \text{V})} = 0.722 \qquad \text{or} \qquad 72\%$$

EXAMPLE 6.16 **OPEN CIRCUIT VOLTAGE AND ILLUMINATION** A solar cell under an illumination of 500 W m^{-2} has a short circuit current I_{sc} of 150 mA and an open circuit output voltage V_{oc} of 0.530 V. What are the short circuit current and open circuit voltage when the light intensity is doubled? Assume $\eta = 1.5$, a typical value for various Si pn junctions.

SOLUTION

The general I–V characteristic under illumination is given by Equation 6.68. Setting $I = 0$ for open circuit,

Open circuit condition

$$I = -I_{ph} + I_o\left[\exp\left(\frac{eV_{oc}}{\eta kT}\right) - 1\right] = 0$$

Assuming that $V_{oc} \gg \eta kT/e$, rearranging the above equation we can find V_{oc},

Open circuit output voltage

$$V_{oc} = \frac{\eta kT}{e}\ln\left(\frac{I_{ph}}{I_o}\right)$$

The photocurrent I_{ph} depends on the light intensity $\mathcal{I}$ via $I_{ph} = K\mathcal{I}$, where K is a constant. Thus, at a given temperature, the change in V_{oc} is

Open circuit voltage and light intensity

$$V_{oc2} - V_{oc1} = \frac{\eta kT}{e}\ln\left(\frac{I_{ph2}}{I_{ph1}}\right) = \frac{\eta kT}{e}\ln\left(\frac{\mathcal{I}_2}{\mathcal{I}_1}\right)$$

The short circuit current is the photocurrent, so at double the intensity this is

$$I_{sc2} = I_{sc1}\left(\frac{\mathcal{I}_2}{\mathcal{I}_1}\right) = (150\text{ mA})(2) = 300\text{ mA}$$

Assuming $\eta = 1.5$, the new open circuit voltage is

$$V_{oc2} = V_{oc1} + \frac{\eta kT}{e}\ln\left(\frac{\mathcal{I}_2}{\mathcal{I}_1}\right) = 0.530\text{ V} + (1.5)(0.026)\ln(2) = 0.557\text{ V}$$

This is a 5 percent increase compared with the 100 percent increase in illumination and the short circuit current.

6.10.2 SERIES AND SHUNT RESISTANCE

Practical solar cells can deviate substantially from the ideal *pn* junction solar cell behavior depicted in Figure 6.53 due to a number of reasons. Consider an illuminated *pn* junction driving a load resistance R_L and assume that photogeneration takes place in the depletion region. As shown in Figure 6.55, the photogenerated electrons have to traverse a surface semiconductor region to reach the nearest finger electrode. All these electron paths in the n-layer surface region to finger electrodes introduce an **effective series resistance** R_s into the photovoltaic circuit. If the finger electrodes are thin, then the resistance of the electrodes themselves will further increase R_s. There is also a series resistance due to the neutral p-region, but this is generally small compared with the resistance of the electron paths to the finger electrodes.

Figure 6.56a shows the equivalent circuit of an ideal *pn* junction solar cell. The photogeneration process is represented by a *constant current generator* I_{ph}, which generates a current that is proportional to the light intensity. The flow of photogenerated carriers across the junction gives rise to a photovoltaic voltage difference V across the junction, and this voltage leads to the normal diode current $I_d = I_o[\exp(eV/\eta kT) - 1]$. This diode current I_d is represented by an ideal *pn* junction diode in the circuit as shown in Figure 6.56a. As apparent, I_{ph} and I_d are in opposite directions (I_{ph} is "up"

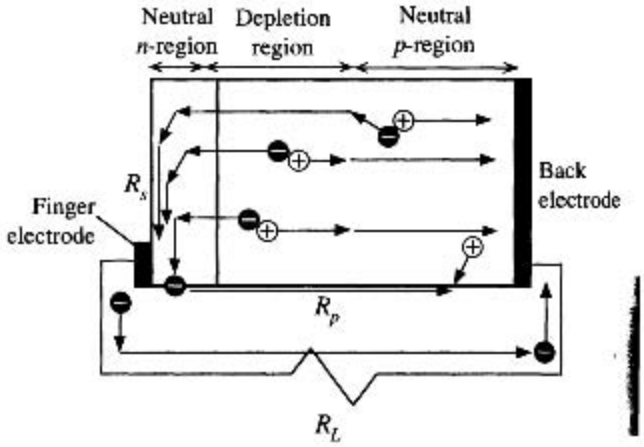


Figure 6.55 Series and shunt resistances and various fates of photogenerated EHPs.

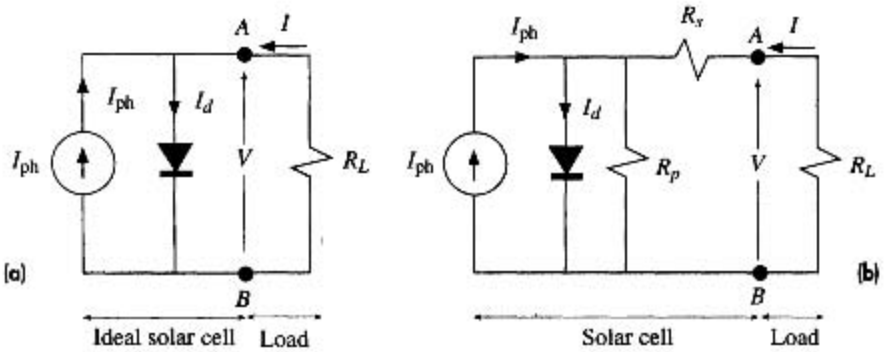


Figure 6.56 The equivalent circuit of a solar cell.
(a) Ideal *pn* junction solar cell.
(b) Parallel and series resistances R_s and R_p.

and I_d is "down"), so in an open circuit the photovoltaic voltage is such that I_{ph} and I_d have the same magnitude and cancel each other. By convention, positive current I at the output terminal is normally taken to flow into the terminal and is given by Equation 6.68. (In reality, of course, the solar cell current is negative, as in Figure 6.53, which represents a current that is flowing out into the load.)

Figure 6.56b shows the equivalent circuit of a more practical solar cell. The **series resistance** R_s in Figure 6.56b gives rise to a voltage drop and therefore prevents the ideal photovoltaic voltage from developing at the output between A and B when a current is drawn. A fraction (usually small) of the photogenerated carriers can also flow through the crystal surfaces (edges of the device) or through *grain boundaries in polycrystalline devices* instead of flowing though the external load R_L. These effects that prevent photogenerated carriers from flowing in the external circuit can be represented by an effective internal **shunt** or **parallel resistance** R_p that diverts the photocurrent away from the load R_L. Typically R_p is less important than R_s in overall

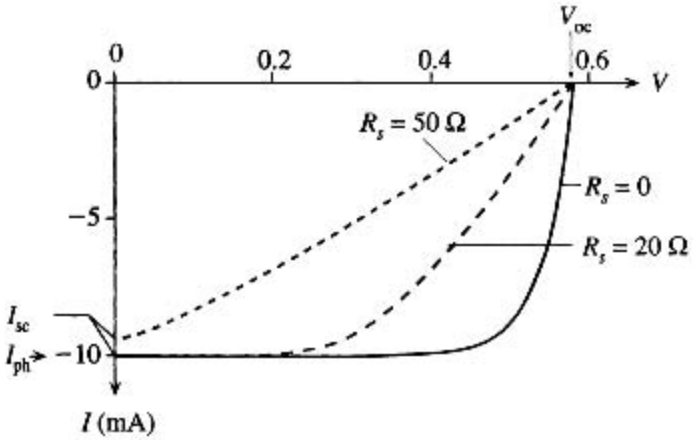


Figure 6.57 The series resistance broadens the *I–V* curve and reduces the maximum available power and hence the overall efficiency of the solar cell.

The example is a Si solar cell with $\eta \approx 1.5$ and $I_o \approx 3 \times 10^{-6}$ mA. Illumination is such that the photocurrent $I_{ph} = 10$ mA.

device behavior, unless the device is highly polycrystalline and the current component flowing through grain boundaries is not negligible.

The series resistance R_s can significantly deteriorate the solar cell performance as illustrated in Figure 6.57 where $R_s = 0$ is the best solar cell case. It is apparent that the available maximum output power decreases with the series resistance which therefore reduces the cell efficiency. Notice also that when R_s is sufficiently large, it limits the short circuit current. Similarly, low shunt resistance values, due to extensive defects in the material, also reduce the efficiency. The difference is that although R_s does not affect the open circuit voltage V_{oc}, low R_p leads to a reduced V_{oc}.

6.10.3 SOLAR CELL MATERIALS, DEVICES, AND EFFICIENCIES

Most solar cells use crystalline silicon because silicon-based semiconductor fabrication is now a mature technology that enables cost-effective devices to be manufactured. Typical Si-based solar cell efficiencies range from about 18 percent for polycrystalline to 22–24 percent in high-efficiency single-crystal devices that have special structures to absorb as many of the incident photons as possible. Solar cells fabricated by making a *pn* junction in the same crystal are called *homojunctions*. The best Si homojunction solar cell efficiencies are about 24 percent for expensive single-crystal passivated emitter rear locally diffused (PERL) cells.[12] The PERL and similar cells have a textured surface that is an array of "inverted pyramids" etched into the surface to capture as much of the incoming light as possible as depicted in Figure 6.58. Normal reflections from a flat crystal surface lead to a loss of light, whereas reflections inside the pyramid allow a second or even a third chance for absorption. Further, after refraction, photons would be entering the semiconductor at oblique angles which means that they will be absorbed in the useful photogeneration volume, that is, within the electron diffusion length of the depletion layer as shown in Figure 6.58.

[12] Much of the pioneering work for high-efficiency PERL solar cells was done by Martin Green and coworkers at the University of New South Wales.

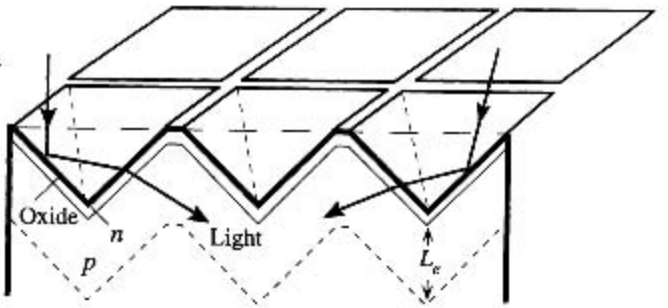


Figure 6.58 An inverted pyramid textured surface substantially reduces reflection losses and increases absorption probability in the device.

Table 6.3 summarizes some typical characteristics of various solar cells. GaAs and Si solar cells have comparable efficiencies though theoretically GaAs with a higher bandgap is supposed to have a better efficiency. The largest factors reducing the efficiency of a Si solar cell are the unabsorbed photons with $h\nu < E_g$ and short wavelength photons absorbed near the surface. Both these factors are improved if tandem cell structures or heterojunctions are used.

There are a number of III–V semiconductor alloys that can be prepared with different bandgaps but with the same lattice constant. Heterojunctions (junctions between different materials) from these semiconductors have negligible interface defects. AlGaAs has a wider bandgap than GaAs and would allow most solar photons to pass through. If we use a thin AlGaAs layer on a GaAs pn junction, as shown in Figure 6.59, then this layer passivates the surface defects normally present in a homojunction GaAs cell. The AlGaAs window layer therefore overcomes the surface recombination limitation and improves the cell efficiency (such cells have efficiencies of about 24 percent).

Table 6.3 Typical characteristics of various solar cells at room temperature under AM1.5 illumination of 1000 W m^{-2}

Semiconductor	E_g (eV)	V_{oc} (V)	J_{sc} (mA cm^{-2})	FF	η (%)	Comments
Si, single crystal	1.1	0.5–0.7	42	0.7–0.8	16–24	*Single crystal, PERL*
Si, polycrystalline	1.1	0.5–0.65	38	0.7–0.8	12–19	
Amorphous Si:Ge:H film					8–13	Amorphous film with tandem structure, convenient large-area fabrication
GaAs, single crystal	1.42	1.02	28	0.85	24–25	
GaAlAs/GaAs, tandem		1.03	27.9	0.864	24.8	Different bandgap materials in tandem increases absorption efficiency
GaInP/GaAs, tandem		2.5	14	0.86	25–30	Different bandgap materials in tandem increases absorption efficiency
CdTe, thin film	1.5	0.84	26	0.75	15–16	
InP, single crystal	1.34	0.87	29	0.85	21–22	
$CuInSe_2$	1.0				12–13	

NOTE: AM1.5 refers to a solar illumination of "Air Mass 1.5," which represents solar radiation falling on the Earth's surface with a total intensity (or irradiance) of 1000 W m^{-2}. AM1.5 is widely used for comparing solar cells.

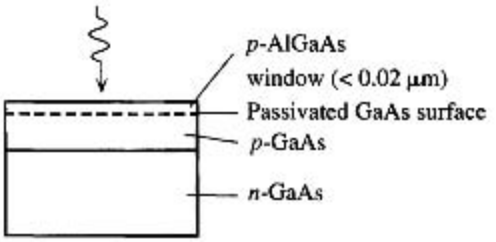


Figure 6.59 AlGaAs window layer on GaAs passivates the surface states and thereby increases the photogeneration efficiency.

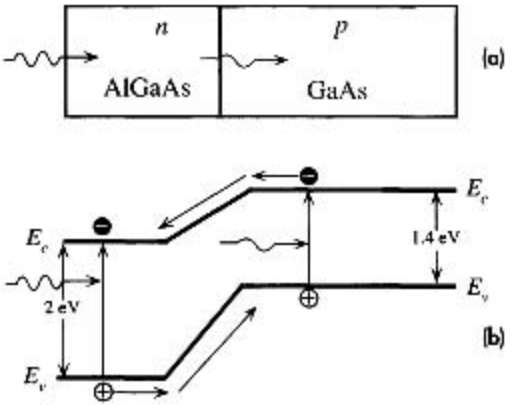


Figure 6.60 A heterojunction solar cell between two different bandgap semiconductors (GaAs and AlGaAs).

Heterojunctions between different bandgap III–V semiconductors that are lattice matched offer the potential of developing high-efficiency solar cells. The simplest single heterojunction example, shown in Figure 6.60, consists of a *pn* junction using a wider bandgap *n*-AlGaAs with *p*-GaAs. Energetic photons ($h\nu > 2$ eV) are absorbed in AlGaAs, whereas those with energies less than 2 eV but greater than 1.4 eV are absorbed in the GaAs layer. In more sophisticated cells, the bandgap of AlGaAs is graded slowly from the surface by varying the composition of the AlGaAs layer.

Tandem or cascaded cells use two or more cells in tandem or in cascade to increase the absorbed photons from the incident light as illustrated in Figure 6.61. The first cell is made from a wider bandgap (E_{g1}) material and only absorbs photons with $h\nu > E_{g1}$. The second cell with bandgap E_{g2} absorbs photons that pass the first cell and have $h\nu > E_{g2}$. The whole structure can be grown within a single crystal by using lattice-matched crystalline layers leading to a monolithic tandem cell. If, in addition, light concentrators are also used, the efficiency can be further increased. For example, a GaAs–GaSb tandem cell operating under a 100-sun condition, that is, 100 times that of ordinary sunlight, have exhibited an efficiency of about 34 percent. Tandem cells have been used in thin-film a-Si:H (hydrogenated amorphous Si) *pin* (*p*-type, intrinsic, and *n*-type structure) solar cells to obtain efficiencies up to about 12 percent. These tandem cells have a-Si:H and a-Si:Ge:H cells and are easily fabricated in large areas.

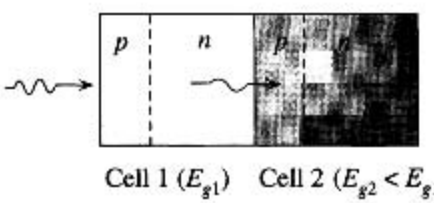


Figure 6.61 A tandem cell.
Cell 1 has a wider bandgap and absorbs energetic photons with $h\nu > E_{g1}$. Cell 2 absorbs photons that pass through cell 1 and have $h\nu > E_{g2}$.

ADDITIONAL TOPICS

6.11 *pin* DIODES, PHOTODIODES, AND SOLAR CELLS

The *pin* Si diode is a device that has a structure with three distinct layers: a heavily doped thin p^+-type layer, a relatively thick intrinsic (i-Si) layer, and a heavily doped thin n^+-type layer, as shown in Figure 6.62a. For simplicity we will assume that the i-layer is truly intrinsic, or at least doped so lightly compared with p^+ and n^+ layers that it behaves almost as if intrinsic. The intrinsic layer is much wider than the p^+ and n^+ regions, typically 5–50 μm depending on the particular application. When the structure is first formed, holes diffuse from the p^+-side and electrons from the n^+-side into the i-Si layer where they recombine and disappear. This leaves behind a thin layer of exposed negatively charged acceptor ions in the p^+-side and a thin layer of exposed positively charged donor ions in the n^+-side as shown in Figure 6.22b. The two charges are separated by the i-Si layer of thickness W. There is a uniform built-in field $\mathcal{E}_o$ in the i-Si layer from the exposed positive ions to the exposed negative ions as illustrated in Figure 6.22c. (Since there is no net space charge in the i-layer, from $d\mathcal{E}/dx = \rho/\varepsilon_o\varepsilon_r = 0$, the field must be uniform.) In contrast, the built-in field in the depletion layer of a *pn* junction is not uniform. With no applied bias, the equilibrium is maintained by the built-in field $\mathcal{E}_o$ which prevents further diffusion of majority carriers from the p^+ and n^+ layers into the i-Si layer. A hole that manages to diffuse from the p^+-side into the i-layer is drifted back by $\mathcal{E}_o$, so the net current is zero. As in the *pn* junction, there is also a built-in potential V_o from the edge of the p^+-side depletion region to the edge of the n^+-side depletion region. V_o (like $\mathcal{E}_o$) provides a potential barrier against further net diffusion of holes and electrons into the i-layer and maintains the equilibrium in the open circuit (net current being zero) as in the *pn* junction. It is apparent from Figure 6.62c that, in the absence of an applied voltage, $\mathcal{E}_o = V_o/W$.

One of the distinct advantages of *pin* diodes is that the depletion layer capacitance is very small and independent of the voltage. The separation of two very thin layers of negative and positive charges by a fixed distance, width W of the i-Si layer, is the same as that in a parallel plate capacitor. The **junction or depletion layer capacitance** of the *pin* diode is simply given by

Junction capacitance of pin

$$C_{\text{dep}} = \frac{\varepsilon_o\varepsilon_r A}{W} \qquad [6.70]$$

where A is the cross-sectional area and $\varepsilon_o\varepsilon_r$ is the permittivity of the semiconductor (Si), respectively. Further, since the width W of the i-Si layer is fixed by the structure, the junction capacitance does not depend on the applied voltage in contrast to that of the *pn* junction. C_{dep} is typically of the order of a picofarad in fast *pin* photodiodes, so with a 50 Ω resistor, the RC_{dep} time constant is about 50 ps.

When a reverse bias voltage V_r is applied across the *pin* device, it drops almost entirely across the width of the i-Si layer. The depletion layer widths of the thin sheets of acceptor and donor charges in the p^+ and n^+ sides are negligible compared with W. The reverse bias V_r increases the built-in voltage to $V_o + V_r$ as shown in Figure 6.62d. The field $\mathcal{E}$ in the i-Si layer is still uniform and increases to

Reverse-biased pin

$$\mathcal{E} = \mathcal{E}_o + \frac{V_r}{W} \approx \frac{V_r}{W} \qquad (V_r \gg V_o) \qquad [6.71]$$

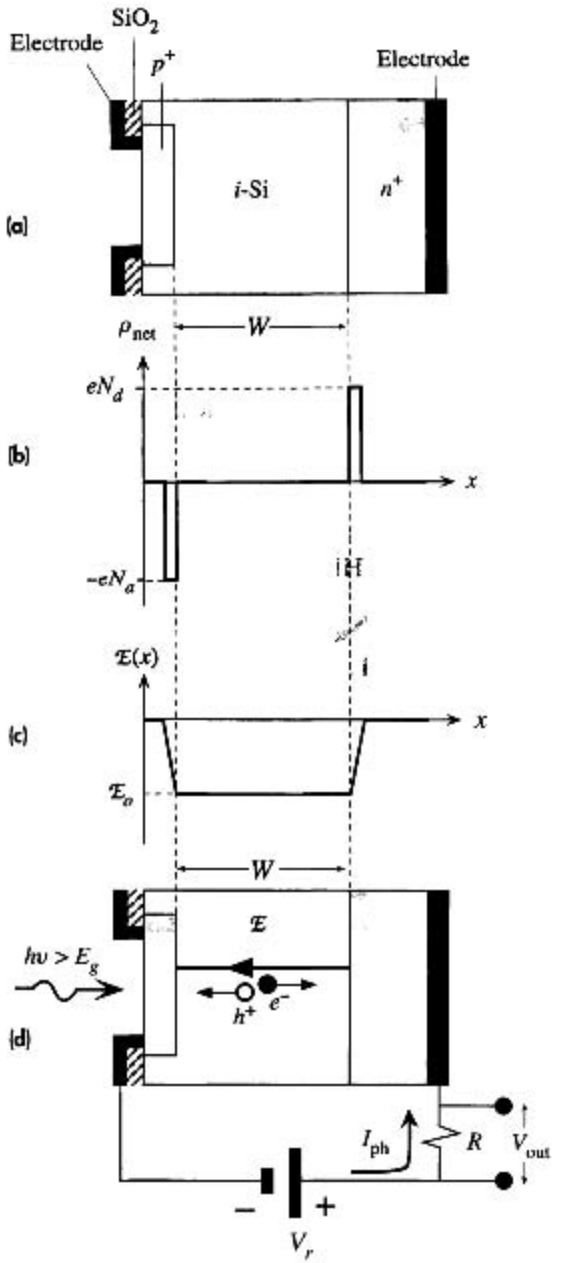


Figure 6.62
(a) The schematic structure of an idealized *pin* photodiode.
(b) The net space charge density across the photodiode.
(c) The built-in field across the diode.
(d) The *pin* photodiode in photodetection is reverse-biased.

Since the width of the i-layer in a *pin* device is typically much larger than the depletion layer width in an ordinary *pn* junction, the *pin* devices usually have higher breakdown voltages, which makes them useful where high breakdown voltages are required.

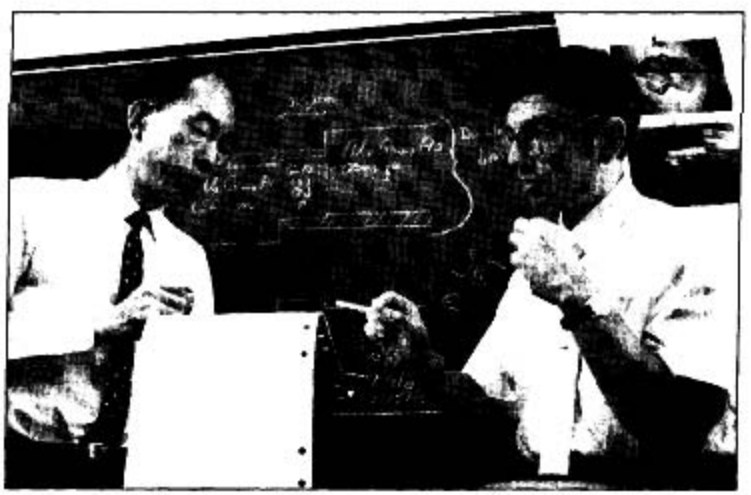
In *pin* photodetectors, the *pin* structure is designed so that photon absorption occurs primarily over the i-Si layer. The photogenerated electron–hole pairs (EHPs) in the i-Si layer are then separated by the field $\mathcal{E}$ and drifted toward the n^+ and p^+ sides,

respectively, as illustrated in Figure 6.62d. While the photogenerated carriers are drifting through the i-Si layer, they give rise to an external photocurrent which is easily detected as a voltage across a small sampling resistor R in Figure 6.62d (or detected by a current-to-voltage converter). The response time of the *pin* photodiode is determined by the transit times of the photogenerated carriers across the width W of the i-Si layer. Increasing W allows more photons to be absorbed, which increases the output signal per input light intensity, but it slows down the speed of response because carrier transit times become longer.

The simple *pn* junction photodiode has two major drawbacks. Its junction or depletion layer capacitance is not sufficiently small to allow photodetection at high modulation frequencies. This is an RC time constant limitation. Secondly, its depletion layer is at most a few microns. This means that at long wavelengths where the penetration depth is greater than the depletion layer width, the majority of photons are absorbed outside the depletion layer where there is no field to separate the EHPs and drift them. The photodetector efficiency is correspondingly low at these long wavelengths. These problems are substantially reduced in the *pin* photodiode.[13] The *pin* photovoltaic devices, such as a-Si:H solar cells, are designed to have the photogeneration occur in the i-layer as in the case of photodetectors. Obviously, there is no external applied bias, and the built-in field $\mathcal{E}_o$ separates the EHPs and drives the photocurrent.

6.12 SEMICONDUCTOR OPTICAL AMPLIFIERS AND LASERS

All practical semiconductor laser diodes are double heterostructures (DH) whose energy band diagrams are similar to the LED diagram in Figure 6.46. The energy band diagram of a forward biased DH laser diode is shown in Figure 6.63a and b.



Izuo Hayashi and Morton Panish at Bell Labs (1971) were able to design the first semiconductor laser that operated continuously at room temperature. (Notice the similarity of the energy band diagram on the chalkboard with that in Figure 6.63.)

SOURCE: Courtesy of Bell Labs, Lucent Technologies.

[13] The *pin* photodiode was invented by J. Nishizawa and his research group in Japan in 1950.

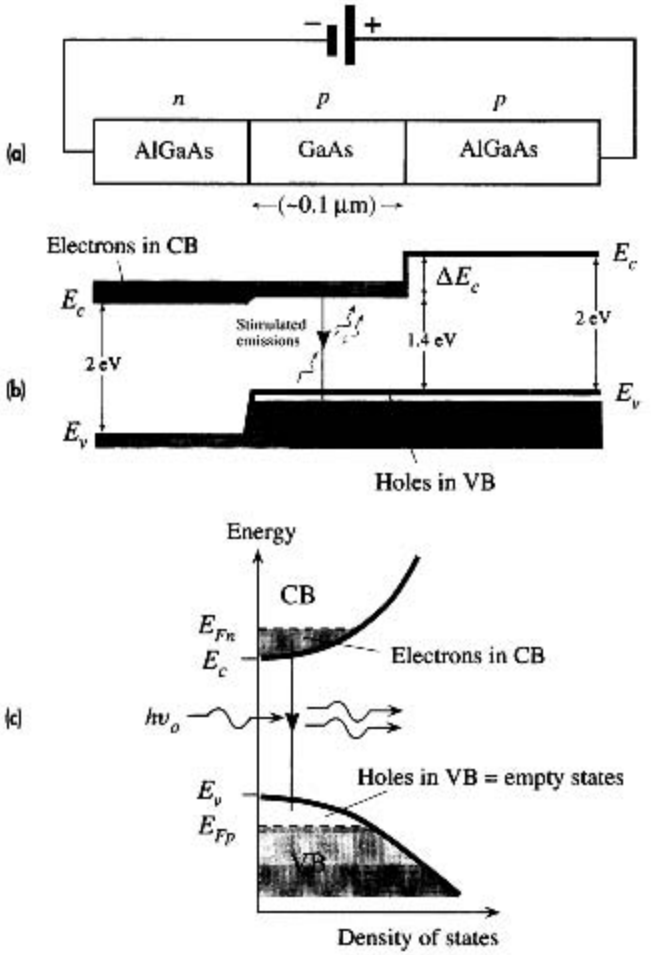


Figure 6.63

(a) A double heterostructure diode has two junctions which are between two different bandgap semiconductors (GaAs and AlGaAs).

(b) Simplified energy band diagram under a large forward bias. Lasing recombination takes place in the *p*-GaAs layer, the *active layer*.

(c) The density of states and energy distribution of electrons and holes in the conduction and valence bands in the active layer.

In this case the semiconductors are AlGaAs with $E_g \approx 2$ eV and GaAs with $E_g \approx 1.4$ eV. The p-GaAs region is a thin layer, typically 0.1–0.2 μm, and constitutes the **active layer** in which stimulated emissions take place. Both p-GaAs and p-AlGaAs are heavily p-type doped and are degenerate with the Fermi level E_{Fp} in the valence band. When a sufficiently large forward bias is applied, E_c of n-AlGaAs moves very close to the E_c of p-GaAs which leads to a large injection of electrons in the CB of n-AlGaAs into p-GaAs as shown in Figure 6.63b. In fact, with a sufficient large forward bias, E_c of AlGaAs can be moved above the E_c of GaAs, which causes an enormous electron injection from n-AlGaAs into the CB of p-GaAs. These injected electrons, however, are *confined* to the CB of p-GaAs since there is a barrier ΔE_c between p-GaAs and p-AlGaAs due to the change in the bandgap.

The p-GaAs layer is degenerately doped. Thus, the top of its valence band (VB) is full of holes, or it has all the electronic states *empty* above the Fermi level E_{Fp}

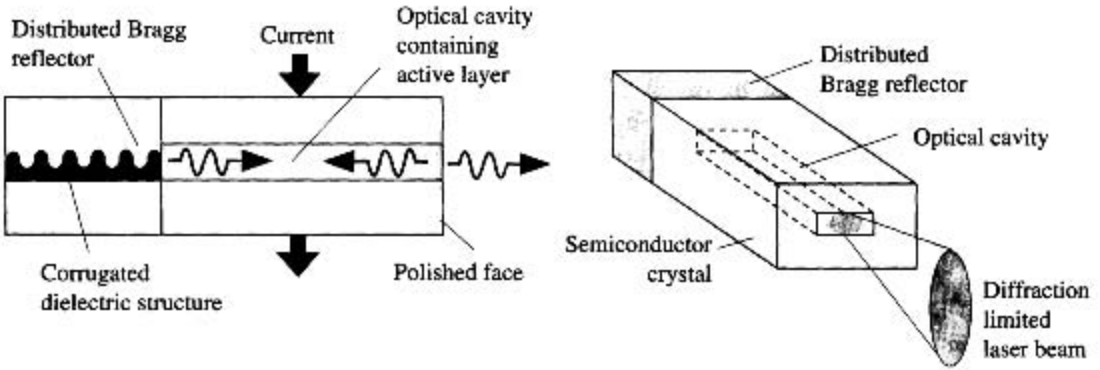


Figure 6.64 Semiconductor lasers have an optical cavity to build up the required electromagnetic oscillations. In this example, one end of the cavity has a Bragg distributed reflector, a reflection grating, that reflects only certain wavelengths back into the cavity.

in this layer. The large forward bias injects a very large concentration of electrons from n-AlGaAs into the conduction band of p-GaAs. Consequently, as shown in Figure 6.63c, there is a large concentration of electrons in the CB and totally empty states at the top of the VB, which means that there is a *population inversion*. An incoming photon with an energy $h\nu_o$ just above E_g can stimulate a conduction electron in the p-GaAs layer to fall down from the CB to the VB and emit a photon by *stimulated emission* as depicted in Figure 6.63c. Such a transition is a photon-stimulated electron–hole recombination, or a lasing recombination. Thus, an avalanche of stimulated emissions in the active layer provides an **optical amplification** of photons with $h\nu_o$ in this layer. The amplification depends on the extent of population inversion and hence on the diode forward current. The device operates as a **semiconductor optical amplifier** which amplifies an optical signal that is passed through the active layer. There is a threshold current below which there is no stimulated emission and no optical amplification.

To construct a **semiconductor laser** with a self-sustained lasing emission we have to incorporate the active layer into an *optical cavity* just as in the case of the HeNe laser in Chapter 3. The optical cavity with reflecting ends, reflects the coherent photons back and forward and encourages their constructive interference within the cavity as depicted in Figure 6.64. This leads to a buildup of high-energy electromagnetic oscillations in the cavity. Some of this electromagnetic energy in the cavity is tapped out as output radiation by having one end of the cavity as partially reflecting. For example, one type of optical cavity, as shown in Figure 6.64, has a special reflector, called a **Bragg distributed reflector** (BDR), at one end to reflect only certain wavelengths back into the cavity.[14] A BDR is a periodic corrugated

[14] Partial reflections of waves from the corrugations in the DBR can interfere constructively and constitute a reflected wave only for certain wavelengths, called *Bragg wavelengths*, that are related to the periodicity of the corrugations. A DBR acts like a reflection grating in optics.

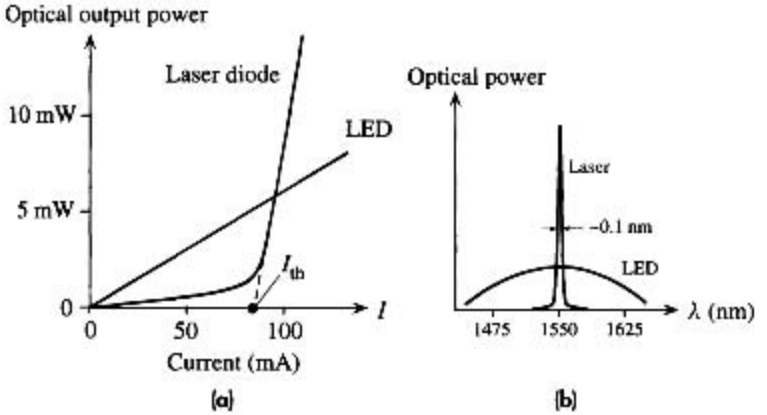


Figure 6.65
(a) Typical optical power output versus forward current for a laser diode and an LED.
(b) Comparison of spectral output characteristics.

structure, like a reflection grating, etched in a semiconductor that reflects only certain wavelengths that are related to the corrugation periodicity. This Bragg reflector has a corrugation periodicity such that it reflects only one desirable wavelength that falls within the optical gain of the active region. This wavelength selective reflection leads to only one possible electromagnetic radiation mode existing in the cavity, which leads to a very narrow output spectrum: a *single-mode output,* that is, only one peak in the output spectrum shown in Figure 3.43. Semiconductor lasers that operate with only one mode in the radiation output are called **single-mode** or **single-frequency lasers;** the spectral linewidth of a single-mode laser output is typically ~ 0.1 nm, which should be compared with an LED spectral width of 150 nm operating at a 1550 nm emission.

The double heterostructure has further advantages. Wider bandgap semiconductors generally have lower refractive indices, which means AlGaAs has a lower refractive index than that of GaAs. The change in the refractive index defines an optical dielectric waveguide that confines the photons to the active region of the optical cavity and thereby reduces photon losses and increases the photon concentration. This increase in the photon concentration increases the rate of stimulated emissions and the efficiency of the laser.

To achieve the necessary stimulated emissions from a laser diode and build up the necessary optical oscillations in the cavity (to overcome all the optical losses) the current must exceed a certain **threshold current** I_{th} as shown in Figure 6.65a. The optical power output at a current I is then very roughly proportional to $I - I_{th}$. There is still some weak optical power output below I_{th}, but this is simply due to spontaneous recombinations of injected electrons and holes in the active layer; the laser diode behaves like a "poor" LED below I_{th}. The output light from an LED however increases almost in proportion to the diode current. Figure 6.65b compares the output spectrum from the two devices. Remember that the output light from the laser diode is *coherent radiation,* whereas that from an LED is a stream of incoherent photons.

CD Selected Topics and Solved Problems

Selected Topics

The *pn* Junction: Diffusion or Drift? Fick or Ohm?
Shot Noise Generated by the *pn* Junction
Voltage Drift in Semiconductor Devices due to Thermoelectric Effects
Transistor Switches: Why the Saturated Collector–Emitter Voltage is 0.2 V
Semiconductor Device Fabrication (Overview)
Photolithography and Minimum Line Width in Semiconductor Fabrication
Depletion MOSFET Fundamentals
High-Frequency Small-Signal BJT Model

Solved Problems

pn Junction: The Shockley Model
Recombination Current and I–V Characteristics of a *pn* Junction Diode
Design of a *pn* Junction Diode
Bipolar Junction Transistors at Low Frequencies: Principles and Solved Problems
BJT and Nonuniform Base Doping Effect
Junction Field Effect Transistor (JFET)
Enhancement MOSFET and CS Amplifier
LED Emission Wavelength and Temperature

DEFINING TERMS

Accumulation occurs when an applied voltage to the gate (or metal electrode) of a MOS device causes the semiconductor under the oxide to have a greater number of majority carriers than the equilibrium value. Majority carriers have been accumulated at the surface of the semiconductor under the oxide.

Active device is a device that exhibits gain (current or voltage, or both) and has a directional electronic function. Transistors are active devices, whereas resistors, capacitors, and inductors are passive devices.

Antireflection coating reduces light reflection from a surface.

Avalanche breakdown is the enormous increase in the reverse current in a *pn* junction when the applied reverse field is sufficiently high to cause the generation of electron–hole pairs by impact ionization in the space charge layer.

Base width modulation (the Early effect) is the modulation of the base width by the voltage appearing across the base–collector junction. An increase in the base to collector voltage increases the collector junction depletion layer width, which results in the narrowing of the base width.

Bipolar junction transistor (BJT) is a transistor whose normal operation is based on the injection of carriers from the emitter into the base region, where they become minority carriers, and their subsequent diffusion to the collector, where they give rise to a collector current. The voltage between the base and the emitter controls the collector current.

Built-in field is the internal electric field in the depletion region of a *pn* junction that is maximum at the metallurgical junction. It is due to exposed negative acceptors on the *p*-side and positive donors on the *n*-side of the junction.

Built-in voltage (V_o) is the voltage across a *pn* junction, going from a *p*- to *n*-type semiconductor, in an open circuit.

Channel is the conducting strip between the source and drain regions of a MOSFET.

Chip is a piece (or a volume) of a semiconductor crystal that contains many integrated active and passive components to implement a circuit.

Collector junction is the metallurgical junction between the base and the collector of a bipolar transistor.

Critical electric field is the field in the space charge (or depletion) region at reverse breakdown (avalanche or Zener).

Depletion layer (or **space charge layer, SCL**) is a region around the metallurgical junction where recombination of electrons and holes has depleted this region of its large number of equilibrium majority carriers.

Depletion (space charge) layer capacitance is the incremental capacitance (dQ/dV) due to the change in the exposed dopant charges in the depletion layer as a result of the change in the voltage across the *pn* junction.

Diffusion is the flow of particles of a given species from high- to low-concentration regions by virtue of their random thermal motions.

Diffusion (storage) capacitance is the *pn* junction capacitance due to the diffusion and storage of minority carriers in the neutral regions when a forward bias is applied.

Dynamic (incremental) resistance r_d of a diode is the change in the voltage across the diode per unit change in the current through the diode $r_d = dV/dI$. It is the low-frequency ac resistance of the diode. *Dynamic conductance* g_d is the reciprocal dynamic resistance: $g_d = 1/r_d$.

Emitter junction is the metallurgical junction between the emitter and the base.

Enhancement MOSFET is a MOSFET device that needs a gate to source voltage above the threshold voltage to form a conducting channel between the source and the drain. In the absence of a gate voltage, there is no conduction between the source and drain. In its usual mode of operation, the gate voltage enhances the conductance of the source to drain inversion layer and increases the drain current.

Epitaxial layer is a thin layer of crystal that has been grown on the surface of another crystal which is usually a substrate, a mechanical support for the new crystal layer. The atoms of the new layer bond to follow the crystal pattern of the substrate, so the crystal structure of the epitaxial layer is matched with the crystal structure of the substrate.

External quantum efficiency is the optical power emitted from a light emitting device per unit electric input power.

Field effect transistor (FET) is a transistor whose normal operation is based on controlling the conductance of a channel between two electrodes by the application of an external field. The effect of the applied field is to control the current flow. The current is due to majority carrier drift from the source to the drain and is controlled by the voltage applied to the gate.

Fill factor (FF) is a figure of merit for a solar cell that represents, as a percentage, the maximum power $I_m V_m$ available to an external load as a fraction of the *ideal* theoretical power determined by the product of the short circuit current I_{sc} and the open circuit voltage V_{oc}: $\text{FF} = (I_m V_m)/(I_{sc} V_{oc})$.

Forward bias is the application of an external voltage to a *pn* junction such that the positive terminal is connected to the *p*-side and the negative to the *n*-side. The applied voltage reduces the built-in potential.

Heterojunction is a junction between different semiconductor materials, for example, between GaAs and AlGaAs ternary alloy. There may or may not be a change in the doping.

Homojunction is a junction between differently doped regions of the same semiconducting material, for example, a *pn* junction in the same silicon crystal; there is no change in the bandgap energy E_g.

Impact ionization is the process by which a high electric field accelerates a free charge carrier (electron in the CB), which then impacts with a Si–Si bond to generate a free electron–hole pair. The impact excites an electron from E_v to E_c.

Integrated circuit (IC) is a chip of a semiconductor crystal in which many active and passive components have been miniaturized and integrated together to form a sophisticated circuit.

Inversion occurs when an applied voltage to the gate (or metal electrode) of a MOS device causes the semiconductor under the oxide to develop a conducting layer (or a channel) at the surface of the semiconductor. The conducting layer has opposite polarity carriers to the bulk semiconductor and hence is termed an inversion layer.

Ion implantation is a process that is used to bombard a sample in a vacuum with ions of a given species of

atom. First the dopant atoms are ionized in a vacuum and then accelerated by applying voltage differences to impinge on a sample to be doped. The sample is grounded to neutralize the implanted ions.

Isoelectronic impurity atom has the same valency as the host atom.

Law of the junction relates the injected minority carrier concentration just outside the depletion layer to the applied voltage. For holes in the n-side, it is

$$p_n(0) = p_{no} \exp\left(\frac{eV}{kT}\right)$$

where $p_n(0)$ is the hole concentration just outside the depletion layer.

Linewidth is the width of the intensity versus wavelength spectrum, usually between the half-intensity points, emitted from a light emitting device.

Long diode is a pn junction with neutral regions longer than the minority carrier diffusion lengths.

Metallurgical junction is where there is an effective junction between the p-type and n-type doped regions in the crystal. It is where the donor and acceptor concentrations are equal or where there is a transition from n- to p-type doping.

Metal-oxide-semiconductor transistor (MOST) is a field effect transistor in which the conductance between the source and drain is controlled by the voltage supplied to the gate electrode, which is insulated from the channel by an oxide layer.

Minority carrier injection is the flow of electrons into the p-side and holes into the n-side of a pn junction when a voltage is applied to reduce the built-in voltage across the junction.

MOS is short for a metal-insulator-semiconductor structure in which the insulator is typically silicon oxide. It can also be a different type of dielectric; for example, it can be the nitride Si_3N_4.

NMOS is an enhancement type n-channel MOSFET.

Passive device or component is a device that exhibits no gain and no directional function. Resistors, capacitors, and inductors are passive components.

Photocurrent is the current generated by a light-receiving device when it is illuminated.

Pinch-off voltage is the gate to source voltage needed to just pinch off the conducting channel between the source and drain with no source to drain voltage applied. It is also the source to drain voltage that just pinches off the channel when the gate and source are shorted. Beyond pinch-off, the drain current is almost constant and controlled by V_{GS}.

PMOS is an enhancement type p-channel MOSFET.

Poly-Si gate is short for a polycrystalline and highly doped Si gate.

Recombination current flows under forward bias to replenish the carriers recombining in the space charge (depletion) layer. Typically, it is described by $I = I_{ro}[\exp(eV/2kT) - 1]$.

Reverse bias is the application of an external voltage to a pn junction such that the positive terminal is connected to the n-side and the negative to the p-side. The applied voltage increases the built-in potential.

Reverse saturation current is the reverse current that would flow in a reverse-biased ideal pn junction obeying the Shockley equation.

Shockley diode equation relates the diode current to the diode voltage through $I = I_o[\exp(eV/kT) - 1]$. It is based on the injection and diffusion of injected minority carriers by the application of a forward bias.

Short diode is a pn junction in which the neutral regions are shorter than the minority carrier diffusion lengths.

Small-signal equivalent circuit of a transistor replaces the transistor with an equivalent circuit that consists of resistances, capacitances, and dependent sources (current or voltage). The equivalent circuit represents the transistor behavior under small-signal ac conditions. The batteries are replaced with short circuits (or their internal resistances). Small signals imply small variations about dc values.

Substrate is a single mechanical support that carries active and passive devices. For example, in integrated circuit technology, typically, many integrated circuits are fabricated on a single silicon crystal wafer that serves as the substrate.

Thermal generation current is the current that flows in a reverse-biased pn junction as a result of the thermal

generation of electron–hole pairs in the depletion layer that become separated and swept across by the built-in field.

Threshold voltage is the gate voltage needed to establish a conducting channel between the source and drain of an enhancement MOST (metal-oxide-semiconductor transistor).

Transistor is a three-terminal solid-state device in which a current flowing between two electrodes is controlled by the voltage between the third and one of the other terminals or by a current flowing into the third terminal.

Turn-on, or **cut-in,** voltage of a diode is the voltage beyond which there is a substantial increase in the current. The turn-on voltage of a Si diode is about 0.6 V whereas it is about 1 V for a GaAs LED. The turn-on voltage of a *pn* junction diode depends on the bandgap of the semiconductor and the device structure.

Zener breakdown is the enormous increase in the reverse current in a *pn* junction when the applied voltage is sufficient to cause the tunneling of electrons from the valence band in the *p*-side to the conduction band in the *n*-side. Zener breakdown occurs in *pn* junctions that are heavily doped on both sides so that the depletion layer width is narrow.

QUESTIONS AND PROBLEMS

6.1 **The *pn* junction** Consider an abrupt Si pn^+ junction that has 10^{15} acceptors cm^{-3} on the *p*-side and 10^{19} donors on the *n*-side. The minority carrier recombination times are $\tau_e = 490$ ns for electrons in the *p*-side and $\tau_h = 2.5$ ns for holes in the *n*-side. The cross-sectional area is 1 mm^2. Assuming a long diode, calculate the current I through the diode at room temperature when the voltage V across it is 0.6 V. What are V/I and the incremental resistance (r_d) of the diode and why are they different?

***6.2** **The Si *pn* junction** Consider a long *pn* junction diode with an acceptor doping N_a of 10^{18} cm^{-3} on the *p*-side and donor concentration of N_d on the *n*-side. The diode is forward-biased and has a voltage of 0.6 V across it. The diode cross-sectional area is 1 mm^2. The minority carrier recombination time τ depends on the dopant concentration $N_{dopant}(cm^{-3})$ through the following approximate relation

$$\tau = \frac{5 \times 10^{-7}}{(1 + 2 \times 10^{-17} N_{dopant})}$$

a. Suppose that $N_d = 10^{15}$ cm^{-3}. Then the depletion layer extends essentially into the *n*-side and we have to consider minority carrier recombination time τ_h in this region. Calculate the diffusion and recombination contributions to the total diode current. What is your conclusion?

b. Suppose that $N_d = N_a = 10^{18}$ cm^{-3}. Then W extends equally to both sides and, further, $\tau_e = \tau_h$. Calculate the diffusion and recombination contributions to the diode current. What is your conclusion?

6.3 **Junction capacitance of a *pn* junction** The capacitance (C) of a reverse-biased abrupt Si p^+n junction has been measured as a function of the reverse bias voltage V_r as listed in Table 6.4. The *pn* junction cross-sectional area is 500 μm × 500 μm. By plotting $1/C^2$ versus V_r, obtain the built-in potential V_o and the donor concentration N_d in the *n*-region. What is N_a?

Table 6.4 Capacitance at various values of reverse bias (V_r)

V_r (V)	1	2	3	5	10	15	20
C (pF)	38.3	30.7	26.4	21.3	15.6	12.9	11.3

6.4 Temperature dependence of diode properties

a. Consider the reverse current in a *pn* junction. Show that

$$\frac{\delta I_{rev}}{I_{rev}} \approx \left(\frac{E_g}{\eta kT}\right)\frac{\delta T}{T}$$

where $\eta = 2$ for Si and GaAs, in which thermal generation in the depletion layer dominates the reverse current, and $\eta = 1$ for Ge, in which the reverse current is due to minority carrier diffusion to the depletion layer. It is assumed that $E_g \gg kT$ at room temperature. Order the semiconductors Ge, Si, and GaAs according to the sensitivity of the reverse current to temperature.

b. Consider a forward-biased *pn* junction carrying a *constant* current I. Show that the change in the voltage across the *pn* junction per unit change in the temperature is given by

$$\frac{dV}{dT} = -\left(\frac{V_g - V}{T}\right)$$

where $V_g = E_g/e$ is the energy gap expressed in volts. Calculate typical values for dV/dT for Ge, Si, and GaAs assuming that, typically, $V = 0.2$ V for Ge, 0.6 V for Si, and 0.9 V for GaAs. What is your conclusion? Can one assume that, typically, $dV/dT \approx -2$ mV °C^{-1} for these diodes?

6.5 Avalanche breakdown Consider a Si p^+n junction diode that is required to have an avalanche breakdown voltage of 25 V. Given the breakdown field $\mathcal{E}_{bc}$ in Figure 6.19, what should be the donor doping concentration?

6.6 Design of a *pn* junction diode Design an abrupt Si pn^+ junction that has a reverse breakdown voltage of 100 V and provides a current of 10 mA when the voltage across it is 0.6 V. Assume that, if N_{dopant} is in cm^{-3}, the minority carrier recombination time is given by

$$\tau = \frac{5 \times 10^{-7}}{(1 + 2 \times 10^{-17} N_{dopant})}$$

Mention any assumptions made.

6.7 Minority carrier profiles (the hyperbolic functions) Consider a *pnp* BJT under normal operating conditions in which the EB junction is forward-biased and the BC junction is reverse-biased. The field in the neutral base region outside the depletion layers can be assumed to be negligibly small. The continuity equation for holes $p_n(x)$ in the *n*-type base region is

$$D_h \frac{d^2 p_n}{dx^2} - \frac{p_n - p_{no}}{\tau_h} = 0 \qquad [6.71]$$

where $p_n(x)$ is the hole concentration at x from just outside the depletion region and p_{no} and τ_h are the equilibrium hole concentration and hole recombination lifetime in the base.

a. What are the boundary conditions at $x = 0$ and $x = W_B$, just outside the collector region depletion layer? (Consider the law of the junction.)

b. Show that the following expression for $p_n(x)$ is a solution of the continuity equation

$$p_n(x) = p_{no}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]\left[\frac{\sinh\left(\frac{W_B - x}{L_h}\right)}{\sinh\left(\frac{W_B}{L_h}\right)}\right] + p_{no}\left[1 - \frac{\sinh\left(\frac{x}{L_h}\right)}{\sinh\left(\frac{W_B}{L_h}\right)}\right] \qquad [6.72]$$

where $V = V_{EB}$ and $L_h = \sqrt{D_h \tau_h}$.

c. Show that Equation 6.72 satisfies the boundary conditions.

***6.8 The *pnp* bipolar transistor** Consider a *pnp* transistor in a common base configuration and under normal operating conditions. The emitter–base junction is forward-biased and the base–collector junction is reverse-biased. The emitter, base, and collector dopant concentrations are $N_{a(E)}$, $N_{d(B)}$,

and $N_{a(C)}$, respectively, where $N_{a(E)} \gg N_{d(B)} \geq N_{a(C)}$. For simplicity, assume uniform doping in all the regions. The base and emitter widths are W_B and W_E, respectively, both much shorter than the minority carrier diffusion lengths, L_h and L_e. The minority carrier lifetime in the base is the hole recombination time τ_h. The minority carrier mobility in the base and emitter are denoted by μ_h and μ_e, respectively.

The minority carrier concentration profile in the base can be represented by Equation 6.72.

a. Assuming that the emitter injection efficiency is unity show that

1. $I_E \approx \dfrac{eAD_h n_i^2 \coth\left(\dfrac{W_B}{L_h}\right)}{L_h N_{d(B)}} \exp\left(\dfrac{eV_{EB}}{kT}\right)$

2. $I_C \approx \dfrac{eAD_h n_i^2 \operatorname{cosech}\left(\dfrac{W_B}{L_h}\right)}{L_h N_{d(B)}} \exp\left(\dfrac{eV_{EB}}{kT}\right)$

3. $\alpha \approx \operatorname{sech}\left(\dfrac{W_B}{L_h}\right)$

4. $\beta \approx \dfrac{\tau_h}{\tau_t}$ where $\tau_t = \dfrac{W_B^2}{2D_h}$ is the base transit time.

b. Consider the total emitter current I_E through the EB junction, which has diffusion and recombination components as follows:

$$I_E = I_{E(so)} \exp\left(\frac{eV_{EB}}{kT}\right) + I_{E(ro)} \exp\left(\frac{eV_{EB}}{2kT}\right)$$

Only the hole component of the diffusion current (first term) can contribute to the collector current. Show that when $N_{a(E)} \gg N_{d(B)}$, the emitter injection efficiency γ is given by

$$\gamma \approx \left[1 + \frac{I_{E(ro)}}{I_{E(so)}} \exp\left(-\frac{eV_{EB}}{2kT}\right)\right]^{-1}$$

How does $\gamma < 1$ modify the expressions derived in part (*a*)? What is your conclusion (consider small and large emitter currents, or $V_{EB} = 0.4$ and 0.7 V)?

6.9 **Characteristics of an *npn* Si BJT** Consider an idealized silicon *npn* bipolar transistor with the properties in Table 6.5. Assume uniform doping in each region. The emitter and base widths are between metallurgical junctions (not neutral regions). The cross-sectional area is 100 μm × 100 μm. The transistor is biased to operate in the normal active mode. The base–emitter forward bias voltage is 0.6 V and the reverse bias base–collector voltage is 18 V.

Table 6.5 Properties of an *npn* BJT

Emitter Width	Emitter Doping	Hole Lifetime in Emitter	Base Width	Base Doping	Electron Lifetime in Base	Collector Doping
10 μm	1×10^{18} cm^{-3}	10 ns	5 μm	1×10^{16} cm^{-3}	200 ns	1×10^{16} cm^{-3}

a. Calculate the depletion layer width extending from the collector into the base and also from the emitter into the base. What is the width of the neutral base region?

b. Calculate α and hence β for this transistor, assuming unity emitter injection efficiency. How do α and β change with V_{CB}?

c. What is the emitter injection efficiency and what are α and β, taking into account that the emitter injection efficiency is not unity?

d. What are the emitter, collector, and base currents?

e. What is the collector current when $V_{CB} = 19$ V but $V_{EB} = 0.6$ V? What is the incremental collector output resistance defined as $\Delta V_{CB}/\Delta I_C$?

***6.10 Bandgap narrowing and emitter injection efficiency** Heavy doping in semiconductors leads to what is called *bandgap narrowing* which is an effective narrowing of the bandgap E_g. If ΔE_g is the reduction in the bandgap, then for an n-type semiconductor, according to Lanyon and Tuft (1979),

Bandgap narrowing

$$\Delta E_g(\text{meV}) = 22.5\left(\frac{n}{10^{18}}\right)^{1/2}$$

where n (in cm^{-3}) is the concentration of majority carriers which is equal to the dopant concentration if they are all ionized (for example, at room temperature). The new effective intrinsic concentration n_{ieff} due to the reduced bandgap is given by

Bandgap narrowing

$$n_{ieff}^2 = N_c N_v \exp\left[-\frac{(E_g - \Delta E_g)}{kT}\right] = n_i^2 \exp\left(\frac{\Delta E_g}{kT}\right)$$

where n_i is the intrinsic concentration in the absence of emitter bandgap narrowing.

Mass action law with bandgap narrowing

The equilibrium electron and hole concentrations n_{no} and p_{no}, respectively, obey

$$n_{no} p_{no} = n_{ieff}^2$$

where $n_{no} = N_d$ since nearly all donors would be ionized at room temperature.

Consider a Si *npn* bipolar transistor operating under normal active conditions with the base–emitter forward biased, and the base–collector reverse biased. The transistor has narrow emitter and base regions. The emitter neutral region width W_E is 1 μm, and the donor doping is 10^{19} cm^{-3}. The width W_B of the neutral base region is 1 μm, and the acceptor doping is 10^{17} cm^{-3}. Assume that W_E and W_B are less than the minority carrier diffusion lengths in the emitter and the base.

a. Obtain an expression for the emitter injection efficiency taking into account the emitter bandgap narrowing effect above.

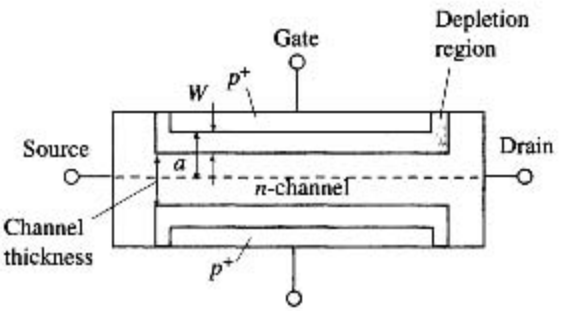
b. Calculate the emitter injection efficiency with and without the emitter bandgap narrowing.

c. Calculate the common emitter current gain β with and without the emitter bandgap narrowing effect given a perfect base transport factor ($\alpha_T = 1$).

6.11 The JFET pinch-off voltage Consider the symmetric n-channel JFET shown in Figure 6.66. The width of each depletion region extending into the n-channel is W. The thickness, or depth, of the channel, defined between the two metallurgical junctions, is $2a$. Assuming an abrupt pn junction and $V_{DS} = 0$, show that when the gate to source voltage is $-V_p$ the channel is pinched off where

$$V_p = \frac{a^2 e N_d}{2\varepsilon} - V_o$$

Figure 6.66 A symmetric JFET.



where V_o is the built-in potential between p^+n junction and N_d is the donor concentration of the channel.

Calculate the pinch-off voltage of a JFET that has an acceptor concentration of 10^{19} cm^{-3} in the p^+ gate, a channel donor doping of 10^{16} cm^{-3}, and a channel thickness (depth) $2a$ of 2 μm.

6.12 **The JFET** Consider an n-channel JFET that has a symmetric p^+n gate–channel structure as shown in Figures 6.27a and 6.66. Let L be the gate length, Z the gate width, and $2a$ the channel thickness. The pinch-off voltage is given by Question 6.11. The drain saturation current I_{DSS} is the drain current when $V_{GS} = 0$. This occurs when $V_{DS} = V_{DS(sat)} = V_P$ (Figure 6.29), so $I_{DSS} = V_P G_{ch}$, where G_{ch} is the conductance of the channel between the source and the pinched-off point (Figure 6.30). Taking into account the shape of the channel at pinch-off, if G_{ch} is about one-third of the conductance of the free or unmodulated (rectangular) channel, show that

$$I_{DSS} = V_P \left[\frac{1}{3} \frac{(e\mu_e N_d)(2a)Z}{L} \right]$$

A particular n-channel JFET with a symmetric p^+n gate–channel structure has a pinch-off voltage of 3.9 V and an I_{DSS} of 5.5 mA. If the gate and channel dopant concentrations are $N_a = 10^{19}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}, respectively, find the channel thickness $2a$ and Z/L. If $L = 10$ μm, what is Z? What is the gate–source capacitance when the JFET has no voltage supplies connected to it?

6.13 **The JFET amplifier** Consider an n-channel JFET that has a pinch-off voltage (V_P) of 5 V and $I_{DSS} = 10$ mA. It is used in a common source configuration as in Figure 6.34a in which the gate to source bias voltage (V_{GS}) is −1.5 V. Suppose that $V_{DD} = 25$ V.

a. If a small-signal voltage gain of 10 is needed, what should be the drain resistance (R_D)? What is V_{DS}?

b. If an ac signal of 3 V peak-to-peak is applied to the gate in series with the dc bias voltage, what will be the ac output voltage peak-to-peak? What is the voltage gain for positive and negative input signals? What is your conclusion?

6.14 **The enhancement NMOSFET amplifier** Consider an n-channel Si enhancement NMOS transistor that has a gate width (Z) of 150 μm, channel length (L) of 10 μm, and oxide thickness (t_{ox}) of 500 Å. The channel has $\mu_e = 700$ cm^2 V^{-1} s^{-1} and the threshold voltage (V_{th}) is 2 V ($\varepsilon_r = 3.9$ for SiO_2).

a. Calculate the drain current when $V_{GS} = 5$ V and $V_{DS} = 5$ V and assuming $\lambda = 0.01$.

b. What is the small-signal voltage gain if the NMOSFET is connected as a common source amplifier, as shown in Figure 6.67, with a drain resistance R_D of 2.2 kΩ, the gate biased at 5 V with respect to

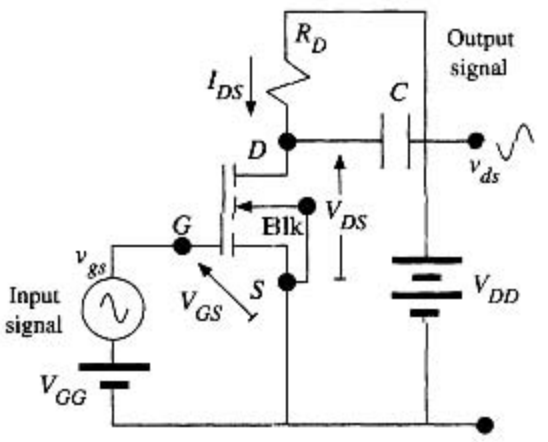


Figure 6.67 NMOSFET amplifier.

source ($V_{GG} = 5$ V) and V_{DD} is such that $V_{DS} = 5$ V? What is V_{DD}? What will happen if the drain supply is smaller?

c. Estimate the most positive and negative input signal voltages that can be amplified if V_{DD} is fixed at the above value in part (*b*).

d. What factors will lead to a higher voltage amplification?

***6.15 Ultimate limits to device performance**

a. Consider the speed of operation of an *n*-channel FET-type device. The time required for an electron to transit from the source to the drain is $\tau_t = L/v_d$, where L is the channel length and v_d is the drift velocity. This transit time can be shortened by shortening L and increasing v_d. As the field increase, the drift velocity eventually saturates at about $v_{dsat} = 10^5$ m s^{-1} when the field in the channel is equal to $\mathcal{E}_c \approx 10^6$ V m^{-1}. A short τ_t requires a field that is at least $\mathcal{E}_c$.

1. What is the change in the PE of an electron when it traverses the channel length L from source to drain if the voltage difference is V_{DS}?
2. This energy must be greater than the energy due to thermal fluctuations, which is of the order of kT. Otherwise, electrons would be brought in and out of the drain due to thermal fluctuations. Given the minimum field and V_{DS}, what is the minimum channel length and hence the minimum transit time?

b. Heisenberg's uncertainty principle relates the energy and the time duration in which that energy is possessed through a relationship of the form (Chapter 3) $\Delta E\ \Delta t > \hbar$. Given that during the transit of the electron from the source to the drain its energy changes by eV_{DS}, what is the shortest transit time τ satisfying Heisenberg's uncertainty principle? How does it compare with your calculation in part (*a*)?

c. How does electron tunneling limit the thickness of the gate oxide and the channel length in a MOSFET? What would be typical distances for tunneling to be effective? (Consider Example 3.10.)

6.16 Energy distribution of electrons in the conduction band of a semiconductor and LED emission spectrum

a. Consider the energy distribution of electrons $n_E(E)$ in the conduction band (CB). Assuming that the density of state $g_{cb}(E) \propto (E - E_c)^{1/2}$ and using Boltzmann statistics $f(E) \approx \exp[-(E - E_F)/kT]$, show that the energy distribution of the electrons in the CB can be written as

$$n_x(x) = Cx^{1/2}\exp(-x)$$

where $x = (E - E_c)/kT$ is electron energy in terms of kT measured from E_c, and C is a temperature-dependent constant (independent of E).

b. Setting arbitrarily $C = 1$, plot n_x versus x. Where is the maximum, and what is the full width at half maximum (FWHM), *i.e.*, between half maximum points?

c. Show that the average electron energy in the CB is $\frac{3}{2}kT$, by using the definition of the average,

$$x_{\text{average}} = \frac{\int_0^\infty x n_x\, dx}{\int_0^\infty n_x\, dx}$$

where the integration is from $x = 0$ (E_c) to say $x = 10$ (far away from E_c where $n_x \to 0$). You need to use a numerical integration.

d. Show that the maximum in the energy distribution is at $x = \frac{1}{2}$ or at $E_{max} = \frac{1}{2}kT$ above E_c.

e. Consider the recombination of electrons and holes in GaAs. The recombination involves the emission of a photon. Given that both electron and hole concentrations have energy distributions in the conduction and valence bands, respectively, sketch schematically the expected light

intensity emitted from electron and hole recombinations against the photon energy. What is your conclusion?

6.17 **LED output spectrum** Given that the width of the relative light intensity between half-intensity points versus photon energy spectrum of an LED is typically $\sim 3kT$, what is the linewidth $\Delta\lambda$ in the output spectrum in terms of the peak emission wavelength? Calculate the spectral linewidth $\Delta\lambda$ of the output radiation from a green LED emitting at 570 nm at 300 K.

6.18 **LED output wavelength variations** Show that the change in the emitted wavelength λ with temperature T from an LED is approximately given by

$$\frac{d\lambda}{dT} \approx -\frac{hc}{E_g^2}\left(\frac{dE_g}{dT}\right)$$

where E_g is the bandgap. Consider a GaAs LED. The bandgap of GaAs at 300 K is 1.42 eV which changes (decreases) with temperature as $dE_g/dT = -4.5 \times 10^{-4}$ eV K^{-1}. What is the change in the emitted wavelength if the temperature change is 10 °C?

6.19 **Linewidth of direct recombination LEDs** Experiments carried out on various direct bandgap semiconductor LEDs give the output spectral linewidth (between half-intensity points) listed in Table 6.6. Since wavelength $\lambda = hc/E_{ph}$, where $E_{ph} = h\nu$ is the photon energy, we know that the spread in the wavelength is related to a spread in the photon energy,

$$\Delta\lambda \approx \frac{hc}{E_{ph}^2}\Delta E_{ph}$$

Suppose that we write $E_{ph} = hc/\lambda$ and $\Delta E_{ph} = \Delta(h\nu) \approx nkT$ where n is a numerical constant. Show that,

$$\Delta\lambda = \frac{nkT}{hc}\lambda^2$$

LED output spectrum linewidth

and by appropriately plotting the data in Table 6.6 find n.

Table 6.6 Linewidth $\Delta\lambda_{1/2}$ between half-points in the output spectrum (intensity versus wavelength) of GaAs and AlGaAs LEDs

	Peak wavelength of emission λ (nm)							
	650	**810**	**820**	**890**	**950**	**1150**	**1270**	**1500**
$\Delta\lambda_{1/2}$ (nm)	22	36	40	50	55	90	110	150
Material (direct E_g)	AlGaAs	AlGaAs	AlGaAs	GaAs	GaAs	InGaAsP	InGaAsP	InGaAsP

6.20 **AlGaAs LED emitter** An AlGaAs LED emitter for use in a local optical fiber network has the output spectrum shown in Figure 6.68. It is designed for peak emission at 820 nm at 25 °C.

a. What is the linewidth $\Delta\lambda$ between half power points at temperatures −40 °C, 25 °C, and 85 °C? Given these three temperatures, plot $\Delta\lambda$ and T (in K) and find the empirical relationship between $\Delta\lambda$ and T. How does this compare with $\Delta(h\nu) \approx 2.5kT$ to $3kT$?

b. Why does the peak emission wavelength increase with temperature?

c. What is the bandgap of AlGaAs in this LED?

d. The bandgap E_g of the ternary alloys $Al_xGa_{1-x}As$ follows the empirical expression

$$E_g(\text{eV}) = 1.424 + 1.266x + 0.266x^2$$

What is the composition of the AlGaAs in this LED?

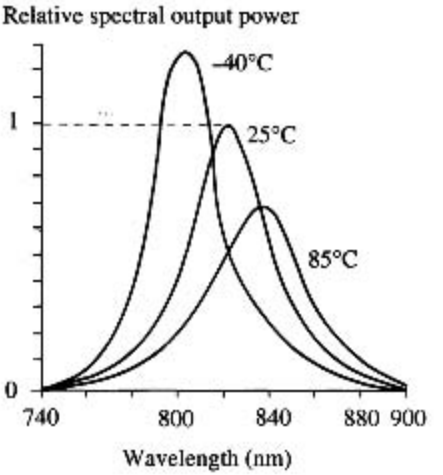


Figure 6.68 The output spectrum from an AlGaAs LED.
Values are normalized to peak emission at 25 °C.

6.21 Solar cell driving a load

a. A Si solar cell of area 2.5 cm × 2.5 cm is connected to drive a load R as in Figure 6.54a. It has the I–V characteristics in Figure 6.53. Suppose that the load is 2 Ω and it is used under a light intensity of 800 W m^{-2}. What are the current and voltage in the circuit? What is the power delivered to the load? What is the efficiency of the solar cell in this circuit?

b. What should the load be to obtain maximum power transfer from the solar cell to the load at 800 W m^{-2} illumination? What is this load at 500 W m^{-2}?

c. Consider using a number of such solar cells to drive a calculator that needs a minimum of 3 V and draws 50 mA at 3–4 V. It is to be used at a light intensity of about 400 W m^{-2}. How many solar cells would you need and how would you connect them?

6.22 Open circuit voltage A solar cell under an illumination of 1000 W m^{-2} has a short circuit current I_{sc} of 50 mA and an open circuit output voltage V_{oc} of 0.65 V. What are the short circuit current and open circuit voltages when the light intensity is halved?

6.23 Maximum power from a solar cell Suppose that the power delivered by a solar cell, $P = IV$, is maximum when $I = I_m$ and $V = V_m$. Suppose that we define normalized voltage and current for maximum power as

Normalized solar cell voltage and current

$$v = \frac{V_m}{\eta V_T} \qquad \text{and} \qquad i = \frac{I_m}{I_{sc}}$$

where η is the ideality factor, $V_T = kT/e$ is called the thermal voltage (0.026 V at 300 K), and $I_{sc} = -I_{ph}$. Suppose that $v_{oc} = V_{oc}/(\eta V_T)$ is the normalized open circuit voltage. Under illumination with the solar cell delivering power with $V > \eta V_T$,

Power delivered by solar cell

$$P = IV = \left[-I_{ph} + I_o \exp\left(\frac{V}{\eta V_T}\right)\right] V$$

One can differentiate $P = IV$ with respect to V, set it to zero for maximum power, and find expressions for I_m and V_m for maximum power. One can then use the open circuit condition ($I = 0$) to relate V_{oc} to I_o. Show that maximum power occurs when

Maximum power delivery

$$v = v_{oc} - \ln(v + 1) \qquad \text{and} \qquad i = 1 - \exp[-(v_{oc} - v)]$$

Consider a solar cell with $\eta = 1.5$, $V_{oc} = 0.60$ V, and $I_{ph} = 35$ mA, with an area of 1 cm^2. Find i and v, and hence the current I_m and voltage V_m for maximum power. (Note: Solve the first equation numerically or graphically to find $v \approx 12.76$.) What is the fill factor?

6.24 **Series resistance** The series resistance causes a voltage drop when a current is drawn from a solar cell. By convention, the positive current is taken to flow into the device. (If calculations yield a negative value, it means that, physically, the current is flowing out, which is the actual case under illumination.) If V is the actual voltage across the solar cell output (accessed by the user), then the voltage across the diode is $V - IR_s$. The solar cell equation becomes

$$I = -I_{ph} + I_d = -I_{ph} + I_o \exp\left(\frac{e(V - IR_s)}{\eta kT}\right)$$

Solar cell with series resistance

Plot I versus V for a Si solar cell that has $\eta = 1.5$ and $I_o = 3 \times 10^{-6}$ mA, for an illumination such that $I_{ph} = 10$ mA for $R_s = 0$, 20 and 50 Ω. What is your conclusion?

6.25 **Shunt resistance** Consider the shunt resistance R_p of a solar cell. Whenever there is a voltage V at the terminals of the solar cell, the shunt resistance draws a current V/R_p. Thus, the total current as seen at the terminals (and flowing in by convention) is

$$I = -I_{ph} + I_d + \frac{V}{R_p} = -I_{ph} + I_o \exp\left(\frac{eV}{\eta kT}\right) + \frac{V}{R_p} = 0$$

Solar cell with shunt resistance

Plot I versus V for a polycrystalline Si solar cell that has $\eta = 1.5$ and $I_o = 3 \times 10^{-6}$ mA, for an illumination such that $I_{ph} = 10$ mA. Use $R_p = \infty$, 1000, 100 Ω. What is your conclusion?

***6.26** **Series connected solar cells** Consider two identical solar cells connected in series. There are two R_s in series and two *pn* junctions in series. If I is the total current through the devices, then the voltage across one *pn* junction is $V_d = \frac{1}{2}[V - I(2R_s)]$ so that the current I flowing into the combined solar cells is

$$I \approx -I_{ph} + I_o \exp\left[\frac{V - I(2R_s)}{2\eta V_T}\right] \qquad V_d > \eta\left(\frac{kT}{e}\right)$$

Two solar cells in series

where $V_T = kT/e$ is the thermal voltage. Rearranging, for two cells in series,

$$V = 2\eta V_T \ln\left(\frac{I + I_{ph}}{I_o}\right) + 2R_s I$$

Two solar cells in series

whereas for one cell,

$$V = \eta V_T \ln\left(\frac{I + I_{ph}}{I_o}\right) + R_s I$$

One solar cell

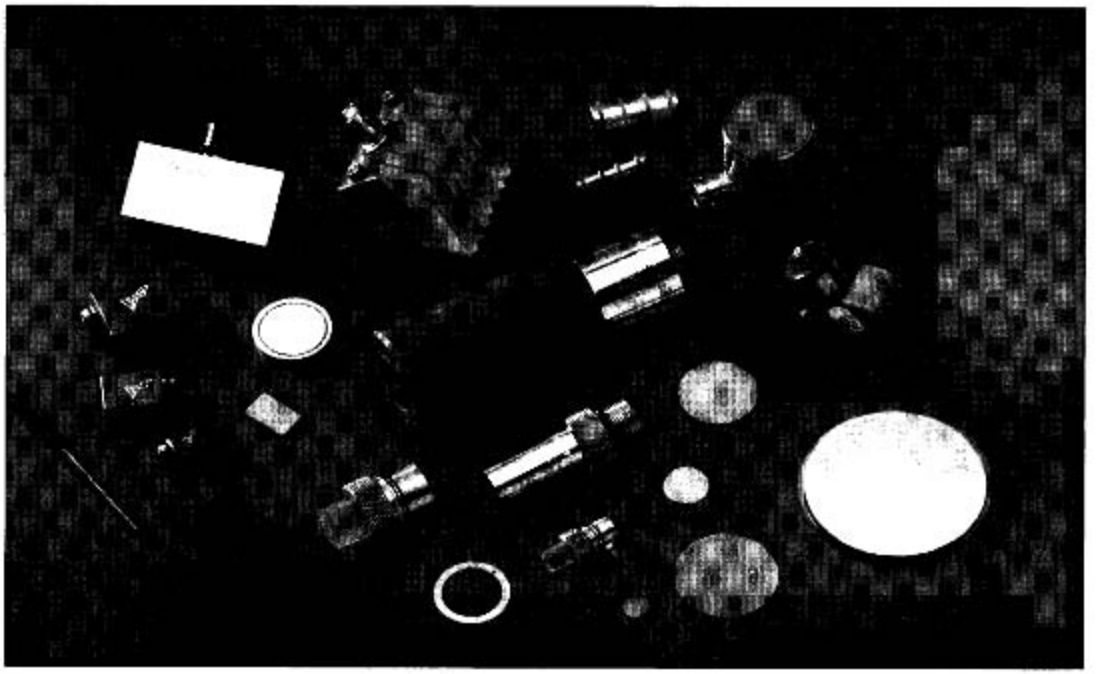
Suppose that the cells have the properties $I_o = 25 \times 10^{-6}$ mA, $\eta = 1.5$, $R_s = 20$ Ω, and both are subjected to the same illumination so that $I_{ph} = 10$ mA. Plot the individual I–V characteristics and the I–V characteristics of the two cells in series. Find the maximum power that can be delivered by one cell and two cells in series. Find the corresponding voltage and current at the maximum power point.

6.27 **A solar cell used in Eskimo Point** The intensity of light arriving at a point on Earth, where the solar latitude is α can be approximated by the Meinel and Meinel equation:

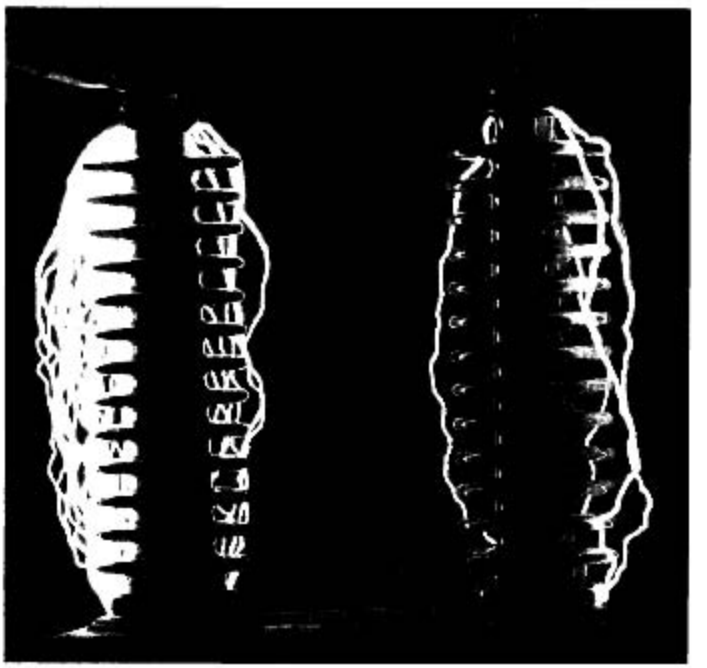
$$\mathcal{I} = 1.353(0.7)^{(\text{cosec}\alpha)^{0.678}} \quad \text{kW m}^{-2}$$

where cosec $\alpha = 1/(\sin\alpha)$. The solar latitude α is the angle between the sun's rays and the horizon. Around September 23 and March 22, the sun's rays arrive parallel to the plane of the equator. What is the maximum power available for a photovoltaic device panel of area 1 m^2 if its efficiency of conversion is 10 percent?

A manufacturer's characterization tests on a particular Si *pn* junction solar cell at 27 °C specifies an open circuit output voltage of 0.45 V and a short circuit current of 400 mA when illuminated directly with a light of intensity 1 kW m^{-2}. The fill factor for the solar cell is 0.73. This solar cell is to be used in a portable equipment application near Eskimo Point (Canada) at a geographical latitude (ϕ) of 63°. Calculate the open circuit output voltage and the maximum available power when the solar cell is used at noon on September 23 when the temperature is around −10 °C. What is the maximum current this solar cell can supply to an electronic equipment? What is your conclusion? (Note: $\alpha + \phi = \pi/2$)



A selection of ultrasonic transducers (piezoelectric effect devices).
SOURCE: Courtesy of Valpey Fisher.



An HV capacitor bushing being subjected to mains-frequency overvoltage. The photo is one of prolonged exposure, recording multiple surface flashovers.
SOURCE: Courtesy of Dr. Simon Rowland, UMIST, England.