Chapter 7
Bipolar Junction Transistors

OBJECTIVES

1. Analyze BJT band diagrams and understand current gain, base transport factor, and emitter injection efficiency
2. Discuss the Ebers Moll coupled diode model, equivalent circuits, cutoff, saturation, and active regions
3. Develop the Gummel–Poon charge control model for second-order effects—early effect, high-level injection
4. Study HBTs and other advanced BJTs

We begin this chapter with a qualitative discussion of charge transport in a *bipolar junction transistor (BJT)*, to establish a sound physical understanding of its operation. Then we shall investigate carefully the charge distributions in the transistor and relate the three terminal currents to the physical characteristics of the device. Our aim is to gain a solid understanding of the current flow and control of the transistor and to discover the most important secondary effects that influence its operation. We shall discuss the properties of the transistor with proper biasing for amplification and then consider the effects of more general biasing, as encountered in switching circuits.

In this chapter we shall use the p-n-p transistor for most illustrations. The main advantage of the p-n-p for discussing transistor action is that hole flow and current are in the same direction. This makes the various mechanisms of charge transport somewhat easier to visualize in a preliminary explanation. Once these basic ideas are established for the p-n-p device, it is simple to relate them to the more widely used transistor, the n-p-n.

7.1 FUNDAMENTALS OF BJT OPERATION

The bipolar transistor is basically a simple device, and this section is devoted to a simple and largely qualitative view of BJT operation. We will deal with the details of these transistors in following sections, but first we must define some terms and gain physical understanding of how carriers are transported through the device. Then we can discuss how the current through two terminals can be controlled by small changes in the current at a third terminal.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 336

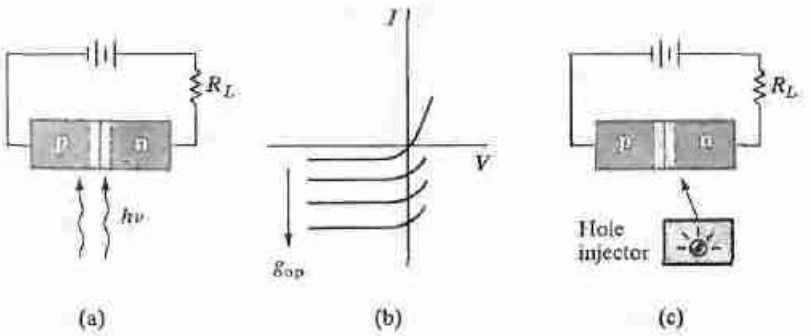


Figure 7–1
External control of the current in a reverse-biased p-n junction: (a) optical generation; (b) junction I–V characteristics as a function of EHP generation; (c) minority carrier injection by a hypothetical device.

Let us begin the discussion of bipolar transistors by considering the reverse-biased p-n junction diode of Fig. 7–1. According to the theory of Chapter 5, the reverse saturation current through this diode depends on the rate at which minority carriers are generated in the neighborhood of the junction. We found, for example, that the reverse current due to holes being swept from n to p is essentially independent of the size of the junction's field and hence is independent of the reverse bias. The reason given was that the hole current depends on how often minority holes are generated by EHP creation within a diffusion length of the junction—not upon how fast a particular hole is swept across the depletion layer by the field. As a result, it is possible to increase the reverse current through the diode by increasing the rate of EHP generation (Fig. 7–1b). One convenient method for accomplishing this is optical excitation of EHPs with light ($h\nu > E_g$), as in Section 4.3. With steady photoexcitation the reverse current will still be essentially independent of bias voltage, and if the dark saturation current is negligible, the reverse current is directly proportional to the optical generation rate g_{op}.

The example of external control of current through a junction by optical generation raises an interesting question: Is it possible to inject minority carriers in to the neighborhood of the junction *electrically* instead of optically? If so, we could control the junction reverse current simply by varying the rate of minority carrier injection. For example, let us consider a hypothetical *hole injection device* as in Fig. 7–1c. If we can inject holes at a predetermined rate into the n side of the junction, the effect on the junction current will resemble the effects of optical generation. The current from n to p will depend on the hole injection rate and will be essentially independent of the bias voltage. There are several obvious advantages to such external control of a current; for example, the current through the reverse-biased junction would vary very little if the load resistor R_L were changed, since the magnitude of the junction voltage is relatively unimportant. Therefore, such an arrangement should be a good approximation to a controllable constant current source.

A convenient hole injection device is a forward-biased p$^+$-n junction. According to Section 5.3.2, the current in such a junction is due primarily to holes injected from the p$^+$ region into the n material. If we make the n side

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 337

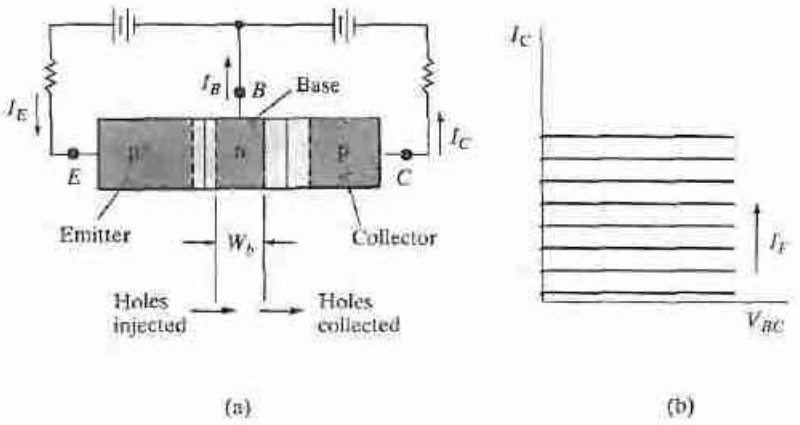


Figure 7–2
A p-n-p transistor: (a) schematic representation of a p-n-p device with a forward-biased emitter junction and a reverse-biased collector junction; (b) *I–V* characteristics of the reverse-biased n-p junction as a function of emitter current.

of the forward-biased junction the same as the n side of the reverse-biased junction, the p^+-n-p structure of Fig. 7–2 results. With this configuration, injection of holes from the p^+-n junction into the center n region supplies the minority carrier holes to participate in the reverse current through the n-p junction. Of course, it is important that the injected holes do not recombine in the n region before they can diffuse to the depletion layer of the reverse-biased junction. Thus we must make the n region narrow compared with a hole diffusion length.

The structure we have described is a p-n-p bipolar junction transistor. The forward-biased junction which injects holes into the center n region is called the *emitter junction*, and the reverse-biased junction which collects the injected holes is called the *collector junction*. The p^+ region, which serves as the source of injected holes, is called the *emitter*, and the p region into which the holes are swept by the reverse-biased junction is called the *collector*. The center n region is called the *base*, for reasons which will become clear in Section 7.3, when we discuss the historical development of transistor fabrication. The biasing arrangement in Fig. 7–2 is called the *common base* configuration, since the base electrode B is common to the emitter and collector circuits.

To have a good p-n-p transistor, we would prefer that almost all the holes injected by the emitter into the base be collected. Thus the n-type base region should be narrow, and the hole lifetime τ_p should be long. This requirement is summed up by specifying $W_b \ll L_p$, where W_b is the length of the *neutral* n material of the base (measured between the depletion regions of the emitter and collector junctions), and L_p is the diffusion length for holes in the base $(D_p\tau_p)^{1/2}$. With this requirement satisfied, an average hole injected at the emitter junction will diffuse to the depletion region of the collector junction without recombination in the base. A second requirement is that the current I_E crossing the emitter junction should be composed almost entirely of holes injected into the base, rather than electrons crossing from base to emitter. This requirement is satisfied by doping the base region lightly compared with the emitter, so that the p^+-n emitter junction of Fig. 7–2 results.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 338

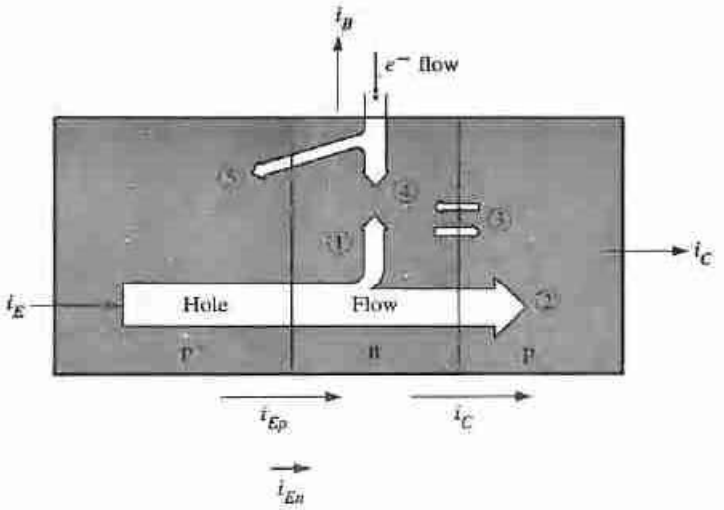


Figure 7–3
Summary of hole and electron flow in a p-n-p transistor with proper biasing: (1) injected holes lost to recombination in the base; (2) holes reaching the reverse-biased collector junction; (3) thermally generated electrons and holes making up the reverse saturation current of the collector junction; (4) electrons supplied by the base contact for recombination with holes; (5) electrons injected across the forward-biased emitter junction.

It is clear that current I_E flows into the emitter of a properly biased p-n-p transistor and that I_C flows out at the collector, since the direction of hole flow is from emitter to collector. However, the base current I_B requires a bit more thought. In a good transistor the base current will be very small since I_E is essentially hole current, and the collected hole current I_C is almost equal to I_E. There must be some base current, however, due to requirements of electron flow into the n-type base region (Fig. 7–3). We can account for I_B physically by three dominant mechanisms:

(a) There must be some recombination of injected holes with electrons in the base, even with $W_b \ll L_p$. The electrons lost to recombination must be resupplied through the base contact.

(b) Some electrons will be injected from n to p in the forward-biased emitter junction, even if the emitter is heavily doped compared to the base. These electrons must also be supplied by I_B.

(c) Some electrons are swept into the base at the reverse-biased collector junction due to thermal generation in the collector. This small current reduces I_B by supplying electrons to the base.

The dominant sources of base current are (a) recombination in the base and (b) injection into the emitter region. Both of these effects can be greatly reduced by device design, as we shall see. In a well-designed transistor, I_B will be a very small fraction (perhaps one-hundredth) of I_E.

In an n-p-n transistor the three current directions are reversed, since electrons flow from emitter to collector and holes must be supplied to the base. The physical mechanisms for operation of the n-p-n can be understood simply by reversing the roles of electrons and holes in the p-n-p discussion.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 339

7.2 AMPLIFICATION WITH BJTS

In this section we shall discuss rather simply the various factors involved in transistor amplification. Basically, the transistor is useful in amplifiers because the currents at the emitter and collector are controllable by the relatively small base current. The essential mechanisms are easy to understand if various secondary effects are neglected. We shall use total current (d-c plus a-c) in this discussion, with the understanding that the simple analysis applies only to d-c and to small-signal a-c at low frequencies. We can relate the terminal currents of the transistor i_E, i_B, and i_C by several important factors. In this introduction we shall neglect the saturation current at the collector (Fig. 7-3, component 3) and such effects as recombination in the transition regions. Under these assumptions, the collector current is made up entirely of those holes injected at the emitter which are not lost to recombination in the base. Thus i_C is proportional to the hole component of the emitter current i_{Ep}:

$$i_C = Bi_{Ep} \tag{7-1}$$

The proportionality factor B is simply the fraction of injected holes which make it across the base to the collector; B is called the *base transport factor*. The total emitter current i_E is made up of the hole component i_{Ep} and the electron component i_{En}, due to electrons injected from base to emitter (component 5 in Fig. 7-3). The *emitter injection efficiency* γ is

$$\boxed{\gamma = \frac{i_{Ep}}{i_{En} + i_{Ep}}} \tag{7-2}$$

For an efficient transistor we would like B and γ to be very near unity; that is, the emitter current should be due mostly to holes ($\gamma \approx 1$), and most of the injected holes should eventually participate in the collector current ($B \approx 1$). The relation between the collector and emitter currents is

$$\frac{i_C}{i_E} = \frac{Bi_{Ep}}{i_{En} + i_{Ep}} = B\gamma \equiv \alpha \tag{7-3}$$

The product $B\gamma$ is defined as the factor α, called the *current transfer ratio*, which represents the emitter-to-collector current amplification. There is no real amplification between these currents, since α is smaller than unity. On the other hand, the relation between i_C and i_B is more promising for amplification.

In accounting for the base current, we must include the rates at which electrons are lost from the base by injection across the emitter junction (i_{En}) and the rate of electron recombination with holes in the base. In each case, the lost electrons must be resupplied through the base current i_B. If the fraction of injected holes making it across the base *without* recombination is B, then it follows that $(1 - B)$ is the fraction *recombining* in the base. Thus the base current is

$$i_B = i_{En} + (1 - B)i_{Ep} \tag{7-4}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 340

neglecting the collector saturation current. The relation between the collector and base currents is found from Eqs. (7-1) and (7-4):

$$\frac{i_C}{i_B} = \frac{B i_{Ep}}{i_{En} + (1 - B) i_{Ep}} = \frac{B[i_{Ep}/(i_{En} + i_{Ep})]}{1 - B[i_{Ep}/(i_{En} + i_{Ep})]} \tag{7-5}$$

$$\boxed{\frac{i_C}{i_B} = \frac{B\gamma}{1 - B\gamma} = \frac{\alpha}{1 - \alpha} = \beta} \tag{7-6}$$

The factor β relating the collector current to the base current is the *base-to-collector current amplification factor.*[1] Since α is near unity, it is clear that β can be large for a good transistor, and the collector current is large compared with the base current.

It remains to be shown that the collector current i_C can be controlled by variations in the small current i_B. In the discussion to this point, we have indicated control of i_C by the emitter current i_E, with the base current characterized as a small side effect. In fact, we can show from space charge neutrality arguments that i_B can indeed be used to determine the magnitude of i_C. Let us consider the transistor of Fig. 7-4, in which i_B is determined by a biasing circuit. For simplicity, we shall assume unity emitter injection efficiency and negligible collector saturation current. Since the n-type base region is electrostatically neutral between the two transition regions, the presence of excess holes in transit from emitter to collector calls for compensating excess

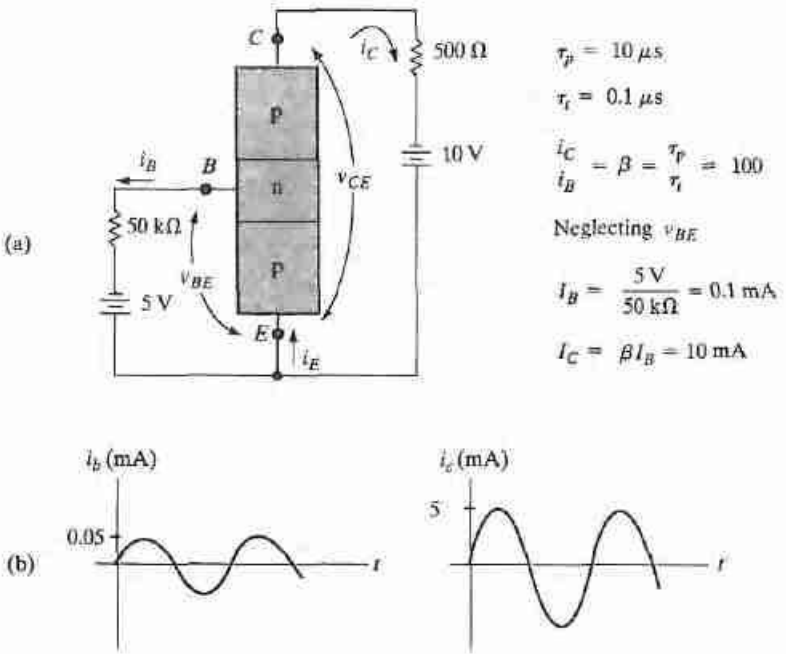


Figure 7-4
Example of amplification in a common-emitter transistor circuit: (a) biasing circuit; (b) addition of an a-c variation of base current i_b to the d-c value of I_B, resulting in an a-c component i_c.

[1] α is also called the *common-base current gain*; β is also called the *common-emitter current gain*.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 341

electrons from the base contact. However, there is an important difference in the times which electrons and holes spend in the base. The average excess hole spends a time τ_t, defined as the *transit time* from emitter to collector. Since the base width W_b is made small compared with L_p, this transit time is much less than the average hole lifetime τ_p in the base.[2] On the other hand, an average excess electron supplied from the base contact spends τ_p seconds in the base supplying space charge neutrality during the lifetime of an average excess hole. While the average electron waits τ_p seconds for recombination, many individual holes can enter and leave the base region, each with an average transit time τ_t. In particular, for each electron entering from the base contact, τ_p/τ_t holes can pass from emitter to collector while maintaining space charge neutrality. Thus the ratio of collector current to base current is simply

$$\frac{i_C}{i_B} = \beta = \frac{\tau_p}{\tau_t} \tag{7.7}$$

for $\gamma = 1$ and negligible collector saturation current.

If the electron supply to the base (i_B) is restricted, the traffic of holes from emitter to base is correspondingly reduced. This can be argued simply by supposing that the hole injection does continue despite the restriction on electrons from the base contact. The result would be a net buildup of positive charge in the base and a loss of forward bias (and therefore a loss of hole injection) at the emitter junction. Clearly, the supply of electrons through i_B can be used to raise or lower the hole flow from emitter to collector.

The base current is controlled independently in Fig. 7–4. This is called a *common-emitter* circuit, since the emitter electrode is common to the base and collector circuits. The emitter junction is clearly forward biased by the battery in the base circuit. The voltage drop in the forward-biased emitter junction is small, however, so that almost all of the voltage from collector to emitter appears across the reverse-biased collector junction. Since v_{BE} is small for the forward-biased junction, we can neglect it and approximate the base current as 5 V/50 kΩ = 0.1 mA. If $\tau_p = 10$ μs and $\tau_t = 0.1$ μs, β for the transistor is 100 and the collector current I_C is 10 mA. It is important to note that i_t is determined by β and the base current, rather than by the battery and resistor in the collector circuit (as long as these are of reasonable values to maintain a reverse-biased collector junction). In this example 5 V of the collector circuit battery voltage appears across the 500 Ω resistor, and 5 V serves to reverse bias the collector junction.

[2]This difference between average hole lifetime before recombination (τ_p) and the average time a hole spends in transit across the base (τ_t) may be confusing at first. How can the lifetime be longer than the time a hole actually spends in transit? The answer depends on the fact that holes are indistinguishable in the recombination kinetics. Think of an analogy with a shooting gallery, in which a good marksman fires slowly at a line of quickly moving ducks. Although many individual ducks make it across the firing line without being hit, the lifetime of an *average* duck within the firing line is determined by the time between shots. We can speak of the lifetime of an average duck because they are essentially indistinguishable. Similarly, the rate of recombination in the base (and therefore i_B) depends on the average lifetime τ_p and the distribution of the indistinguishable holes in the base region.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 342

If a small a-c current i_b is superimposed on the steady state base current of Fig. 7–4a, a corresponding a-c current i_c appears in the collector circuit. The time-varying portion of the collector current will be i_b multiplied by the factor β, and current gain results.

We have neglected a number of important properties of the transistor in this introductory discussion, and many of these properties will be treated in detail below. We have established, however, the fundamental basis of operation for the bipolar transistor and have indicated in a simplified way how it can be used to produce current gain in an electronic circuit.

EXAMPLE 7–1

(a) Show that Eq. (7-7) is valid from arguments of the steady state replacement of stored charge. Assume that $\tau_n = \tau_p$.

(b) What is the steady state charge $Q_n = Q_p$ due to excess electrons and holes in the neutral base region for the transistor of Fig. 7–4?

SOLUTION

(a) In steady state there are excess electrons and holes in the base. The charge in the electron distribution Q_n is replaced every τ_n seconds. Thus $i_B = Q_n/\tau_n$. The charge in the hole distribution Q_p is collected every τ_t seconds, and $i_C = Q_p/\tau_t$. For space charge neutrality, $Q_n = Q_p$, and

$$\frac{i_C}{i_B} = \frac{Q_n/\tau_t}{Q_n/\tau_p} = \frac{\tau_p}{\tau_t}$$

(b) $Q_n = Q_p = i_C\tau_t = i_B\tau_p = 10^{-9}$ C.

7.3 BJT FABRICATION

The first transistor invented by Bardeen and Brattain in 1947 was the *point contact* transistor. In this device two sharp metal wires, or "cat's whiskers," formed an "emitter" of carriers and a "collector" of carriers. These wires were simply pressed onto a slab of Ge which provided a "base" or mechanical support, through which the injected carriers flowed. This basic invention rapidly led to the BJT, in which charge injection and collection was achieved using two p-n junctions in close proximity to each other. The p-n junctions in BJTs can be formed in a variety of ways using thermal diffusion, but modern devices are generally made using ion implantation (Section 5.1.4).

Let us review a simplified version of how to make a double polysilicon, self-aligned n-p-n Si BJT. This is the most commonly used, state-of-the-art technique for making BJTs for use in an IC. Use of n-p-n transistors is more popular than p-n-p devices because of the higher mobility of electrons compared to holes. The process steps are shown in cross-sectional view in Fig. 7–5. A p-type Si substrate is oxidized, windows are defined using photolithography and etched in the oxide. Using the photoresist and oxide as an implant mask, a donor with very small diffusivity in Si, such as As or Sb, is implanted into the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 343

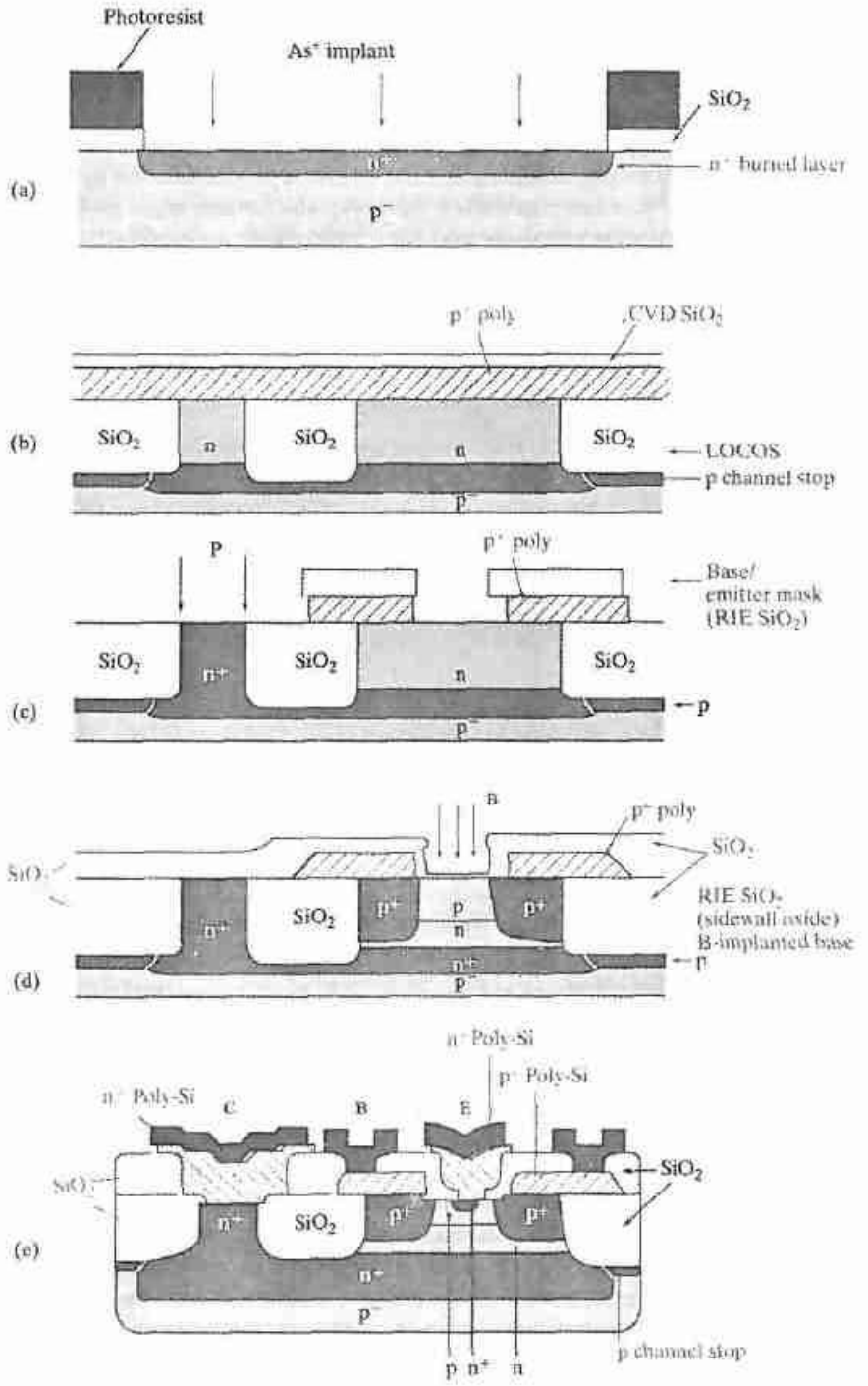


Figure 7–5 Process flow for double polysilicon, self-aligned n-p-n BJT: (a) n^+ buried layer formation; (b) n epitaxy followed by LOCOS isolation; (c) base-emitter window definition and (optional) masked "sinker" implant (P) into collector contact region; (d) intrinsic base implant using self-aligned oxide sidewall spacers; (e) self-aligned formation of n^+ emitter, as well as n^+ collector contact.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 344

open window to form a highly conductive n^+ layer (Fig. 7–5a). Subsequently, the photoresist and the oxide are removed, and a lightly doped n-type epitaxial layer is grown. During this high temperature growth, the implanted n^+ layer diffuses only slightly towards the surface and becomes a conductive *buried collector* (also called a *sub-collector*). The n^+ sub-collector layer guarantees a low collector series resistance when it is connected subsequently to the collector ohmic contact, sometimes through the use of an optional, masked deep n^+ "sinker" implant or diffusion only in the collector contact region (Fig. 7–5c). The lightly doped n-type collector region above the n^+ sub-collector in the part of the BJT where the base and emitter are formed ensures a high base-collector reverse breakdown voltage. (It turns out that wherever the sub-collector is formed, and subsequently the epitaxial layer is grown on top, there is a notch or step in the substrate surface. This notch is not explicitly shown in Fig. 7–5a. This notch is very useful as a marker of the location of the sub-collectors because subsequently, we have to align the LOCOS isolation mask with respect to the sub-collector.)

For integrated circuits involving not just discrete BJTs, but many interconnected transistors, there are issues involving electrical isolation of adjacent BJTs in order to ensure that there is no electrical cross-talk between them. As described in Section 6.4.1, such isolation can be achieved by LOCOS to form field or isolation oxides after a B channel stop implant (Fig. 7–5b). Another isolation scheme that is particularly well suited for high-density bipolar circuits involves the formation of shallow trenches by RIE, backfilled with oxide and polysilicon (Section 9.3.1). In this process a nitride layer is patterned and used as an etch mask for an anisotropic etch of the silicon to form the trench. Using reactive ion etching, a narrow trench about 1 μm deep can be formed with very straight sidewalls. Oxidation inside the trench forms an insulating layer, and the trench is then filled with oxide by Low-Pressure Chemical Vapor Deposition (LPCVD).

A polysilicon layer is deposited by LPCVD, and doped heavily p^+ with B either during deposition or subsequently by ion implantation. An oxide layer is deposited next by LPCVD. Using photolithography with the base/emitter mask, a window is etched in the polysilicon/oxide stack by RIE (Fig. 7–5c). A heavily doped "extrinsic" p^+ base is formed by diffusion of B from the doped polysilicon layer into the substrate in order to provide a low-resistance, high-speed base ohmic contact. An oxide layer is then deposited by LPCVD, which has the effect of closing up the base window that was etched previously, and B is implanted into this window (Fig. 7–5d). This base implant forms a more lightly p doped "intrinsic" base through which most of the current flows from the emitter to the collector. The more heavily doped extrinsic base forms a collar around the intrinsic base, and serves to reduce the base series resistance. It is critical that the base be enclosed well within the collector because otherwise it would be shorted to the p^- substrate. Finally, another LPCVD oxide layer is deposited to close up the base window further, and the oxide is etched all the way to the Si substrate by RIE, leaving oxide *spacers* on the sidewalls. Heavily n^+ doped (typically with As) polysilicon is then deposited on the substrate, patterned and etched

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 345

forming *polysilicon emitter (polyemitter)* and collector contacts, as shown in Fig. 7-5e. (The use of two LPCVD polysilicon layers explains why this process is referred to as the double-polysilicon process.) Arsenic from the polysilicon is diffused into the substrate to form the n^+ emitter region nested within the base in a *self-aligned* manner, as well as the n^+ collector contact. Self alignment refers to the fact that a separate lithography step is not required to form the n^+ emitter region. We cleverly made use of the oxide sidewall spacers to ensure that the n^+ emitter region lies within the intrinsic p-type base. This is critical because otherwise the emitter gets shorted to the collector; we also want a gap between the n^+ emitter and the p^+ extrinsic base, because otherwise the emitter-base junction capacitance becomes too high. In the vertical direction, the difference between the emitter-base junction and the base-collector junction determines the base width. This is made very narrow in high gain, high speed BJTs.

Finally, an oxide layer is deposited by CVD, windows are etched in it corresponding to the emitter (E), base (B), and collector (C) contacts, and a suitable contact metal such as Al is sputter deposited to form the ohmic contacts. The Al is patterned photolithographically using the interconnect mask, and etched using RIE. The many ICs that are made simultaneously on the wafer are then separated into individual dies by sawing, mounted on suitable packages, and the various contacts are wire bonded to the external leads of the package.

7.4 MINORITY CARRIER DISTRIBUTIONS AND TERMINAL CURRENTS

In this section we examine the operation of a BJT in more detail. We begin our analysis by applying the techniques of previous chapters to the problem of hole injection into a narrow n-type base region. The mathematics is very similar to that used in the problem of the narrow base diode (Prob. 5.35). Basically, we assume holes are injected into the base at the forward-biased emitter, and these holes diffuse to the collector junction. The first step is to solve for the excess hole distribution in the base, and the second step is to evaluate the emitter and collector currents (I_E, I_C) from the gradient of the hole distribution on each side of the base. Then the base current (I_B) can be found from a current summation or from a charge control analysis of recombination in the base.

We shall at first simplify the calculations by making several assumptions:

1. Holes diffuse from emitter to collector; drift is negligible in the base region.
2. The emitter current is made up entirely of holes; the emitter injection efficiency is $\gamma = 1$.
3. The collector saturation current is negligible.
4. The active part of the base and the two junctions are of uniform cross-sectional area A; current flow in the base is essentially one-dimensional from emitter to collector.
5. All currents and voltages are steady state.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 346

In later sections we shall consider the implications of imperfect injection efficiency, drift due to nonuniform doping in the base, structural effects such as different areas for the emitter and collector junctions, and capacitance and transit time effects in a-c operation.

7.4.1 Solution of the Diffusion Equation in the Base Region

Since the injected holes are assumed to flow from emitter to collector by diffusion, we can evaluate the currents crossing the two junctions by techniques used in Chapter 5. Neglecting recombination in the two depletion regions, the hole current entering the base at the emitter junction is the current I_E, and the hole current leaving the base at the collector is I_C. If we can solve for the distribution of excess holes in the base region, it is simple to evaluate the gradient of the distribution at the two ends of the base to find the currents. We shall consider the simplified geometry of Fig. 7-6a, in which the base width is W_b between the two depletion regions, and the uniform cross-sectional area is A. In equilibrium, the Fermi level is flat, and the band diagram corresponds to that for two back-to-back p-n junctions. But, for a forward-biased emitter and a reverse-biased collector (normal active mode), the Fermi level splits up into quasi-Fermi levels, as shown in Fig. 7-6b. The barrier at the emitter-base junction is reduced by the forward bias, and that at the collector-base junction is increased by the reverse bias. The excess hole concentration at the edge of the emitter depletion region Δp_E and the corresponding concentration on the collector side of the base Δp_C are found from Eq. (5-29):

$$\Delta p_E = p_n(e^{qV_{EB}/kT} - 1) \tag{7-8a}$$

$$\Delta p_C = p_n(e^{qV_{CB}/kT} - 1) \tag{7-8b}$$

If the emitter junction is strongly forward biased ($V_{EB} \gg kT/q$) and the collector junction is strongly reverse biased ($V_{CB} \ll 0$), these excess concentrations simplify to

$$\Delta p_E \simeq p_n e^{qV_{EB}/kT} \tag{7-9a}$$

$$\Delta p_C \simeq -p_n \tag{7-9b}$$

We can solve for the excess hole concentration as a function of distance in the base $\delta p(x_n)$ by using the proper boundary conditions in the diffusion equation, Eq. (4-34b):

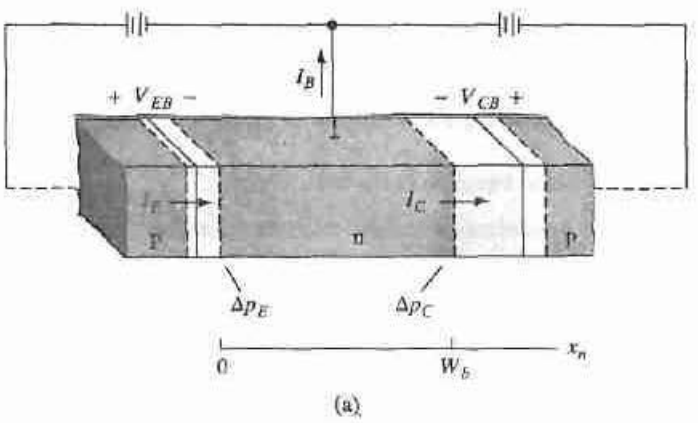
$$\frac{d^2\delta p(x_n)}{dx_n^2} = \frac{\delta p(x_n)}{L_p^2} \tag{7-10}$$

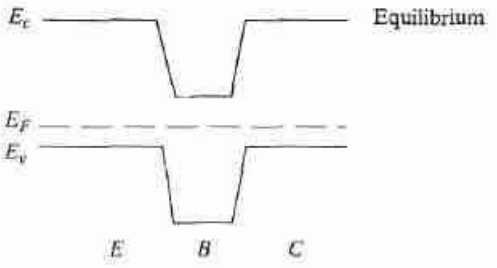
The solution of this equation is

$$\delta p(x_n) = C_1 e^{x_n/L_p} + C_2 e^{-x_n/L_p} \tag{7-11}$$

where L_p is the diffusion length of holes in the base region. Unlike the simple problem of injection into a long n region, we cannot eliminate one of the constants by assuming the excess holes disappear for large x_n. In fact, since $W_b \ll L_p$ in a properly designed transistor, most of the injected holes reach

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 347





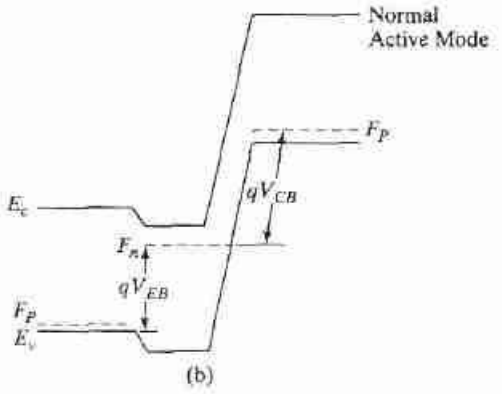


Figure 7–6
(a) Simplified p-n-p transistor geometry used in the calculations; (b) a p-n-p transistor at equilibrium (flat Fermi level) and with normal active bias. The quasi-Fermi levels are separated by the applied voltage times q.

the collector at W_b. The solution is very similar to that of the narrow base diode problem. In this case the appropriate boundary conditions are

$$\delta p(x_n = 0) = C_1 + C_2 = \Delta p_E \tag{7-12a}$$

$$\delta p(x_n = W_b) = C_1 e^{W_b/L_P} + C_2 e^{-W_b/L_P} = \Delta p_C \tag{7-12b}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 348

Solving for the parameters C_1 and C_2 we obtain

$$C_1 = \frac{\Delta p_C - \Delta p_E e^{-W_b/L_p}}{e^{W_b/L_p} - e^{-W_b/L_p}} \tag{7-13a}$$

$$C_2 = \frac{\Delta p_E e^{W_b/L_p} - \Delta p_C}{e^{W_b/L_p} - e^{-W_b/L_p}} \tag{7-13b}$$

These parameters applied to Eq. (7-11) give the full expression for the excess hole distribution in the base region. For example, if we assume that the collector junction is strongly reverse biased [Eq. (7-9b)] and the equilibrium hole concentration p_n is negligible compared with the injected concentration Δp_E, the excess hole distribution simplifies to

$$\delta p(x_n) = \Delta p_E \frac{e^{W_b/L_p} e^{-x_n/L_p} - e^{-W_b/L_p} e^{x_n/L_p}}{e^{W_b/L_p} - e^{-W_b/L_p}} \quad (\text{for } \Delta p_C \simeq 0) \tag{7-14}$$

The various terms in Eq. (7-14) are sketched in Fig. 7-7a, and the corresponding excess hole distribution in the base region is demonstrated for a moderate value of W_b/L_p. Note that $\delta p(x_n)$ varies almost linearly between the emitter and collector junction depletion regions. As we shall see, the slight deviation from linearity of the distribution indicates the small value of I_B caused by recombination in the base region.

The minority carrier electron concentrations in the emitter and collector corresponding to a forward-biased emitter and reverse-biased collector are shown in Fig. 7-7b. Here, the excess electron concentration in the p^+ emitter is shown to decay exponentially to zero, corresponding to a *long* diode. This is because, at high emitter doping levels, the minority carrier electron diffusion length is often shorter than the thin emitter region. Otherwise, the narrow diode expressions must be used in the emitter region. For polyemitter structures, the carrier concentration profile is more complicated, as discussed later. The exact form of i_{En} and the emitter injection efficiency will therefore depend on these details.

7.4.2 Evaluation of the Terminal Currents

Having solved for the excess hole distribution in the base region, we can evaluate the emitter and collector currents from the gradient of the hole concentration at each depletion region edge. From Eq. (4-22b) we have

$$I_p(x_n) = -qAD_p \frac{d\delta p(x_n)}{dx_n} \tag{7-15}$$

This expression evaluated at $x_n = 0$ gives the hole component of the emitter current,

$$I_{Ep} = I_p(x_n = 0) = qA\frac{D_p}{L_p}(C_2 - C_1) \tag{7-16}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 349

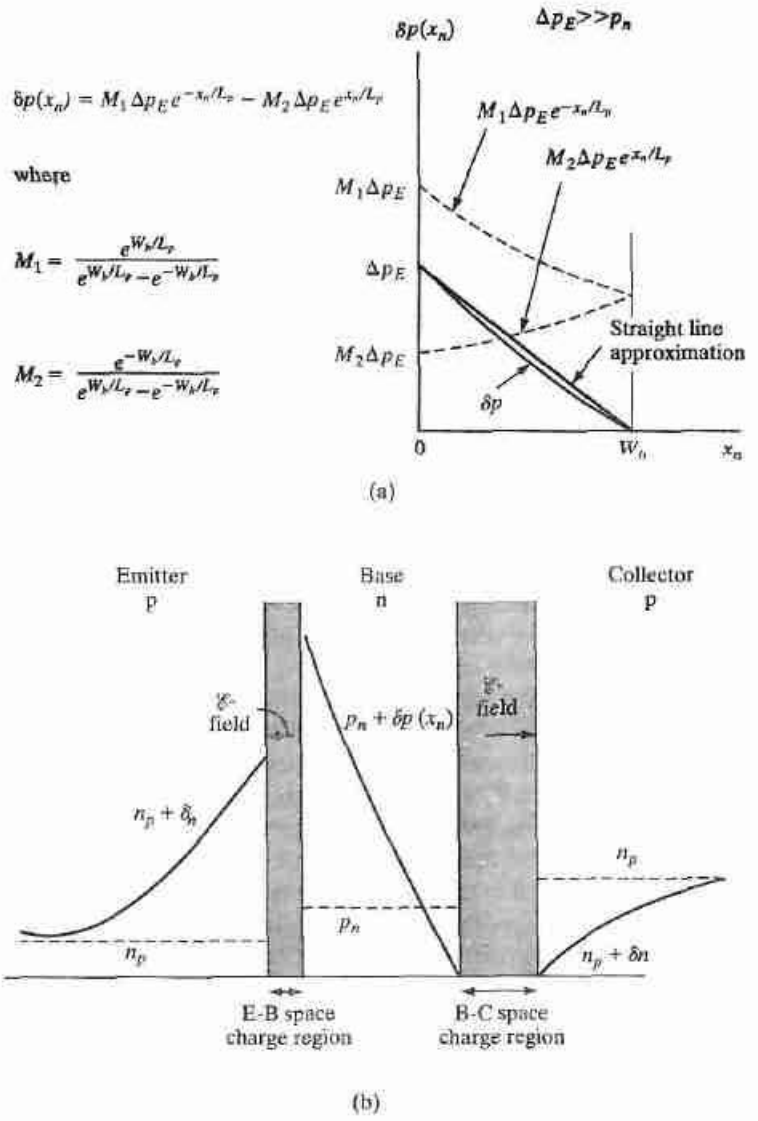


Figure 7–7
(a) Sketch of the terms in Eq. (7–14), illustrating the linearity of the hole distribution in the base region. In this example, $W_b/L_p = 1/2$. (b) Electron distributions in the emitter and collector.

Similarly, if we neglect the electrons crossing from collector to base in the collector reverse saturation current, I_C is made up entirely of holes entering the collector depletion region from the base. Evaluating Eq. (7–15) at $x_n = W_b$ we have the collector current

$$I_C = I_p(x_n = W_b) = qA\frac{D_p}{L_p}(C_2 e^{-W_b/L_p} - C_1 e^{W_b/L_p}) \qquad (7\text{–}17)$$

When the parameters C_1 and C_2 are substituted from Eqs. (7–13), the emitter and collector currents take a form that is most easily written in terms of

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 350

hyperbolic functions:

$$I_{Ep} = qA\frac{D_p}{L_p}\left[\frac{\Delta p_E(e^{W_b/L_p} + e^{-W_b/L_p}) - 2\Delta p_C}{e^{W_b/L_p} - e^{-W_b/L_p}}\right]$$

$$\boxed{I_{Ep} = qA\frac{D_p}{L_p}\left(\Delta p_E \operatorname{ctnh}\frac{W_b}{L_p} - \Delta p_C \operatorname{csch}\frac{W_b}{L_p}\right)} \quad (7\text{-}18a)$$

$$\boxed{I_C = qA\frac{D_p}{L_p}\left(\Delta p_E \operatorname{csch}\frac{W_b}{L_p} - \Delta p_C \operatorname{ctnh}\frac{W_b}{L_p}\right)} \quad (7\text{-}18b)$$

Now we can obtain the value of I_B by a current summation, noting that the sum of the base and collector currents leaving the device must equal the emitter current entering. If $I_E = I_{Ep}$ for $\gamma = 1$,

$$I_B = I_E - I_C = qA\frac{D_p}{L_p}\left[(\Delta p_E + \Delta p_C)\left(\operatorname{ctnh}\frac{W_b}{L_p} - \operatorname{csch}\frac{W_b}{L_p}\right)\right]$$

$$\boxed{I_B = qA\frac{D_p}{L_p}\left[(\Delta p_E + \Delta p_C)\tanh\frac{W_b}{2L_p}\right]} \quad (7\text{-}19)$$

By using the techniques of Chapter 5 we have evaluated the three terminal currents of the transistor in terms of the material parameters, the base width, and the excess concentrations Δp_E and Δp_C. Furthermore, since these excess concentrations are related in a straightforward way to the emitter and collector junction bias voltages by Eq. (7-8), it should be simple to evaluate the transistor performance under various biasing conditions. It is important to note here that Eqs. (7-18) and (7-19) are not restricted to the case of the usual transistor biasing. For example, Δp_c may be $-p_n$ for a strongly reverse-biased collector, or it may be a significant positive number if the collector is positively biased. The generality of these equations will be used in Section 7.5 in considering the application of transistors to switching circuits.

EXAMPLE 7-2

(a) Find the expression for the current I for the transistor connection shown if $\gamma = 1$.

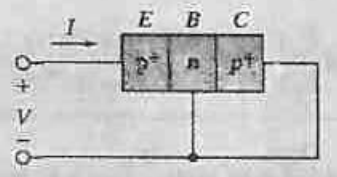
(b) How does the current I divide between the base lead and the collector lead?

SOLUTION

(a) Since $V_{CB} = 0$, Eq. (7-8b) gives $\Delta p_C = 0$. Thus from Eq. (7-18a),

$$I_E = I = \frac{qAD_p}{L_p}\Delta p_E \operatorname{ctnh}\frac{W_b}{L_p}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 351



(b) $$I_C = \frac{qAD_p}{L_p}\,\Delta p_E \operatorname{csch}\frac{W_b}{L_p}$$

$$I_B = \frac{qAD_p}{L_p}\,\Delta p_E \tanh\frac{W_b}{2L_p}$$

where I_C and I_B are the components in the collector lead and base lead, respectively. Note that these results are analogous to those of Probs. 5.35 and 5.36 for the narrow base diode.

7.4.3 Approximations of the Terminal Currents

The general equations of the previous section can be simplified for the case of normal transistor biasing, and such simplification allows us to gain insight into the current flow. For example, if the collector is reverse biased, $\Delta p_C = -p_n$ from Eq. (7-9b). Furthermore, if the equilibrium hole concentration p_n is small (Fig. 7-8a), we can neglect the terms involving Δp_C. For $\gamma \simeq 1$, the terminal currents reduce to those of Example 7-2:

$$I_E \simeq qA\frac{D_p}{L_p}\,\Delta p_E \operatorname{ctnh}\frac{W_b}{L_p} \tag{7-20a}$$

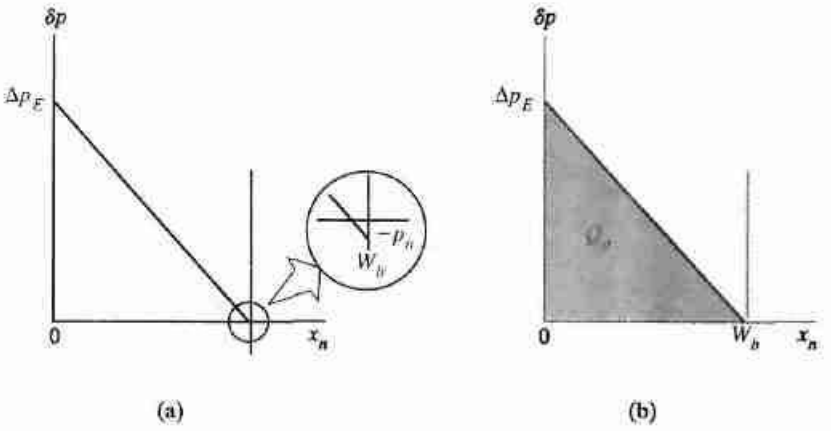


Figure 7–8 Approximate excess hole distributions in the base: (a) forward-biased emitter, reverse-biased collector; (b) triangular distribution for $V_{CB} = 0$ or for negligible p_n.

$$I_C \simeq qA\frac{D_p}{L_p}\Delta p_E \operatorname{csch}\frac{W_b}{L_p} \tag{7-20b}$$

$$I_B \simeq qA\frac{D_p}{L_p}\Delta p_E \tanh\frac{W_b}{2L_p} \tag{7-20c}$$

Series expansions of the hyperbolic functions are given in Table 7-1. For small values of W_b/L_p, we can neglect terms above the first order of the argument. It is clear from this table and Eq. (7-20) that I_C is only slightly smaller than I_E, as expected. The first-order approximation of tanh y is simply y, so that the base current is

$$I_B \simeq qA\frac{D_p}{L_p}\Delta p_E\frac{W_b}{2L_p} = \frac{qAW_b\Delta p_E}{2\tau_p} \tag{7-21}$$

The same approximate expression for the base current is found from the difference in the first-order approximations to I_E and I_C:

$$\begin{aligned} I_B &= I_E - I_C \\ &= qA\frac{D_p}{L_p}\Delta p_E\left[\left(\frac{1}{W_b/L_p}+\frac{W_b/L_p}{3}\right)-\left(\frac{1}{W_b/L_p}-\frac{W_b/L_p}{6}\right)\right] \\ &\simeq \frac{qAD_pW_b\Delta p_E}{2L_p^2} = \frac{qAW_b\Delta p_E}{2\tau_p} \end{aligned} \tag{7-22}$$

This expression for I_B accounts for recombination in the base region. We must include injection into the emitter in many BJT devices, as discussed in Section 7.4.4.

If recombination in the base dominates the base current, I_B can be obtained from the charge control model, assuming an essentially straight-line hole distribution in the base (Fig. 7-8b). Since the hole distribution diagram appears as a triangle in this approximation, we have

$$Q_p \simeq \tfrac{1}{2}qA\,\Delta p_E W_b \tag{7-23}$$

Table 7-1 Expansions of several pertinent hyperbolic functions.

$\operatorname{sech} y = 1 - \frac{y^2}{2} + \frac{5y^4}{24} - \cdots$
$\operatorname{ctnh} y = \frac{1}{y} + \frac{y}{3} - \frac{y^3}{45} + \cdots$
$\operatorname{csch} y = \frac{1}{y} - \frac{y}{6} + \frac{7y^3}{360} - \cdots$
$\tanh y = y - \frac{y^3}{3} + \cdots$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 353

If we consider that this stored charge must be replaced every τ_p seconds and relate the recombination rate to the rate at which electrons are supplied by the base current, I_B becomes

$$I_B \simeq \frac{Q_p}{\tau_p} = \frac{qAW_b\Delta p_E}{2\tau_p} \tag{7-24}$$

which is the same as that found in Eqs. (7-21) and (7-22).

Since we have neglected the collector saturation current and have assumed $\gamma = 1$ in these approximations, the difference between I_E and I_C is accounted for by the requirements of recombination in the base. In Eq. (7-24) we have a clear demonstration that the base current is reduced for small W_b and large τ_p. We can increase τ_p by using light doping in the base region, which of course also improves the emitter injection efficiency.

The straight-line approximation of the excess hole distribution (Fig. 7-8) is fairly accurate in calculating the base current. On the other hand, it does not give a valid picture of I_E and I_C. If the distribution were perfectly straight, the slope would be the same at each end of the base region. This would imply zero base current, which is not the case. There must be some "droop" to the distribution, as in the more accurate curve of Fig. 7-7. This slight deviation from linearity gives a steeper slope at $x_n = 0$ than at $x_n = W_b$, and the value of I_E is larger than I_C by the amount I_B. The reason we can use the straight-line approximation in the charge control calculation of base current is that the area under the hole distribution curve is essentially the same in the two cases.

7.4.4 Current Transfer Ratio

The value of I_E calculated thus far in this section is more properly designated I_{Ep}, since we have assumed that $\gamma = 1$ (the emitter current due entirely to hole injection). Actually, there is always some electron injection across the forward-biased emitter junction in a real transistor, and this effect is important in calculating the current transfer ratio. As shown in Example 7-3, the emitter injection efficiency of a p-n-p transistor can be written in terms of the emitter and base properties:

$$\gamma = \left[1 + \frac{L_p^n n_n \mu_n^p}{L_n^p p_p \mu_p^n} \tanh \frac{W_b}{L_p^n}\right]^{-1} \simeq \left[1 + \frac{W_b n_n \mu_n^p}{L_n^p p_p \mu_p^n}\right]^{-1} \tag{7-25}$$

In this equation we use superscripts to indicate which side of the emitter-base junction is referred to. For example, L_p^n is the hole diffusion length in the n-type base region and μ_n^p is the electron mobility in the p-type emitter region. In an n-p-n the superscripts and subscripts would be changed along with the majority carrier symbols. Using Eq. (7-20a) for I_{Ep}, and Eq. (7-20b)

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 354

for I_C, the base transport factor B is

$$B = \frac{I_C}{I_{Ep}} = \frac{\operatorname{csch} W_b/L_p}{\operatorname{ctnh} W_b/L_p} = \operatorname{sech} \frac{W_b}{L_p} \tag{7-26}$$

and the current transfer ratio α is the product of B and γ as in Eq. (7-3).

EXAMPLE 7-3

Extend Eq. (7-20a) to include the effects of nonunity emitter injection efficiency ($\gamma < 1$). Derive Eq. (7-25) for γ. Assume that the emitter region is long compared with an electron diffusion length.

SOLUTION

Equation (7-20a) is actually I_{Ep}.

$$I_{En} = \frac{qAD_n^p}{L_n^p} n_p e^{qV_{EB}/kT} \text{ for } V_{EB} \gg kT/q$$

Thus, the total emitter current is

$$I_E = I_{Ep} + I_{En} = qA\left[\frac{D_p^n}{L_p^n} p_n \operatorname{ctnh} \frac{W_b}{L_p^n} + \frac{D_n^p}{L_n^p} n_p\right] e^{qV_{EB}/kT}$$

$$\gamma = \frac{I_{Ep}}{I_E} = \left[1 + \frac{I_{En}}{I_{Ep}}\right]^{-1} = \left[1 + \frac{\frac{D_n^p}{L_n^p} n_p}{\frac{D_p^n}{L_p^n} p_n} \tanh \frac{W_b}{L_p^n}\right]^{-1}$$

Using $\frac{n_p}{p_n} = \frac{n_n}{p_p}, \frac{D_n^p}{D_p^n} = \frac{\mu_n^p}{\mu_p^n}$, and $\frac{D}{\mu} = \frac{kT}{q}$ yields

$$\gamma = \left[1 + \frac{L_p^n n_n \mu_n^p}{L_n^p p_p \mu_p^n} \tanh \frac{W_b}{L_p^n}\right]^{-1}$$

Since we have assumed that the emitter is long compared with the electron diffusion length in the emitter, the long-diode expression applies to I_{En}. On the other hand, for a very narrow emitter width W_e, L_n^p would be replaced by W_e, corresponding to the narrow-diode expression.

7.5 GENERALIZED BIASING

The expressions derived in Section 7.4 describe the terminal currents of the transistor, if the device geometry and other factors are consistent with the assumptions. Real transistors may deviate from these approximations, as we shall see in Section 7.7. The collector and emitter junctions may differ in area, saturation current, and other parameters, so that the proper description of the terminal currents may be more complicated than Eqs. (7-18) and (7-19) suggest. For example, if the roles of emitter and collector are reversed, these equations

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 355

predict that the behavior of the transistor is symmetrical. Real transistors, on the other hand, are generally not symmetrical between emitter and collector. This is a particularly important consideration when the transistor is not biased in the usual way. We have discussed normal biasing (sometimes called the *normal active* mode), in which the emitter junction is forward biased and the collector is reverse biased. In some applications, particularly in switching, this normal biasing rule is violated. In these cases it is important to account for the differences in injection and collection properties of the two junctions. In this section we shall develop a generalized approach which accounts for transistor operation in terms of a coupled-diode model, valid for all combinations of emitter and collector bias. This model involves four measurable parameters that can be related to the geometry and material properties of the device. Using this model in conjunction with the charge control approach, we can describe the physical operation of a transistor in switching circuits and in other applications.

7.5.1 The Coupled-Diode Model

If the collector junction of a transistor is forward biased, we cannot neglect Δp_C; instead, we must use a more general hole distribution in the base region. Figure 7-9a illustrates a situation in which the emitter and collector junctions are both forward biased, so that Δp_E and Δp_C are positive numbers. We can handle this situation with Eqs. (7-18) and (7-19) for the symmetrical transistor. It is interesting to note that these equations can be considered as linear superpositions of the effects of injection by each junction. For example, the straight line hole distribution of Fig. 7-9a can be broken into the two components of Figs. 7-9b and 7-9c. One component (Fig. 7-9b) accounts for the holes injected by the emitter and collected by the collector. We can call the resulting currents (I_{EN} and I_{CN}) the *normal mode* components, since they are due to injection from emitter to collector. The component of the hole distribution illustrated by Fig. 7-9c results in currents I_{EI} and I_{CI}, which describe injection in the *inverted mode* of injection from collector to emitter.[3] Of

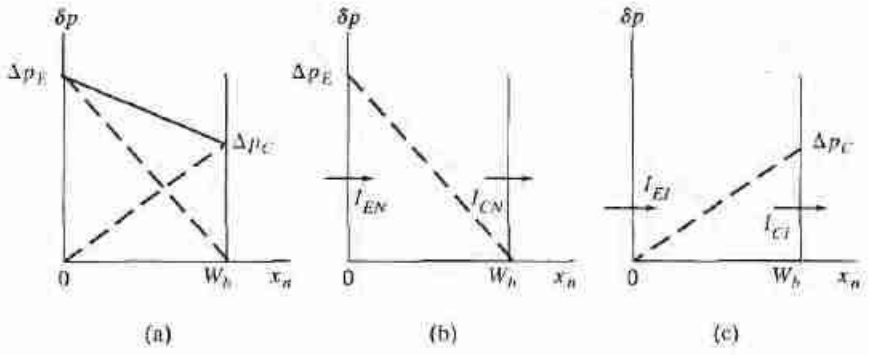


Figure 7-9 Evaluation of a hole distribution in terms of components due to normal and inverted modes: (a) approximate hole distribution in the base with emitter and collector junctions forward biased; (b) component due to injection and collection in the normal mode; (c) component due to the inverted mode.

[3]Here the words *emitter* and *collector* refer to physical regions of the device rather than to the functions of injection and collection of holes.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 356

course, these inverted components will be negative, since they account for hole flow opposite to our original definitions of I_E and I_C.

For the symmetrical transistor, these various components are described by Eqs. (7-18). Defining $a \equiv (qAD_p/L_p)$ ctnh (W_b/L_p) and $b \equiv (qAD_p/L_p)$ csch (W_b/L_p), we have

$$I_{EN} = a\Delta p_E \quad \text{and} \quad I_{CN} = b\Delta p_E \quad \text{with } \Delta p_C = 0 \tag{7-27a}$$

$$I_{EI} = -b\Delta p_C \quad \text{and} \quad I_{CI} = -a\Delta p_C \quad \text{with } \Delta p_E = 0 \tag{7-27b}$$

The four components are combined by linear superposition in Eq. (7-18):

$$\begin{aligned} I_E &= I_{EN} + I_{EI} = a\Delta p_E - b\Delta p_C \\ &= \mathsf{A}(e^{qV_{EB}/kT} - 1) - \mathsf{B}(e^{qV_{CB}/kT} - 1) \end{aligned} \tag{7-28a}$$

$$\begin{aligned} I_C &= I_{CN} + I_{CI} = b\Delta p_E - a\Delta p_C \\ &= \mathsf{B}(e^{qV_{EB}/kT} - 1) - \mathsf{A}(e^{qV_{CB}/kT} - 1) \end{aligned} \tag{7-28b}$$

where $\mathsf{A} \equiv ap_n$ and $\mathsf{B} \equiv bp_n$.

We can see from these equations that a linear superposition of the normal and inverted components does give the result we derived previously for the symmetrical transistor. To be more general, however, we must relate the four components of current by factors which allow for asymmetry in the two junctions. For example, the emitter current in the normal mode can be written

$$I_{EN} = I_{ES}(e^{qV_{EB}/kT} - 1), \quad \Delta p_C = 0 \tag{7-29}$$

where I_{ES} is the magnitude of the emitter saturation current in the normal mode. Since we specify $\Delta p_C = 0$ in this mode, we imply that $V_{CB} = 0$ in Eq. (7-8b). Thus we shall consider I_{ES} to be the magnitude of the emitter saturation current with the collector junction short circuited. Similarly, the collector current in the inverted mode is

$$I_{CI} = -I_{CS}(e^{qV_{CB}/kT} - 1), \quad \Delta p_E = 0 \tag{7-30}$$

where I_{CS} is the magnitude of the collector saturation current with $V_{EB} = 0$. As before, the minus sign associated with I_{CI} simply means that in the inverted mode holes are injected opposite to the defined direction of I_C.

The corresponding collected currents for each mode of operation can be written by defining a new α for each case:

$$I_{CN} = \alpha_N I_{EN} = \alpha_N I_{ES}(e^{qV_{EB}/kT} - 1) \tag{7-31a}$$

$$I_{EI} = \alpha_I I_{CI} = -\alpha_I I_{CS}(e^{qV_{CB}/kT} - 1) \tag{7-31b}$$

where α_N and α_I are the ratios of collected current to injected current in each mode. We notice that in the inverted mode the injected current is I_{CI} and the collected current is I_{EI}.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 357

The total current can again be obtained by superposition of the components:

$$I_E = I_{EN} + I_{EI} = I_{ES}(e^{qV_{EB}/kT} - 1) - \alpha_I I_{CS}(e^{qV_{CB}/kT} - 1) \quad (7\text{-}32a)$$

$$I_C = I_{CN} + I_{CI} = \alpha_N I_{ES}(e^{qV_{EB}/kT} - 1) - I_{CS}(e^{qV_{CB}/kT} - 1) \quad (7\text{-}32b)$$

These relations were derived by J. J. Ebers and J. L. Moll and are referred to as the *Ebers–Moll equations.*[4] While the general form is the same as Eqs. (7–28) for the symmetrical transistor, these equations allow for variations in I_{ES}, I_{CS}, α_I, and α_N due to asymmetry between the junctions. Although we shall not prove it here, it is possible to show by reciprocity arguments that

$$\alpha_N I_{ES} = \alpha_I I_{CS} \quad (7\text{-}33)$$

even for nonsymmetrical transistors.

An interesting feature of the Ebers–Moll equations is that I_E and I_C are described by terms resembling diode relations (I_{EN} and I_{CI}), plus terms which provide coupling between the properties of the emitter and collector (I_{EI} and I_{CN}). This *coupled-diode* property is illustrated by the equivalent circuit of Fig. 7–10. In this figure we take advantage of Eq. (7–8) to write the Ebers–Moll equations in the form

$$I_E = I_{ES}\frac{\Delta p_E}{p_n} - \alpha_I I_{CS}\frac{\Delta p_C}{p_n} = \frac{I_{ES}}{p_n}(\Delta p_E - \alpha_N \Delta p_C) \quad (7\text{-}34a)$$

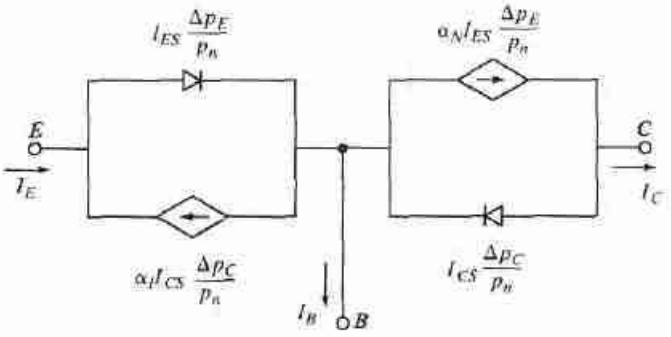


Figure 7–10 An equivalent circuit synthesizing the Ebers–Moll equations.

$$I_B = (1 - \alpha_N) I_{ES}\frac{\Delta p_E}{p_n} + (1 - \alpha_I) I_{CS}\frac{\Delta p_C}{p_n}$$

[4] J. J. Ebers and J. L. Moll, "Large-Signal Behavior of Junction Transistors," *Proceedings of the IRE* 42, pp. 1761–72 (December 1954). In the original paper and in many texts, the terminal currents are all defined as flowing *into* the transistor. This introduces minus signs into the expressions for I_C and I_B as we have developed them here.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 358

$$I_C = \alpha_N I_{ES}\frac{\Delta p_E}{p_n} - I_{CS}\frac{\Delta p_C}{p_n} = \frac{I_{CS}}{p_n}(\alpha_I \Delta p_E - \Delta p_C) \qquad (7\text{-}34b)$$

It is often useful to relate the terminal currents to each other as well as to the saturation currents. We can eliminate the saturation current from the coupling term in each part of Eq. (7-32). For example, by multiplying Eq. (7-32a) by α_N and subtracting the resulting expression from Eq. (7-32b), we have

$$I_C = \alpha_N I_E - (1 - \alpha_N\alpha_I)I_{CS}(e^{qV_{CB}/kT} - 1) \qquad (7\text{-}35)$$

Similarly, the emitter current can be written in terms of the collector current:

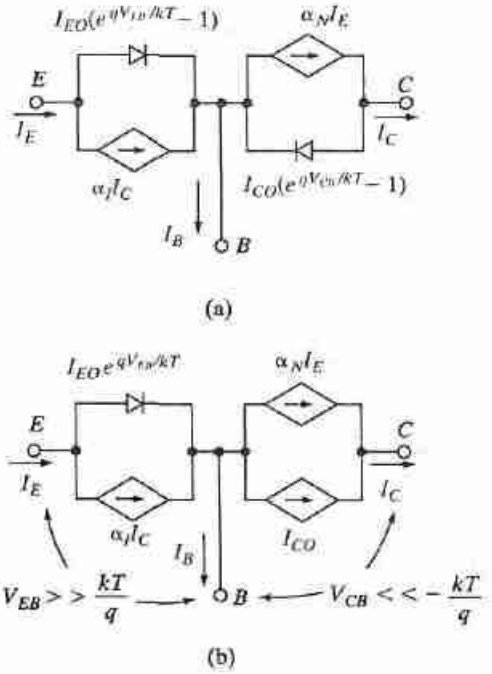
$$I_E = \alpha_I I_C + (1 - \alpha_N\alpha_I)I_{ES}(e^{qV_{EB}/kT} - 1) \qquad (7\text{-}36)$$

The terms $(1 - \alpha_N\alpha_I)I_{CS}$ and $(1 - \alpha_N\alpha_I)I_{ES}$ can be abbreviated as I_{CO} and I_{EO}, respectively, where I_{CO} is the magnitude of the collector saturation current with the emitter junction *open* ($I_E = 0$), and I_{EO} is the magnitude of the emitter saturation current with the collector open. The Ebers-Moll equations then become

$$I_E = \alpha_I I_C + I_{EO}(e^{qV_{EB}/kT} - 1) \qquad (7\text{-}37a)$$

$$I_C = \alpha_N I_E - I_{CO}(e^{qV_{CB}/kT} - 1) \qquad (7\text{-}37b)$$

and the equivalent circuit is shown in Fig. 7-11a. In this form the equations describe both the emitter and collector currents in terms of a simple diode



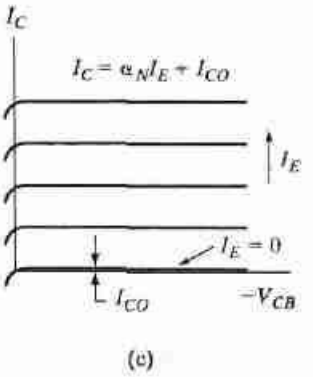


Figure 7-11
Equivalent circuits of the transistor in terms of the terminal currents and the open-circuit saturation currents: (a) synthesis of Eqs. (7-37); (b) equivalent circuit with normal biasing; (c) collector characteristics with normal biasing.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 359

characteristic plus a current generator proportional to the other current. For example, under normal biasing the equivalent circuit reduces to the form shown in Fig. 7-11b. The collector current is α_N times the emitter current plus the collector saturation current, as expected. The resulting collector characteristics of the transistor appear as a series of reverse-biased diode curves, displaced by increments proportional to the emitter current (Fig. 7-11c).

EXAMPLE 7-4 A symmetrical p^+-n-p^+ bipolar transistor has the following properties:

	Emitter	Base
$A = 10^{-4}$ cm^2	$N_a = 10^{17}$	$N_d = 10^{15}$ cm^{-3}
$W_b = 1$ μm	$\tau_n = 0.1$ μs	$\tau_p = 10$ μs
	$\mu_p = 200$	$\mu_n = 1300$ cm^2/V-s
	$\mu_n = 700$	$\mu_p = 450$

(a) Calculate the saturation current $I_{ES} = I_{CS}$.

(b) With $V_{EB} = 0.3$ V and $V_{CB} = -40$ V, calculate the base current I_B, assuming perfect emitter injection efficiency.

(c) Calculate the base transport factor B, emitter injection efficiency γ, and amplification factor β, assuming that the emitter region is long compared with L_n.

SOLUTION In the base,

$$p_n = n_i^2/n_a = (1.5 \times 10^{10})^2/10^{15} = 2.25 \times 10^5$$

$$D_p = 450(0.0259) = 11.66,\ L_p = (11.66 \times 10^{-5})^{1/2} = 1.08 \times 10^{-2}$$

$$W_b/L_p = 10^{-4}/1.08 \times 10^{-2} = 9.26 \times 10^{-3}$$

$$I_{ES} = I_{CS} = qA(D_p/L_p)p_n \operatorname{ctnh}(W_b/L_p)$$

$$= (1.6 \times 10^{-19})(10^{-4})(11.66/1.08 \times 10^{-2})$$

$$(2.25 \times 10^5) \operatorname{ctnh} 9.26 \times 10^{-3}$$

$$= 4.2 \times 10^{-13}\ \text{A}$$

$$\Delta p_E = p_n e^{qV_{EB}/kT},\ \Delta p_C \approx 0$$

$$\Delta p_E = 2.25 \times 10^5 \times e^{(0.3/0.0259)} = 2.4 \times 10^{10}$$

$$I_B = qA(D_p/L_p)\Delta p_E \tanh(W_b/2L_p)$$

or

$$I_B = \frac{Q_b}{\tau_p} = qAW_b\Delta p_E/2\tau_p = \mathbf{1.9 \times 10^{-12}\ A}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 360

In the emitter,

$$D_n = 700(0.0259) = 18.13$$

$$L_n = (18.13 \times 10^{-7})^{1/2} = 1.35 \times 10^{-3}$$

$$I_{En} = \frac{qAD_n^E}{L_n^E} n_p^E e^{qV_{EB}/kT}$$

$$I_{Ep} = \frac{qAD_p^B}{L_p^B} p_n^B \operatorname{ctnh}\frac{W_b}{L_p^B} e^{qV_{EB}/kT}$$

$$\gamma = \frac{I_{Ep}}{I_{En} + I_{Ep}} = \left[1 + \frac{I_{En}}{I_{Ep}}\right]^{-1}$$

$$= \left[1 + \frac{D_n^E/L_n^E}{D_p^B/L_p^B}\frac{n_p^E}{p_n^B}\tanh\frac{W_b}{L_p}\right]^{-1} \qquad \left(\text{use } \frac{n_p^E}{p_n^B} = \frac{n_n^B}{p_p^E}\right)$$

$$\gamma = \left[1 + \frac{18.13 \times 1.08 \times 10^{-2} \times 10^{15}}{11.66 \times 1.35 \times 10^{-3} \times 10^{17}}\tanh 9.26 \times 10^{-3}\right]^{-1} = \mathbf{0.99885}$$

$$B = \operatorname{sech}\frac{W_b}{L_p} = \operatorname{sech} 9.26 \times 10^{-3} = \mathbf{0.99996}$$

$$\alpha = B\gamma = (0.99885)(0.99996) = \mathbf{0.9988}$$

$$\beta = \frac{\alpha}{1 - \alpha} = \frac{0.9988}{0.0012} = \mathbf{832}$$

7.5.2 Charge Control Analysis

The charge control approach is useful in analyzing the transistor terminal currents, particularly in a-c applications. Considerations of transit time effects and charge storage are revealed easily by this method. Following the techniques of the previous section, we can separate an arbitrary excess hole distribution in the base into the normal and inverted distributions of Fig. 7-9. The charge stored in the normal distribution will be called Q_N and the charge under the inverted distribution will be called Q_I. Then we can evaluate the currents for the normal and inverted modes in terms of these stored charges. For example, the collected current in the normal mode I_{CN} is simply the charge Q_N divided by the mean time required for this charge to be collected. This time is the transit time for the normal mode τ_{tN}. On the other hand, the emitter current must support not only the rate of charge collection by the collector but also the recombination rate in the base Q_N/τ_{pN}. Here we use a subscript N with the transit time and lifetime in the normal mode in contrast to the inverted mode, to allow for possible asymmetries due to imbalance in

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 361

the transistor structure. With these definitions, the normal components of current become

$$I_{CN} = \frac{Q_N}{\tau_{tN}}, \quad I_{EN} = \frac{Q_N}{\tau_{tN}} + \frac{Q_N}{\tau_{pN}} \tag{7-38a}$$

Similarly, the inverted components are

$$I_{EI} = -\frac{Q_I}{\tau_{tI}}, \quad I_{CI} = -\frac{Q_I}{\tau_{tI}} - \frac{Q_I}{\tau_{pI}} \tag{7-38b}$$

where the I subscripts on the stored charge and on the transit and recombination times designate the inverted mode. Combining these equations as in Eq. (7-32) we have the terminal currents for general biasing:

$$I_E = Q_N\left(\frac{1}{\tau_{tN}} + \frac{1}{\tau_{pN}}\right) - \frac{Q_I}{\tau_{tI}} \tag{7-39a}$$

$$I_C = \frac{Q_N}{\tau_{tN}} - Q_I\left(\frac{1}{\tau_{tI}} + \frac{1}{\tau_{pI}}\right) \tag{7-39b}$$

It is not difficult to show that these equations correspond to the Ebers-Moll relations [Eq. (7-34)], where

$$\begin{aligned} \alpha_N &= \frac{\tau_{pN}}{\tau_{tN} + \tau_{pN}}, & \alpha_I &= \frac{\tau_{pI}}{\tau_{tI} + \tau_{pI}} \\ I_{ES} &= q_N\left(\frac{1}{\tau_{tN}} + \frac{1}{\tau_{pN}}\right), & I_{CS} &= q_I\left(\frac{1}{\tau_{tI}} + \frac{1}{\tau_{pI}}\right) \\ Q_N &= q_N\frac{\Delta p_E}{p_n}, & Q_I &= q_I\frac{\Delta p_C}{p_n} \end{aligned} \tag{7-40}$$

The base current in the normal mode supports recombination, and the base-to-collector current amplification factor β_N takes the form predicted by Eq. (7-7):

$$I_{BN} = \frac{Q_N}{\tau_{pN}}, \quad \beta_N = \frac{I_{CN}}{I_{BN}} = \frac{\tau_{pN}}{\tau_{tN}} \tag{7-41}$$

This expression for β_N is also obtained from $\alpha_N/(1 - \alpha_N)$. Similarly, I_{BI} is Q_I/τ_{pI}, and the total base current is

$$I_B = I_{BN} + I_{BI} = \frac{Q_N}{\tau_{pN}} + \frac{Q_I}{\tau_{pI}} \tag{7-42}$$

This expression for the base current is substantiated by $I_E - I_C$ from Eq. (7-39).

The effects of time dependence of stored charge can be included in these equations by the methods introduced in Section 5.5.1. We can include

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 362

the proper dependencies by adding a rate of change of stored charge to each of the injection currents I_{EN} and I_{CI}:

$$i_E = Q_N\left(\frac{1}{\tau_{tN}} + \frac{1}{\tau_{pN}}\right) - \frac{Q_I}{\tau_{tI}} + \frac{dQ_N}{dt} \tag{7-43a}$$

$$i_C = \frac{Q_N}{\tau_{tN}} - Q_I\left(\frac{1}{\tau_{tI}} + \frac{1}{\tau_{pI}}\right) - \frac{dQ_I}{dt} \tag{7-43b}$$

$$i_B = \frac{Q_N}{\tau_{pN}} + \frac{Q_I}{\tau_{pI}} + \frac{dQ_N}{dt} + \frac{dQ_I}{dt} \tag{7-43c}$$

We shall return to these equations in Section 7.8, when we discuss the use of transistors at high frequencies.

7.6 SWITCHING

In a switching operation a transistor is usually controlled in two conduction states, which can be referred to loosely as the "on" state and the "off" state. Ideally, a switch should appear as a short circuit when turned on and an open circuit when turned off. Furthermore, it is desirable to switch the device from one state to the other with no lost time in between. Transistors do not fit this ideal description of a switch, but they can serve as a useful approximation in practical electronic circuits. The two states of a transistor in switching can be seen in the simple common-emitter example of Fig. 7-12. In this figure the collector current i_C is controlled by the base current i_B over most of the family of characteristic curves. The load line specifies the locus of allowable (i_C, v_{CE}) points for the circuit, in analogy with Fig. 6-2. If i_B is such that the operating point lies somewhere between the two end points of the load line (Fig. 7-12b), the transistor operates in the normal active mode. That is, the emitter junction is forward biased and the collector is reverse biased, with a reasonable value of i_B flowing out of the base. On the other hand, if the base current is zero or negative, the point C is reached at the bottom end of the load line, and the collector current is negligible. This is the "off" state of the transistor, and the device is said to be operating in the *cutoff* regime. If the base current is positive and sufficiently large, the device is driven to the *saturation* regime, marked S. This is the "on" state of the transistor, in which a large value of i_C flows with only a very small voltage drop v_{CE}. As we shall see below, the beginning of the saturation regime corresponds to the loss of reverse bias across the collector junction. In a typical switching operation the base current swings from positive to negative, thereby driving the device from saturation to cutoff, and vice versa. In this section we shall explore the nature of conduction in the cutoff and saturation regimes; also we shall investigate the factors affecting the speed with which the transistor can be switched between the two states.

The various regions of operation of a BJT are illustrated in Fig. 7-12c. If the emitter junction is forward biased and the collector reverse biased,

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 363

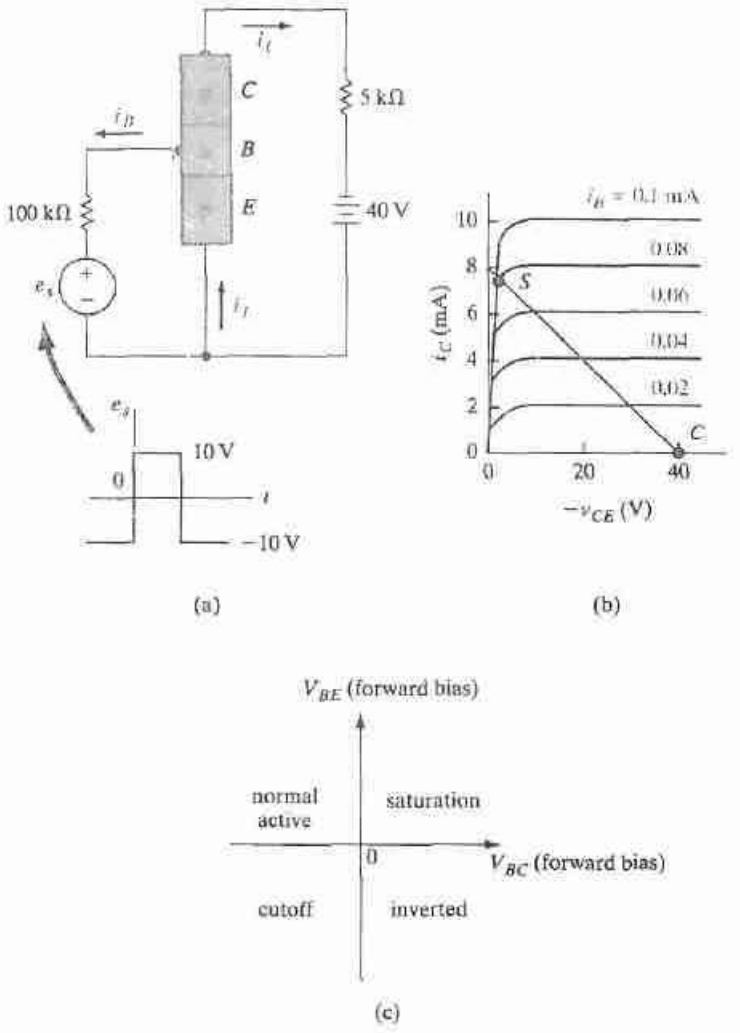


Figure 7–12 Simple switching circuit for a transistor in the common-emitter configuration: (a) biasing circuit; (b) collector characteristics and load line for the circuit, with cutoff and saturation indicated; (c) operating regimes of a BJT.

we have the normal active mode; the opposite gives us the inverted mode of operation. If both junctions are reverse biased, we get cutoff, leading to a very high impedance state of the BJT; if both junctions are forward biased, we have saturation and a low impedance state.

7.6.1 Cutoff

If the emitter junction is reverse biased in the cutoff regime (negative i_B), we can approximate the excess hole concentrations at the edges of the reverse-biased emitter and collector junctions as

$$\frac{\Delta p_E}{p_n} \simeq \frac{\Delta p_C}{p_n} \simeq -1 \tag{7-44}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 364

which implies $p(x_n) = 0$. The excess hole distribution in the base is approximately constant at $-p_n$, with some slope to the distribution at each edge to account for the reverse saturation current in the junctions. The base current i_B can be approximated for a symmetrical transistor on a charge storage basis as $-qAp_nW_b/\tau_p$. In this calculation a negative excess hole concentration corresponds to *generation* in the same way that a positive distribution indicates recombination. This expression is also obtained by applying Eq. (7-44) to Eq. (7-19) with an approximation from Table 7-1. Physically, a small saturation current flows from n to p in each reverse-biased junction, and this current is supplied by the base current i_B (which is negative when flowing into the base of a p-n-p device according to our definitions). A more general evaluation of the currents can be obtained from the Ebers-Moll equations by applying Eq. (7-44) to Eq. (7-34):

$$i_E = -I_{ES} + \alpha_I I_{CS} = -(1 - \alpha_N)I_{ES} \tag{7-45a}$$

$$i_C = -\alpha_N I_{ES} + I_{CS} = (1 - \alpha_I)I_{CS} \tag{7-45b}$$

$$i_B = i_E - i_C = -(1 - \alpha_N)I_{ES} - (1 - \alpha_I)I_{CS} \tag{7-45c}$$

If the short-circuit saturation currents I_{ES} and I_{CS} are small and α_N and α_I are both near unity, these currents will be negligible and the cutoff regime will closely approximate the "off" condition of an ideal switch. The equivalent circuit corresponding to Eq. (7-45) is illustrated in Fig. 7-13b.

7.6.2 Saturation

The saturation regime begins when the reverse bias across the collector junction is reduced to zero, and it continues as the collector becomes forward biased. The excess hole distribution in this case is illustrated in Fig. 7-14. The device is saturated when $\Delta p_C = 0$, and forward bias of the collector junction

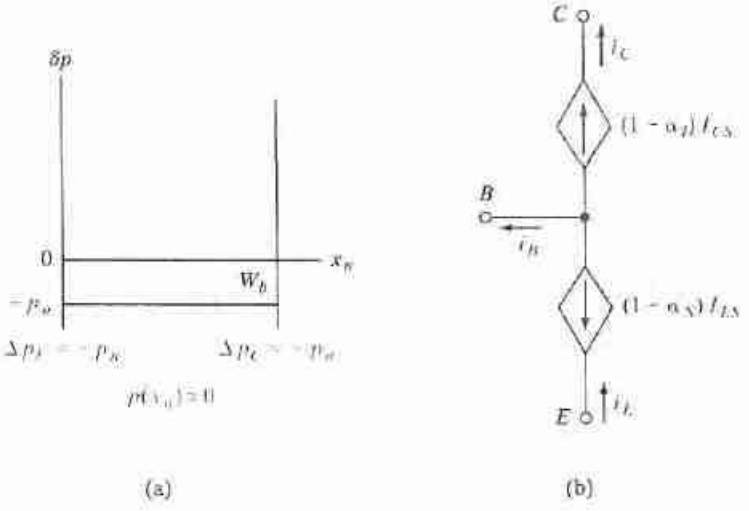


Figure 7-13
The cutoff regime of a p-n-p transistor: (a) excess hole distribution in the base region with emitter and collector junctions reverse biased; (b) equivalent circuit corresponding to Eq. (7-45).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 365

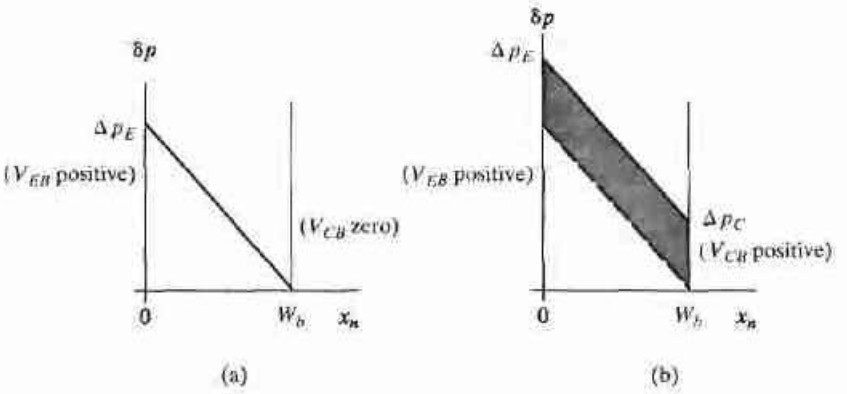


Figure 7–14
Excess hole distribution in the base of a saturated transistor: (a) the beginning of saturation; (b) oversaturation.

(Fig. 7–14b) leads to a positive Δp_C, driving the device further into saturation. With the load line fixed by the battery and the 5-kΩ resistor in Fig. 7–12, saturation is reached by increasing the base current i_B. We can see how a large value of i_B leads to saturation by applying the reasoning of charge control to Fig. 7–14. Since a certain amount of stored charge is required to accommodate a given i_B (and vice versa), an increase in i_B calls for an increase in the area under the $\delta p(x_n)$ distribution.

In Fig. 7–14a the device has just reached saturation, and the collector junction is no longer reverse biased. The implication of this condition for the circuit of Fig. 7–12 is easy to state. Since the emitter junction is forward biased and the collector junction has zero bias, very little voltage drop appears across the device from collector to emitter. The magnitude of $-v_{CE}$ is only a fraction of a volt. Therefore, almost all of the battery voltage appears across the resistor, and the collector current is approximately 40 V/5 kΩ = 8 mA. As the device is driven deeper into saturation (Fig. 7–14b), the collector current stays essentially constant while the base current increases. In this saturation condition the transistor approximates the "on" state of an ideal switch.

Whereas the degree of "oversaturation" (indicated by the shaded area in Fig. 7–14b) does not affect the value of i_C significantly, it is important in determining the time required to switch the device from one state to the other. For example, from previous experience we expect the turn-off time (from saturation to cutoff) to be longer for larger values of stored charge in the base. We can calculate the various charging and delay times from Eq. (7–43). Detailed calculations are somewhat involved, but we can simplify the problem greatly with approximations of the type used in Chapter 5 for transient effects in p-n junctions.

7.6.3 The Switching Cycle

The various mechanisms of a switching cycle are illustrated in Fig. 7–15. If the device is originally in the cutoff condition, a step increase of base current to I_B causes the hole distribution to increase approximately as illustrated in

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 366

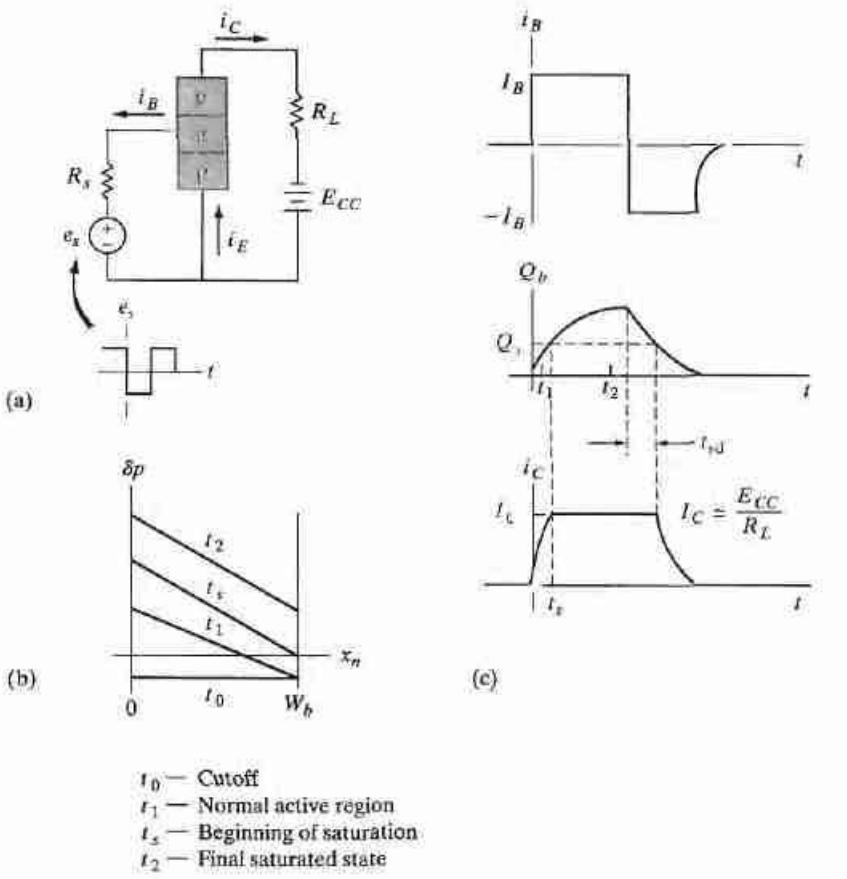


Figure 7–15 Switching effects in a common-emitter transistor circuit: (a) circuit diagram; (b) approximate hole distributions in the base during switching from cutoff to saturation; (c) base current, stored charge, and collector current during a turn-on and a turn-off transient.

Fig. 7-15b. As in the transient analysis of Chapter 5, we assume for simplicity of calculation that the distribution maintains a simple form in each time interval of the transient. At time t_s the device enters saturation, and the hole distribution reaches its final state at t_2. As the stored charge in the base Q_b increases, there is an increase in the collector current i_C. The collector current does not increase beyond its value at the beginning of saturation t_s, however. We can approximate this saturated collector current as $I_C \simeq E_{CC}/R_L$, where E_{CC} is the value of the collector circuit battery and R_L is the load resistor ($I_C \simeq 8$ mA for the example of Fig. 7-12). There is an essentially exponential increase in the collector current while Q_b rises to its value Q_s at t_s; this rise time serves as one of the limitations of the transistor in a switching application. Similarly, when the base current is switched negative (e.g., to the value $-I_B$), the stored charge must be withdrawn from the base before cutoff is reached. While Q_b is larger than Q_s, the collector current remains at the value I_C, fixed by the battery and resistor. Thus there is a storage delay time t_{sd} after the base current is switched and before i_C begins to fall toward zero. After the stored charge is reduced below Q_s, i_C drops exponentially with the characteristic fall

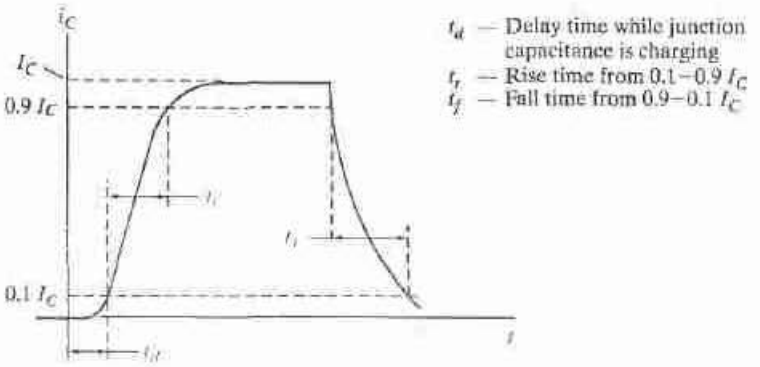


Figure 7-16 Collector current during switching transients, including the delay time required for charging the junction capacitance; definitions of the rise time and fall time.

time. Once the stored charge is withdrawn, the base current cannot be maintained any longer at its large negative value and must decay to the small cutoff value described by Eq. (7-45c).

7.6.4 Specifications for Switching Transistors

We can determine t_s and t_{sd} by solving for the time-dependent base current, $i_B(t)$, given by an expression similar to Eq. (5-47). We must also not neglect the charging time of the emitter junction capacitance in going from cutoff to saturation. Since the emitter junction is reverse biased in cutoff, it is necessary for the emitter space charge layer to be charged to the forward-bias condition before collector current can flow. Therefore, we should include a *delay time* t_d as in Fig. 7-16 to account for this effect. Typical values of t_d are given in the specification information of most switching transistors, along with a *rise time* t_r, defined as the time required for the collector current to rise from 10 to 90% of its final value. A third specification is the *fall time* t_f required for i_C to fall through a similar fraction of its turn-off excursion.

7.7 OTHER IMPORTANT EFFECTS

The approach we have taken in analyzing the properties of transistors has involved a number of simplifying assumptions. Some of the assumptions must be modified in dealing with practical devices. In this section we investigate some common deviations from the basic theory and indicate situations in which each effect is important. Since the various effects discussed here involve modifications of the more straightforward theory, they are often labeled "secondary effects." This does not imply that they are unimportant; in fact, the effects described in this section can dominate the conduction in transistors under certain conditions of device geometry and circuit application.

In this section we shall consider the effects of nonuniform doping in the base region of the transistor. In particular, we shall find that graded doping can lead to a drift component of charge transport across the base, adding to the diffusion of carriers from emitter to collector. We shall discuss the effects of large reverse bias on the collector junction, in terms of widening the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 368

space charge region about the junction and avalanche multiplication. We shall see that transistor parameters are affected at high current levels by the degree of injection and by heating effects. We shall consider several structural effects that are important in practical devices, such as asymmetry in the areas of the emitter and collector junctions, series resistance between the base contact and the active part of the base region, and nonuniformity of injection at the emitter junction. All these effects are important in understanding the operation of transistors, and proper consideration of their interactions can contribute greatly to the usefulness of practical transistor circuits.

7.7.1 Drift in the Base Region

The assumption of uniform doping in the base breaks down for implanted junction transistors which usually involve an appreciable amount of impurity grading; for example, the implanted transistor of Fig. 7-5 has a doping profile similar to that sketched in Fig. 7-17. In this example there is a fairly sharp discontinuity in the doping profile, when the donor concentration in the base region becomes smaller than the constant p-type background doping in the collector. Similarly, the emitter is assumed to be a heavily doped (p^+) shallow region, providing a second rather sharp boundary for the base. Within the base region itself, however, the net doping concentration ($N_d - N_a \equiv N$) varies along a profile which decreases from the emitter edge to the collector edge. The most likely doping distribution in the base is a portion of a gaussian (see Section 5.1.4); however, we can clearly see the effect of an impurity gradient by assuming for simplicity that $N(x_n)$ varies exponentially within the base region (Fig. 7-17b).

One important result of a graded base region is that a built-in electric field exists from emitter to collector (for a p-n-p), thereby adding a drift component to the transport of holes across the base. We can demonstrate this effect very simply by considering the required balance of drift and diffusion in the base at equilibrium. If the net donor doping of the base is large enough

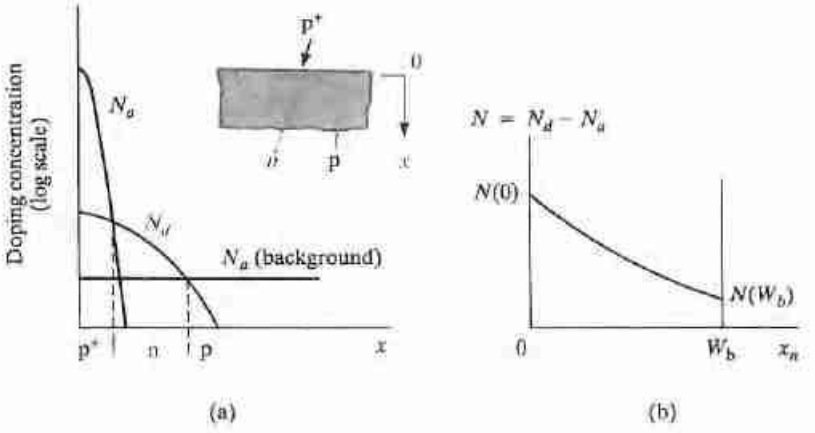


Figure 7-17
Graded doping in the base region of a p-n-p transistor: (a) typical doping profile on a semilog plot; (b) approximate exponential distribution of the net donor concentration in the base region on a linear plot.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 369

to allow the usual approximation $n(x_n) = N(x_n)$, the balance of electron drift and diffusion currents at equilibrium requires

$$I_n(x_n) = qA\mu_n N(x_n)\mathscr{E}(x_n) + qAD_n\frac{dN(x_n)}{dx_n} = 0 \qquad (7\text{-}46)$$

Therefore, the built-in electric field is

$$\mathscr{E}(x_n) = -\frac{D_n}{\mu_n}\frac{1}{N(x_n)}\frac{dN(x_n)}{dx_n} = \frac{kT}{q}\frac{1}{N(x_n)}\frac{dN(x_n)}{dx_n} \qquad (7\text{-}47)$$

For a doping profile $N(x_n)$ that decreases in the positive x_n-direction, this field is positive, directed from emitter to collector.

For the example of an exponential doping profile, the electric field $\mathscr{E}(x_n)$ turns out to be constant with position in the base. We can represent an exponential distribution as

$$N(x_n) = N(0)e^{-ax_n/W_b} \qquad \text{where } a = \ln\frac{N(0)}{N(W_b)} \qquad (7\text{-}48)$$

Taking the derivative of this distribution and substituting in Eq. (7-47), we obtain the constant field

$$\mathscr{E}(x_n) = \frac{kT}{q}\frac{a}{W_b} \qquad (7\text{-}49)$$

Since this field aids the transport of holes across the base region from emitter to collector, the transit time τ_t is reduced below that of a comparable uniform base transistor. Similarly, electron transport in an n-p-n is aided by the built-in field in the base. This shortening of the transit time can be very important in high-frequency devices (Section 7.8.2). Another approach for obtaining a built-in field is to vary the alloy composition x (and therefore E_g) in a base made of an alloy such as $Si_{1-x}Ge_x$ or $In_xGa_{1-x}As$. We will discuss this further in Section 7.9.

7.7.2 Base Narrowing

In the discussion of transistors thus far, we have assumed that the effective base width W_b is essentially independent of the bias voltages applied to the collector and emitter junctions. This assumption is not always valid; for example, the p^+-n-p^+ transistor of Fig. 7-18 is affected by the reverse bias applied to the collector. If the base region is lightly doped, the depletion region at the reverse-biased collector junction can extend significantly into the n-type base region. As the collector voltage is increased, the space charge layer takes up more of the metallurgical width of the base L_b, and as a result, the effective base width W_b is decreased. This effect is variously called *base narrowing, base-width modulation*, and the *Early effect* after J. M. Early, who first interpreted it. The effects of base narrowing are apparent in the collector characteristics for the common-emitter configuration (Fig. 7-18b).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 370

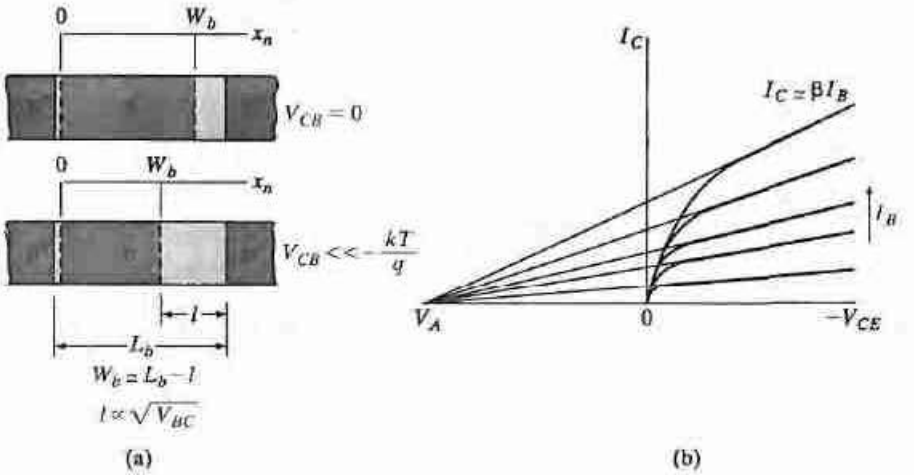


Figure 7–18
The effects of base narrowing on the characteristics of a p⁺-n-p⁺ transistor: (a) decrease in the effective base width as the reverse bias on the collector junction is increased; (b) common-emitter characteristics showing the increase in I_C with increased collector voltage. The black lines in (b) indicate the extrapolation of the curves to the Early voltage V_A.

The decrease in W_b causes β to increase. As a result, the collector current I_C increases with collector voltage rather than staying constant as predicted from the simple treatment. The slope introduced by the Early effect is almost linear with I_C, and the common-emitter characteristics extrapolate to an intersection with the voltage axis at V_A, called the Early voltage.

For the p⁺-n-p⁺ device of Fig. 7–18 we can approximate the length l of the collector junction depletion region in the n material from Eq. (5–23b) with V_0 replaced by $V_0 - V_{CB}$ and V_{CB} taken to be large and negative:

$$l = \left(\frac{2\epsilon V_{BC}}{qN_d}\right)^{1/2} \tag{7–50}$$

If the reverse bias on the collector junction is increased far enough, it is possible to decrease W_b to the extent that the collector depletion region essentially fills the entire base. In this punch-through condition holes are swept directly from the emitter region to the collector, and transistor action is lost. Punch-through is a breakdown effect that is generally avoided in circuit design. In most cases, however, avalanche breakdown of the collector junction occurs before punch-through is reached. We shall discuss the effects of avalanche multiplication in the following section.

In devices with graded base doping, base narrowing is of less importance. For example, if the donor concentration in the base region of a p-n-p increases with position from the collector to the emitter, the intrusion of the collector space charge region into the base becomes less important with increased bias as more donors are available to accommodate the space charge.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 371

7.7.3 Avalanche Breakdown

Before punch-through occurs in most transistors, avalanche multiplication at the collector junction becomes important (see Section 5.4.2). As Fig. 7-19 indicates, the collector current increases sharply at a well-defined breakdown voltage BV_{CBO} for the common-base configuration. For the common-emitter case, however, there is a strong influence of carrier multiplication over a fairly broad range of collector voltage. Furthermore, the breakdown voltage in the common-emitter case BV_{CEO} is significantly smaller than BV_{CBO}. We can understand these effects by considering breakdown for the condition $I_E = 0$ in the common-base case and for $I_B = 0$ in the common-emitter case. These conditions are implied by the O in the subscripts of BV_{CEO} and BV_{CBO}. In each case the terminal current I_C is the current entering the collector depletion region multiplied by the factor M. Including multiplication due to impact ionization, Eq.(7-37b) becomes

$$I_C = (\alpha_N I_E + I_{CO})M = (\alpha_N I_E + I_{CO})\frac{1}{1 - (V_{BC}/BV_{CBO})^n} \quad (7\text{-}51)$$

where for M we have used the empirical expression given in Eq. (5-44).

For the limiting common-base case of $I_E = 0$ (the lowest curve in Fig. 7-23a), I_C is simply MI_{CO}, and the breakdown voltage is well defined, as in an isolated junction. The term BV_{CBO} signifies the collector junction breakdown voltage in common-base with the emitter open. In the common-emitter case the situation is somewhat more complicated. Setting $I_B = 0$, and therefore, $I_C = I_E$ in Eq. (7-51), we have

$$I_C = \frac{MI_{CO}}{1 - M\alpha_N} \quad (7\text{-}52)$$

We notice that in this case the collector current increases indefinitely when $M\alpha_N$ approaches unity. By contrast, M must approach infinity in the common-base case before BV_{CBO} is reached. Since α_N is close to unity in most transistors, M need be only slightly larger than unity for Eq. (7-52) to approach

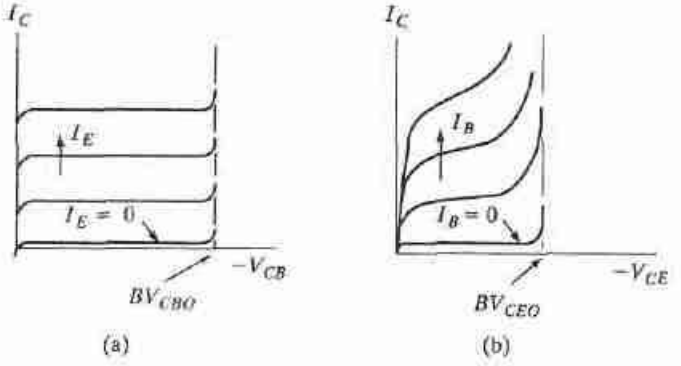


Figure 7-19 Avalanche breakdown in a transistor: (a) common-base configuration; (b) common-emitter configuration.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 372

breakdown. Avalanche multiplication thus dominates the current in a common-emitter transistor well below the breakdown voltage of the isolated collector junction. The sustaining voltage for avalanching in the common-emitter case BV_{CEO} is therefore smaller than BV_{CBO}.

We can understand physically why multiplication is so important in the common-emitter case by considering the effect of M on the base current. When an ionizing collision occurs in the collector junction depletion region, a secondary hole and electron are created. The primary and secondary holes are swept into the collector in a p-n-p, but the electron is swept into the base by the junction field. Therefore, the supply of electrons to the base is increased, and from our charge control analysis we conclude that hole injection at the emitter must increase to maintain space charge neutrality. This is a regenerative process, in which an increased injection of holes from the emitter causes an increased multiplication current at the collector junction; this in turn increases the rate at which secondary electrons are swept into the base, calling for more hole injection. Because of this regenerative effect, it is easy to understand why the multiplication factor M need be only slightly greater than unity to start the avalanching process.

7.7.4 Injection Level; Thermal Effects

In discussions of transistor characteristics we have assumed that α and β are independent of carrier injection level. Actually, the parameters of a practical transistor may vary considerably with injection level, which is determined by the magnitude of I_E or I_C. For very low injection, the assumption of negligible recombination in the junction depletion regions is invalid (see Section 5.6.2). This is particularly important in the case of recombination in the emitter junction, where any recombination tends to degrade the emitter injection efficiency γ. Thus we expect that α and β should decrease for low values of I_C, causing the curves of the collector characteristics to be spaced more closely for low currents than for higher currents.

As I_C is increased beyond the low injection level range, α and β increase but fall off again at very high injection. The primary cause of this fall-off is the increase of majority carriers at high injection levels (see Section 5.6.1). For example, as the concentration of excess holes injected into the base becomes large, the matching excess electron concentration can become greater than the background n_{B0}. This base conductivity modulation effect results in a decrease in γ as more electrons are injected across the emitter junction into the emitter region.

Large values of I_C may be accompanied by significant power dissipation in the transistor and therefore heating of the device. In particular, the product of I_C and the collector voltage V_{BC} is a measure of the power dissipated at the collector junction. This dissipation is due to the fact that carriers swept through the collector junction depletion region are given increased kinetic energy, which in turn is given up to the lattice in scattering collisions. It is very important that the transistor be operated in a range such that $I_C V_{BC}$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 373

does not exceed the maximum power rating of the device. In devices designed for high power capability, the transistor is mounted on an efficient heat sink, so that thermal energy can be transferred away from the junction.

If the temperature of the device is allowed to increase due to power dissipation or thermal environment, the transistor parameters change. The most important parameters dependent on temperature are the carrier lifetimes and diffusion coefficients. In Si or Ge devices the lifetime τ_p increases with temperature for most cases, due to thermal reexcitation from recombination centers. This increase in τ_p tends to increase β for the transistor. On the other hand, the mobility decreases with increasing temperature in the lattice-scattering range, varying approximately as $T^{-3/2}$ (see Fig. 3-22). Thus from the Einstein relation, we expect D_p to decrease as the temperature increases, thereby causing a drop in β due to an increasing transit time τ_t. Of these competing processes, the effect of increasing lifetime with temperature usually dominates, and β becomes larger as the device is heated. It is clear from this effect that *thermal runaway* can occur if the circuit is not designed to prevent it. For example, a large power dissipation in the device can cause an increase in T; this results in a large β and therefore a large I_C for a given base current; the large I_C causes more collector dissipation and the cycle continues. This type of runaway of the collector current can result in overheating and destruction of the device.

7.7.5 Base Resistance and Emitter Crowding

A number of structural effects are important in determining the operation of a transistor. For example, the emitter and collector areas are considerably different in the implanted transistor of Fig. 7-20a. This and most other structural effects can be accounted for by differences in α_N, α_I, and other parameters in the Ebers-Moll model. Several effects caused by the structural arrangement of real transistors deserve special attention, however. One of the most important of these effects is the fact that base current must pass from the active part of the base region to the base contacts B. Thus, to be accurate, we should include a resistance r_b in equivalent models for the transistor to account for voltage drops which may occur between B and the active part of the base. Because of r_b, it is common to contact the base with the metallization pattern on both sides of the emitter, as in Fig. 7-20c.

If the transistor is designed so that the n-type regions leading from the base to the contacts are large in cross-sectional area, the base resistance r_b may be negligible. On the other hand, the distributed resistance r_b' along the thin base region is almost always important.[5] Since the width of the base between emitter and collector is very narrow, this distributed resistance is usually quite high. Therefore, as base current flows from points within the base region toward each end, a voltage drop occurs along r_b'. In this case the forward bias across the emitter-base junction is not uniform, but instead varies with position according to the voltage drop in the distributed base resistance.

[5]The distributed resistance r_b' is often called the *base spreading resistance*.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 374

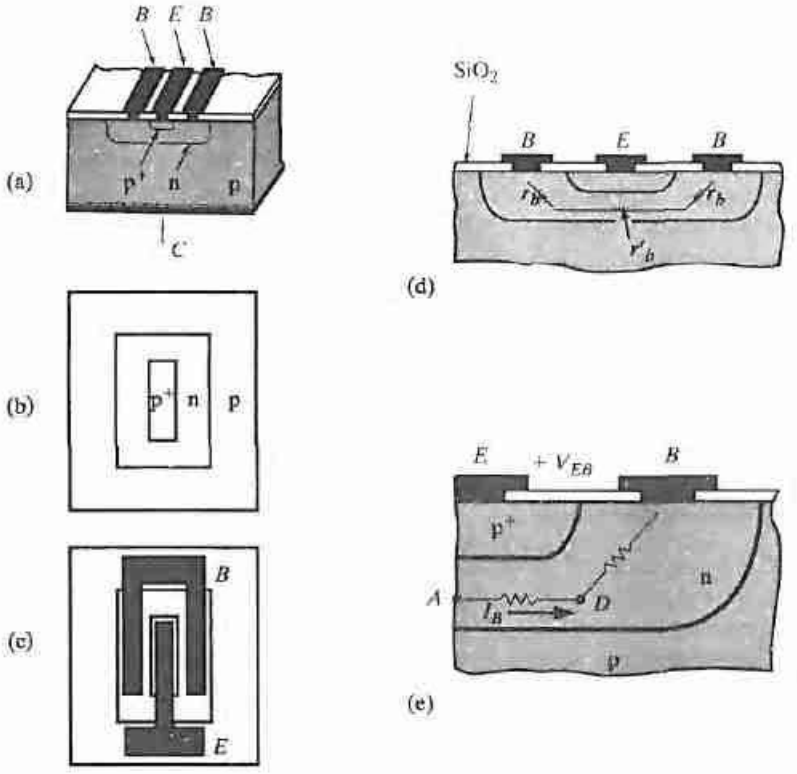


Figure 7–20
Effects of a base resistance: (a) cross section of an implanted transistor; (b) and (c) top view, showing emitter and base areas and metallized contacts; (d) illustration of base resistance; (e) expanded view of distributed resistance in the active part of the base region.

In particular, the forward bias of the emitter junction is largest at the corner of the emitter region near the base contact. We can see that this is the case by considering the simplified example of Fig. 7-20e. Neglecting variations in the base current along the path from point A to the contact B, the forward bias of the emitter junction above point A is approximately

$$V_{EA} = V_{EB} - I_B(R_{AD} + R_{DB}) \tag{7-53}$$

Actually, the base current is not uniform along the active part of the base region, and the distributed resistance of the base is more complicated than we have indicated. But this example does illustrate the point of nonuniform injection. Whereas the forward bias at A is approximately described by Eq. (7-53), the emitter bias voltage at point D is

$$V_{ED} = V_{EB} - I_B R_{DB} \tag{7-54}$$

which can be significantly closer to the applied voltage V_{EB}.

Since the forward bias is largest at the edge of the emitter, it follows that the injection of holes is also greatest there. This effect is called *emitter crowding*, and it can strongly affect the behavior of the device. The most important result of emitter crowding is that high-injection effects described in the previous section can become dominant locally at the corners of the emitter

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 375

before the overall emitter current is very large. In transistors designed to handle appreciable current, this is a problem which must be dealt with by proper structural design. The most effective approach to the problem of emitter crowding is to distribute the emitter current along a relatively large emitter edge, thereby reducing the current density at any one point. Clearly, what is needed is an emitter region with a large perimeter compared with its area. A likely geometry to accomplish this is a long thin stripe for the emitter, with base contacts on each side (Fig. 7-20b and c). With this geometry the total emitter current I_E is spread out along a rather long edge on each side of the stripe. An even better geometry is several emitter stripes, connected electrically by the metallization and separated by interspersing base contacts (Fig. 7-21). Many such thin emitter and base contact "fingers" can be interlaced to provide for handling large current in a power transistor. This is often called very descriptively an *interdigitated* geometry.

7.7.6 Gummel-Poon Model

The Ebers-Moll model runs into problems if a high degree of accuracy is required for very small BJTs, or where second order effects become important. The Gummel-Poon model, which is a charge control model, incorporates much more physics. We present a simplified version of the model here.

As discussed in Section 7.7.1, for typical graded doping profiles in the base, there is a built-in electric field that causes drift of minority carriers in the same direction they are diffusing from emitter to collector. This forms the starting point for the derivation of the Gummel-Poon model. As mentioned in Section 7.7.1, this electric field aids the motion of the minority carriers in the base, so the current can be written as

$$I_{Ep} = qA\mu_p p(x_n)\mathcal{E} - qAD_p \frac{dp(x_n)}{dx_n} \tag{7-55}$$

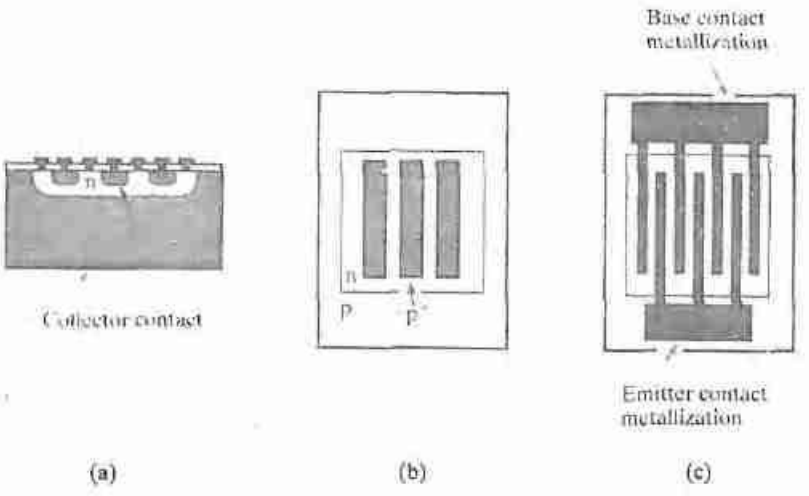


Figure 7-21
An interdigitated geometry to compensate for the effects of emitter crowding in a power transistor: (a) cross section; (b) top view of implanted regions; (c) top view with metallized contacts. The metal interconnections are isolated from the device by an oxide layer except where they contact the appropriate base and emitter regions at "windows" in the oxide.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 376

We can replace the electric field in the base in Eq. (7-55) by the expression in Eq. (7-47), assuming $n(x_n) = N_d(x_n)$

$$I_{Ep} = qA\mu_p p\left(\frac{-kT}{q}\frac{1}{n}\frac{dn}{dx_n}\right) - qAD_p\frac{dp}{dx_n}$$

$$= -\frac{qAD_p}{n}\left(p\frac{dn}{dx_n} + n\frac{dp}{dx_n}\right) \tag{7-56}$$

where the Einstein relation is used. We recognize the expression in parenthesis as the derivative of the pn product.

$$I_{Ep} = \frac{-qAD_p}{n}\frac{d(pn)}{dx_n} \tag{7-57a}$$

$$\frac{-I_{Ep}n}{qAD_p} = \frac{d(pn)}{dx_n} \tag{7-57b}$$

We integrate both sides of Eq. (7-57b) from the emitter-base junction (0) to the base-collector junction (W_b), keeping in mind that the current I_{Ep}, flowing from the emitter to the collector, is more or less constant in the narrow base (so that it can be pulled out of the integral).

$$-I_{Ep}\int_0^{W_b}\frac{ndx_n}{qAD_p} = \int_0^{W_b}\frac{d(pn)}{dx_n}dx_n = p(W_b)n(W_b) - p(0)n(0) \tag{7-58}$$

Now, as described in Sections 5.2.2 and 5.3.2, the pn product changes from the equilibrium value

$$pn = n_i^2 \tag{7-59a}$$

to the nonequilibrium expression

$$pn = n_i^2 e^{\frac{F_n - F_p}{kT}} = n_i^2 e^{\frac{qV}{kT}} \tag{7-59b}$$

where the separation of the Fermi levels is determined by the applied bias across the junction. Applying this to Eq. 7-58, we get

$$p(W_b)n(W_b) = n_i^2 e^{\frac{qV_{CB}}{kT}} \tag{7-60a}$$

$$p(0)n(0) = n_i^2 e^{\frac{qV_{EB}}{kT}} \tag{7-60b}$$

$$I_{Ep} = \frac{-qAD_p n_i^2\left(e^{\frac{qV_{CB}}{kT}} - e^{\frac{qV_{EB}}{kT}}\right)}{\int_0^{W_b} ndx_n} \tag{7-61}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 377

We have assumed a constant hole diffusivity in the base, D_p. The integral in the denominator corresponds to the integrated majority carrier charge in the base, and is known as the *base Gummel number*, Q_B. In the normal active mode of operation, where the collector-base junction is reverse biased (V_{CB} is negative), and the emitter-base junction is forward biased, the emitter hole current flowing to the collector (which is the dominant current) becomes

$$I_{Ep} = \frac{qAD_p n_i^2 e^{\frac{qV_{EB}}{kT}}}{Q_B} \tag{7-62a}$$

We can similarly write the base electron current flowing back into the emitter as

$$I_{En} = \frac{qAD_n n_i^2 e^{\frac{qV_{EB}}{kT}}}{Q_E} \tag{7-62b}$$

where Q_E is the integrated majority carrier charge in the emitter, known as the *emitter Gummel number*. The crux of the Gummel–Poon model is that the currents are expressed in terms of the net integrated charges in the base and emitter regions, and can easily handle nonuniform doping. Also, since we have obtained expressions for I_{Ep} and I_{En} in terms of Q_B and Q_E, we can write down BJT parameters such as the emitter injection efficiency γ [see Eq. (7–2)], in terms of the Gummel numbers.

We can also modify the Gummel–Poon model to handle several second-order effects such as the Early effect and high-level injection in the base, simply by writing the expression for the base Gummel number, Q_B, more precisely as follows:

$$Q_B = \int_{0(V_{EB})}^{W_b(V_{CB})} n(x_n)dx_n \tag{7-63}$$

where we explicitly account for the fact that the integration limits, the base-emitter junction ($x_n = 0$) and the base-collector junction (W_b) are bias dependent. This is, of course, the Early effect (Section 7.7.2).

Furthermore, we see that under high-level injection (Sections 5.6.1 and 7.7.4), the integrated majority carrier charge becomes greater than the integrated base dopant charge:

$$\int_0^{W_b} n(x_n)dx_n > \int_0^{W_b} N_D(x_n)dx_n \tag{7-64}$$

Clearly, from Eq. (7–61), this will cause the current from the emitter-to-collector, I_{Ep}, to increase less rapidly with emitter-base voltage at high biases. Based on what we learned in Section 5.6.1 about high-level injection in a diode, the emitter-to-collector current for high-level injection in the base increases as

$$I_C \propto I_{Ep} \propto e^{\frac{qV_{EB}}{2kT}} \tag{7-65a}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 378

On the other hand, because the emitter doping is typically higher than the base doping, one does not see high-level injection effects in the emitter, and the base current injected into the emitter scales as

$$I_B \propto I_{En} \propto e^{\frac{qV_{EB}}{kT}} \tag{7-65b}$$

Hence, for high V_{EB},

$$\beta = \frac{I_C}{I_B} \propto \frac{e^{\frac{qV_{EB}}{2kT}}}{e^{\frac{qV_{EB}}{kT}}} \propto e^{\frac{-qV_{EB}}{2kT}} \propto I_C^{-1}$$

This result shows that the common-emitter gain decreases at high injection levels due to excess majority carriers in the base.

The Gummel–Poon model also accounts for generation–recombination effects in the base-emitter depletion region at low current levels. As discussed in Section 5.6.2, such effects are accounted for by the diode ideality factor, **n**. Hence the base current injected into the emitter can be written as

$$I_B \propto I_{En} \propto e^{\frac{qV_{EB}}{nkT}} \tag{7-67a}$$

On the other hand, the generally large emitter current injected into the base is not likely to be affected by generation–recombination. Therefore,

$$I_{Ep} \propto e^{\frac{qV_{EB}}{kT}} \tag{7-67b}$$

Thus, for low V_{EB} or low I_C, the current gain

$$\beta = \frac{I_C}{I_B} \propto \frac{e^{\frac{qV_{EB}}{kT}}}{e^{\frac{qV_{EB}}{nkT}}} \propto e^{\frac{qV_{EB}}{kT}\left(1-\frac{1}{n}\right)} \propto I_C^{\left(1-\frac{1}{n}\right)} \tag{7-68}$$

The transistor collector current I_C and the base current I_B are plotted on a semilog scale as a function of V_{EB} in Fig. 7–22a. This is referred to as a *Gummel* plot. The current gain β is shown as a function of I_C in Fig. 7–22b. We see the dependence of β on I_C in the different bias regions, as described by the Gummel–Poon model. At low injection levels β is degraded by poor emitter injection efficiency [Eq. (7–68)], and at high currents β decreases due to excess majority charge in the base, which degrades γ. It can be shown that if one neglects the Early effect (i.e., we assume an infinite Early voltage) and the drop of β at high current levels, the Gummel–Poon model simplifies to the Ebers–Moll model.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 379

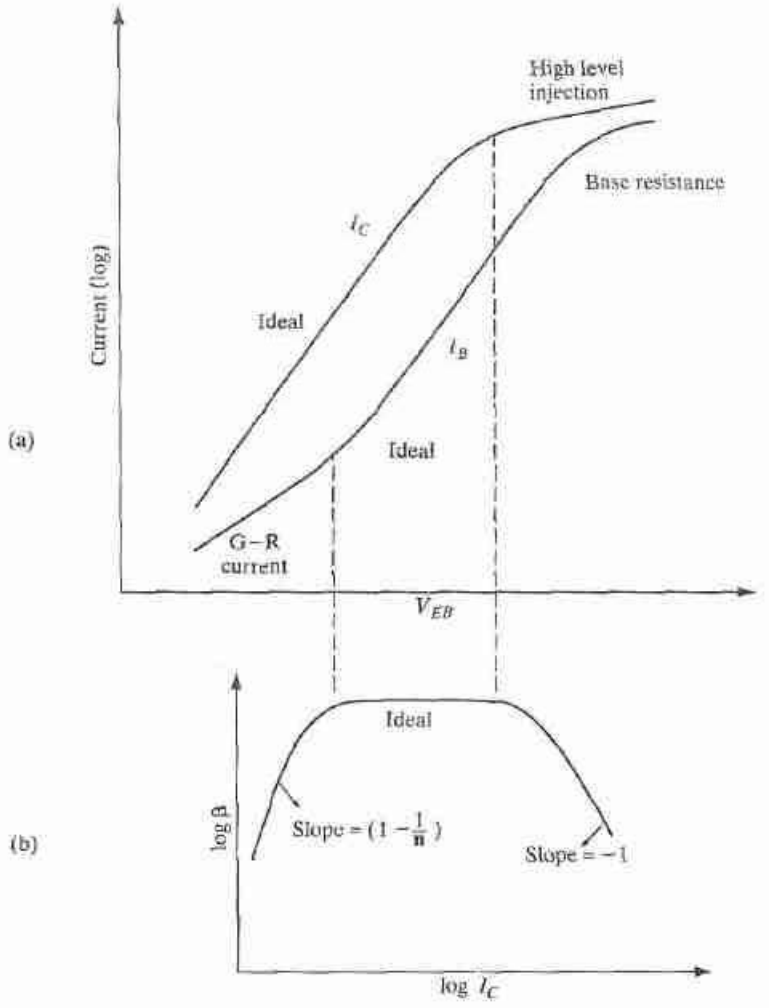


Figure 7–22 Current–voltage characteristics of BJTs: (a) Gummel plot of log of collector and base currents as a function of emitter-base forward bias; (b) d-c common-emitter current gain ($=I_C/I_B$) as a function of I_C. For the intermediate range, we have ideal behavior where both I_C and I_B increase exponentially with forward bias with ideality factor $n = 1$, leading to a current-independent gain. For very low forward biases, the generation–recombination (G–R) current increases I_B and the gain drops. For high forward biases, high-level injection effects cause the I_C to increase more slowly ($n = 2$) than I_B, and the gain drops as I_C^{-1}.

7.7.7 Kirk Effect

The current gain drops at high collector currents due to yet another mechanism known as the *Kirk effect*. This involves an effective widening of the neutral base due to modification of the depletion space charge distribution at the reverse-biased base-collector junction. This is caused by the buildup of mobile carriers due to increased current flow from the emitter to the collector. This is illustrated in Fig. 7-23 for a p-n-p BJT. Notice that the polarity of these mobile charges adds to the fixed donor charges on the base side of the base-collector depletion region, but subtracts from the fixed acceptor charges

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 380

on the collector side of the junction (Fig. 7-23c). Therefore, fewer uncompensated donors (and thus a smaller depletion width) are needed to maintain the reverse voltage V_{CB} across this junction. As a result, the neutral base width increases from W_b in Fig. 7-23b to W_b' in Fig. 7-23c. Also, the depletion region extends more into the collector side. This is tantamount to moving the base-collector junction deeper into the collector. This leads to an effective widening of the neutral base region (the Kirk effect) and to a drop of the current gain and an increase of the base transit time.

The electric field profile in the collector depletion region in the presence of uncompensated dopant charges and mobile carriers (due to the current flow) is given by Poisson's equation.

$$\frac{d\mathscr{E}}{dx} = \frac{1}{\epsilon}\left[q(N_d^+ - N_a^-) + \frac{I_c}{Av_d}\right] \tag{7-69}$$

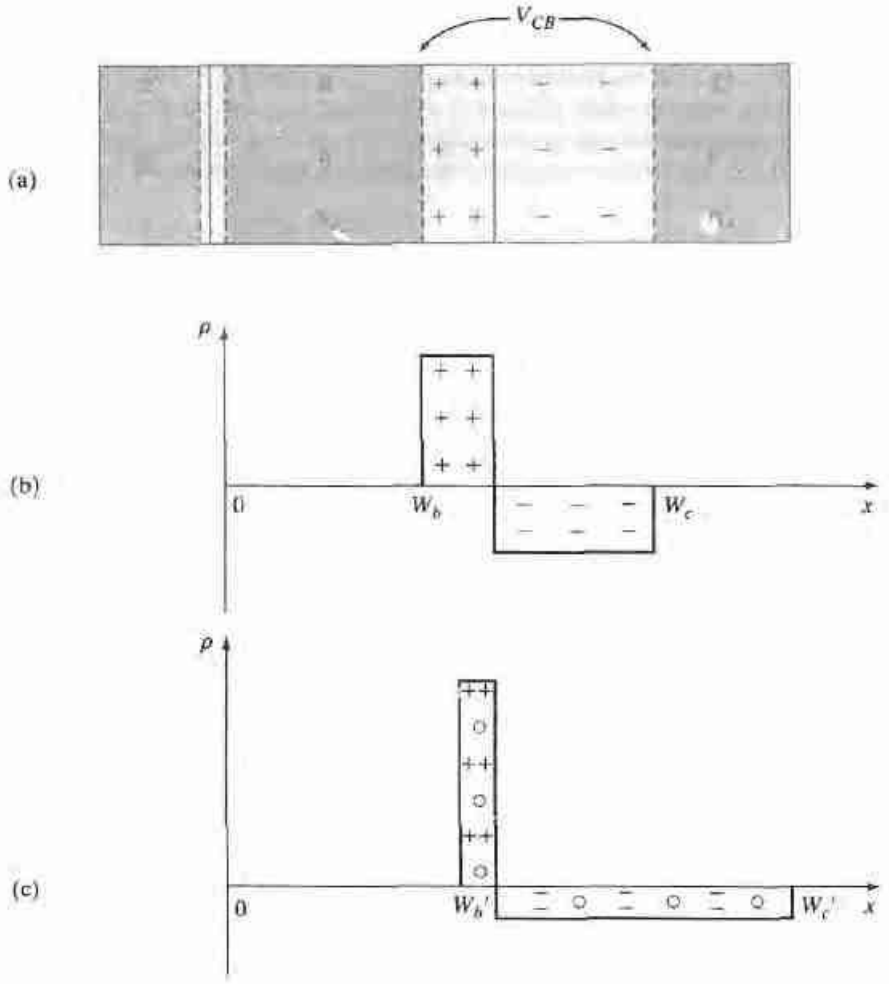


Figure 7-23
Kirk effect: (a) cross section of p-n-p BJT; (b) space-charge distribution in the base-collector reverse-biased junction for very low currents; (c) space-charge distribution at the base-collector junction for higher current levels. We see that the injected mobile holes (shown in color) add to the space charge of the immobile donors on the base side of the depletion region, but subtract from the space charge of the immobile acceptors on the collector side. This leads to a widening of the neutral base width from W_b to W_b'.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 381

where the mobile carrier charge concentration is given by the last term, and v_d is the drift velocity of the carriers.

The voltage across the reverse-biased collector-base junction, V_{CB}, is related to the electric field profile by:

$$V_{CB} = -\int_{W_b}^{W_c} \mathscr{E} dx \tag{7-70}$$

Assuming that V_{CB} is fixed, and I_C increases, the last term in Eq. (7-69) becomes more important with respect to the ionized dopant charges. In Poisson's equation [Eq. (7-69)], the extra holes injected into the depletion region have the same effect as if the doping level on the base side were increased and that on the collector side decreased. Since the integral of this field with respect to distance is fixed at V_{CB} [Eq. (7-70)], this implies that the depletion region on the base side collapses.

Although we have chosen to illustrate the Kirk effect for a p-n-p BJT, similar results are obtained for the n-p-n transistor. Obviously, the treatment is identical except for the polarity of the various charges. From a more detailed analysis, for n-p-n devices with an n^+ sub-collector, it can be shown that the base widening can extend at even higher current levels all the way through the lightly doped collector region to the heavily doped buried sub-collector.

7.8 FREQUENCY LIMITATIONS OF TRANSISTORS

In this section we discuss the properties of bipolar transistors under high-frequency operation. Some of the frequency limitations are junction capacitance, charging times required when excess carrier distributions are altered, and transit time of carriers across the base region. Our aim here is not to attempt a complete analysis of high-frequency operation, but rather to consider the physical basis of the most important effects. Therefore, we shall include the dominant capacitances and charging times and discuss the effects of the transit time on high-frequency devices.

7.8.1 Capacitance and Charging Times

The most obvious frequency limitation of transistors is the presence of junction capacitance at the emitter and collector junctions. We have considered this type of capacitance in Chapter 5, and we can include junction capacitors C_{je} and C_{jc} in circuit models for the transistor (Fig. 7-24a). If there is some equivalent resistance r_b between the base contact and the active part of the base region, we can also include it in the model, along with r_c to account for a series collector resistance.[6] Clearly, the combinations of r_b with C_{je} and r_c with C_{jc} can introduce important time constants into a-c circuit applications of the device.

[6]Since elements such as r_b and r_c in Fig. 7-24 are added to the basic transistor model we have previously analyzed, it is most convenient to refer here to the terminal voltages and currents as v_{be}, i_c, and so on. In this way we can use previously derived expressions involving the internal quantities (i_b and v_{eb} in Eq. (7-75), for example). In most circuits texts the primes are instead used for the internal quantities, just the opposite method from that used here.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 382

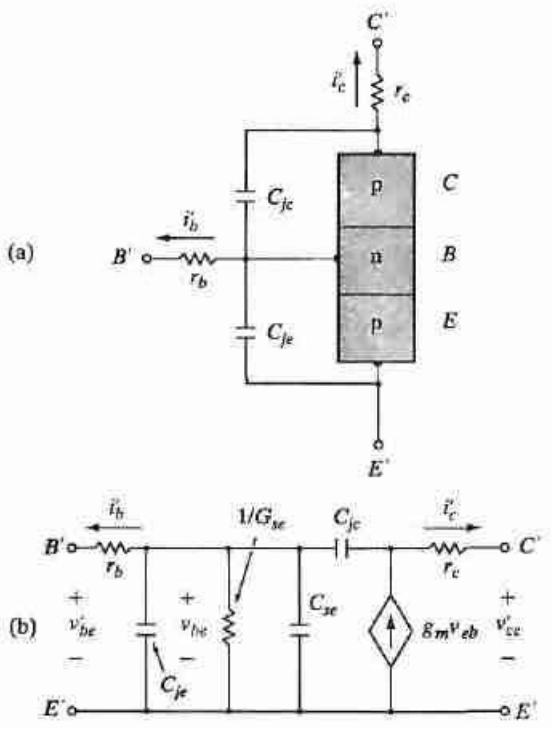


Figure 7–24
Models for a-c operation: (a) inclusion of base and collector resistances and junction capacitances; (b) hybrid-pi model synthesizing Eqs. (7–75) and (7–76).

From Section 5.5.4 we recall that capacitive effects can arise from the requirements of altering the carrier distributions during time-varying injection. In a-c circuits the transistor is usually biased to a certain steady state operating point characterized by the d-c quantities V_{BE}, V_{CE}, I_C, I_B, and I_E; and then a-c signals are superimposed upon these steady state values. We shall call the a-c terms v_{be}, v_{ce}, i_c, i_b, and i_e. Total (a-c + d-c) quantities will be lowercase with capitalized subscripts.

If a small a-c signal is applied to the emitter p-n junction along with a d-c level, we can show that

$$\Delta p_E(t) \approx \Delta p_E(\text{d-c})\left(1 + \frac{qv_{eb}}{kT}\right) \tag{7-71}$$

We can relate this time-varying excess hole concentration to the stored charge in the base region, and then use Eq. (7–43) to determine the resulting currents. For simplicity we shall assume the device is biased in the normal active mode and use only $Q_N(t)$. Assuming an essentially triangular excess hole distribution in the base, Eq. (7–23) gives

$$Q_N(t) = \tfrac{1}{2}qAW_b\Delta p_E(t) = \tfrac{1}{2}qAW_b\Delta p_E(\text{d-c})\left[1 + \frac{qv_{eb}}{kT}\right] \tag{7-72}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 383

The terms outside the brackets constitute the d-c stored charge $I_B\tau_p$:

$$Q_n(t) = I_B\tau_p\left(1 + \frac{qv_{eb}}{kT}\right) \tag{7-73}$$

Now that we have a simple relation for the time-dependent stored charge, we can use Eq. (7-43c) to write the total base current as

$$i_B(t) = \frac{Q_N(t)}{\tau_p} + \frac{dQ_N(t)}{dt} \tag{7-74a}$$

As discussed in Section 5.5.4, we must be careful about boundary conditions in determining where the stored charges are extracted or "reclaimed" in a diode. For the emitter-base diode in a BJT, we have a "short" diode where Eq. (5-64) is applicable. The result is that only 2/3 of the stored charge is reclaimed. Hence, we obtain

$$i_B(t) = I_B + \frac{q}{kT}I_B v_{eb} + \frac{2}{3}\frac{q}{kT}I_B\tau_p\frac{dv_{eb}}{dt} \tag{7-74b}$$

The a-c component of the base current is

$$i_b = G_{se}v_{eb} + C_{se}\frac{dv_{eb}(t)}{dt} \tag{7-75}$$

where

$$G_{se} \equiv \frac{q}{kT}I_B \quad \text{and} \quad C_{se} \equiv \frac{2}{3}\frac{q}{kT}I_B\tau_p = \frac{2}{3}G_{se}\tau_p$$

Thus, as in the case of the simple diode, an a-c conductance and capacitance are associated with the emitter-base junction due to charge storage effects. From Eq. (7-43b) we have

$$i_C(t) = \frac{Q_N(t)}{\tau_t} = \beta I_B + \frac{q}{kT}\beta I_B v_{eb}$$

$$i_c = g_m v_{eb} \quad \text{where } g_m \equiv \frac{q}{kT}\beta I_B = \frac{3}{2}\frac{C_{se}}{\tau_t} \tag{7-76}$$

The quantity g_m is an a-c *transconductance*, which is evaluated at the steady state value of collector current $I_C = \beta I_B$. We can synthesize Eqs. (7-75) and (7-76) in an equivalent a-c circuit as in Fig. 7-24b. In this equivalent circuit the voltage v_{be} used in the calculations appears "inside" the device, so that a new applied voltage v_{be} must be used external to r_b to refer to the voltage applied between the contacts, and similarly for v'_{ce}. This equivalent model is discussed in detail in most electronic circuits texts; it is often called

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 384

From Fig. 7-24b it is clear that several charging times are important in the a-c operation of a transistor; the most important are the time required to charge the emitter and collector depletion regions and the delay time in altering the charge distribution in the base region. Other delay times included in a complete analysis of high-frequency transistors are the transit time through the collector depletion region and the charge storage time in the collector region. If all of these are included in a single delay time τ_d, we can estimate the upper frequency limit of the device. This is usually defined as the *cutoff frequency* for the transistor $f_T \equiv (2\pi\tau_d)^{-1}$. It is possible to show that f_T represents the frequency at which the a-c amplification for the device $[\beta(a\text{-}c) \equiv h_{fe} = \partial i_c/\partial i_b]$ drops to unity.

7.8.2 Transit Time Effects

In high-frequency transistors the ultimate limitation is often the transit time across the base. For example, in a p-n-p device the time τ_t required for holes to diffuse from emitter to collector can determine the maximum frequency of operation for the device. We can calculate τ_t for a transistor with normal biasing and $\gamma = 1$ from Eq. (7-20) and the relation $\beta \simeq \tau_p/\tau_t$:

$$\beta \simeq \frac{\operatorname{csch} W_b/L_p}{\tanh W_b/2L_p} = \frac{2L_p^2}{W_b^2} = \frac{2D_p\tau_p}{W_b^2} = \frac{\tau_p}{\tau_t}$$

$$\tau_t = \frac{W_b^2}{2D_p} \tag{7-77}$$

Another instructive way of calculating τ_t is to consider that the diffusing holes *seem* to have an average velocity $\langle v(x_n)\rangle$ (actually the individual hole motion is completely random, as discussed in Section 4.4.1). The hole current $i_p(x_n)$ is then given by

$$i_p(x_n) = qAp(x_n)\langle v(x_n)\rangle \tag{7-78}$$

The transit time is

$$\tau_t = \int_0^{W_n} \frac{dx_n}{\langle v(x_n)\rangle} = \int_0^{W_n} \frac{qAp(x_n)}{i_p(x_n)} dx_n \tag{7-79}$$

For a triangular distribution as in Fig. 7-8b, the diffusion current is almost constant at $i_p = qAD_p\Delta p_E/W_b$, and τ_t becomes

$$\tau_t = \frac{qA\,\Delta p_E W_b/2}{qAD_p\Delta p_E/W_b} = \frac{W_b^2}{2D_p} \tag{7-80}$$

as before. The average velocity concept should not be pushed too far in the case of diffusion, but it does serve to illustrate the point that a delay time exists between the injection and collection of holes.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 385

We can estimate the transit time for a typical device by choosing a value of W_b, say 0.1 μm (10^{-5} cm). For Si, a typical number for D_p is about 10 cm²/s; then for this transistor $\tau_t = 0.5 \times 10^{-11}$ s. Approximating the upper frequency limit as $(2\pi\tau_t)^{-1}$, we can use the transistor to about 30 GHz. Actually, this estimate is too optimistic because of other delay times. The transit time can be reduced by making use of field-driven currents in the base. For the implanted transistor of Fig. 7–17, the holes drift in the built-in field from emitter to collector over most of the base region. By increasing the doping gradient in the base, we can reduce the transit time and thereby increase the maximum frequency of the transistor.

7.8.3 Webster Effect

While the transit time expression [Eq. (7-80)] is valid for low-level injection, τ_t is reduced by up to a factor of 2 under high-level injection. This occurs because the majority carrier concentration increases significantly above its equilibrium value in the base, to match the injected minority carrier concentration. Since the minority carrier concentration decreases from the base-emitter junction to the base-collector junction (see Fig. 7–8), so does the majority carrier concentration. This tends to create a diffusion of the majority carriers from emitter to base. Such majority carrier diffusion would upset the drift-diffusion balance required to maintain a quasi-equilibrium distribution in the base. Therefore, a built-in electric field develops in the base to create an opposing majority carrier drift current. The direction of this induced field then *aids* the minority carrier transport from the emitter to the collector, reducing the transit time τ_t in Eq. (7-80).

This is known as the Webster effect. It is interesting to note that this effect is similar to the drift field effects in the base region due to nonuniform base doping (Section 7.7.1). For the Webster effect, the induced field is not due to nonuniform doping but rather the nonuniform majority carrier concentration induced by carrier injection.

7.8.4 High-Frequency Transistors

The most obvious generality we can make about the fabrication of high-frequency transistors is that the physical size of the device must be kept small. The base width must be narrow to reduce the transit time, and the emitter and collector areas must be small to reduce junction capacitance. Unfortunately, the requirement of small size generally works against the requirements of power rating for the device. Since we usually require a trade-off between frequency and power, the dimensions and other design features of the transistor must be tailored to the specific circuit requirements. On the other hand, many of the fabrication techniques useful for power devices can be adapted to increase the frequency range. For example, the method of interdigitation (Fig. 7–21) provides a means of increasing the useful emitter edge length while keeping the overall emitter area to a minimum. Therefore, some form

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 386

of interdigitation is generally used in transistors designed for high frequency and reasonable power requirements.

Another set of parameters that must be considered in the design of a high-frequency device is the effective resistance associated with each region of the transistor. Since the emitter, base, and collector resistances affect the various RC charging times, it is important to keep them to a minimum. Therefore, the metallization patterns contacting the emitter and base regions must not present significant series resistance. Furthermore, the semiconductor regions themselves must be designed to reduce resistance. For example, the series base resistance r_b of an n-p-n device can be reduced greatly by performing a p^+ diffusion between the contact area on the surface and the active part of the base region. Further reduction of base resistance by heavy doping of the base requires the use of a heterojunction (Section 7.9) to maintain γ at an acceptable value.

In Si, n-p-n transistors are usually preferred, since the electron mobility and diffusion coefficient are higher than for holes. It is common to fabricate n-p-n transistors in n-type epitaxial material grown on an n^+ substrate. The heavily doped substrate provides a low-resistance contact to the collector region, while maintaining low doping in the epitaxial collector material to ensure a high breakdown voltage of the collector junction (see Section 7.3). It is important, however, to keep the collector depletion region as small as possible to reduce the transit time of carriers drifting through the collector junction. This can be accomplished by making the lightly doped collector region narrow so that the depletion region under bias extends to the n^+ substrate.

As described in Section 7.3, an n^+ polysilicon layer is generally deposited on top of the single-crystal n^+ emitter. The reason for this can be understood from Fig. 7-25. If an ohmic contact were formed directly on top of the narrow n^+ emitter, the excess minority carrier hole concentration profile would have a steep gradient because the excess minority carrier concentrations are zero at the ohmic contacts. This leads to a large back-injected base current

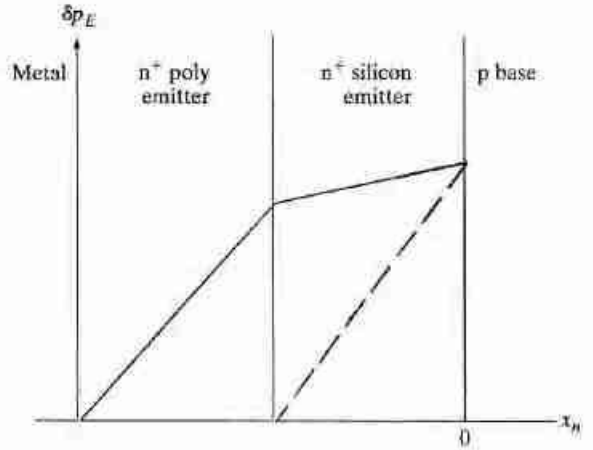


Figure 7–25
Excess minority hole concentration profile in the emitter region of an n-p-n transistor with and without the use of poly-emitters. The carrier profile if an ohmic contact is formed directly on the single-crystal Si emitter is shown by the dashed line.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 387

I_{Ep} into the emitter and a degraded emitter injection efficiency. The use of a polysilicon emitter causes the hole concentration profile to have a gentler slope near the base-emitter junction, thereby leading to an improved γ. The steeper slope of the hole concentration profile in the polysilicon emitter simply reflects the fact that the grain boundary defects in polysilicon lead to a shorter diffusion length.

In addition to the various parameters of the device itself, the transistor must be packaged properly to avoid parasitic resistance, inductance, or capacitance at high frequencies. We shall not attempt to describe the many techniques for mounting and packaging transistors here, since methods vary greatly among manufacturers.

7.9 HETEROJUNCTION BIPOLAR TRANSISTORS

In Section 7.4.4 we saw that the emitter injection efficiency of a bipolar transistor is limited by the fact that carriers can flow from the base into the emitter region, over the emitter junction barrier, which is reduced by the forward bias. According to Eq. (7-25) it is necessary to use lightly doped material for the base region and heavily doped material for the emitter to maintain a high value of γ and, therefore, α and β. Unfortunately, the requirement of light base doping results in undesirably high base resistance. This resistance is particularly noticeable in transistors with very narrow base regions. Furthermore, degenerate doping can led to a slight shrinkage of E_g in the emitter as the donor states merge with the conduction band. This can decrease the emitter injection efficiency. Therefore, a more suitable BJT for high frequency would have a heavily doped base and a lightly doped emitter. This is just the opposite of the traditional BJT discussed thus far in this chapter. To accomplish such a radically different transistor design, we need some other mechanism instead of doping to control the relative amount of injection of electrons and holes across the emitter junction.

If transistors are made in materials that allow heterojunctions to be used, the emitter injection efficiency can be increased without strict requirements on doping. In Fig. (7-26) an n-p-n transistor made in a single material (*homojunction*) is contrasted with a *heterojunction bipolar transistor (HBT)*, in which the emitter is a wider band gap semiconductor. It is possible in such a structure for the barrier for electron injection (qV_n) to be smaller than the hole barrier (qV_p). Since carrier injection varies exponentially with the barrier height, even a small difference in these two barriers can make a very large difference in the transport of electrons and holes across the emitter junction. Neglecting differences in carrier mobilities and other effects, we can approximate the dependence of carrier injection across the emitter as

$$\frac{I_n}{I_p} \propto \frac{N_d^E}{N_a^B} e^{\Delta E_g/kT} \tag{7-81}$$

In this expression, the ratio of electron current I_n to hole current I_p crossing the emitter junction is proportional to the ratio of the doping in the emitter

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 388

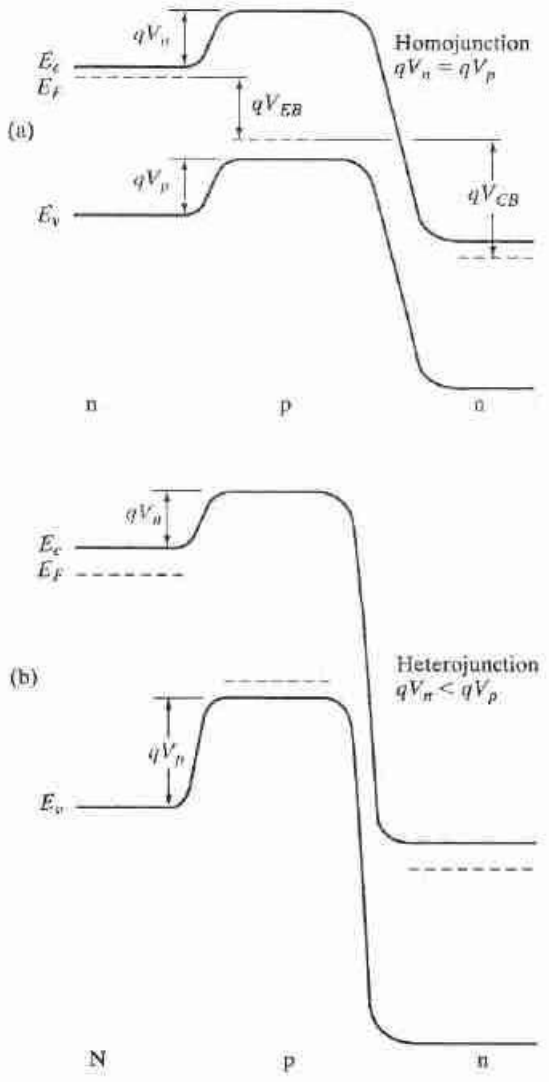


Figure 7–26
Contrast of carrier injection at the emitter of (a) a homojunction BJT and (b) a heterojunction bipolar transistor (HBT). In the forward-biased homojunction emitter, the electron barrier qV_n and the hole barrier qV_p are the same. In the HBT with a wide band gap emitter, the electron barrier is smaller than the hole barrier, resulting in the preferential injection of electrons across the emitter junction.

N_d^E and the base N_a^B. In the homojunction BJT this doping ratio is all we have to work with in designing a useful emitter junction. However, in the HBT there is an additional factor in which the band gap difference ΔE_g between the wide band gap emitter and the narrow band gap base appears in an exponential factor. As a result, a relatively small value of ΔE_g in the exponential term can dominate Eq. (7-81). This allows us to choose the doping terms for lower base resistance and emitter junction capacitance. In particular, we can choose a heavily doped base to reduce the base resistance and a lightly doped emitter to reduce junction capacitance.

The heterojunction shown in Fig. 7–26 has a smooth barrier, without the spike and notch commonly observed for heterojunctions (see Fig. 5-46).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 389

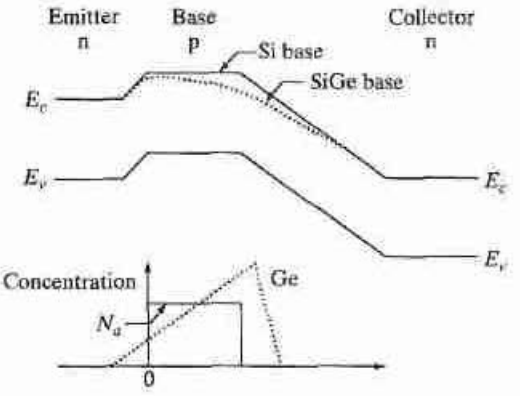


Figure 7–27
SiGe graded base in a Si n-p-n transistor. The built-in drift field in the conduction band in the base aids electron transport.

It is possible to smooth out such discontinuities in the bands by grading the composition of the ternary or quaternary alloy between the two materials (Fig. 7-27). Clearly, grading out the conduction band spike improves the electron injection by reducing the barrier that electrons must overcome. There are some HBT designs, however, that make use of the spike as a "launching ramp" to inject hot electrons into the base.

Materials commonly used in HBTs obviously include the AlGaAs/GaAs system because of its wide range of lattice-matched composition. In addition, the InGaAsP system (including $In_{0.53}Ga_{0.47}As$) grown on InP has become popular in HBT design. It is also possible to make HBTs using elemental semiconductor heterostructures such as $Si/Si_{1-x}Ge_x$. In this material system, the band gap difference ΔE_g between the Si emitter and the narrower band gap $Si_{1-x}Ge_x$ base occurs primarily in the valence band. As a result, a rather small addition of Ge to the base results in higher electron injection efficiency than is possible in homojunction silicon bipolar transistors.

If the SiGe alloy composition in the base of an n-p-n is varied such that E_g decreases slightly from the emitter side to the collector side of the base, a built-in electric field accelerates electrons through the base region (Fig. 7-27). The resulting field-aided base transport is a major advantage of the HBT. Such graded Ge profiles make the structure more manufacturable, because one does not have to worry about a mismatch between the p-n junction and the heterojunction.

SUMMARY

7.1 A bipolar junction transistor (BJT) consists of an emitter-base (EB) diode connected *back to back* to a collector–base (CB) diode. In the *normal active* mode of operation, the EB junction is forward biased such that carriers injected from the emitter can diffuse across the *narrow* base and be collected by the reverse-biased CB junction.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 390

7.2 The terminal currents can be determined by solving the drift–diffusion and continuity equations for a narrow-base diode with suitable boundary conditions. The fraction of the total emitter current that is due to injected carriers from emitter to base is the *emitter injection efficiency* and is made close to unity by doping the emitter much higher than the base. Another elegant way to increase emitter injection efficiency is to use a larger band gap emitter than in the base, as in heterojunction bipolar transistors (*HBT*).

7.3 The fraction of the carriers that survives diffusion across the narrow base is the *base transport factor*, which is made close to unity by making the base width narrow compared with the diffusion length. The product of emitter injection efficiency and base transport factor gives the current transfer ratio (I_C/I_E), from which one can determine the common-emitter current gain (I_C/I_B). *Active devices such as BJTs provide power gain.*

7.4 Equivalent-circuit models such as *Ebers–Moll* treat the BJT as *two coupled back-to-back diodes*, in parallel with controlled current sources. These diodes can be forward or reverse biased for switching applications. If they are both forward biased, the BJT is in a low impedance state (saturation); if they are both reverse biased, it is in a high impedance state (cutoff).

7.5 A *charge control analysis* of BJTs is possible (as for other devices) in terms of the integrated carrier concentration in the base. I_C is given by the base charge divided by the base transit time, while I_B is the base charge divided by the carrier lifetime (for unity emitter injection efficiency).

7.6 The *Gummel–Poon* charge control model gives the current components in terms of the integrated base and emitter dopant charges (Gummel numbers). It is more accurate and can handle some second-order effects better than Ebers–Moll.

7.7 Second-order effects include I_C increase due to neutral base narrowing for increasing CB reverse bias (*Early* effect), current gain drop at high I_C (*Kirk* effect), I_E crowding at the emitter edges due to series resistance effects, and BJT breakdown at high biases.

PROBLEMS

7.1 Given the data of Prob. 5.2, plot the doping profiles $N_a(x)$ and $N_d(x)$ for the following double-diffused transistor: The starting wafer is n-type Si with $N_d = 10^{16}$ cm^{-3}; $N_s = 5 \times 10^{19}$ cm^{-2} boron atoms are deposited on the surface, and these atoms are diffused into the wafer at 1100°C for 1 hr ($D = 3 \times 10^{-13}$ cm^2/sec for B in Si at 1100°C); then the wafer is placed in a phosphorus diffusion furnace at 1000°C for 15 min ($D = 3 \times 10^{-14}$ cm^2/sec for P in Si at 1000°C). During the emitter diffusion, the surface concentration is held constant at 5×10^{20} cm^{-3}. You may assume that the base doping profile does not change appreciably during the emitter diffusion, which takes place at a lower temperature and for a shorter time. Find the width of the base region from plots of $N_a(x)$ and $N_d(x)$. *Hint:* Use five-cycle semilog paper and let x vary from zero to about 1.5 μm in steps that are chosen to be simple multiples of $2\sqrt{Dt}$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 391

7.2 Sketch the ideal collector characteristics (i_C, $-v_{CE}$) for the transistor of Fig. 7-4; let i_B vary from zero to 0.2 mA in increments of 0.02 mA, and let $-v_{CE}$ vary from 0 to 10 V. Draw a load line on the resulting characteristics for the circuit of Fig. 7-4, and find the steady state value of $-V_{CE}$ graphically for $I_B = 0.1$ mA.

7.3 Calculate and plot the excess hole distribution $\delta p(x_n)$ in the base of a p-n-p transistor from Eq. (7-14), assuming $W_b/L_p = 0.5$. The calculations are simplified if the vertical scale is measured in units of $\delta p/\Delta p_E$ and the horizontal scale in units of x_n/L_p. In good transistors, W_b/L_p is much smaller than 0.5; however, $\delta p(x_n)$ is quite linear even for this rather large base width.

7.4 Redraw Fig. 7-3 for an n⁺-p-n BJT, and explain the various components of current flow and current directions for the normal active mode of operation. Draw the energy band diagram for equilibrium and this bias condition.

7.5 Assume that a p-n-p transistor is doped such that the emitter doping is 10 times that in the base, the minority carrier mobility in the emitter is one-half that in the base, and the base width is one-tenth the minority carrier diffusion length. The carrier lifetimes are equal. Calculate α and β for this transistor.

7.6 (a) In a BJT, we increase the base doping by a factor of 10 and halve the base width. *Calculate approximately* by what factor the collector current changes in the normal active mode, assuming that everything else stays the same. (b) In a certain BJT, the emitter doping is 100 times greater than the base doping, the emitter width is 0.1 times the base width, and we can assume both base and emitter widths to be much shorter than the carrier diffusion lengths L_n and L_p. What is the emitter injection efficiency? What is the base transport factor? (c) Suppose $L_n = L_p$. Now assuming that both emitter and base are much wider than the diffusion lengths, what is the emitter injection efficiency and base transport factor?

7.7 The symmetrical p⁺-n-p⁺ transistor of Fig. P7-7 is connected as a diode in the four configurations shown. Assume that $V \gg kT/q$. Sketch $\delta p(x_n)$ in the base region for each case. Which connection seems most appropriate for use as a diode? Why?

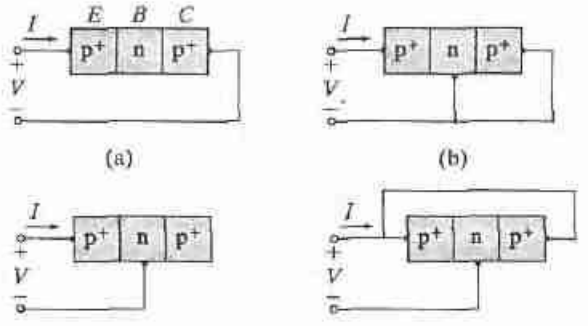


Figure P7-7

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 392

7.8 Derive Eq. (7–19) from the charge control approach by integrating Eq. (7–11) across the base region and applying Eq. (7–13). Now apply the charge control approach to the linear distribution in Fig. 7–8b to find the base transport factor B. Use Eq. (7–80) for the transit time. Is the result the same as Eq. (7–26) for small W_b/L_p?

7.9 (a) Find the expression for the current I for the transistor connection of Fig. P7–7b; compare the result with the narrow-base diode problem (Prob. 5.35).

(b) How does the current I divide between the base lead and the collector lead?

7.10 For a p-n-p BJT with $N_E > N_B > N_C$, show the dominant current components, with proper arrows, for directions in the normal active mode. If $I_{Ep} = 10$ mA, $I_{En} = 100$ μA, $I_{Cp} = 9.8$ mA, and $I_{Cn} = 1$ μA, calculate the base transport factor, emitter injection efficiency, common-base current gain, common-emitter current gain, and I_{CBO}. If the minority stored base charge is 4.9×10^{-11} C, calculate the base transit time and lifetime.

7.11 A p-n-p BJT has emitter (N_E), base (N_B), and collector (N_C) doping of 10^{20} cm^{-3}, 10^{18} cm^{-3}, and 10^{17} cm^{-3}, respectively, and a base width of 0.5 micron. Calculate the peak electric field at the CB junction, and the CB depletion capacitance per unit area for the normal active mode of operation with a $V_{CB} = 50$ V. How much is the neutral base width narrowing at this voltage, ignoring the EB depletion region? What impact does this have on the output characteristics of the BJT, and what is this effect called?

7.12 It is obvious from Eqs. (7–35) and (7–36) that I_{EO} and I_{CO} are the saturation currents of the emitter and collector junctions, respectively, with the opposite junction open circuited.

(a) Show that this is true from Eq. (7–32).

(b) Find expressions for the following excess concentrations: Δp_C with the emitter junction forward biased and the collector open; Δp_E with the collector junction forward biased and the emitter open.

(c) Sketch $\delta p(x_n)$ in the base for the two cases of part (b).

7.13 (a) Show that the definitions of Eq. (7–40) are correct; what does q_N represent?

(b) Show that Eqs. (7–39) correspond to Eqs. (7–34), using the definitions of Eq. (7–40).

7.14 (a) How is it possible that the average time an injected hole spends in transit across the base τ_t is shorter than the hole lifetime in the base τ_p?

(b) Explain why the turn-on transient of a BJT is faster when the device is driven into oversaturation.

7.15 Use Example 5–7 to design an n-p-n heterojunction bipolar transistor with reasonable γ and base resistance.

7.16 The current amplification factor β of a BJT is very sensitive to the base width as well as to the ratio of the base doping to the emitter doping. Calculate and plot β for a p-n-p BJT with $L_p^n = L_n^p$, for:

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 393

(a) $n_n = p_p$, $W_b/L_p^n = 0.01$ to 1;

(b) $W_b = L_p^n$, $n_n/p_p = 0.01$ to 1.

Neglect mobility variations ($\mu_n^p = \mu_p^n$).

7.17 (a) How much charge (in coulombs) due to excess holes is stored in the base of the transistor shown in Fig. 7–4 at the d-c bias given?

(b) Why is the base transport factor B different in the normal and inverted modes for the transistor shown in Fig. 7–5?

7.18 A Si p-n-p transistor has the following properties at room temperature:

$\tau_n = \tau_p = 0.1\ \mu s$

$D_n = D_p = 10\ cm^2/s$

$N_E = 10^{19}\ cm^{-3}$ = emitter concentration

$N_B = 10^{16}\ cm^{-3}$ = base concentration

$N_C = 10^{15}\ cm^{-3}$ = collector concentration

W_E = emitter width = 3 μm

W = metallurgical base width = 1.5 μm = distance between base-emitter junction and base-collector junction

A = cross-sectional area = $10^{-5}\ cm^2$

Calculate the neutral base width W_b for $V_{CB} = 0$ and $V_{EB} = 0.2$ V. Repeat for 0.6 V.

7.19 For the BJT in Prob. 7.18, calculate the base transport factor and the emitter injection efficiency for $V_{EB} = 0.2$ and 0.6 V.

7.20 For the BJT in Prob. 7.18, calculate α, β, I_E, I_B, and I_C for the two values of V_{EB}. What is the base Gummel number in each case?

7.21 A Si p-n-p BJT has the following parameters at room temperature.

Emitter	Base	Collector
$N_a = 5 \times 10^{18}\ cm^{-3}$	$N_d = 10^{16}$	$N_a = 10^{15}$
$\tau_n = 100$ ps	$\tau_p = 2500$ ps	$\tau_n = 2\ \mu s$
$\mu_n = 150\ cm^2/V\text{-}s$	$\mu_n = 1500$	$\mu_n = 1500$
$\mu_p = 100\ cm^2/V\text{-}s$	$\mu_p = 400$	$\mu_p = 450$

Base width $W_b = 0.2\ \mu m$

Area = $10^{-4}\ cm^2$

Calculate the β of the transistor from B and γ, and using the charge control model. Comment on the results.

7.22 For the BJT in Prob. 7.21, calculate the charge stored in the base when $V_{CB} = 0$ and $V_{EB} = 0.7$ V. If the base transit time is the dominant delay component for this BJT, what is the f_T?

7.23 Assume the transit time for electrons across the base of an n-p-n transistor is 100 ps, and electrons cross the 1-μm depletion region of the collector junction at their scattering limited velocity. The emitter-base junction charging time is

30 ps and the collector capacitance and resistance are 0.1 pF and 10 Ω, respectively. Find the cutoff frequency f_T.

7.24 An n-p-n Si transistor has an emitter doping of 10^{18} donors/cm^3 and a base doping of 10^{16} acceptors/cm^3. At what forward bias of the emitter junction does high-level injection occur (injected electrons equal to the base doping)? Comment on the emitter injection efficiency for electrons.

READING LIST

Very useful applets for understanding BJT operation are available at http://jas.eng.buffalo.edu/

Bardeen, J., and W. H. Brattain. "The Transistor, a Semiconductor Triode." *Phys. Rev.* 74 (1948), 230.

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

Neamen, D. A. *Semiconductor Physics and Devices: Basic Principles.* Homewood, IL: Irwin, 2003.

Neudeck, G. W. *Modular Series on Solid State Devices: Vol. III. The Bipolar Junction Transistor.* Reading, MA: Addison-Wesley, 1983.

Shockley, W. "The Path to the Conception of the Junction Transistor." *IEEE Trans. Elec. Dev.* ED-23 (1976), 597.

Sze, S. M. *Physics of Semiconductor Devices.* New York: Wiley, 1981.

Taur, Y., and T. H. Ning. *Fundamentals of Modern VLSI Devices.* Cambridge: Cambridge University Press, 1998.

SELF QUIZ

Question 1

Consider the following bipolar junction transistor (BJT) circuit and somewhat idealized transistor characteristics where, in particular, the voltage drop across the forward-biased base-emitter junction is assumed to be constant and equal to 1 V, for simplicity.

(a) What is the (common-emitter) gain β?

(b) Draw the load line on the transistor characteristics.

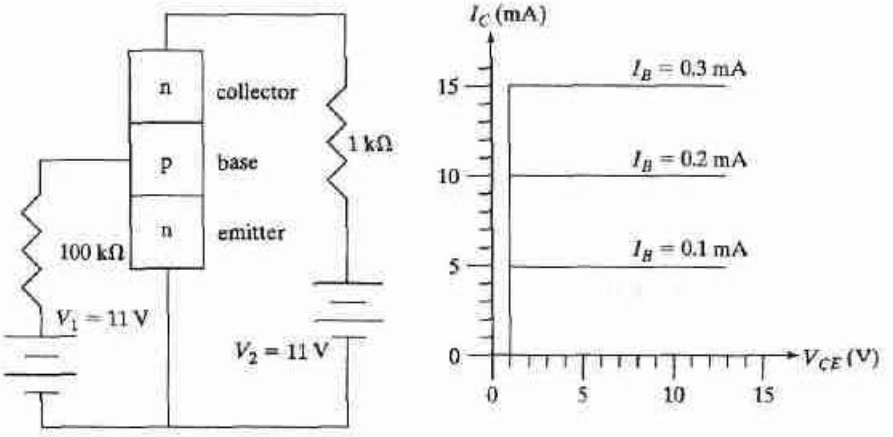
(c) What is the collector-emitter voltage drop in this circuit within half a volt?

(d) If voltage V_1 could be changed, what value of V_1 would drive the BJT in this circuit to the edge of saturation?

(e) For the given base current of 0.1 mA, consistent with your answer to part (b), what is the collector-emitter voltage drop, V_{CE}, in this circuit within half a volt?

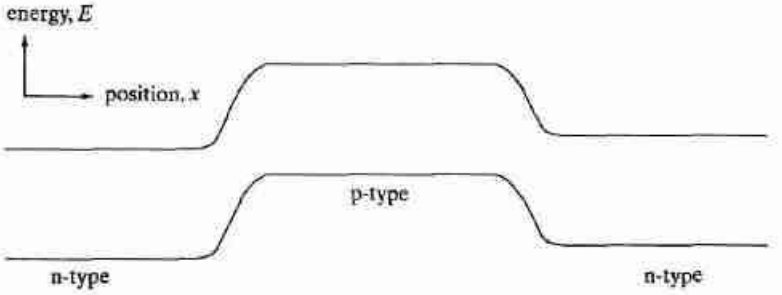
(f) What minimum value of I_B (assuming V_1 to be variable) would drive the BJT transistor into saturation?

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 395



Question 2

The following is a band diagram within an n-p-n bipolar junction transistor (BJT) at equilibrium. Sketch in the Fermi level as a function of position. Qualitative accuracy is sufficient.



Question 3

Would decreasing the base width of a BJT *increase*, *decrease*, or leave *unchanged* the following, assuming that the device remained unchanged otherwise? (Circle your choice for each.)

a) the emitter injection efficency γ? increase / unchanged / decrease

b) the base transport factor B? increase / unchanged / decrease

c) the (common-emitter) gain β? increase / unchanged / decrease

d) the *magnitude* of the Early voltage V_A? increase / unchanged / decrease

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 396

Question 4

Sketch the cross section of an n-p-n BJT and point out the dominant current components on it showing the correct directions of the various current vectors. If we increase the base doping, qualitatively explain how the various components change.

Question 5

Would decreasing the base doping of the BJT *increase*, *decrease*, or leave essentially *unchanged* (circle the correct answers):

(a) emitter injection efficiency γ? **increase / decrease / unchanged**

(b) base transport factor B? **increase / decrease / unchanged**

(c) magnitude of the Early voltage V_A? **increase / decrease / unchanged**

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 397

Chapter 8

Optoelectronic Devices

OBJECTIVES

1. Understand solar cells
2. Study photodetectors such as APDs
3. Study incoherent light sources (LEDs) and coherent light sources (lasers)

So far we have primarily concentrated on electronic devices. There is also a wide variety of very interesting and useful device functions involving the interaction of photons with semiconductors. These devices provide the optical sources and detectors that allow broadband telecommunications and data transmission over optical fibers. This important area of device applications is called *optoelectronics*. In this chapter we will discuss devices that detect photons and those that emit photons. Devices that convert optical energy into electrical energy include photodiodes and solar cells. Emitters of photons include incoherent sources such as light-emitting diodes (LEDs) and coherent sources in the form of lasers.

8.1 PHOTODIODES

In Section 4.3.4 we saw that bulk semiconductor samples can be used as photoconductors by providing a change in conductivity proportional to an optical generation rate. Often, junction devices can be used to improve the speed of response and sensitivity of detectors of optical or high-energy radiation. Two-terminal devices designed to respond to photon absorption are called *photodiodes*. Some photodiodes have extremely high sensitivity and response speed. Since modern electronics often involves optical as well as electrical signals, photodiodes serve important functions as electronic devices. In this section, we shall investigate the response of p-n junctions to optical generation of EHPs and discuss a few typical *photodiode detector* structures. We shall also consider the very important use of junctions as *solar cells*, which convert absorbed optical energy into useful electrical power.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 398

8.1.1 Current and Voltage in an Illuminated Junction

In Chapter 5 we identified the current due to drift of minority carriers across a junction as a generation current. In particular, carriers generated within the depletion region W are separated by the junction field, electrons being collected in the n region and holes in the p region. Also, minority carriers generated thermally within a diffusion length of each side of the junction diffuse to the depletion region and are swept to the other side by the electric field. If the junction is uniformly illuminated by photons with $h\nu > E_g$, an added generation rate g_{op} (EHP/cm^3-s) participates in this current (Fig. 8-1). The number of holes created per second within a diffusion length of the transition region on the n side is $AL_p g_{op}$. Similarly $AL_n g_{op}$ electrons are generated per second within L_n of x_{j0} and AWg_{op} carriers are generated within W. The resulting current due to collection of these optically generated carriers by the junction is

$$\boxed{I_{op} = qAg_{op}(L_p + L_n + W)} \tag{8-1}$$

If we call the thermally generated current described in Eq. (5-37b) I_{th}, we can add the optical generation of Eq. (8-1) to find the total reverse current with illumination. Since this current is directed from n to p, the diode equation [Eq. (5-36)] becomes

$$I = I_{th}(e^{qV/kT} - 1) - I_{op}$$

$$I = qA\left(\frac{L_p}{\tau_p}p_n + \frac{L_n}{\tau_n}n_p\right)(e^{qV/kT} - 1) - qAg_{op}(L_p + L_n + W) \tag{8-2}$$

Thus the I–V curve is lowered by an amount proportional to the generation rate (Fig. 8-1c). This equation can be considered in two parts—the current described by the usual diode equation, and the current due to optical generation.

When the device is short circuited ($V = 0$), the terms from the diode equation cancel in Eq. (8-2), as expected. However, there is a short-circuit current from n to p equal to I_{op}. Thus the I–V characteristics of Fig. 8-1c cross

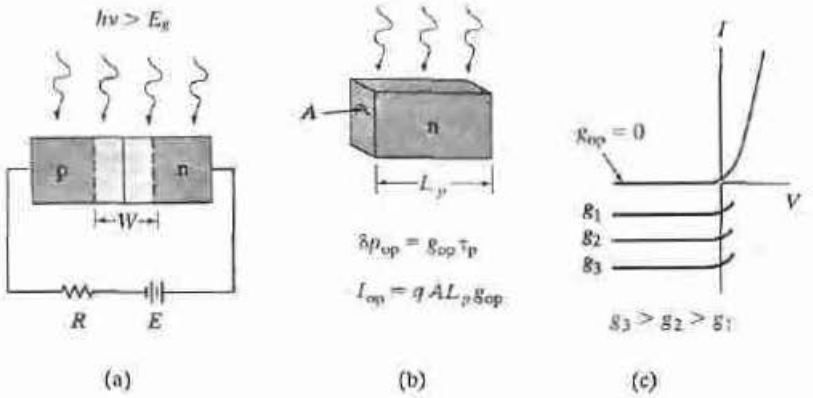


Figure 8-1
Optical generation of carriers in a p-n junction: (a) absorption of light by the device; (b) current I_{op} resulting from EHP generation within a diffusion length of the junction on the n side; (c) I–V characteristics of an illuminated junction.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 399

the I-axis at negative values proportional to g_{op}. When there is an open circuit across the device, $I = 0$ and the voltage $V = V_{oc}$ is

$$V_{oc} = \frac{kT}{q}\ln[I_{op}/I_{th} + 1]$$

$$= \frac{kT}{q}\ln\left[\frac{L_p + L_n + W}{(L_p/\tau_p)p_n + (L_n/\tau_n)n_p} \cdot g_{op} + 1\right] \qquad (8\text{-}3a)$$

For the special case of a symmetrical junction, $p_n = n_p$ and $\tau_p = \tau_n$, we can rewrite Eq. (6-5) in terms of the thermal generation rate $p_n/\tau_n = g_{th}$ and the optical generation rate g_{op}. Neglecting generation within W:

$$V_{oc} \simeq \frac{kT}{q}\ln\frac{g_{op}}{g_{th}} \quad \text{for } g_{op} \gg g_{th} \qquad (8\text{-}3b)$$

Actually, the term $g_{th} = p_n/\tau_n$ represents the *equilibrium* thermal generation–recombination rate. As the minority carrier concentration is increased by optical generation of EHPs, the lifetime τ_n becomes shorter, and p_n/τ_n becomes larger (p_n is fixed, for a given N_d and T). Therefore, V_{oc} cannot increase indefinitely with increased generation rate; in fact, the limit on V_{oc} is the equilibrium contact potential V_0 (Fig. 8-2). This result is to be expected, since the contact potential is the maximum forward bias that can appear across a junction. The appearance of a forward voltage across an illuminated junction is known as the *photovoltaic effect.*

Depending on the intended application, the photodiode of Fig. 8-1 can be operated in either the third or fourth quarters of its I-V characteristic. As Fig. 8-3 illustrates, power is delivered to the device from the external circuit when the current and junction voltage are both positive or both negative (first or third quadrants). In the fourth quadrant, however, the junction voltage is positive and the current is negative. In this case power is delivered from the junction to the external circuit (notice that in the fourth quadrant the current flows from the negative side of V to the positive side, as in a battery).

If power is to be extracted from the device, the fourth quadrant is used; on the other hand, in applications as a photodetector we usually reverse bias the junction and operate it in the third quadrant. We shall investigate these applications more closely in the discussion to follow.

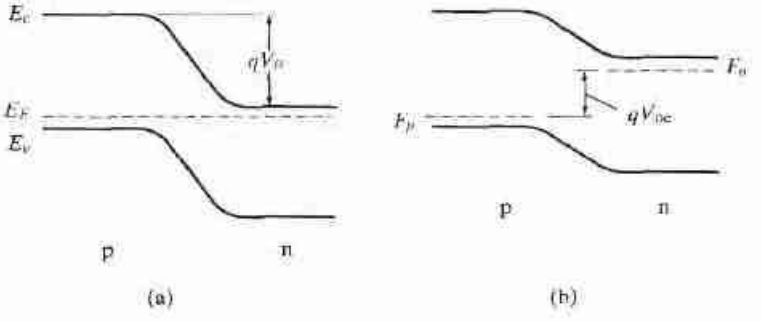


Figure 8-2 Effects of illumination on the open-circuit voltage of a junction: (a) junction at equilibrium; (b) appearance of a voltage V_{oc} with illumination.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 400

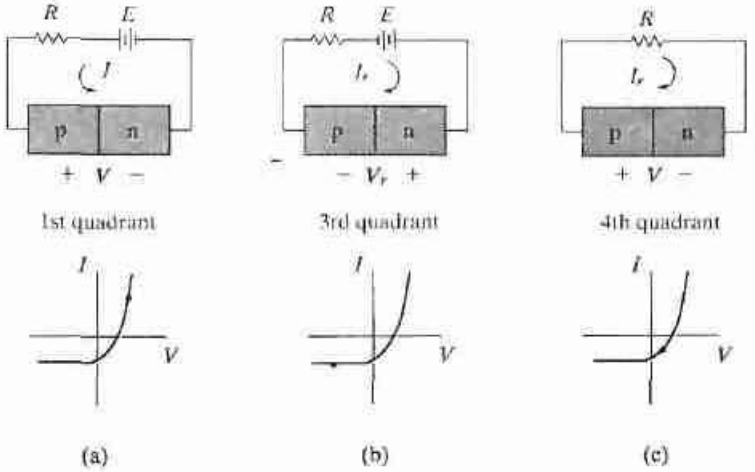


Figure 8–3 Operation of an illuminated junction in the various quadrants of its *I–V* characteristic; in (a) and (b), power is delivered to the device by the external circuit; in (c) the device delivers power to the load.

EXAMPLE 8–1

For steady state optical excitation, we can write the hole diffusion equation as

$$D_p \frac{d^2\delta p}{dx^2} = \frac{\delta p}{\tau_p} - g_{op}$$

Assume that a long p^+-n diode is uniformly illuminated by an optical signal, resulting in g_{op} EHP/cm^3-s. Calculate the hole diffusion current $I_p(x_n)$ and evaluate it at $x_n = 0$. Compare the result with Eq. (8-2) evaluated for a p^+-n junction.

SOLUTION

$$\frac{d^2\delta p}{dx_n^2} = \frac{\delta p}{L_p^2} - \frac{g_{op}}{D_p}$$

$$\delta p(x_n) = Be^{-x_n/L_p} + \frac{g_{op}L_p^2}{D_p}$$

At $x_n = 0$, $\delta p(0) = \Delta p_n$. Thus, $B = \Delta p_n - \dfrac{g_{op}L_p^2}{D_p}$

(a) $$\delta p(x_n) = [p_n(e^{qV/kT} - 1) - g_{op}L_p^2/D_p]e^{-x_n/L_p} + g_{op}L_p^2/D_p$$

$$\frac{d\delta p}{dx_n} = -\frac{1}{L_p}[\Delta p_n - g_{op}L_p^2/D_p]e^{-x_n/L_p}$$

(b) $$I_p(x_n) = -qAD_p\frac{d\delta p}{dx_n} = \frac{qAD_p}{L_p}[\Delta p_n - g_{op}L_p^2/D_p]e^{-x_n/L_p}$$

$$I_p(x_n = 0) = \frac{qAD_p}{L_p} p_n(e^{qV/kT} - 1) - qAL_p g_{op}$$

The last equation corresponds to Eq. (8–2) for $n_p \ll p_n$, except that the component due to generation on the p side is not included.

8.1.2 Solar Cells

Since power can be delivered to an external circuit by an illuminated junction, it is possible to convert solar energy into electrical energy. If we consider the fourth quadrant of Fig. 8-3c, it appears doubtful that much power can be delivered by an individual device. The voltage is restricted to values less than the contact potential, which in turn is generally less than the band gap voltage E_g/q. For Si the voltage V_{oc} is less than about 1 V. The current generated depends on the illuminated area, but typically I_{op} is in the 10–100 mA range for a junction with an area of about 1 cm^2. However, if many such devices are used, the resulting power can be significant. In fact, arrays of p-n junction solar cells are currently used to supply electrical power for many space satellites. Solar cells can supply power for the electronic equipment aboard a satellite over a long period of time, which is a distinct advantage over batteries. The array of junctions can be distributed over the surface of the satellite or can be contained in solar cell "paddles" attached to the main body of the satellite (Fig. 8–4).

To utilize a maximum amount of available optical energy, it is necessary to design a solar cell with a large area junction located near the surface of the device (Fig. 8–5). The planar junction is formed by diffusion or ion implantation, and the surface is coated with appropriate materials to reduce reflection and to decrease surface recombination. Many compromises must be made in solar cell design. In the device shown in Fig. 8-5, for example, the junction depth d must be less than L_p in the n material to allow holes generated near the surface to diffuse to the junction before they recombine; similarly, the thickness of the p region must be such that electrons generated in this region can diffuse to the junction before recombination takes place. This requirement implies a proper match between the electron diffusion length L_n, the thickness of the p region, and the mean optical penetration depth $1/\alpha$ [see Eq. (4–2)]. It is desirable to have a large contact potential V_0 to obtain a large photovoltage, and therefore heavy doping is indicated; on the other hand, long lifetimes are desirable and these are reduced by doping too heavily. It is important that the series resistance of the device be very small so that power is not lost to heat due to ohmic losses in the device itself. A series resistance of only a few ohms can seriously reduce the output power of a solar cell (Prob. 8.4). Since the area is large, the resistance of the p-type body of the device can be made small. However, contacts to the thin n region require special design. If this region is contacted at the edge, current must flow along the thin n region to the contact, resulting in a large series resistance.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 402

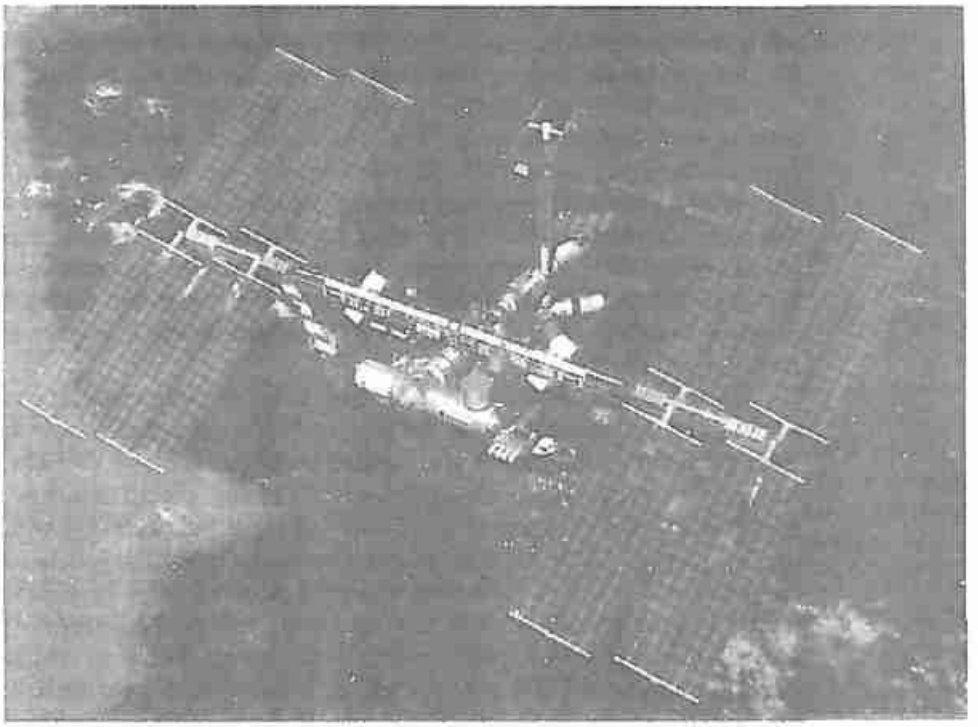


Figure 8–4
Solar cell arrays attached to the International Space Station. The solar array wings measure 74 m tip to tip. Each wing contains 32,800 solar cells and can produce 62 kW of power for the station. (Provided through the courtesy of the National Aeronautics and Space Administration.)

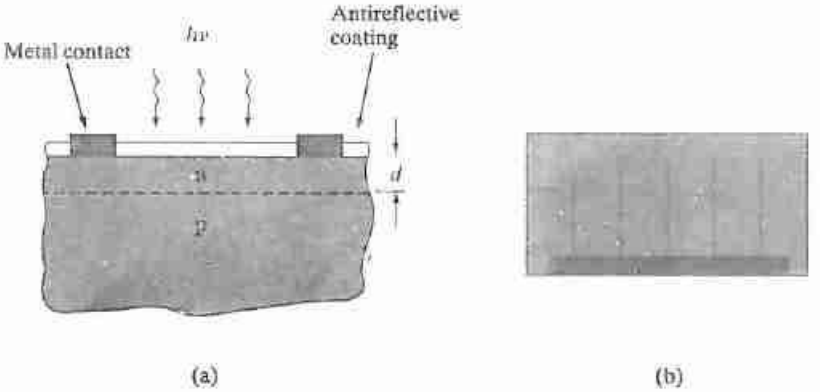


Figure 8–5
Configuration of a solar cell: (a) enlarged view of the planar junction; (b) top view, showing metal contact "fingers."

To prevent this effect, the contact can be distributed over the n surface by providing small contact fingers as in Fig. 8-5b. These narrow contacts serve to reduce the series resistance without interfering appreciably with the incoming light.

Figure 8-6 shows the fourth-quadrant portion of a solar cell characteristic, with I_r plotted upward for convenience of illustration. The open-circuit voltage V_{oc} and the short-circuit current I_{sc} are determined for a given light level by the cell properties. The maximum power delivered to a load by this solar cell occurs when the product VI_r is a maximum. Calling these values of voltage and current V_m and I_m, we can see that the maximum delivered power illustrated by the shaded rectangle in Fig. 8-6 is less than the $I_{sc}V_{oc}$ product. The ratio $I_mV_m/I_{sc}V_{oc}$ is called the *fill factor*, and is a figure of merit for solar cell design.

Applications of solar cells are not restricted to outer space. It is possible to obtain useful power from the sun in terrestrial applications using solar cells, even though the solar intensity is reduced by the atmosphere. Current worldwide total power generation capability is ~15 TW and corresponds to an annual energy usage of ~500 quads (1 quad = 10^{15}, or a quadrillion, BTU), with an annual increase of 1–2%. Of this, ~80% comes from fossil fuels (oil, natural gas, and coal), which will be exhausted in several hundred years. In addition, CO_2 emissions and global warming have spurred a renewed interest in "green" sources of energy such as photovoltaics, even though the current installed capacity is only about 1 GW. About 1 kW/m^2 is available in a particularly sunny location (translating to about 600 TW potentially available worldwide), but not all of this solar power can be converted to electricity. Much of the photon flux is at energies less than the cell band gap and is not absorbed. High-energy photons are strongly absorbed, and the resulting EHPs may recombine at the surface. A well-made single-crystal Si cell can have about 25% efficiency for solar energy conversion, providing approximately 250 W/m^2 of electrical power under full illumination. This is a modest amount of power per unit solar cell area, considering the cost and effort involved in fabricating a large area of Si cells. Amorphous Si thin-film solar cells can be made more cheaply, but have lower efficiencies (~10%) because of the defects in the material.

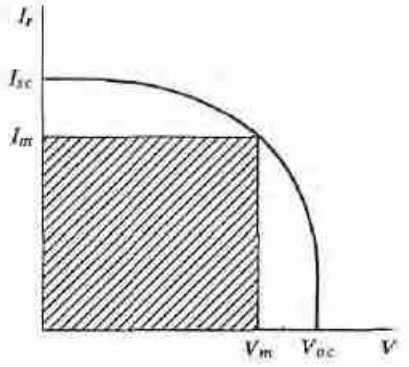


Figure 8–6
I–V characteristics of an illuminated solar cell. The maximum power rectangle is shaded.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 404

The cost and scalability of photovoltaic technology are of paramount importance for terrestrial applications. Currently, it costs only about 3 cents per kWh for electricity generation from fossil fuels, but about 10 times that amount from amorphous Si solar cells, and the time to recover the investment in photovoltaics is about four years. In terms of scalability, at 10% cell efficiency, approximately 3% of the land area would have to be covered with solar cells to meet U.S. energy needs, which would, of course, create other environmental problems. One approach to obtaining more power per cell is to focus considerable light onto the cell with the use of mirrors. Although Si cells lose efficiency at the resulting high temperatures, GaAs and related compounds can be used at 100°C or higher. In such solar concentrator systems, more effort and expense can be put into the solar cell fabrication, since fewer cells are required. For example, a GaAs–AlGaAs heterojunction cell provides good conversion efficiency and operates at the elevated temperatures that are common in solar concentrator systems. In such systems, the area required is shifted from the cell to the concentrator.

EXAMPLE 8–2

A Si solar cell has a short-circuit current of 100 mA and an open-circuit voltage of 0.8 V under full solar illumination. The fill factor is 0.7. What is the maximum power delivered to a load by this cell?

SOLUTION

$$P_{max} = (f.f.)\, I_{sc} V_{oc} = (0.8)(100)(0.7) = \mathbf{56\ mW}$$

8.1.3 Photodetectors

When the photodiode is operated in the third quadrant of its I–V characteristic (Fig. 8-3b), the current is essentially independent of voltage but is proportional to the optical generation rate. Such a device provides a useful means of measuring illumination levels or of converting time-varying optical signals into electrical signals.

In most optical detection applications the detector's speed of response, or bandwidth, is critical. For example, if the photodiode is to respond to a series of light pulses 1 ns apart, the photogenerated minority carriers must diffuse to the junction and be swept across to the other side in a time much less than 1 ns. The carrier diffusion step in this process is time consuming and should be eliminated if possible. Therefore, it is desirable that the width of the depletion region W be large enough so that most of the photons are absorbed within W rather than in the neutral p and n regions. When an EHP is created in the depletion region, the electric field sweeps the electron to the n side and the hole to the p side. Since this carrier drift occurs in a very short time, the response of the photodiode can be quite fast. When the carriers are generated primarily within the depletion layer W, the detector is called a *depletion layer photodiode*. Obviously, it is desirable to dope at least one side of the junction lightly so that W can be made large. The appropriate width

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 405

for W is chosen as a compromise between sensitivity and speed of response. If W is wide, most of the incident photons will be absorbed in the depletion region, leading to a high sensitivity. Also, a wide W results in a small junction capacitance [see Eq. (5-62)], thereby reducing the RC time constant of the detector circuit. On the other hand, W must not be so wide that the time required for drift of photogenerated carriers out of the depletion region is excessive, leading to low bandwidth.

One convenient method of controlling the width of the depletion region is to build a *p-i-n photodetector* (Fig. 8–7). The "i" region need not be truly intrinsic, as long as the resistivity is high. It can be grown epitaxially on the n-type substrate, and the p region can be obtained by implantation. When this device is reverse biased, the applied voltage appears almost entirely across the i region. If the carrier lifetime within the i region is long compared with the drift time, most of the photogenerated carriers will be collected by the n and p regions. An important figure of merit for a photodetector is the external quantum efficiency η_Q, defined as the number of carriers that are collected for every photon impinging on the detector. For a photocurrent density J_{op}, we collect J_{op}/q carriers per unit area per second. For an incident optical power density of P_{op}, the number of photons shining on the detector per unit area per second is $P_{op}/h\nu$. Therefore,

$$\eta_Q = (J_{op}/q)/(P_{op}/h\nu) \tag{8-4}$$

For a photodiode that has no current gain, the maximum η_Q is unity. If low-level optical signals are to be detected, it is often desirable to operate the photodiode in the avalanche region of its characteristic. In this mode each photogenerated carrier results in a significant change in the current because of avalanche multiplication, leading to *gain* and external quantum efficiencies of greater than 100%. *Avalanche photodiodes* (APDs) are useful as detectors in fiber-optic systems (Section 8.2.2).

The type of photodiode described here is sensitive to photons with energies near the band gap energy (*intrinsic* detectors). If $h\nu$ is less than E_g, the photons will not be absorbed; on the other hand, if the photons are much more energetic than E_g, they will be absorbed very near the surface, where the

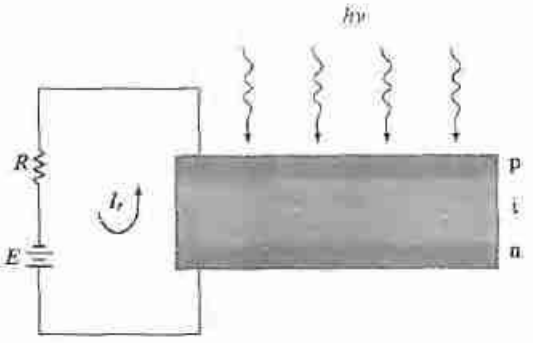


Figure 8–7
Schematic representation of a p-i-n photodiode.

recombination rate is high. Therefore, it is necessary to choose a photodiode material with a band gap corresponding to a particular region of the spectrum. Detectors sensitive to longer wavelengths can be designed such that photons can excite electrons into or out of impurity levels (*extrinsic* detectors). However, the sensitivity of such extrinsic detectors is much less than intrinsic detectors, where electron–hole pairs are generated by excitation across the band gap.

By using lattice-matched multilayers of compound semiconductors, the band gap of the absorbing region can be tailored to match the wavelength of light being detected. Wider band gap material can then be used as a window through which the light is transmitted to the absorbing region (Fig. 8-8). For example, we saw in Fig. 1-13 that InGaAs with an In mole fraction of 53% can be grown epitaxially on InP with excellent lattice-matching. This composition of InGaAs has a band gap of about 0.75 eV, which is sensitive to a useful wavelength for fiber-optic systems (1.55 μm), as we shall see in Section 8.2.2. In making a photodiode using InGaAs as the active material, it is possible to bring the light through the wider band gap $In_{0.52}Al_{0.48}As$ (also lattice-matched to InP), thus greatly reducing surface recombination effects. In the case of avalanche photodiodes requiring narrow-band-gap material, it is often advantageous to absorb the light in the narrow-gap semiconductor (e.g., InGaAs) and transport the resulting carriers to a junction made in wider band gap material (e.g., InAlAs), where the avalanche multiplication takes place at high fields. Such separation of the absorption and multiplication (SAM) APDs avoids the excessive leakage currents typical of reverse-biased junctions in narrow-gap materials. In this particular structure, there is also a doped InAlAs *charge* layer, leading to a SACM APD, which helps optimize (decrease) the electric field further between the multiplication and absorption regions. In certain APDs, one can grade the alloy composition between the high-band-gap multiplication regions and the lower-band-gap absorption region to avoid any band-edge discontinuities, which can trap photogenerated carriers. The photocurrent and dark current both increase with bias because of avalanche multiplication (Fig. 8-8b). One obviously wants to maximize the difference ΔI between the photocurrent I_p and the dark current I_d. The ratio of ΔI at different voltages to that at a low reference voltage is defined as the *gain* of the APD.

8.1.4 Gain, Bandwidth, and Signal-to-Noise Ratio of Photodetectors

In optical communication systems, the sensitivity (which depends on gain) of the photodetector and its response time (bandwidth) are of critical importance. Typical gain–bandwidth characteristics of such a SACM APD are shown in Fig. 8-8c and are limited by the transit time of carriers through the structure. Unfortunately, designs that increase gain tend to decrease bandwidth and vice versa. It is common to express the *gain–bandwidth product* as a figure of merit for detectors. In a p-i-n diode, there is no gain mechanism, since at most one electron–hole pair is collected by the junction for each photon absorbed. Thus, the gain is essentially unity, and the gain–bandwidth

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 407

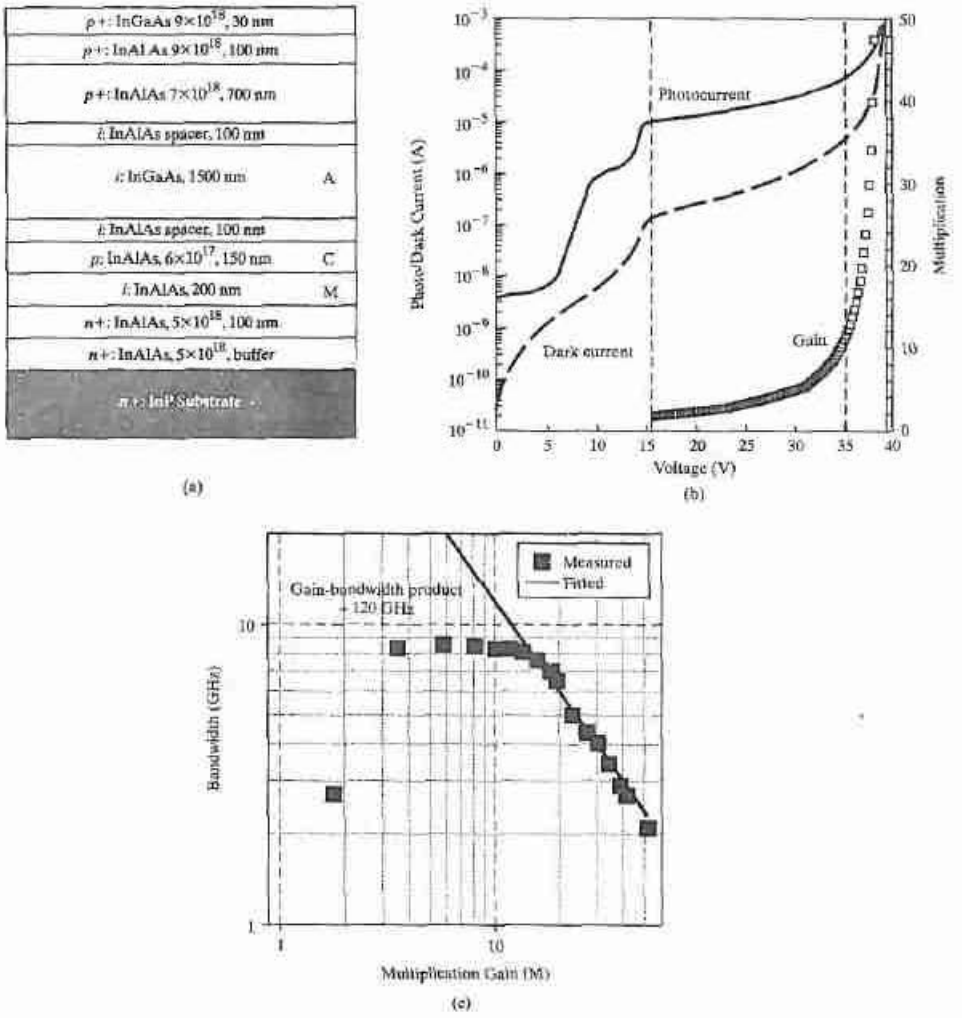


Figure 8–8
Use of multilayer heterojunctions to enhance the photodiode operation: (a) an avalanche photodiode in which light near 1.55 μm is absorbed in a narrow-band-gap material (InGaAs, E_g = 0.75 eV) after passing through a wider-gap material (InP and InAlAs); holes are swept to an InAlAs junction, where avalanche multiplication takes place. The *i* regions are lightly doped; (b) photocurrent, dark current, and gain increasing as a function of bias because of avalanche multiplication; (c) typical gain–bandwidth characteristics of such a SACM APD. [After X. Zheng, J. Hsu, J. Hurst, X. Li, S. Wang, X. Sun, A. Holmes, J. Campbell, A. Huntington, and L. Coldren, *IEEE J. Quant. Elec.*, 40(8), pp. 1068–1073, Aug. 2004.]

product is determined by the bandwidth, or frequency response. In a p-i-n, the response time is dependent on the width of the depletion region.

Another important property of detectors is the *signal-to-noise ratio*, which is the amount of usable information compared with the background noise in the detector. In the case of photoconductors, the major source of noise is random thermal motion of the carriers, leading to fluctuations in the dark current (called *Johnson noise*). The noise current increases with temperature ($\sim kT$) and with the conductance of the material in the dark. Therefore, the photoconductor noise at a given temperature can be reduced by increasing the dark resistance. Yet another source of noise at low frequencies is *1/f* or flicker noise, due to carrier trapping and detrapping at defects.

In a p-i-n diode, the dark current is smaller and the dark resistance much higher than in a photoconductor, and the main source of noise is random thermal generation–recombination of EHPs within this region (called *shot noise*). The shot noise is ultimately due to the quantization of the charge on electrons and holes. The noise in a p-i-n device is considerably lower than that in a photoconductor, as well as in APDs, for reasons described next.

Avalanche photodiodes have the advantage of providing gain through the avalanche multiplication effect. The disadvantage is increased noise relative to the p-i-n, due to random fluctuations in the avalanche process. This noise is reduced if the impact ionization in the high-field region is due to only one type of carrier, since more fluctuations in the ionization process occur when both electrons and holes participate. In Si, the ability of electrons to create EHPs in an impact ionization event is much higher than for holes. Therefore, Si APDs can be operated with high gain and relatively low noise. Unfortunately, Si APDs cannot be used for most fiber-optic transmission because Si is transparent at the wavelengths of low loss and low dispersion (λ = 1.55 and 1.3 μm) for optical fibers. For these longer wavelengths, $In_{0.53}Ga_{0.47}As$ has become the material of choice. However, the ionization rates of electrons and holes in most compound semiconductors are comparable, which degrades their noise and frequency response relative to Si APDs.

The various sources of noise determine the signal-to-noise (S/N) ratio in a photodetector. One quantifies it as a noise-equivalent power (NEP), which is the minimum detectable signal that would produce the same rms output as the noise. The detectivity of the photodetector is then defined as D = 1/NEP.

The NEP depends on the area of the photodetector as well as the bandwidth. The specific detectivity, D^*, is then defined as that for a detector of unit area and a bandwidth of 1 Hz. Clearly, once the bandwidth requirements are met, it is desirable to choose a photodetector with the highest D^*.

Another approach that has demonstrated excellent performance in terms of both high sensitivity and bandwidth utilizes a waveguide structure (Fig. 8-9). Here, unlike the situation with the p-i-n photodiode (Fig. 8-7) or the APD (Fig. 8-8), the light strikes the photodiode perpendicular to the current transport. The advantage is that now the absorption region can be made quite long, along the photon path, leading to high sensitivity, while at

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 409

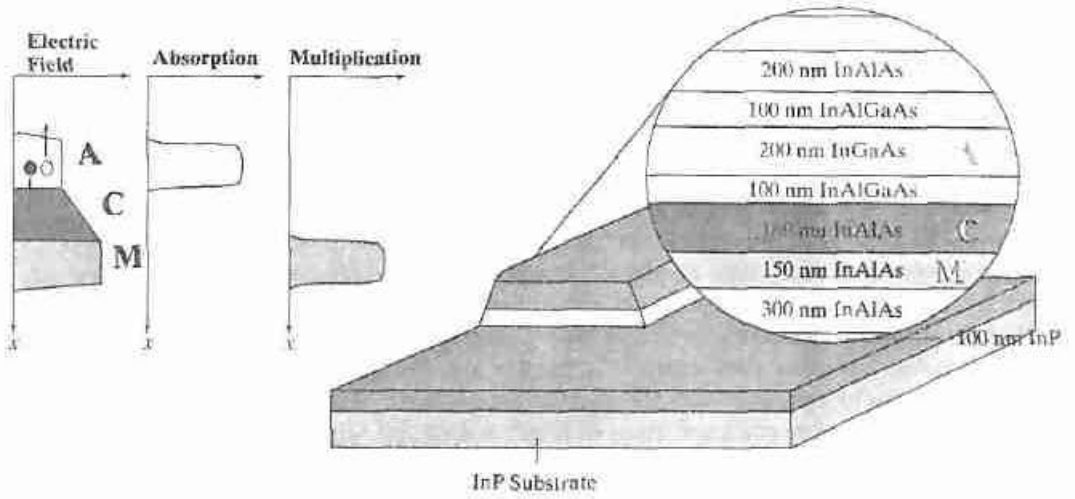


Figure 8–9
Schematic of a waveguide photodiode. The photons are strongly absorbed in the narrow-band-gap InGaAs region A, and carriers are multiplied by the avalanche process in region M. The charge region C helps optimize the electric field profile between A and M.

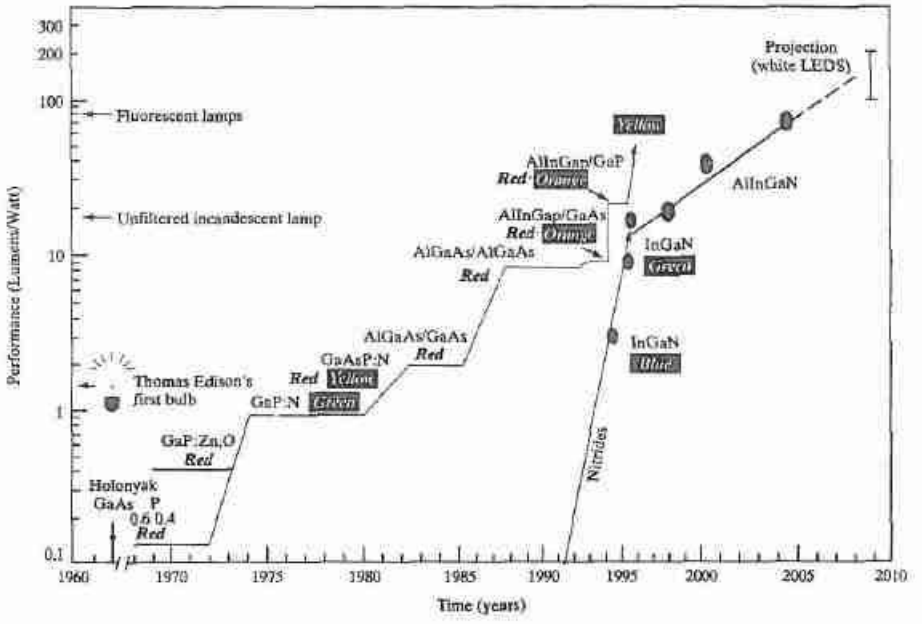


Figure 8–10
Improvement of luminous intensity of LEDs over time. [Modified from M. G. Craford, *IEEE Circuits and Devices*, p. 24, Sept. 1992.]

the same time the photogenerated carriers have to traverse a short distance in the perpendicular direction, leading to short transit times and high bandwidth.

8.2 LIGHT-EMITTING DIODES

When carriers are injected across a forward-biased junction, the current is usually accounted for by recombination in the transition region and in the neutral regions near the junction. In a semiconductor with an indirect band gap, such as Si or Ge, the recombination releases heat to the lattice. On the other hand, in a material characterized by direct recombination, considerable light may be given off from the junction under forward bias. This effect, called *injection electroluminescence* (Section 4.2.2), provides an important application of diodes as generators of light. The use of light-emitting diodes (LEDs) in digital displays is well known. There are also other important applications, in traffic and automotive signals and in illumination. Another important device making use of radiative recombination in a forward-biased p-n junction is the *semiconductor laser*. As we shall see in Section 8.4, lasers emit coherent light in much narrower wavelength bands than LEDs, with more collimation (directionality), and are very useful for fiber-optic communication systems, as described in Section 8.2.2.

For LEDs, the frequency (color) of the photon is governed by the band gap of the semiconductor as given by the Planck relation, $h\nu = E_g$, which, in appropriate units, can be expressed as E_g (eV) = 1.24/λ (μm). A very important metric of an LED is the external quantum efficiency η_{ext}, which is defined as the light output divided by the electrical input power:

$$\eta_{ext} = (\text{Internal radiative efficiency}) \times (\text{Extraction efficiency}) \quad (8\text{–}5)$$

The internal efficiency is a function of the quality of the material and the structure and composition of the layer. Defects in the material will clearly lead to nonradiative recombination. However, even if the internal efficiency is high, not all emitted photons are extracted from the LED. The emitted photons from an LED have a wide angular distribution, unlike those in a laser. For example, if an LED had a planar surface, the photons incident on the semiconductor–air interface at angles greater than a critical angle would undergo total internal reflection and ultimately be lost via absorption within the semiconductor. Hence, typically, LEDs are made with a dome-type encapsulation, which acts as a lens so that more of the photons can be extracted.

8.2.1 Light-Emitting Materials

The band gaps of various binary compound semiconductors are illustrated in Fig. 4–4 relative to the spectrum. There is a wide variation in band gaps and, therefore, in available photon energies, extending from the ultraviolet (GaN, 3.4 eV) into the infrared (InSb, 0.18 eV). In fact, by utilizing ternary

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 411

and quaternary compounds the number of available energies can be increased significantly (see Figs. 1–13 and 3–6). A good example of the variation in photon energy obtainable from the compound semiconductors is the ternary alloy gallium arsenide–phosphide, which is illustrated in Fig. 8–11. When the percentage of As is reduced and P is increased in this material, the resulting band gap varies from the direct 1.43-eV gap of GaAs (infrared) to the indirect 2.26-eV gap of GaP (green). The band gap of $GaAs_{1-x}P_x$ varies almost linearly with x until the 0.45 composition is reached, and electron–hole recombination is direct over this range. The most common alloy composition used in LED displays is $x \simeq 0.4$. For this composition the band gap is direct, since the Γ minimum (at $\mathbf{k} = 0$) is the lowest part of the conduction

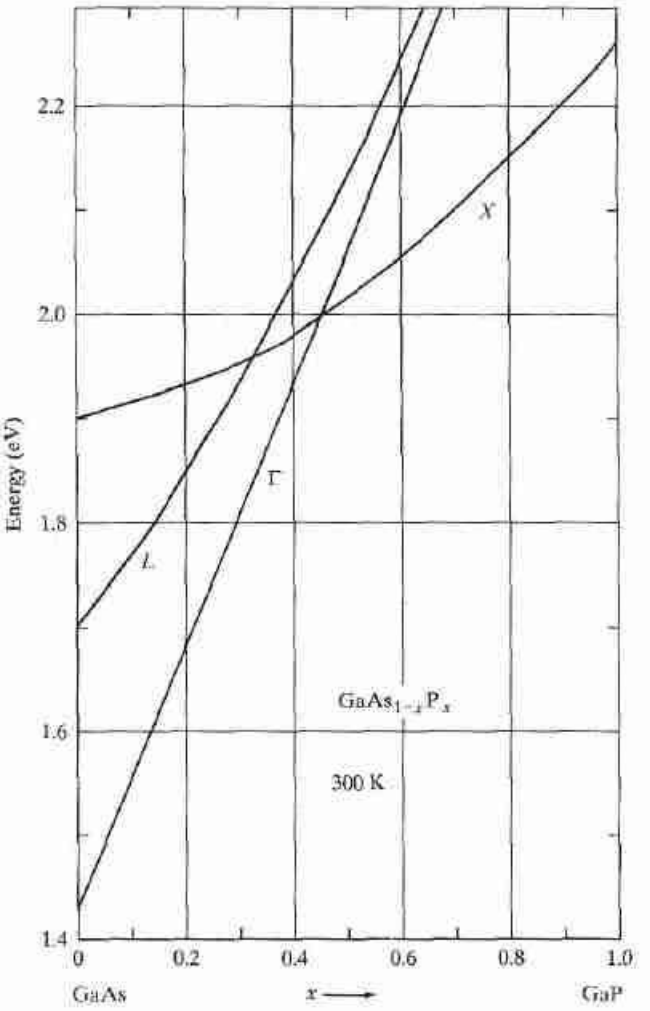


Figure 8–11
Conduction band energies as a function of alloy composition for $GaAs_{1-x}P_x$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 412

band. This results in efficient radiative recombination, and the emitted photons (~1.9 eV) are in the red portion of the spectrum.

For $GaAs_{1-x}P_x$ with P concentrations above 45 percent, the band gap is due to the indirect X minimum. Radiative recombination in such indirect materials is generally unlikely, because electrons in the conduction band have different momentum from holes in the valence band (see Fig. 3–5). Interestingly, however, indirect $GaAs_{1-x}P_x$ (including GaP, $x = 1$) doped with nitrogen can be used in LEDs with light output in the yellow to green portions of the spectrum. This is possible because the nitrogen impurity binds an electron very tightly. This confinement in real space (Δx) means that the electron momentum is spread out in momentum space Δp by the Heisenberg uncertainty principle [see Eq. (2–18)]. As a result, the momentum conservation rules, which generally prevent radiative recombination in indirect materials, are circumvented. Thus nitrogen doping of $GaAs_{1-x}P_x$ is not only useful technologically, but also provides an interesting and practical illustration of the uncertainty principle.

In many applications light from a laser or an LED need not be visible to the eye. Infrared emitters such as GaAs, InP, and mixed alloys of these compounds are particularly well suited to fiber-optical communication systems or TV remote controls. For example, a laser or light-emitting diode can be used in conjunction with a photodiode or other photosensitive device to transmit information optically between locations. By varying the current through the diode, the light output can be modulated such that analog or digital information appears in the optical signal directed at the detector. Alternatively, the information may be introduced between the source and detector. For example, a semiconductor laser-photodetector arrangement can be used in a compact disc or DVD system for reading digital information from the spinning disc. A light emitter and a photodiode form an *optoelectronic pair*, which provides complete electrical isolation between input and output, since the only link between the two devices is optical. In an *optoelectronic isolator*, both devices may be mounted on a ceramic substrate and packaged together to form a unit that passes information while maintaining isolation.

In view of the broad range of applications requiring semiconductor lasers and LEDs with visible and infrared wavelengths, the wide variety of available III–V materials is extremely useful. In addition to the AlGaAs and GaAsP systems shown in Figs. 3–6 and 8–11, the InAlGaP system is useful for red, yellow, and orange wavelengths, and AlGaInN is a strong emitter in the blue and green. Why is there so much interest in short wavelength emitters such as blue-green LEDs? As shown in Fig. 8–10, moderate efficiency (~10 lumens/watt) red, green, and yellow-green LEDs have existed for a long time in the GaAsP system, based on concepts such as N isoelectronic doping (Section 8.2.1). More recently, in the mid-1990s, much higher brightness and efficiency (~30 lumens/watt) red-orange-yellow LEDs based on the InAlGaP system have been developed. The higher-band-gap InAlGaN system has direct band gaps over the entire alloy composition range and hence offers highly efficient light emission in the blue and green part of the spectrum. A major goal of the optoelectronics community has been the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 413

achievement of high-efficiency red, green, and blue emitters, because those colors are the three primary colors of the spectrum. By combining these color LEDs (or in conjunction with suitable phosphors), one can form intense white light sources (~500 lumens) with luminous efficiencies ~2 times greater than those of conventional incandescent lightbulbs. Since general-purpose illumination typically requires kilolumens of light, it may be practical to combine a few of these high-brightness LEDs for that purpose. They also have much longer lifetimes (5,000–100,000 hours versus 2,000 hours for conventional lightbulbs) and much higher energy efficiencies. If the costs of these LEDs can be reduced by ~100 times, they will become competitive, leading to a significant reduction in global energy demand. Arrays of red, green, and blue emitters are being used in outdoor displays and TV screens, as well as in automotive taillights and signal lights. Red, yellow, and green LEDs are being used for traffic lights because they have a much higher reliability and lifetime than conventional lightbulbs and they save energy.

8.2.2 Fiber-Optic Communications

The transmission of optical signals from source to detector can be greatly enhanced if an *optical fiber* is placed between the light source and the detector. An optical fiber is essentially a "light pipe" or waveguide for optical frequencies. The fiber is typically drawn from a boule of glass to a diameter of ~25 μm. The fine glass fiber is relatively flexible and can be used to guide optical signals over distances of kilometers without the necessity of perfect alignment between source and detector. This significantly increases the applications of optical communication in areas such as telephone and data transmission.

One type of optical fiber has an outer layer of very pure fused silica (SiO_2), with a core of germanium doped glass having a higher index of refraction (Fig. 8-12a).[1] Such a *step-index* fiber maintains the light beam primarily in the central core with little loss at the surface. The light is transmitted along the length of the fiber by internal reflection at the step in the refractive index.

Losses in the fiber at a given wavelength can be described by an attenuation coefficient α [similar to the absorption coefficient of Eq. (4-3)]. The intensity of the signal at a distance x along the fiber is then related to the starting intensity by the usual expression.

$$I(x) = I_0 e^{-\alpha x} \tag{8-6}$$

The attenuation is not the same for all wavelengths, however, and it is therefore important to choose a signal wavelength carefully. A plot of α vs. λ for a typical silica glass fiber is shown in Fig. 8-13. It is clear that dips in α near 1.3 and 1.55 μm provide "windows" in the attenuation, which can be exploited to reduce the degradation of signals. The overall decrease in attenuation with

[1]The *index of refraction* (or *refractive index*) n compares the velocity of light **v** in the material with its velocity c in a vacuum: $n = c/v$. Thus, if $n_1 > n_2$ in Fig. 8–12a, the velocity of light is greater in material 2 than in 1. The value of n varies somewhat with the wavelength of light.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 414

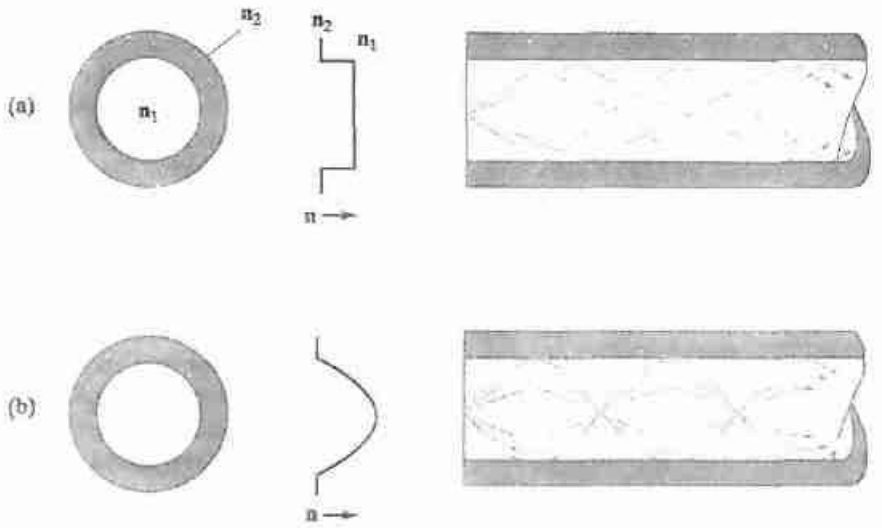


Figure 8–12
Two examples of multimode fibers: (a) *a step index* having a core with a slightly larger refractive index n_1; (b) a *graded index* having, in this case, a parabolic grading of n_1 in the core. The figure illustrates the cross section (left) of the fiber, its index-of-refraction profile (center), and typical mode patterns (right).

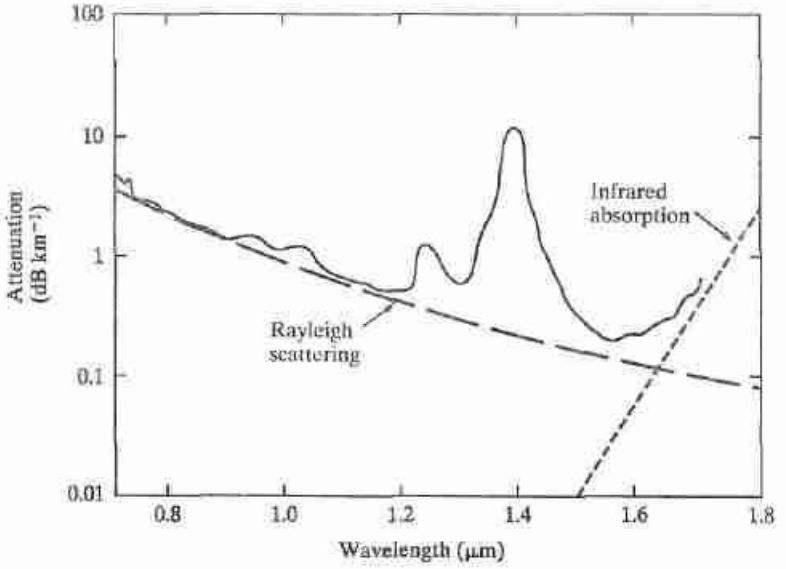


Figure 8–13
Typical plot of attenuation coefficient α vs. wavelength λ for a fused silica optical fiber. Peaks are due primarily to OH^- impurities.

increasing wavelength is due to the reduced scattering from small random inhomogeneities which result in fluctuations of the refractive index on a scale comparable to the wavelength. This type of attenuation, called *Rayleigh scattering*, decreases with the fourth power of wavelength. This effect is observed at sunrise and sunset, when attenuation of short wavelength blue and green light results in red and orange sunlight. Obviously, Rayleigh

scattering encourages operation at long wavelengths in fiber-optic systems. However, a competing process of infrared absorption dominates for wavelengths longer than about 1.7 μm, due to vibrational excitation of the atoms making up the glass. Therefore, a useful minimum in absorption for silica fibers occurs at about 1.55 μm, where epitaxial layers in the (In, Ga) (As, P) system can be grown lattice-matched to InP substrates (see Fig. 1-13).

Another consideration in choice of operating wavelength is the *pulse dispersion*, or spreading of data pulses as they propagate down the fiber. This effect can be caused by the wavelength dependence of the refractive index, causing different optical frequencies to travel down the fiber with slightly different velocities. This effect, called *chromatic dispersion*, is much less pronounced at the 1.3 μm window in Fig. 8-13. Another cause of dispersion is the fact that different modes propagate with different path lengths (Fig. 8-12a). This type of dispersion can be reduced by grading the refractive index of the core (Fig. 8-12b) such that various modes are continually refocused, reducing the differences in path lengths.

The light source in a fiber-optic system today is a laser, because the light is made up of essentially a single frequency and allows a very large information bandwidth. (Semiconductor lasers suitable for fiber-optic communications will be discussed in Section 8.4.) In early optoelectronic systems for fiber optics, it was most convenient to use the well-established GaAs–AlGaAs system for making lasers and LEDs. These light sources are highly efficient, and good detectors can be made using Si p-i-n or APDs. However, these sources operate in the wavelength range near 0.9 μm, where the attenuation is greater than that for longer wavelengths. Modern systems, therefore, operate near the 1.3- and 1.55-μm minima shown in Fig. 8-13. At these wavelengths, lasers can be made by using InGaAs or InGaAsP grown on InP, and detectors can be made of the same materials. (See Fig. 8-8.)

Multimode fibers are larger (~25 μm in diameter) than *single-mode* fibers (~5 μm) and transmit a coherent laser beam. By forming numerous optical fibers into a bundle, with an appropriate jacket for mechanical strength, an enormous amount of information can be transmitted over long distances.[2] Depending upon the losses in the fibers, repeater stations may be required periodically along the path. Thus, many photodetectors and laser sources are required in a fiber-optic system. Semiconductor device development, including appropriate binary, ternary, and quaternary compounds for both emitters and detectors, is therefore crucial to the successful implementation of such optical communications systems. A schematic of a fiber-optic transmission system is shown in Fig. 8-14.

[2]Transmission rates of 40 G-bit/s are becoming standard, but even higher rates of 400 G-bit/s have been achieved using ultradense-wavelength-division multiplexing (UDWDM). This approach uses slightly different wavelengths or colors to carry different channels of information along the same fiber. As a convenient calibration of the 40-G-bit/s rate, it is worth noting that the human eye is able to transmit about 1 G-bit/s to the brain.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 416

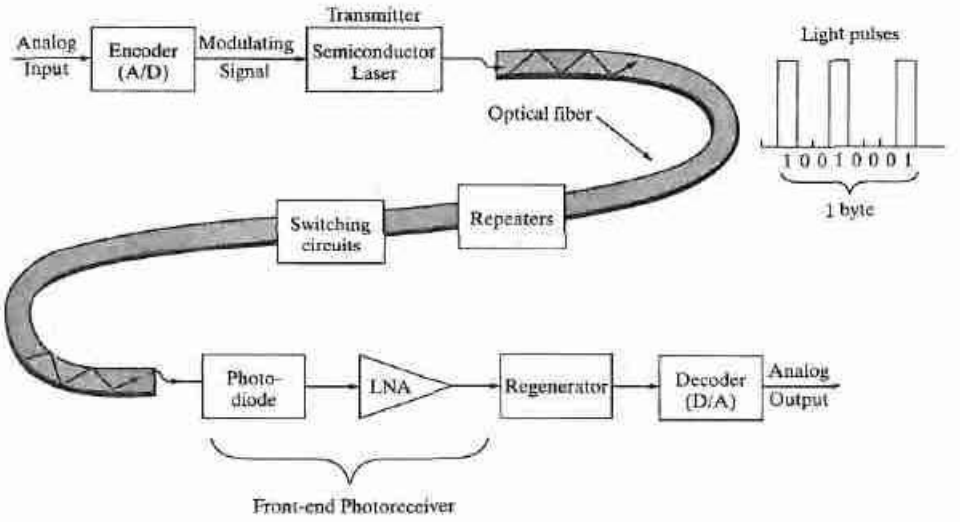


Figure 8–14
Schematic of a fiber-optic communication system illustrating the transmission of analog signals, such as those in telephony or TV. After the signal is converted to a digital electrical signal, it is used to modulate the laser light output as light pulses that are transmitted down the fiber, with periodic amplification of the signal with repeaters to compensate for fiber loss. Switching circuits route signals appropriately. After the optical signal is transformed to an electrical output by a photodetector and a low-noise preamplifier (LNA), it is converted to an analog signal, once distortions of the digital signals due to propagation through the fiber have been corrected with the use of a "regenerator."

EXAMPLE 8–3

What composition x of $Al_xGa_{1-x}As$ would emit red light at 680 nm? What composition of $GaAs_{1-x}P_x$?

SOLUTION

The band gap E_g (eV) = 1.24/λ (μm) = 1.24/0.68 = 1.82 eV.
From Fig. 3–6, we get $Al_{0.32}Ga_{0.68}As$.
From Fig. 8–11, we get $GaAs_{0.68}P_{0.32}$.

8.3 LASERS

The word LASER is an acronym for *light amplification by stimulated emission of radiation*, which sums up the operation of an important optical and electronic device. The laser is a source of highly directional, monochromatic, coherent light, and as such it has revolutionized some longstanding optical problems and has created some new fields of basic and applied optics. The light from a laser, depending on the type, can be a continuous beam of low or medium power, or it can be a short burst of intense light delivering millions of watts. Light has always been a primary communications link between

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 417

humans and the environment, but until the invention of the laser, the light sources available for transmitting information and performing experiments were generally neither monochromatic nor coherent, and were of relatively low intensity. Thus the laser is of great interest in optics; but it is equally important in optoelectronics, particularly in fiber-optic communications. The last three letters in the word *laser* are intended to imply how the device operates: by the *stimulated emission* of *radiation*. In Chapter 2 we discussed the emission of radiation when excited electrons fall to lower energy states; but generally, these processes occur randomly and can therefore be classed as *spontaneous emission*. This means that the rate at which electrons fall from an upper level of energy E_2 to a lower level E_1 is at every instant proportional to the number of electrons remaining in E_2 (the *population* of E_2). Thus if an initial electron population in E_2 were allowed to decay, we would expect an exponential emptying of the electrons to the lower energy level, with a mean decay time describing how much time an average electron spends in the upper level. An electron in a higher or excited state need not wait for spontaneous emission to occur, however; if conditions are right, it can be *stimulated* to fall to the lower level and emit its photon in a time much shorter than its mean spontaneous decay time. The stimulus is provided by the presence of photons of the proper wavelength. Let us visualize an electron in state E_2 waiting to drop spontaneously to E_1 with the emission of a photon of energy $h\nu_{12} = E_2 - E_1$ (Fig. 8–15). Now we assume that this electron in the upper state is immersed in an intense field of photons, each having energy $h\nu_{12} = E_2 - E_1$, and in phase with the other photons. The electron is induced to drop in energy from E_2 to E_1, contributing a photon whose wave is *in phase* with the radiation field. If this process continues and other electrons are stimulated to emit photons in the same fashion, a large radiation field can build up. This radiation will be *monochromatic* since each photon will have an energy of precisely $h\nu_{12} = E_2 - E_1$ and will be *coherent*, because all the photons released will be in phase and reinforcing. This process of stimulated emission can be described quantum mechanically to relate the probability of emission to the intensity of the radiation field. Without quantum mechanics we can make a few observations here about the relative rates at which the absorption and emission processes occur. Let us assume the instantaneous populations of E_1 and E_2 to be n_1 and n_2, respectively. We know

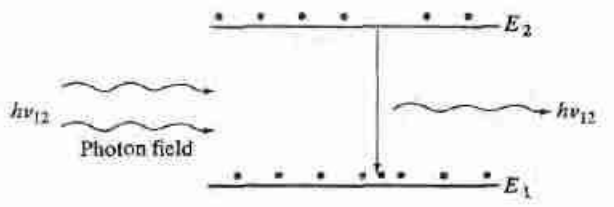


Figure 8–15
Stimulated transition of an electron from an upper state to a lower state, with accompanying photon emission.

from earlier discussions of distributions and the Boltzmann factor that at *thermal equilibrium* the relative population will be

$$\frac{n_2}{n_1} = e^{-(E_2 - E_1)/kT} = e^{-h\nu_{12}/kT} \tag{8-7}$$

if the two levels contain an equal number of available states.

The negative exponent in this equation indicates that $n_2 \ll n_1$ at equilibrium; that is, most electrons are in the lower energy level as expected. If the atoms exist in a radiation field of photons with energy $h\nu_{12}$, such that the energy density of the field is $\rho(\nu_{12})$,[3] then stimulated emission can occur along with absorption and spontaneous emission. The rate of stimulated emission is proportional to the instantaneous number of electrons in the upper level n_2 and to the energy density of the stimulating field $\rho(\nu_{12})$. Thus we can write the stimulated emission rate as $B_{21}n_2\rho(\nu_{12})$, where B_{21} is a proportionality factor. The rate at which the electrons in E_1 absorb photons should also be proportional to $\rho(\nu_{12})$, and to the electron population in E_1. Therefore, the absorption rate is $B_{12}n_1\rho(\nu_{12})$, where B_{12} is a proportionality factor for absorption. Finally, the rate of spontaneous emission is proportional only to the population of the upper level. Introducing still another coefficient, we can write the rate of spontaneous emission as $A_{21}n_2$ For steady state the two emission rates must balance the rate of absorption to maintain constant populations n_1 and n_2 (Fig. 8-16).

$$\begin{array}{lll} B_{12}n_1\rho(\nu_{12}) & = A_{21}n_2 & + B_{21}n_2\rho(\nu_{12}) \\ \text{Absorption} & = \text{spontaneous emission} & + \text{stimulated emission} \end{array} \tag{8-8a}$$

This relation was described by Einstein, and the coefficients B_{12}, A_{21}, B_{21} are called the *Einstein coefficients.* We notice from Eq. (8-8a) that no energy density ρ is required to cause a transition from an upper to a lower state; spontaneous emission occurs without an energy density to drive it. The reverse is not true, however; exciting an electron to a higher state (absorption) requires the application of energy, as we would expect thermodynamically.

At equilibrium, the ratio of the stimulated to spontaneous emission rates is generally very small, and the contribution of stimulated emission is negligible. With a photon field present,

$$\frac{\text{Stimulated emission rate}}{\text{Spontaneous emission rate}} = \frac{B_{21}n_2\rho(\nu_{12})}{A_{21}n_2} = \frac{B_{21}}{A_{21}}\rho(\nu_{12}) \tag{8-8b}$$

As Eq. (8-8b) indicates, the way to enhance the stimulated emission over spontaneous emission is to have a very large photon field energy density $\rho(\nu_{12})$. In the laser, this is encouraged by providing an *optical resonant cavity* in which the

[3]The energy density $\rho(\nu_{12})$ indicates the total energy in the radiation field per unit volume and per unit frequency, due to photons with $h\nu_{12} = E_2E_1$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 419

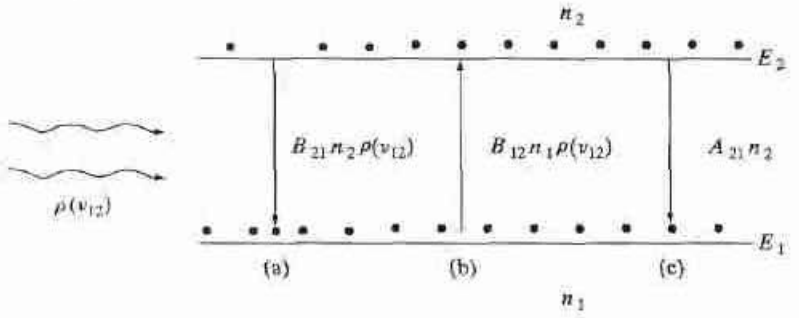


Figure 8–16
Balance of absorption and emission in steady state: (a) stimulated emission; (b) absorption; (c) spontaneous emission.

photon density can build up to a large value through multiple internal reflections at certain frequencies (ν).

Similarly, to obtain more stimulated emission than absorption we must have $n_2 > n_1$:

$$\frac{\text{Stimulated emission rate}}{\text{Absorption rate}} = \frac{B_{21}n_2\rho(\nu_{12})}{B_{12}n_1\rho(\nu_{12})} = \frac{B_{21}}{B_{12}}\frac{n_2}{n_1} \tag{8–9}$$

Thus if stimulated emission is to dominate over absorption of photons from the radiation field, we must have a way of maintaining more electrons in the upper level than in the lower level. This condition is quite unnatural, since Eq. (8–7) indicates that n_2/n_1 is less than unity for any equilibrium case. Because of its unusual nature, the condition $n_2 > n_1$ is called *population inversion.* It is also referred to as a condition of *negative temperature.* This rather startling terminology emphasizes the nonequilibrium nature of population inversion, and refers to the fact that the ratio n_2/n_1 in Eq. (8–7) could be larger than unity only if the temperature were negative. Of course, this manner of speaking does not imply anything about temperature in the usual sense of that word. The fact is that Eq. (8–7) is a thermal equilibrium equation and cannot be applied to the situation of population inversion without invoking the concept of negative temperature.

In summary, Eqs. (8–8b) and (8–9) indicate that if the photon density is to build up through a predominance of stimulated emission over both spontaneous emission and absorption, two requirements must be met. We must provide (1) an optical resonant cavity to encourage the photon field to build up and (2) a means of obtaining population inversion.

An optical resonant cavity can be obtained using reflecting mirrors to reflect the photons back and forth, allowing the photon energy density to build up. One or both of the end mirrors are constructed to be partially transmitting so that a fraction of the light will "leak out" of the resonant system. This transmitted light is the output of the laser. Of course, in designing such a laser one must choose the amount of transmission to be a small perturbation

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 420

on the resonant system. The gain in photons per pass between the end plates must be larger than the transmission at the ends, scattering from impurities, absorption, and other losses. The arrangement of parallel plates providing multiple internal reflections is similar to that used in the Fabry-Perot interferometer;[4] thus the reflecting ends of the laser cavity are often referred to as Fabry-Perot faces. As Fig. 8-17 indicates, light of a particular frequency can be reflected back and forth within the resonant cavity in a reinforcing (coherent) manner if an integral number of half-wavelengths fit between the end mirrors. Thus the length of the cavity for stimulated emission must be

$$L = \frac{\mathbf{m}\lambda}{2} \tag{8-10}$$

where **m** is an integer. In this equation λ is the photon wavelength within the laser material. If we wish to use the wavelength λ_0 of the output light in the atmosphere (often taken as the vacuum value), the index of refraction **n** of the laser material must be considered

$$\lambda_0 = \lambda \mathbf{n} \tag{8-11}$$

In practice, $L \gg \lambda$, and Eq. (8-10) is automatically satisfied over some portion of the mirror. An important exception occurs in the vertical cavity surface-emitting lasers discussed in Section 8.4.4, for which the cavity length is comparable to the wavelength.

There are ways of obtaining population inversion in the atomic levels of many solids, liquids, and gases, and in the energy bands of semiconductors. Thus the possibilities for laser systems with various materials are quite extensive. An early laser system used a ruby rod. In gas lasers, electrons are excited to metastable levels in molecules to achieve population inversion. These are interesting and useful laser systems, but in view of our emphasis on semiconductor devices in this book, we will move to the description of semiconductor lasers.

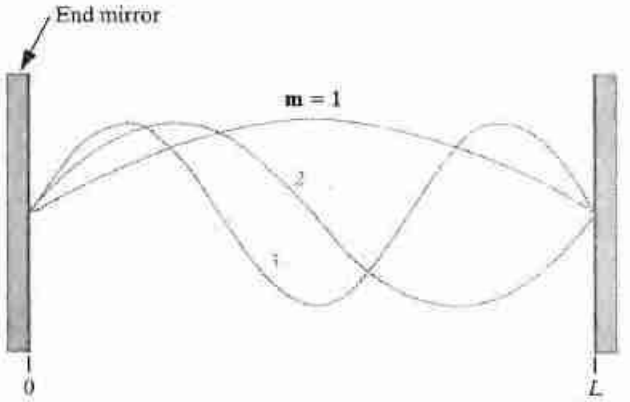


Figure 8–17
Resonant modes within a laser cavity.

[4]Interferometers are discussed in many sophomore physics texts.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 421

8.4 SEMICONDUCTOR LASERS

The laser became an important part of semiconductor device technology in 1962 when the first p-n junction lasers were built in GaAs (infrared)[5] and GaAsP (visible).[6] We have already discussed the incoherent light emission from p-n junctions (LEDs), generated by the spontaneous recombination of electrons and holes injected across the junction. In this section we shall concentrate on the requirements for population inversion due to these injected carriers and the nature of the coherent light from p-n junction lasers. These devices differ from solid, gas, and liquid lasers in several important respects. Junction lasers are remarkably small (typically on the order of $0.1 \times 0.1 \times 0.3$ mm), they exhibit high efficiency, and the laser output is easily modulated by controlling the junction current. Semiconductor lasers operate at low power compared, for example, with ruby or CO_2 lasers; on the other hand, these junction lasers compete with He-Ne lasers in power output. Thus the function of the semiconductor laser is to provide a portable and easily controlled source of low-power coherent radiation. They are particularly suitable for fiber-optic communication systems (Section 8.2.2).

8.4.1 Population Inversion at a Junction

If a p-n junction is formed between degenerate materials, the bands under forward bias appear as shown in Fig. 8-18. If the bias (and thus the current) is large enough, electrons and holes are injected into and across the transition region in considerable concentrations. As a result, the region about the junction is far from being depleted of carriers. This region contains a large concentration of electrons within the conduction band and a large concentration of holes within the valence band. If these population densities are high enough, a condition of population inversion results, and the region about the junction over which it occurs is called an *inversion region.*[7]

Population inversion at a junction is best described by the use of the concept of *quasi-Fermi levels* (Section 4.3.3). Since the forward-biased condition of Fig. 8-18 is a distinctly nonequilibrium state, the equilibrium equations defining the Fermi level are not applicable. In particular, the concentration of electrons in the inversion region (and for several diffusion lengths into the p material) is larger than equilibrium statistics would imply; the same is also true for the injected holes in the n material. We can use Eqs. (4-15) to describe the carrier concentrations in terms of the quasi-Fermi levels for electrons and holes in steady state. Thus

$$n = N_c e^{-(E_c - F_n)/kT} = n_i e^{(F_n - E_i)/kT} \tag{8-12a}$$

$$p = N_v e^{-(F_p - E_v)/kT} = n_i e^{(E_i - F_p)/kT} \tag{8-12b}$$

[5] R. N. Hall et al., *Physical Review Letters* 9, pp. 366–368 (November 1, 1962); M. I. Nathan et al., *Applied Physics Letters* 1, pp. 62–64 (November 1, 1962); T. M. Quist et al., *Applied Physics Letters* 1, pp. 91–92 (December 1, 1962).

[6] N. Holonyak, Jr., and S. F. Bevacqua, *Applied Physics Letters* 1, pp. 82–83 (December 1, 1962).

[7] This is a different meaning of the term from that used in reference to MOS transistors.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 422

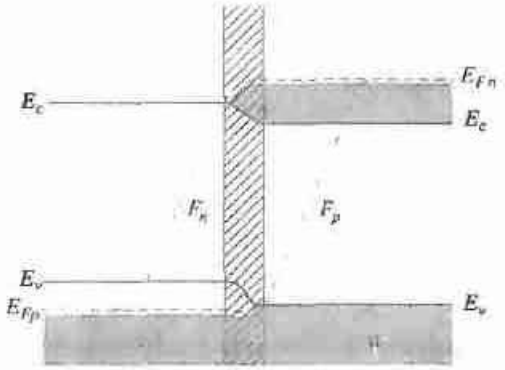


Figure 8–18
Band diagram of a p-n junction laser under forward bias. The crosshatched region indicates the inversion region at the junction.

Using Eqs. (8–12a) and (8–12b), we can draw F_n and F_p on any band diagram for which we know the electron and hole distributions. For example, in Fig. 8–18, F_n in the neutral n region is essentially the same as the equilibrium Fermi level E_{Fn}. This is true to the extent that the electron concentration on the n side is equal to its equilibrium value. However, since large numbers of electrons are injected across the junction, the electron concentration begins at a high value near the junction and decays exponentially to its equilibrium value n_p deep in the p material. Therefore, F_n drops from E_{Fn} as shown in Fig. 8–18. We notice that, deep in the neutral regions, the quasi-Fermi levels are essentially equal. The separation of F_n and F_p at any point is a measure of the departure from equilibrium at that point. Obviously, this departure is considerable in the inversion region, since F_n and F_p are separated by an energy greater than the band gap (Fig. 8–19).

Unlike the case of the two-level system discussed in Section 8.3, the condition for population inversion in semiconductors must take into account the distribution of energies available for transitions between the bands. The basic definition of population inversion holds—for dominance of stimulated emission between two energy levels separated by energy $h\nu$, the electron population of the upper level must be greater than that of the lower level. The unusual aspect of a semiconductor is that bands of levels are available for

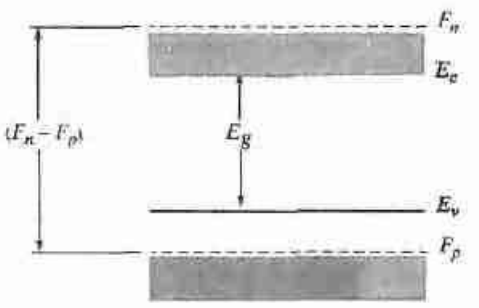


Figure 8–19
Expanded view of the inversion region.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 423

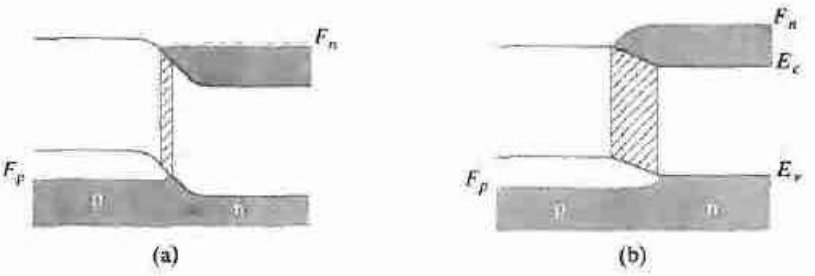


Figure 8–20
Variation of the inversion-region width with forward bias: $V(a) < V(b)$.

such transitions. Population inversion obviously exists for transitions between the bottom of the conduction band E_c and the top of the valence band E_v in Fig. 8-19. In fact, transitions between levels in the conduction band up to F_n and levels in the valence band down to F_p take place under conditions of population inversion. For any given transition energy $h\nu$ in a semiconductor, population inversion exists when

$$(F_n - F_p) > h\nu \tag{8-13a}$$

For band-to-band transitions, the minimum requirement for population inversion occurs for photons with $h\nu = E_c - E_v = E_g$

$$(F_n - F_p) > E_g \tag{8-13b}$$

When F_n and F_p lie within their respective bands (as in Fig. 8-19), stimulated emission can dominate over a range of transitions, from $h\nu = (F_n - F_p)$ to $h\nu = E_g$. As we shall see below, the dominant transitions for laser action are determined largely by the resonant cavity and the strong recombination radiation occurring near $h\nu = E_g$.

In choosing a material for junction laser fabrication, it is necessary that electron–hole recombination occur directly, rather than through trapping processes such as are dominant in Si or Ge. Gallium arsenide is an example of such a "direct" semiconductor. Furthermore, we must be able to dope the material n-type or p-type to form a junction. If an appropriate resonant cavity can be constructed in the junction region, a laser results in which population inversion is accomplished by the bias current applied to the junction (Fig. 8-20).

8.4.2 Emission Spectra for p-n Junction Lasers

Under forward bias, an inversion layer can be obtained along the plane of the junction, where a large population of electrons exists at the same location as a large hole population. A second look at Fig. 8-19 indicates that spontaneous emission of photons can occur due to direct recombination of electrons and holes, releasing energies ranging from approximately $F_n - F_p$ to E_g. That is, an electron can recombine over an energy from F_n to F_p, yielding a photon of energy $h\nu = F_n - F_p$, or an electron can recombine from the bottom of the conduction band to the top of the valence band, releasing a

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 424

photon with $h\nu = E_c - E_v = E_g$. These two energies serve as the approximate outside limits of the laser spectra.

The photon wavelengths which participate in stimulated emission are determined by the length of the resonant cavity as in Eq. (8-10). Figure 8-21 illustrates a typical plot of emission intensity vs. photon energy for a semiconductor laser. At low current levels (Fig. 8-21a), a spontaneous emission spectrum containing energies in the range $E_g < h\nu < (F_n - F_p)$ is obtained. As the current is increased to the point that significant population inversion exists, stimulated emission occurs at frequencies corresponding to the cavity modes as shown in Fig. 8-21b. These modes correspond to successive numbers of integral half-wavelengths fitted within the cavity, as described by Eq. (8-10). Finally, at a still higher current level, a most preferred mode or set of modes will dominate the spectral output (Fig. 8-21c). This very intense mode represents the main laser output of the device; the output light will be composed of almost monochromatic radiation superimposed on a relatively weak radiation background, due primarily to spontaneous emission.

The separation of the modes in Fig. 8-21b is complicated by the fact that the index of refraction **n** for GaAs depends on wavelength λ. From Eq. (8-10) we have

$$\mathbf{m} = \frac{2L\mathbf{n}}{\lambda_0} \tag{8-14}$$

If **m** (the number of half-wavelengths in L) is large, we can use the derivative to find its rate of change with λ_0:

$$\frac{d\mathbf{m}}{d\lambda_0} = -\frac{2L\mathbf{n}}{\lambda_0^2} + \frac{2L}{\lambda_0}\frac{d\mathbf{n}}{d\lambda_0} \tag{8-15}$$

Now reverting to discrete changes in **m** and λ_0, we can write

$$-\Delta\lambda_0 = \frac{\lambda_0^2}{2L\mathbf{n}}\left(1 - \frac{\lambda_0}{\mathbf{n}}\frac{d\mathbf{n}}{d\lambda_0}\right)^{-1}\Delta\mathbf{m} \tag{8-16}$$

If we let $\Delta\mathbf{m} = -1$, we can calculate the change in wavelength $\Delta\lambda_0$ between adjacent modes (i.e., between modes **m** and **m** − 1).

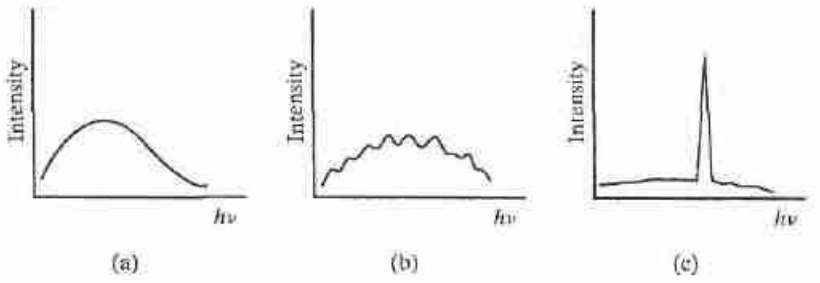


Figure 8-21 Light intensity vs. photon energy $h\nu$ for a junction laser: (a) incoherent emission below threshold; (b) laser modes at threshold; (c) dominant laser mode above threshold. The intensity scales are greatly compressed from (a) to (b) to (c).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 425

8.4.3 The Basic Semiconductor Laser

To build a p-n junction laser, we need to form a junction in a highly doped, direct semiconductor (GaAs, for example), construct a resonant cavity in the proper geometrical relationship to the junction, and make contact to the junction in a mounting which allows for efficient heat transfer. The first lasers were built as shown in Fig. 8-22. Beginning with a degenerate n-type sample, a p region is formed on one side, for example by diffusing Zn into the n-type GaAs. Since Zn is in column II of the periodic table and is introduced substitutionally on Ga sites, it serves as an acceptor in GaAs; therefore, the heavily doped Zn diffused layer forms a p^+ region (Fig. 8-22b). At this point we have a large-area planar p-n junction. Next, grooves are cut or etched along the length of the sample as in Fig. 8-22c, leaving a series of long p regions isolated from each other. These p-n junctions can be cut or broken apart (Fig. 8-22d) and then cleaved into devices of the desired length.

At this point in the fabrication process, the very important requirements of a resonant cavity must be considered. It is necessary that the front and back faces (Fig. 8-22e) be flat and parallel. This can be accomplished by cleaving. If the sample has been oriented so that the long junctions of Fig. 8-22d are perpendicular to a crystal plane of the material, it is possible to cleave the sample along this plane into laser devices, letting the crystal structure itself provide the parallel faces. The device is then mounted on a suitable header, and contact is made to the p region. Various techniques are used to provide adequate heat sinking of the device for large forward current levels.

8.4.4 Heterojunction Lasers

The device described above was the first type used in the early development of semiconductor lasers. Since the device contains only one junction in a single type of material, it is referred to as a *homojunction* laser. To obtain more efficient lasers, and particularly to build lasers that operate at room temperature, it is necessary to use multiple layers in the laser structure. Such devices, called *heterojunction lasers*, can be made to operate continuously at room temperature to satisfy the requirements of optical communications.

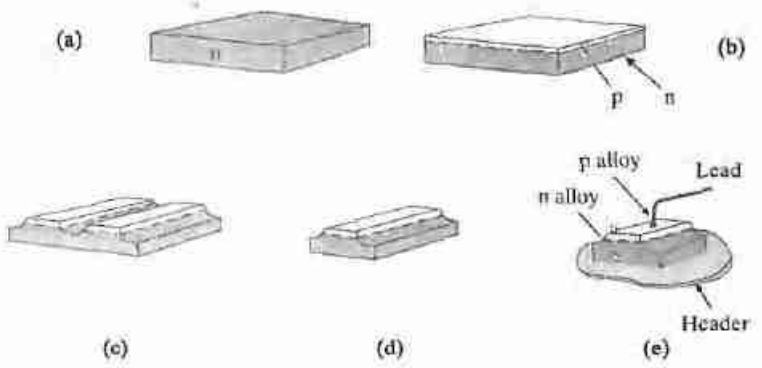


Figure 8–22 Fabrication of a simple junction laser: (a) degenerate n-type sample; (b) diffused p layer; (c) isolation of junctions by cutting or etching; (d) individual junction to be cut or cleaved into devices; (e) mounted laser structure.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 426

An example of a heterojunction laser is shown in Fig. 8–23. In this structure the injected carriers are confined to a narrow region so that population inversion can be built up at lower current levels. The result is a lowering of the *threshold current* at which laser action begins. Carrier confinement is obtained in this single-heterojunction laser by the layer of AlGaAs grown epitaxially on the GaAs.

In GaAs the laser action occurs primarily on the p side of the junction due to a higher efficiency for electron injection than for hole injection. In a normal p-n junction the injected electrons diffuse into the p material such that population inversion occurs for only part of the electron distribution near the junction. However, if the p material is narrow and terminated in a barrier, the injected electrons can be confined near the junction. In Fig. 8–23a, an epitaxial layer of p-type AlGaAs ($E_g \simeq 2$ eV) is grown on top of the thin p-type GaAs region. The wider band gap of AlGaAs effectively terminates the p-type GaAs layer, since injected electrons do not surmount the barrier at the GaAs–AlGaAs heterojunction (Fig. 8–23b). As a result of the confinement of injected electrons, laser action begins at a substantially lower current than for simple p-n junctions. In addition to the effects of carrier confinement, the change of refractive index at the heterojunction provides a waveguide effect for optical confinement of the photons.

A further improvement can be obtained by sandwiching the active GaAs layer between two AlGaAs layers (Fig. 8–24). This *double-heterojunction* structure further confines injected carriers to the active region, and the change in refractive index at the GaAs–AlGaAs boundaries helps to confine the generated light waves. In the double-heterojunction laser shown in Fig. 8–24b the injected current is restricted to a narrow stripe along the lasing direction, to reduce the total current required to drive the device. This type of laser was a major step forward in the development of lasers for fiber-optic communications.

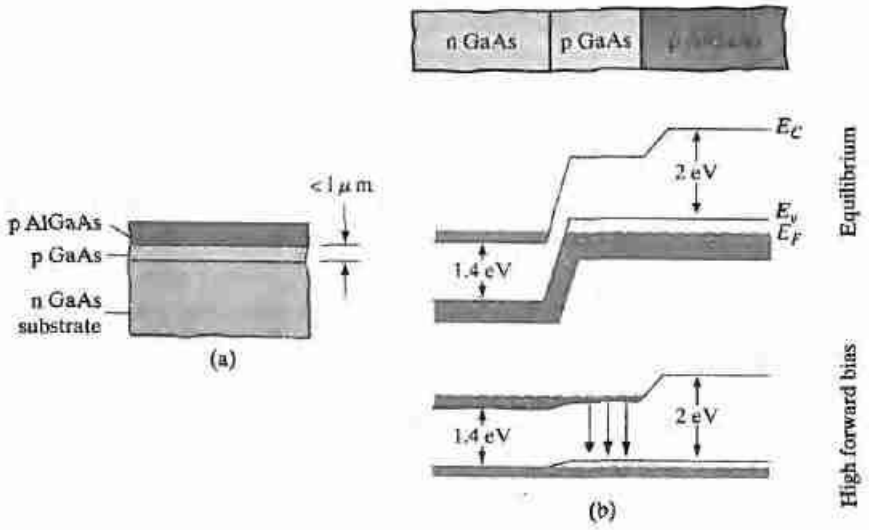


Figure 8–23
Use of a single heterojunction for carrier confinement in laser diodes: (a) AlGaAs heterojunction grown on the thin p-type GaAs layer; (b) band diagrams for the structure of (a), showing the confinement of electrons to the thin p region under bias.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 427

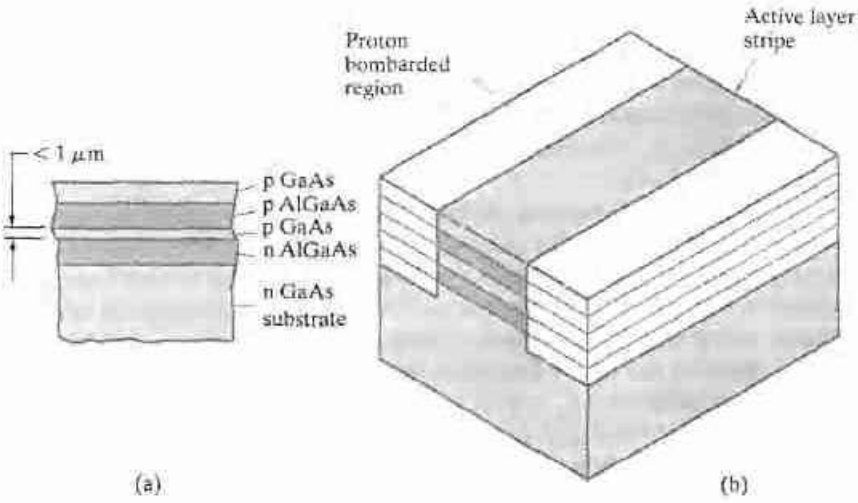


Figure 8–24
A double-heterojunction laser structure: (a) multiple layers used to confine injected carriers and provide waveguiding for the light; (b) a stripe geometry designed to restrict the current injection to a narrow stripe along the lasing direction. One of many methods for obtaining the stripe geometry, this example is obtained by proton bombardment of the shaded regions in (b), which converts the GaAs and AlGaAs to a semi-insulating form.

Separate Confinement and Graded Index Channels. One of the disadvantages of the double-heterostructure laser shown in Fig. 8–24 is the fact that the carrier confinement and the optical waveguiding both depend on the same heterojunctions. It is much better to optimize these two functions by using a narrow confinement region for keeping the carriers in a region of high recombination, and a somewhat wider optical waveguide region. In Fig. 8–25a we show a *separate confinement* laser in which the width of the optical waveguiding region (w) is optimized by using the refractive index step at a separate heterojunction from that used to confine the carriers. For example, in the GaAs–$Al_xGa_{1-x}As$ system the optical confinement (waveguiding) occurs at a boundary with much larger composition x (and therefore smaller refractive index) than is the case for the carrier confinement barrier. By grading the composition of the AlGaAs it is possible to obtain even better waveguiding. For example, in Fig. 8–25b a parabolic grading of the refractive index leads to a waveguide within the laser analogous to that shown in Fig. 8–12 for a fiber. This *graded index separate confinement heterostructure (GRINSCH)* laser also provides built-in fields for better electron confinement.

Vertical Cavity Surface-Emitting Lasers (VCSELs). There are advantages to laser structures in which light is emitted normal to the surface, including ease of device testing on the wafer before packaging. An interesting approach is the VCSEL, in which the cavity mirrors are replaced by DBRs, which use many partial reflectors spaced to reflect light constructively. DBRs can be grown by MBE or OMVPE. In Fig. 8–26 the bottom DBR mirror of a VCSEL is composed of many alternating layers of AlAs and GaAs with thickness one-quarter of a wavelength in each material. The top mirror is composed of deposited dielectric layers (alternating ZnSe and MgF). Current is funneled into the active region from the top contact by using an oxide layer achieved by laterally oxidizing an AlGaAs layer to form an aluminum oxide. The active

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 428

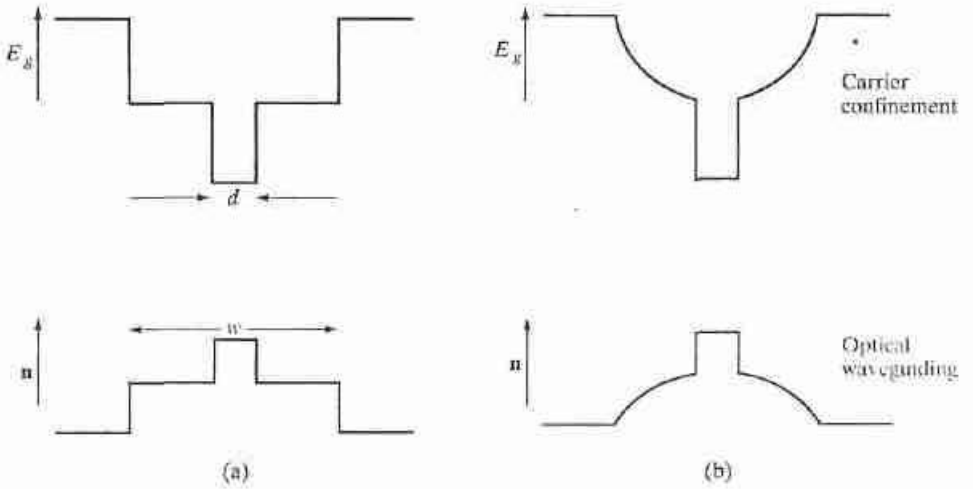


Figure 8–25
Separate confinement of carriers and waveguiding: (a) use of separate changes in AlGaAs alloy composition to confine carriers in the region (d) of the smallest band gap and to obtain waveguiding (w) at the larger step in the refractive index; (b) grading the alloy composition, and therefore the refractive index, for better waveguiding and carrier confinement.

region of the laser employs InGaAs-GaAs quantum wells, and the GaAs cavity between the two DBRs is one wavelength long. The VCSEL can be made with much shorter cavity length than other structures, and as a result of Eq. (8–16) the laser modes are widely separated in wavelength. Thus single-mode laser operation is more easily achieved with the VCSEL. Lasing can be achieved at very low current ($< 50\ \mu A$) with this device.

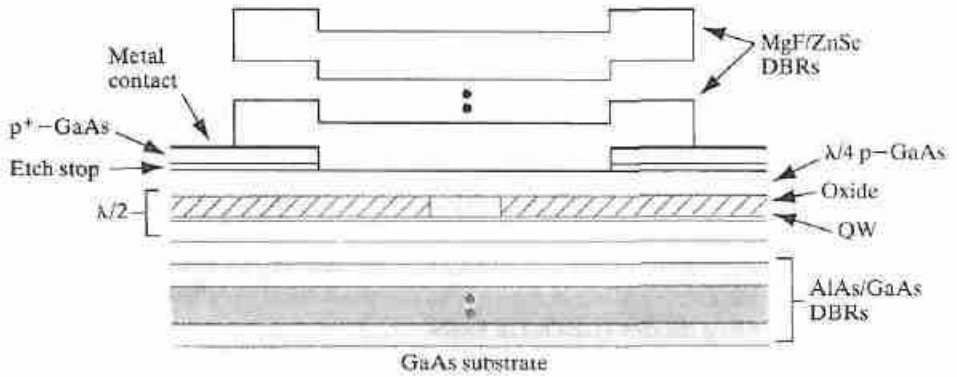


Figure 8–26
Schematic cross section of oxide-confined vertical cavity surface-emitting laser diode. [D. G. Deppe et al., *IEEE J. Selected Topics in Quantum Elec.*, 3(3) (June 1997): 893–904].

8.4.5 Materials for Semiconductor Lasers

We have discussed the properties of the junction laser largely in terms of GaAs and AlGaAs. However, as discussed in Section 8.2.2, the InGaAsP/InP system is particularly well suited for the type of lasers used in fiber-optic communication systems. Lattice-matching (Section 1.4.1) is important in creating heterostructures by epitaxial growth. The fact that the AlGaAs band gap can be varied by choice of composition on the column III sublattice allows the formation of barriers and confining layers such as those shown in Section 8.4.4. The quaternary alloy InGaAsP is particularly versatile in the fabrication of laser diodes, allowing considerable choice of wavelength and flexibility in lattice-matching. By choice of composition, lasers can be made in the infrared range 1.3–1.55 μm required for fiber optics. Since four components can be varied in choosing an alloy composition, InGaAsP allows simultaneous choice of energy gap (and therefore emission wavelength) and lattice constant (for lattice-matched growth on convenient substrates). In many applications, however, other wavelength ranges are required for laser output. For example, the use of lasers in pollution diagnostics requires wavelengths farther in the infrared than are available from InGaAsP and AlGaAs. In this application the ternary alloy PbSnTe provides laser output wavelengths from about 7 μm to more than 30 μm at low temperatures, depending on the material composition. For intermediate wavelengths, the InGaSb system can be used.

Materials chosen for the fabrication of semiconductor lasers must be efficient light emitters and also be amenable to the formation of p-n junctions and in most cases the formation of heterojunction barriers. These requirements eliminate some materials from practical use in laser diodes. For example, semiconductors with indirect band gaps are not sufficiently efficient light emitters for practical laser fabrication. The II–VI compounds, on the other hand, are generally very efficient at emitting light but junctions are difficult to form. By modern crystal growth techniques such as MBE and MOVPE it is possible to grow junctions in ZnS, ZnSe, ZnTe, and alloys of these materials, using N as the acceptor. Lasers can be made in these materials which emit in the green and blue-green regions of the spectrum.

In recent years much progress has been made in the growth of large band gap semiconductors using GaN, and its alloys with InN and AlN. The InAlGaN system has direct band gaps over the entire alloy composition range, and hence offers very efficient light emission. Band gaps range from about 2 eV for InN, to 3.4 eV for GaN and 5 eV for AlN. This covers the wavelength range from about 620 nm to about 248 nm, which is from blue to UV. The resurgence of interest in this field was triggered by the work of Nakamura at Nichia Corporation in Japan who demonstrated very high-efficiency blue light emitting diodes (LEDs) in GaN.

Two of the problems which had stymied progress in this field since pioneering work by Pankove in the 1970s was the absence of a suitable substrate having sufficient lattice match with GaN, and the inability to achieve

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 430

p-type doping in this semiconductor. GaN bulk crystals cannot be grown easily because of the high vapor pressure of the nitrogen-bearing precursor (generally ammonia). This requires growth at high temperature and pressure. This precludes using bulk GaN wafers as substrates for epitaxial growth. However, epitaxial layers can be grown on other substrates with reasonable success, in spite of the lattice mismatch.

GaN exists in the cubic zinc blende form (which is the preferred structure) as well as the hexagonal wurtzite form. It was demonstrated recently that cubic GaN could be grown heteroepitaxially on sapphire, even though it is not lattice-matched to GaN. In fact, sapphire does not even have a cubic crystal structure—it is hexagonal. The lattice constant of GaN is about 4.5 Å, while that of sapphire is 4.8 Å, which is a huge lattice mismatch. Contrary to what would normally be expected, however, high-quality epitaxial GaN films can be grown on sapphire by MOCVD using ammonia and tri-methyl gallium as the precursors. One possible reason for the high quality of the films, as evidenced by blue LEDs and short wavelength lasers fabricated in these nitrides, is that these large-band-gap semiconductors have very high chemical bond strengths. This apparently precludes the easy propagation of dislocation defects from the heterointerface to the active part of the devices, where they would form traps and kill optical efficiency. Yet another lattice-mismatched substrate that has been successfully used for these nitride semiconductors is SiC.

The second breakthrough required in the nitrides was the ability to achieve high p-type doping so that p-n junctions could be formed. It has been demonstrated that Mg (which is a column II element) doping of MOCVD films, followed by high temperature annealing can be used to achieve high acceptor concentrations in these systems.

Why is there so much interest in short wavelength emitters such as blue LEDs and semiconductor lasers? As discussed in Section 8.2.1, high-efficiency red, green, and yellow-green LEDs have existed for a long time in the GaAsP system, using concepts such as N isoelectronic doping. It has been a major goal of the optoelectronics community to achieve high-efficiency blue emitters because, along with red and green, blue completes the list of three primary additive colors. In fact, blue LEDs made in GaN have been combined with the other color LEDs to form very intense white light sources with luminous efficiencies exceeding those of conventional lightbulbs. Arrays of red, green, and blue emitters can be used in outdoor displays and TV screens. Red, yellow, and green LEDs are candidates for traffic lights because they have much higher reliability and lifetime than conventional lightbulbs, and save energy.

Short wavelength emitters such as UV/blue semiconductor lasers are important for storage applications such as digital versatile discs (DVDs), which are higher density versions of compact discs (CDs). The storage density on these discs is inversely proportional to the square of the laser wavelength that is used to read the information. Thus reducing the laser wavelength by a factor of two leads to a four-fold increase of storage density.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 431

Such increased storage capacity opens up entirely new applications for DVDs that were not possible previously with conventional CDs, for example, the storage of full-length movies. A recent example of success in this rapidly progressing field is a 417-nm semiconductor laser made with InGaN multi-quantum-well heterostructures.

SUMMARY

8.1 The *emission and absorption of light by semiconductors* gives rise to useful *optoelectronic* devices. Emission typically occurs in *direct* band gap materials. Indirect semiconductors are less optically efficient because they require the involvement of both photons (which carry energy, but little momentum) and phonons (which carry momentum, but little energy).

8.2 In some optoelectronic devices, *electron–hole pairs* (EHPs) are *generated* by the *absorption* of *light*. These carriers can be collected as electrical energy in solar cells or photodetectors. Generated EHPs add to the thermally generated reverse current in diodes (Eq. 8–2), from which one can obtain the open-circuit voltage or short-circuit current in solar cells, a potential renewable source of energy.

8.3 In photodetectors, there is a trade-off between *speed* (which requires a short absorption region for a short transit time) and *sensitivity* (which requires long absorption regions). Clever structures to get *both* include *waveguide* photodetectors. One can achieve *gain* by using APDs, in which impact ionization can increase the number of EHPs generated for a certain number of absorbed photons. However, this can be at the expense of increased impact ionization *noise* in photodiodes.

8.4 Photodiodes, in conjunction with light emitters, are used with optical fibers for fiber-optic communication, which is the backbone of the Internet and worldwide communication today. Optical fibers are waveguides that work on the basis of the total internal reflection of light from a low-refractive-index outer layer to a high-refractive-index medium (glass).

8.5 Semiconductors can be used not only to *detect* photons, but also to *emit* photons by the *recombination* of EHPs in *direct* band gap semiconductors. Light emitters can be noncoherent, such as LEDs, or coherent, such as lasers. The color of light depends on the semiconductor, because the photon frequency is proportional to the band gap (Planck relationship).

8.6 Lasers are more difficult to make than LEDs, because they require not just EHPs, but *population inversion* (i.e., more carriers in the higher energy levels than in the lower ones). This enables *stimulated* emission of *phase-coherent* radiation (governed by the Einstein B coefficient) to dominate over *phase-incoherent spontaneous* radiation (determined by the Einstein A coefficient). Lasers also require an optical cavity to allow the photon field to build up. Light can come out from the side of the cavity, as in edge-emitting lasers, or from the top surface, as in vertical cavity surface-emitting lasers (VCSELs). Double-heterostructure lasers are useful because they allow both carrier and photon confinement.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 432

PROBLEMS

8.1 For the p-i-n photodiode of Fig. 8–7, (a) explain why this detector does not have gain; (b) explain how making the device more sensitive to low light levels degrades its speed; (c) if this device is to be used to detect light with $\lambda = 0.6$ μm, what material would you use and what substrate would you grow it on?

8.2 Assume that a rectangular bar photoconductor of length L and cross section A has a voltage V applied to it. We shine light on it to generate g_{op} electron–hole pairs uniformly. If the mobility μ_n and lifetime τ_n of the electrons are much higher than those for the holes, derive an expression for the change of conductance upon shining light. Derive an expression for the transit time τ_t of the photogenerated carriers across the photoconductor, assuming (a) low voltage and then (b) high voltage. The saturation velocity in this semiconductor is v_s.

8.3 For a Si photoconductor of length 5 μm, doped n-type at 10^{15} cm^{-3}, calculate the *change* in current density when we shine light on the photoconductor under the following circumstances: We create 10^{20} electron–hole pairs/cm^3-s and carrier-recombination lifetimes, $\tau = 0.1$ μs. The applied voltage is 2.5 V across the photoconductor's length. How about if we increase the voltage to 2500 V? The electron and hole mobilities are 1500 cm^2/V-s and 500 cm^2/V-s, respectively, in the ohmic region for electric fields below 10^3 V/cm. For higher fields, electrons and holes have a saturation velocity of 10^7 cm/s.

8.4 A Si solar cell with a dark saturation current I_{th} of 5 nA is illuminated such that the short-circuit current is 200 mA. Plot the I–V curve for the cell as in Fig. 8–6. (Remember that I is negative, but is plotted positive as I_r.)

8.5 Show schematically and discuss how several semiconductor materials might be used together to obtain a more efficient solar cell.

8.6 A major problem with solar cells is internal resistance, generally in the thin region at the surface, which must be only partially contacted, as in Fig. 8–5. Assume that the cell of Prob. 8.4 has a series resistance of 1 Ω, so that the cell voltage is reduced by the IR drop. Replot the I-V curve for this case and compare with the cell of Prob. 8.4.

8.7 (a) Why must a solar cell be operated in the fourth quadrant of the junction I–V characteristic?

(b) What is the advantage of a quaternary alloy in fabricating LEDs for fiber optics?

(c) Why is a reverse-biased GaAs p-n junction not a good photodetector for light of $\lambda = 1$ μm?

8.8 A Si solar cell 2 cm × 2 cm with $I_{th} = 32$ nA has an optical generation rate of 10^{18} EHP/cm^3 -s within $L_p = L_n = 2$ μm of the junction. If the depletion width is 1 μm, calculate the short-circuit current and the open-circuit voltage for this cell.

8.9 The maximum power delivered by a solar cell can be found by maximizing the I-V product.

(a) Show that maximizing the power leads to the expression

$$\left(1 + \frac{q}{kT}V_{mp}\right)e^{qV_{mp}/kT} = 1 + \frac{I_{sc}}{I_{th}}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 433

where V_{mp} is the voltage for maximum power, I_{sc} is the magnitude of the short-circuit current, and I_{th} is the thermally induced reverse saturation current.

(b) Write this equation in the form $\ln x = C - x$ for the case $I_{sc} \gg I_{th}$, and $V_{mp} \gg kT/q$.

(c) Assume a Si solar cell with a dark saturation current I_{th} of 1.5 nA is illuminated such that the short-circuit current is $I_{sc} = 100$ mA. Use a graphical solution to obtain the voltage V_{mp} at maximum delivered power.

(d) What is the maximum power output of the cell at this illumination?

8.10 For a solar cell, Eq. (8-2) can be rewritten

$$V = \frac{kT}{q}\ln\left(1 + \frac{I_{sc} + I}{I_{th}}\right)$$

Given the cell parameters of Prob. 8.9, plot the I-V curve as in Fig. 8-6 and draw the maximum power rectangle. Remember that I is a negative number but is plotted positive as I, in the figure. I_{th} and I_{sc} are positive magnitudes in the equation.

8.11 Solar cells are severely degraded by unwanted series resistance. For the cell described in Prob. 8.6, include a series resistance R, which reduces the cell voltage by the amount IR. Calculate and plot the fill factor for a series resistance R from 0 to 5 Ω, and comment on the effect of R on cell efficiency.

8.12 If the band gap of a semiconductor is 1.8 eV, what are the frequency and momentum of the photon that is emitted?

8.13 If one makes an LED in a semiconductor with a band gap of 2.5 eV, what wavelength of light will it emit? Can you use it to efficiently detect photons of wavelength 1900 nm? 100 nm?

8.14 The degenerate occupation of bands shown in Fig. 8-19 helps maintain the laser requirement that emission must overcome absorption. Explain how the degeneracy prevents band-to-band absorption at the emission wavelength.

8.15 Assume that the system described by Eq. (8-7) is in thermal equilibrium at an extremely high temperature such that the energy density $\rho(\nu_{12})$ is essentially infinite. Show that $B_{12} = B_{21}$.

8.16 The system described by Eq. (8-7) interacts with a blackbody radiation field whose energy density per unit frequency at ν_{12} is

$$\rho(\nu_{12}) = \frac{8\pi h\nu_{12}^3}{c^3}\left[e^{h\nu_{12}/kT} - 1\right]^{-1}$$

from Planck's radiation law. Given the result of Prob. 8.15, find the value of the ratio A_{21}/B_{12}.

8.17 Assuming equal electron and hole concentrations and band-to-band transitions, calculate the minimum carrier concentration $n = p$ for population inversion in GaAs at 300 K. The intrinsic carrier concentration in GaAs is about 10^6 cm^{-3}.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 434

8.18 For steady state optical excitation, we can write the hole diffusion equation as

$$D_p \frac{d^2\delta p}{dx^2} = \frac{\delta p}{\tau_p} - g_{op}$$

Assume that a long p⁺-n diode is uniformly illuminated by an optical signal, resulting in g_{op} EHP/cm³-s. Show that the excess hole distribution in the n region is

$$\delta p(x_n) = \left[p_n(e^{qV/kT} - 1) - g_{op}\frac{L^2_p}{D_p} \right] e^{-x_n/L_p} + \frac{g_{op}L^2_p}{D_p}$$

READING LIST

Bhattacharya, P. *Semiconductor Optoelectronic Devices.* Englewood Cliffs, NJ: Prentice Hall, 1994.

Campbell, J. C., A. G. Dentai, W. S. Holden, and B. L. Kasper. "High Performance Avalanche Photodiode with Separate Absorption, Grading and Multiplication Regions." *Electronics Letters*, 19 (1983): 818+.

Casey, Jr., H. C., and M. B. Panish. *Heterostructure Lasers: Part A. Fundamental Principles.* New York: Academic Press, 1978.

Cheo, P. K. *Fiber Optics and Optoelectronics,* 2d ed. Englewood Cliffs, NJ: Prentice Hall, 1990.

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

Dupuis, R. D. "AlGaAs-GaAs Lasers Grown by MOCVD—A Review." *Journal of Crystal Growth* 55 (October 1981): 213–222.

Jewell, J. L., and G. R. Olbright. "Surface-Emitting Lasers Emerge from the Laboratory." *Laser Focus World* 28 (May 1992): 217–223.

Palais, J. C. *Fiber Optic Communication*, 3d ed. Englewood Cliffs, NJ: Prentice Hall, 1992.

Pankove, J. I. *Optical Processes in Semiconductors.* Englewood Cliffs, NJ: Prentice Hall, 1971.

Singh, J. *Optoelectronics: An Introduction to Materials and Devices.* New York: McGraw-Hill, 1996.

Verdeyen, J. T. *Laser Electronics,* 3d ed. Englewood Cliffs, NJ: Prentice Hall, 1994.

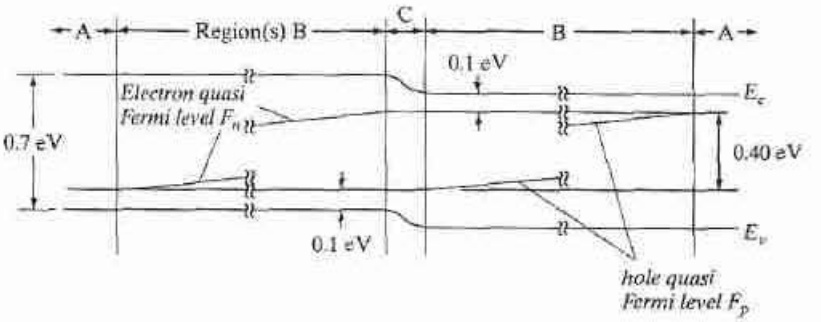
Yamamoto, Y., and R. E. Slusher. "Optical Processes in Microcavities." *Physics Today* 46 (June 1993): 66–73.

SELF QUIZ

Question 1

Consider the following band diagram of a simple light-emitting diode. Assume essentially all recombination is direct and results in light emission. The forward-bias current consists of holes injected from a contact on the left and electrons injected from a contact on the right.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 435



(a) In which region would you expect the optical recombination rate to be the greatest? Circle one.

Region(s) A / Region(s) B / Region C

(b) What is the approximate energy of the emitted photons in eV?

(c) For a steady state current $I = 10$ mA, assuming all photons escape, what is the optical output power consistent with your answer to part (b)? _____ W

(hint) Watts = Amps · eV/q = Amps · Volts

(d) If the voltage drop across the depletion region is 0.4 V, what is the separation of the quasi-Fermi levels in Region C? How can this be less than the total forward bias of 1.4 V?

(e) What is the electrical power consumed? _____ W

(f) In terms of the ratio of optical power out (c) to electrical power in (e), how efficient is this light-emitting diode? _____ %

Question 2

A solar cell has a short-circuit current of 50 mA and an open-circuit voltage of 0.7 V under full illumination. What is the maximum power delivered by this cell if the fill factor is 0.8?

Question 3

If one makes an LED in a semiconductor with a direct band gap of 2.5 eV, what wavelength light will it emit? Can you use it to detect photons of wavelength 0.9 μm? 0.1 μm?

Question 4

What is most attractive about solar cells as a global energy source? Why haven't they been adopted more widely so far?

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 436

Chapter 9
Integrated Circuits

OBJECTIVES

1. Understand IC scaling issues
2. Describe CMOS process integration
3. Understand the basics of logic gates and CCDs
4. Study the operation of different memory cells—SRAMs, DRAMs, flash
5. Understand IC packaging

Just as the transistor revolutionized electronics by offering more flexibility, convenience, and reliability than the vacuum tube, the integrated circuit enables new applications for electronics that were not possible with discrete devices. Integration allows complex circuits consisting of many thousands of transistors, diodes, resistors, and capacitors to be included in a chip of semiconductor. This means that sophisticated circuitry can be miniaturized for use in space vehicles, in large-scale computers, and in other applications where a large collection of discrete components would be impractical. In addition to offering the advantages of miniaturization, the simultaneous fabrication of many ICs on a single Si wafer greatly reduces the cost and increases the reliability of each of the finished circuits. Certainly discrete components have played an important role in the development of electronic circuits; however, most circuits are now fabricated on the Si chip rather than with a collection of individual components. Therefore, the traditional distinctions between the roles of circuit and system designers do not apply to IC development.

In this chapter we shall discuss various types of ICs and the fabrication steps used in their production. We shall investigate techniques for building large numbers of transistors, capacitors, and resistors on a single chip of Si, as well as the interconnection, contacting, and packaging of these circuits in usable form. All the processing techniques discussed here are very basic and general. There would be no purpose in attempting a comprehensive review of all the subtleties of device fabrication in a book of this type. In fact, the only way to keep up with such an expanding field is to study the current literature. Many good reviews are suggested in the reading list at the end of this chapter; more important, current issues of those periodicals cited can be

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 437

consulted for up-to-date information regarding IC technology. Having the background of this chapter, one should be able to read the current literature and thereby keep abreast of the present trends in this very important field of electronics.

9.1 BACKGROUND

In this section we provide an overview of the nature of integrated circuits and the motivation for using them. It is important to realize the reasons, both technical and economic, for the dramatic rise of ICs to their present role in electronics. We shall discuss several main types of ICs and point out some of the applications of each. More specific fabrication techniques will be presented in later sections.

9.1.1 Advantages of Integration

It might appear that building complicated circuits, involving many interconnected components on a single Si substrate, would be risky both technically and economically. In fact, however, modern techniques allow this to be done reliably and relatively inexpensively; in most cases an entire circuit on a Si chip can be produced much more inexpensively and with greater reliability than a similar circuit built up from individual components. The basic reason is that many identical circuits can be built simultaneously on a single Si wafer (Fig. 9-1); this process is called *batch fabrication*. Although the processing steps for the wafer are complex and expensive, the large number of resulting integrated circuits makes the ultimate cost of each fairly low. Furthermore, the processing steps are essentially the same for a circuit containing millions of transistors as for a simpler circuit. This drives the IC industry to build increasingly complex circuits and systems on each chip, and use larger Si wafers (e.g., 8-inch diameter). As a result, the number of components in each circuit increases without a proportional increase in the ultimate cost of the system. The implications of this principle are tremendous for circuit designers; it greatly increases the flexibility of design criteria. Unlike circuits with individual transistors and other components wired together or placed on a circuit board, ICs allow many "extra" components to be included without greatly raising the cost of the final product. Reliability is also improved since all devices and interconnections are made on a single rigid substrate, greatly minimizing failures due to the soldered interconnections of discrete component circuits.

The advantages of ICs in terms of miniaturization are obvious. Since many circuit functions can be packed into a small space, complex electronic equipment can be employed in many applications where weight and space are critical, such as in aircraft or space vehicles. In large-scale computers it is now possible not only to reduce the size of the overall unit but also to facilitate maintenance by allowing for the replacement of entire circuits quickly and easily. Applications of ICs are pervasive in such consumer products as

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 438



Figure 9–1
A 300-mm-diameter (about 12-inch) wafer of integrated circuits. The circuits are tested on the wafer and then sawed apart into individual chips for mounting into packages. (Photograph courtesy of Texas Instruments.)

watches, calculators, automobiles, telephones, television, and appliances. Miniaturization and the cost reduction provided by ICs mean that we all have increasingly more sophisticated electronics at our disposal.

Some of the most important advantages of miniaturization pertain to response time and the speed of signal transfer between circuits. For example, in high-frequency circuits it is necessary to keep the separation of various components small to reduce time delay of signals. Similarly, in very high-speed computers it is important that the various logic and information storage circuits be placed close together. Since electrical signals are ultimately limited by the speed of light (about 1 ft/ns), physical separation of the circuits can be an important limitation. As we shall see in Section 9.5, *large-scale integration* of many circuits on a Si chip has led to major reductions in computer size, thereby tremendously increasing speed and function density. In addition to decreasing the signal transfer time, integration can reduce parasitic capacitance and inductance between circuits. Reduction of these parasitics can provide significant improvement in the operating speed of the system.

We have discussed several advantages of reducing the size of each unit in the batch fabrication process, such as miniaturization, high-frequency and

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 439

switching speed improvements, and cost reduction due to the large number of circuits fabricated on a single wafer. Another important advantage has to do with the percentage of usable devices (often called the *yield*) which results from batch fabrication. Faulty devices usually occur because of some defect in the Si wafer or in the fabrication steps. Defects in the Si can occur because of lattice imperfections and strains introduced in the crystal growth, cutting, and handling of the wafers. Usually such defects are extremely small, but their presence can ruin devices built on or around them. Reducing the size of each device greatly increases the chance for a given device to be free of such defects. The same is true for fabrication defects, such as the presence of a dust particle on a photolithographic mask. For example, a lattice defect or dust particle $\frac{1}{2}\,\mu m$ in diameter can easily ruin a circuit which includes the damaged area. If a fairly large circuit is built around the defect it will be faulty; however, if the device size is reduced so that four circuits occupy the same area on the wafer, chances are good that only the one containing the defect will be faulty and the other three will be good. Therefore, the percentage yield of usable circuits increases over a certain range of decreasing chip area. There is an optimum area for each circuit, above which defects are needlessly included and below which the elements are spaced too closely for reliable fabrication.

9.1.2 Types of Integrated Circuits

There are several ways of categorizing ICs as to their use and method of fabrication. The most common categories are *linear* or *digital*, according to application, and *monolithic* or *hybrid*, according to fabrication.

A linear IC is an IC that performs amplification or other essentially linear operations on signals. Examples of linear circuits are simple amplifiers, operational amplifiers, and analog communications circuits. Digital circuits involve logic and memory, for applications in computers, calculators, microprocessors, and the like. By far the greatest volume of ICs has been in the digital field, since large numbers of such circuits are required. Because digital circuits generally require only the "on–off" operation of transistors, the design requirements for integrated digital circuits are often less stringent than for linear circuits. Although transistors can be fabricated as easily in an integrated form as in a discrete form, passive elements (resistors and capacitors) are usually more difficult to produce to close tolerances in ICs.

Integrated circuits that are included entirely on a single chip of semiconductor (usually Si) are called *monolithic* circuits (Fig. 9-1). The word *monolithic* literally means "one stone" and implies that the entire circuit is contained in a single piece of semiconductor. Any additions to the semiconductor sample, such as insulating layers and metallization patterns, are intimately bonded to the surface of the chip. A *hybrid* circuit may contain one or more monolithic circuits or individual transistors bonded to an insulating substrate with resistors, capacitors, or other circuit elements, together with appropriate interconnections. Monolithic circuits have the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 440

advantage that all of their components are contained in a single rigid structure that can be batch fabricated; that is, hundreds of identical circuits can be built simultaneously on a Si wafer. On the other hand, hybrid circuits offer excellent isolation between components and allow the use of more precise resistors and capacitors. Furthermore, hybrid circuits are often less expensive to build in small numbers.

9.2 EVOLUTION OF INTEGRATED CIRCUITS

The IC was invented in February 1959 by Jack Kilby of Texas Instruments. The planar version of the IC was developed independently by Robert Noyce at Fairchild in July 1959. Since then, the evolution of this technology has been extremely fast paced. One way to gauge the progress of the field is to look at the complexity of ICs as a function of time. Figure 9-2 shows the number of transistors used in MOS *microprocessor* IC chips as a function of time. It is amazing that on this semilog plot, where we have plotted the log of the component count as a function of time, we get a straight line over three decades, indicating that there has been an exponential growth in the complexity of chips. The component count has roughly doubled every 18 months, as was noted early on by Gordon Moore of Intel corporation. This regular doubling has become known as *Moore's law*.

The history of ICs can be described in terms of different eras, depending on the component count. Small-scale integration (SSI) refers to the integration of $1-10^2$ devices, medium-scale integration (MSI) to the integration of 10^2-10^3 devices, large-scale integration (LSI) to the integration of 10^3-10^5 devices, very large-scale integration (VLSI) to the integration of 10^5-10^6 devices, and now ultra large scale integration (ULSI) to the integration of 10^6-10^9 devices. Of course, these boundaries are somewhat fuzzy. The next generation has been dubbed *gigascale integration* (GSI). Wags have suggested that after that we will have RLSI, or "ridiculously large-scale integration."

The main factor that has enabled this increase in complexity is the ability to shrink or scale devices. Typical dimensions or feature sizes of state-of-the-art *dynamic random-access memories (DRAMs)* at different times are also shown as a semilog plot in Fig. 9-3. Once again, we see a straight line, reflecting an exponential decrease in the typical feature sizes with time over three decades. Clearly, one can pack a larger number of components with greater functionality on an IC if they are smaller. As discussed in Section 6.5.9, scaling also has other advantages in terms of faster ICs, which consume less power.

While scaling represents an opportunity, it also presents tremendous technological challenges. The most notable among these challenges lie in lithography and etching, as discussed in Section 5.1. However, since scaling of horizontal dimensions also requires scaling of vertical geometries, as discussed in Section 6.5.9, there are also tremendous challenges in terms of doping, gate dielectrics, and metallization. In addition, small features and large

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 441

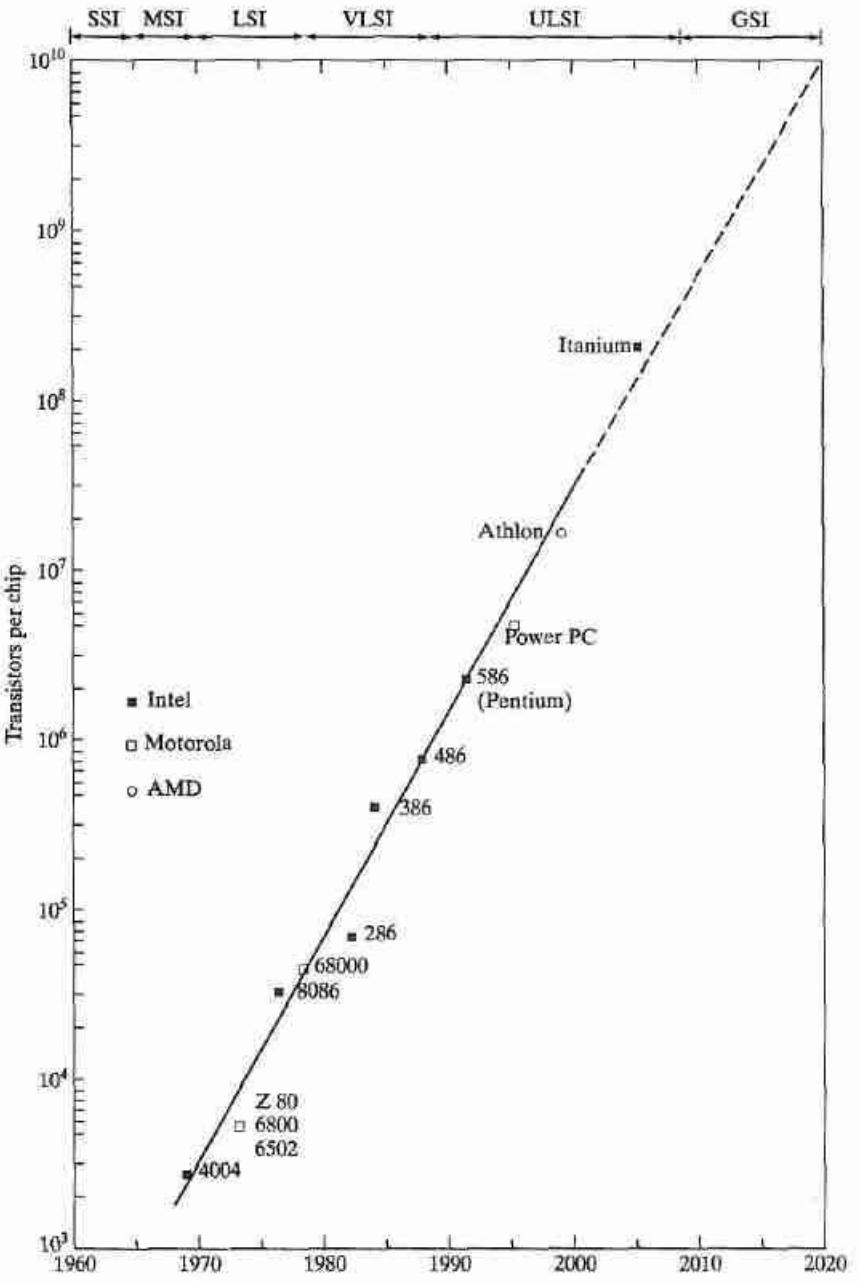


Figure 9–2
Moore's law for integrated circuits: Exponential increase in transistor count as a function of time for different generations of microprocessors. The dashed line indicates projections based on the International Technology Roadmap for Semiconductors (ITRS). Notice that the transistor count in the future may not increase at the same rate as in the past, due to practical constraints such as economics and power dissipation.

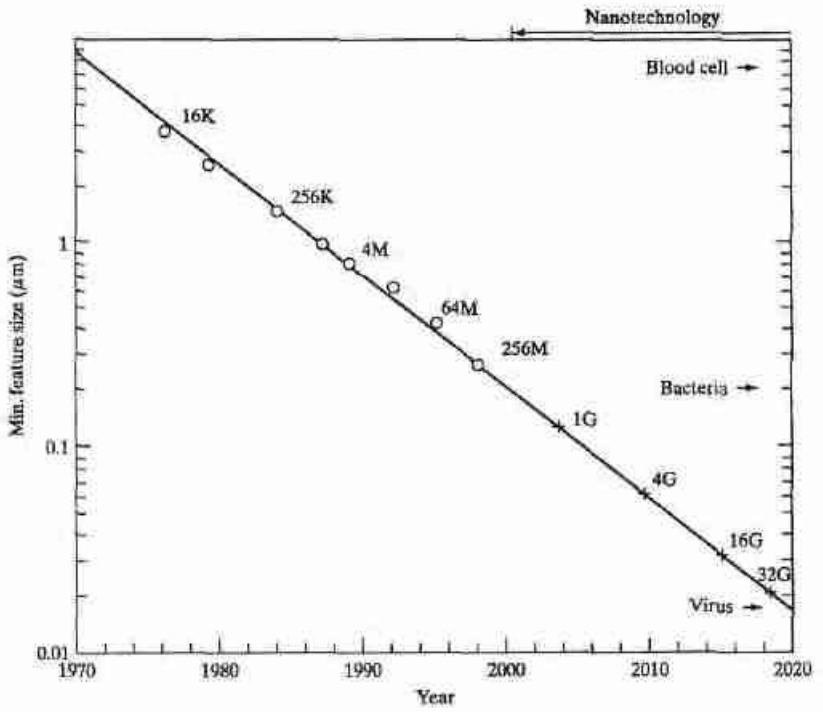


Figure 9–3
Exponential decrease in typical feature size with time for different generations of dynamic random-access memories (16-kb to 32-Gb DRAMs). For reference, sizes of blood cells, bacteria, and viruses are shown on the μm scale. The dashed line indicates projections based on the ITRS. Dimensions below 100 nm are considered to be in the realm of nanotechnology.

chips require device fabrication in extremely clean environments. Particles that may not have caused yield problems in a 1-μm IC technology can have catastrophic effects for a 0.25-μm process, which requires purer chemicals, cleaner equipment, and more stringent clean rooms. In fact, the levels of cleanliness required bypassed the best surgical operating rooms early in the evolution shown in Fig. 9-3. The cleanliness of these facilities is designated by the *class* of the clean room. For instance, a Class 1 clean room, which was state-of-the-art in 2000, has less than 1 particle of size 0.2 μm or larger per cubic foot. There are more of the smaller particles and fewer of the larger ones. Obviously, the lower the class of a clean room, the better it is. A Class 1 clean room is much cleaner than a Class 100 fabrication facility, or "fab." As one might expect, such high levels of cleanliness come with a hefty price tag: A state-of-the-art fab in 2000 comes equipped with a price tag of about 2 billion dollars.

In spite of the costs, the economic payoff for ULSI is tremendous. Just for calibration, let us examine some economic statistics at the dawn of the third millennium. The total annual economic output of all the countries in the world, or the so-called gross world product (GWP), is about 50 trillion US dollars. The US gross national product is about 10 trillion dollars, or about a fifth of the GWP. The worldwide IC industry output is about 200 billion dollars, and that of the entire worldwide electronics industry in which these ICs

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 443

participate is about 1 trillion dollars. As a single industry, electronics is one of the biggest in terms of the dollar amount. It has surpassed, for example, automobiles (worldwide sales of about 50 million cars annually) and petrochemicals. About 100–200 million personal computers are sold annually worldwide.

Perhaps even more dramatic than these raw economic numbers is the growth rate of these markets. If one were to plot IC sales as a function of time, one again finds a more-or-less exponential increase in sales with time over three decades. Of great importance to the consumer, the cost per electronic function has dropped dramatically over the same period. For example, the cost per bit of semiconductor memory (DRAM) has dropped from about 1 cent/bit in 1970 to about 10^{-4} cent per bit today, an improvement of four orders of magnitude in 30 years. There are no parallels in any other industry for this consistent improvement in functionality with such lowered cost.

While ICs started with bipolar processes in the 1960s, they were gradually supplanted by MOS and then CMOS devices, for reasons discussed in Chapters 6 and 7. Currently, about 88% of the IC market is MOS based and about 8% BJT based. Optoelectronic devices based on compound semiconductors are still a relatively small component of the semiconductor market (about 4%), but are expected to grow in the future. Of the MOS ICs, the bulk are digital ICs. Of the entire semiconductor industry, only about 14% are analog ICs. Semiconductor memories such as DRAMs, SRAMs, and nonvolatile flash memories make up approximately 25% of the market, microprocessors about 25%, and other application-specific ICs (ASICs) about 20%.

9.3 MONOLITHIC DEVICE ELEMENTS

Now we shall consider the various elements that make up an integrated circuit, and some of the steps in their fabrication. The basic elements are fairly easy to name—transistors, resistors, capacitors, and some form of interconnection. There are some elements in integrated circuits, however, which do not have simple counterparts in discrete devices. We shall consider one of these, charge transfer devices, in Section 9.4. Discussion of fabrication technology is difficult in a book of this type, since device fabrication engineers seem to make changes faster than typesetters do! Since this important and fascinating field is changing so rapidly, the reader should obtain a basic understanding of device design and processing from this discussion and then search out new innovations in the current literature.

9.3.1 CMOS Process Integration

A particularly useful device for digital applications is a combination of n-channel and p-channel MOS transistors on adjacent regions of the chip. This *complementary MOS* (commonly called *CMOS*) combination is illustrated in

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 444

the basic inverter circuit of Fig. 9-4a. In this circuit the drains of the two transistors are connected together and form the output, while the input terminal is the common connection to the transistor gates. The p-channel device has a negative threshold voltage, and the n-channel transistor has a positive threshold voltage. Therefore, a zero voltage input ($V_{in} = 0$) gives zero gate voltage for the n-channel device, but the voltage between the gate and source of the p-channel device is $-V_{DD}$. Thus the p-channel device is on, the n-channel device is off, and the full voltage V_{DD} is measured at V_{out} (i.e., V_{DD} appears across the nonconducting n-channel transistor). Alternatively, a positive value of V_{in} turns the n-channel transistor on, and the p-channel off. The output voltage measured across the "on" n-channel device is essentially zero. Thus, the circuit operates as an inverter—with a binary "1" at the input, the output is in the "0" state, whereas a "0" input produces a "1" output. The beauty of this circuit is that one of the devices is turned off for either condition. Since the devices are connected in series, no drain current flows, except for a small charging current during the switching process from one state to the other. Since the CMOS inverter uses ultra little power, it is particularly useful in applications such as electronic watch circuits which depend on very low power consumption. CMOS is also advantageous in ultra large-scale integrated circuits (Section 9.5), since even small power dissipation in each transistor becomes a problem when millions of them are integrated on a chip.

The device technology for achieving CMOS circuits consists mainly in arranging for both n- and p-channel devices with similar threshold voltages on

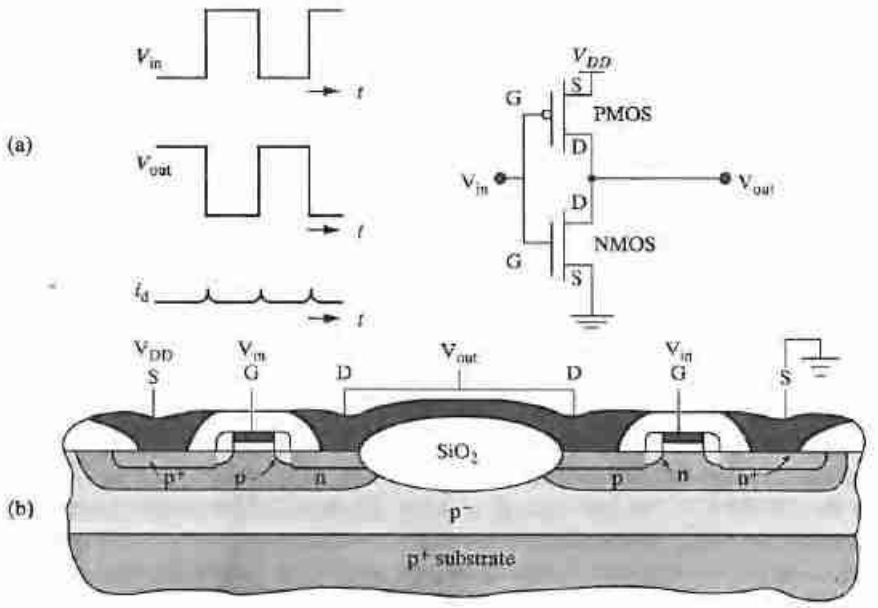


Figure 9–4 Complementary MOS structure: (a) CMOS inverter; (b) formation of p-channel and n-channel devices together.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 445

the same chip. To achieve this goal, a diffusion or implantation must be performed in certain areas to obtain n and p regions for the fabrication of each type of device. These regions are called *tubs, tanks,* or *wells* (Fig. 9–5). The critical parameter of the tub is its net doping concentration, which must be closely controlled by ion implantation. With the tub in place, source and drain implants are performed to make the n-channel and p-channel transistors. Matching of the two transistors is achieved by control of the surface doping in the tub and by threshold adjustment of both transistors by ion implantation.

Including bipolar transistors in the basic CMOS technology allows flexibility in circuit design, particularly for providing drive currents. The combination of bipolar and CMOS (called BiCMOS) provides circuits with increased speed.

Attention must be paid in CMOS designs to the fact that combining n-channel and p-channel devices in close proximity can lead to inadvertent (*parasitic*) bipolar structures. In fact, a p-n-p-n structure can be found in Fig. 9–4b, which can serve as an inefficient but troublesome *thyristor* (see Chapter 10). Under certain biasing conditions the p-n-p part of the structure can supply base current to the n-p-n structure, causing a large current to flow. This process, called *latchup*, can be a serious problem in CMOS circuits. Several methods have been used to eliminate the latchup problem, including using both n-type and p-type tubs, separated by trench isolation (Fig. 9–6). The use of two separate tubs (wells) also allows independent control of threshold voltages in both types of transistor.

We can illustrate most of the common fabrication steps for MOS integrated circuits by studying the flow of a twin-well Self-Aligned siLICIDE (SALICIDE) CMOS process. This process is particularly important because most high-performance digital ICs, including microprocessors, memories, and application-specific ICs (ASICs), are made basically in this way. In order to make enhancement-mode n-channel devices, we need a p-type substrate, and vice versa. Since CMOS requires both, we must start either with an n-type or a p-type wafer and then make selected regions of the substrate have opposite doping by forming wells. For example, Fig. 9–5a shows a lightly doped p-epitaxial layer on a p^+-substrate. We can make n-channel devices in this layer. By implanting n-wells wherever needed, we can make p-channel devices also. This is an n-well CMOS process. Alternatively, if we start with an n-substrate and make p-wells in certain regions, we have a p-well CMOS process. For optimal device performance, however, it is usually desirable to separately implant both the n- and the p-well regions, which is called *twin-well* CMOS. The rationale for this can be appreciated if we keep in mind that for a state-of-the-art IC, typical doping levels are $\sim 10^{18}$ cm^{-3} and junction depths are ~1 μm in these wells. The doping levels have to be high enough to prevent punch-through breakdown due to drain-induced barrier lowering (DIBL) in the MOSFETs, but low enough to keep the threshold voltages acceptably small. If we choose to use the p-substrate for n-channel devices as in Fig. 9–5a, the p-type epitaxial layer must be doped to 10^{18} cm^{-3}, and the implanted n-type layer must be achieved by counter-doping at a level of $\sim 2\times 10^{18}$ cm^{-3},

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 446

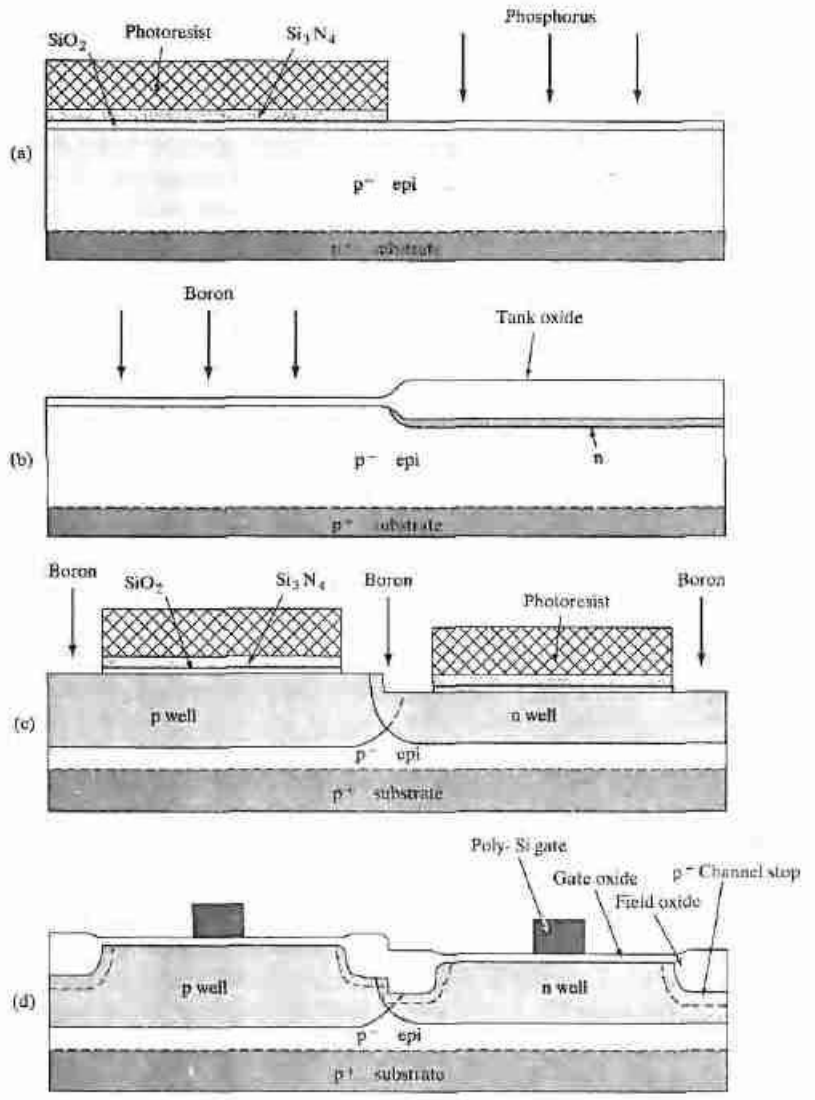


Figure 9–5
Self-aligned twin well process: (a) an n-well formation using P donor implant and a photoresist mask; (b) p-well formation using B acceptor implant. A thick (~200 nm) "tank" oxide layer is grown wherever the silicon nitride–oxide stack is etched off, and the tank oxide is used to block the B implant in the n-wells in a self-aligned manner; (c) an isolation pattern for field transistors showing a B channel stop implant using a photoresist mask; (d) local oxidation of silicon wherever the nitride mask is removed, leading to thick LOCOS field oxide.

resulting in a net n-type doping of $\sim 1\times 10^{18}$ cm^{-3}, but a total doping in this region of $\sim 3\times 10^{18}$ cm^{-3}. Such high levels of total doping are detrimental to carrier transport because they cause excessive ionized impurity scattering. Hence, for high performance ICs, the starting epitaxial doping level is generally very low ($\sim 10^{16}$ cm^{-3}). This layer is grown on a heavily doped substrate ($\sim 10^{19}$ cm^{-3}), to provide a highly conducting electrical ground plane. This helps with noise problems in ICs and helps minimize the problem of *latchup* by bypassing majority carriers (in this case holes) to the p$^+$ substrate.

To form the twin wells in a self-aligned fashion, we first grow thermally a "pad" oxide (~20 nm) on the Si substrate, followed by low-pressure chemical

vapor deposition (LPCVD) of silicon nitride (~20 nm). As shown in Fig. 9–5a this oxide–nitride stack is covered by photoresist, and a window is opened for the n-well. Reactive ion etching (RIE) is then used to etch the oxide–nitride stack. Using the photoresist as an implant mask, we then do an n-type implant using phosphorus. Phosphorus is preferred to As for this purpose because P is lighter and has a higher projected range; also, P diffuses faster. This fast diffusion is needed to drive the dopants fairly deep into the substrate to form the n-well. After the implant, the photoresist is removed, and the patterned wafer is subjected to wet oxidation to grow a "tank" oxide (~200 nm). Note in Fig. 9–5b that the tank oxidation process consumes Si from the substrate, and the resulting oxide swells up. In fact, for every micron of thermally grown oxide, the oxidation consumes 0.44 μm of Si, resulting in a 2.2-fold volume expansion. The oxide does not grow in the regions that are protected by silicon nitride, because nitride blocks the diffusion of oxygen and water molecules (and thereby prevents oxidation of the Si substrate). The pad oxide that is used under the nitride has two roles: It minimizes the thermal-expansion mismatch and concomitant stress between silicon nitride and the substrate; it also prevents chemical bonding of the silicon nitride to the silicon substrate.

Using the tank oxide as a *self-aligned* implant mask (i.e., without actually having to do a separate photolithographic step), one does a p-type well implant using boron (B; Fig. 9–5b). The tank oxide must thus be much thicker than the projected range of the B. The concept of self-alignment is very important and is a recurring theme in IC processing. It is simpler and cheaper to use self-alignment than a separate lithographic step. It also allows a tighter packing density of the twin wells, because it is not required to account for lithographic misalignment during layout. The P and the B are then diffused into the substrate to a well depth of typically a micron by a drive-in diffusion at very high temperatures (~1000°C) for several hours. After this diffusion, the silicon nitride–oxide stack and the tank oxide are etched away. Since the tank oxidation consumes Si from the substrate, etching it off leads to a step in the Si substrate delineating the n-well and p-well regions. The step is important in terms of the alignment of subsequent reticles and is shown in an exaggerated fashion in Fig. 9–5c. The step, however, is disadvantageous from the depth-of-focus point of view during lithography. So, instead of a self-aligned twin-well process as just described, one often uses two separate lithography steps for the two well implants. This approach leads to a more planar structure, but a slightly larger well geometry because the wells are not self-aligned. There are many other variations of CMOS processes used, depending on the application.

Next, we form the isolation regions or the field transistors which guarantee that there will be no electrical cross talk between adjacent transistors, unless they are intentionally interconnected (Fig. 9–5c). This is achieved by ensuring that the threshold voltage of any parasitic transistor that may form in the isolation regions is much higher than the power supply voltage on the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 448

chip, so that the parasitic channel can never turn on under operating conditions. From the threshold voltage expression [Eq. (6–38)], we notice that V_T can be raised by increasing substrate doping and increasing gate oxide thickness. However, a problem with that approach is the subthreshold slope S [Eq. (6–66)], which degrades with increasing substrate doping and gate oxide thickness. One needs to optimize both V_T and S such that the off-state leakage current in the field between transistors is sufficiently low at zero gate bias.

A stack of silicon dioxide-silicon nitride is photolithographically patterned as in Fig. 9-5c and subjected to RIE. A boron *"channel stop"* implant between the twin wells increases the acceptor doping and thus increases the threshold voltage in the p-well between the n-channel transistors (the *field* threshold). However, B will compensate the donor doping on the n-well side, and thus reduce the threshold in the n-well between p-channel devices. The B channel stop dose must thus be optimized to have acceptably high field thresholds in both types of wells.

After the channel stop implant, the photoresist is removed and the wafer with the patterned nitride-oxide stack shown in Fig. 9-5c (without photoresist) is subjected to wet oxidation to selectively grow a field oxide ~300 nm thick. The nitride layer blocks oxidation of the Si substrate in the regions where we plan to make the transistors. This procedure, where Si is oxidized to form SiO_2 in regions not protected by nitride, is called LOCal Oxidation of Silicon (LOCOS) (Section 6.4.1). In this case, LOCOS provides electrical isolation between the two transistors, as shown in Fig. 9-5d.

The volume expansion of 2.2× upon oxidation is an important issue because the selective oxidation occurs in narrow, confined regions. The compressive stress, if excessive, can cause dislocation defects in the substrate. Another issue is the lateral oxidation near the nitride mask edges, which causes the nitride mask to lift up near the edges, forming what is known as a *bird's beak* and causing a *lateral moat encroachment* of ~0.2 μm into each active region, thereby wasting precious Si real estate. There have been various modified LOCOS and other isolation schemes proposed to minimize this lateral encroachment. A notable example is *Shallow Trench Isolation* (STI) which involves using RIE to etch a shallow (~1 μm) trench or groove in the Si substrate after the isolation pattern, filling it up completely by deposition of a dielectric layer of SiO_2 and polysilicon by low-pressure chemical vapor deposition (LPCVD), and then using chemical mechanical polishing (CMP) to planarize the structure (Fig. 9-6). This consumes less Si real estate compared to LOCOS, but gives superior isolation because the sharp corners at the bottom of the trench give rise to potential barriers that block leakage currents (the *corner* effect).

The pad oxide between the nitride and the Si surface minimizes the stress due to the nitride, and prevents bonding of the nitride to the Si, as mentioned above. Any residual nitride on the Si would retard subsequent gate oxide formation, leading to weak spots in the gate region of the MOSFETs. This problem is known as the *white ribbon* effect or the *Kooi* effect, after the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 449

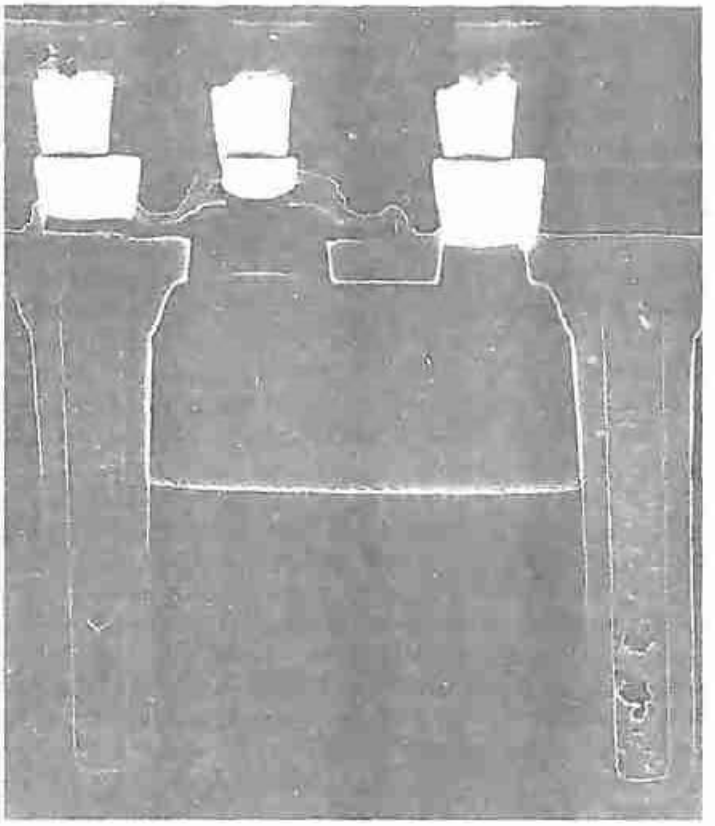


Figure 9–6
Trench isolation: A trench or groove is etched in the substrate with the use of RIE and is refilled with oxide and polysilicon. Trench isolation provides superior electrical isolation compared with LOCOS, and uses less Si real estate.

Dutch scientist who first identified it. The pad oxide mitigates this problem, but does not solve it completely. Therefore, very often a "sacrificial" or "dummy" oxide is grown to consume a layer of Si containing any residual nitride, and this oxide is wet etched prior to the growth of the actual gate oxide.

Next, an ultra-thin (~5–10 nm) gate oxide is grown on the substrate. Since the electrical quality of this oxide and its interface with the Si substrate is of paramount importance to the operation of the MOSFETs, dry oxidation is used for this step. It is common to incorporate some nitrogen at the Si-SiO_2 interface, forming oxy-nitrides which improve the interface quality in terms of hot electron effects. After the oxidation, it is immediately covered with LPCVD polysilicon in order to minimize contamination of the gate oxide. The polysilicon gate layer is doped very heavily (typically n^+ using a phosphorus dopant source, $POCl_3$ in a diffusion furnace) all the way to the polysilicon-oxide interface in order to make it behave electrically like a metal electrode. Alternatively, the LPCVD polysilicon film may also be *in situ* doped during the deposition itself by flowing in an appropriate dopant gas such as phosphine or diborane. Heavy doping of the gate material is very important,

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 450

because otherwise a depletion layer can be formed in the polysilicon gate (the *poly depletion* effect). This could result in a depletion capacitance in series with the gate oxide capacitance, thereby reducing the overall gate capacitance and, therefore, the drive current [see Eq. (6–53)]. The high doping ($\sim 10^{20}$ cm^{-3}) in the polysilicon gate is also important for reducing the resistance of the gate and its *RC* time constant. The uniformly high doping in the polysilicon layer is facilitated by the presence of the grain boundary defects in the film, because diffusivity of dopants along grain boundaries is many orders of magnitude higher than in single-crystal Si.

The doped polysilicon layer is then patterned to form the gates, and etched anisotropically by RIE to achieve vertical sidewalls. That is extremely important because this etched polysilicon gate is used as a self-aligned implant mask for the source/drain implants. As mentioned above, self-aligned processes are always desirable in terms of process simplicity and packing density. It is particularly useful in this case because we thereby guarantee that there will be *some* overlap of the gate with the source/drain but minimal overlap. The overlap is determined by the lateral scattering of the ions and by the lateral diffusion of the dopants during subsequent thermal processing (such as source/drain implant anneals). If there were no overlap, the channel would have to be turned on in this region by the gate fringing fields. The resulting potential barrier in the channel would degrade the device current. On the other hand, if there is too much overlap, it leads to an overlap capacitance between the source or drain and the gate. This is particularly bothersome near the drain end because it leads to the Miller overlap capacitance which causes undesired capacitive feedback between the output drain terminal and the input gate terminal (see Section 6.5.8).

Fabrication steps for the n-channel MOSFETs in the p-well are shown in Fig. 9–7. After the polysilicon gate is etched, we first do a self-aligned n-type source–drain implant, during which the tank masking level is used to protect the PMOS devices with a layer of photoresist. The NMOS source and drain implants are done in two stages. The first implant is a lightly doped drain (LDD) implant (Fig. 9–7a). This is typically a dose of $\sim 10^{13}$–10^{14} cm^{-2}, corresponding to a concentration of 10^{18}–10^{19} cm^{-3}, and an ultra-shallow junction depth of 50–100 nm. When a MOSFET is operated in the saturation region, the drain-channel junction is reverse biased, resulting in a very high electric field in the pinch-off region. As we saw in Section 5.4 for reverse-biased p-n junctions, reducing the doping level increases the depletion width and makes the peak electric field at the junction smaller. As discussed in Section 6.5.9, electrons traveling from the source to the drain in the channel can gain kinetic energy and thereby become hot electrons, which create damage. The low doping in the LDD helps reduce hot carrier effects at the drain end. The shallow junction depths in the LDD are also important for reducing short channel effects such as DIBL and charge sharing (Sections 6.5.10 and 6.5.11). The penalty that we pay with the use of an LDD region is that the source-to-drain series resistance goes up, which degrades the drive current.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 451

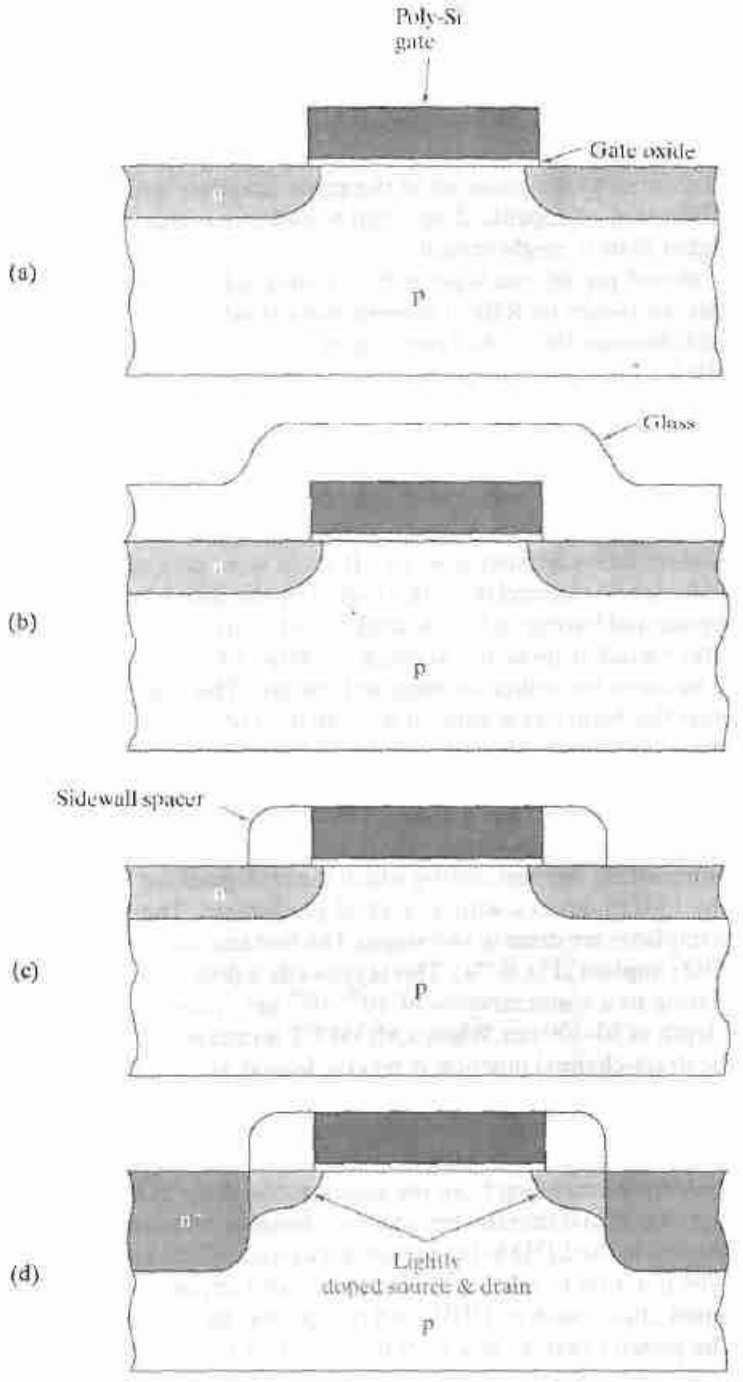


Figure 9–7
Fabrication of the lightly doped drain structure, using sidewall spacers. The polysilicon gate covers the thin gate oxide and masks the first low-dose implant (a). A thick layer is deposited by CVD (b) and is anisotropically etched away to leave only the sidewall spacers (c). These spacers serve as a mask for the second, high-dose implant. After a drive-in diffusion, the LDD structure results (d).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 452

As the technology is evolving towards lower power supply voltages, hot carrier effects are becoming less important. This, along with the need to reduce series resistance, has driven the trend towards increasing doping in the LDD to levels above 10^{19} cm^{-3}. In fact, the use of the term LDD then becomes a bit of a misnomer, and is often replaced by the term *source/drain extension* or *tip*.

After the LDD regions are formed alongside the polysilicon gate, we implant deeper (~200 nm) and more heavily doped (10^{20} cm^{-3}) source and drain junctions farther away from the gate edges (Fig. 9-7d). This more conductive region allows ohmic contacts to the source and drain to be formed more easily than they could be directly to the LDD regions, and reduces the source/drain series resistance. This implant is done using a self-aligned scheme by the formation of *sidewall oxide spacers*. After removing the photoresist covering the PMOS devices, we deposit *conformal* LPCVD oxide (~100-200 nm thick) using an organic precursor called tetra-ethyl-ortho-silicate (TEOS) over the entire wafer at fairly high temperatures (~700°C) (Fig. 9-7b). The term conformal means that the deposited film has the same thickness everywhere, and follows the topography on the wafer. This oxide layer is then subjected to RIE, which is anisotropic (i.e., it etches predominantly in the vertical direction) (Section 5.1.7). If the RIE step is timed to just etch off the deposited oxide on the flat surfaces, it leaves oxide sidewall spacers on the edges of the polysilicon gate, as shown in Fig. 9-7c. This sidewall spacer is used as a self-aligned mask to protect the LDD regions very near the gate during the heavier, deeper n^+ source and drain implants (Fig. 9-7d).

Next, the NMOS devices are masked by photoresist, and a p^+ source and drain implant is done for the PMOSFETs (Fig. 9-8a). It may be noted that an LDD was not used for the PMOS. This is due to the fact that hot hole effects are less problematic than hot electron degradation, partly due to the lower hole mobility and partly due to the higher Si–SiO_2 barrier in the valence band (5 eV) than in the conduction band (3.1 eV). After the source-drain implants are done, the dopants are activated and the ion implant damage is healed by a furnace anneal, or more frequently by using a Rapid Thermal Anneal (RTA). In this anneal we use the minimum acceptable temperature and time combination (the thermal budget) because it is critically important to keep the dopant profiles as compact as possible in ultra-small MOSFETs.

We can now appreciate why most CMOS logic devices are made on p-type substrates, rather than n-type. The n-channel MOSFETs generate a lot more substrate current due to hot carrier effects than PMOSFETs. The holes, thus generated, can more easily flow to ground in a p-type substrate, than in an n-substrate. Also, it is easier to dope substrates p-type with B during Czochralski crystal growth than n-type with Sb. Antimony is the preferred donor, rather than As or P for bulk doping of the Si melt, because Sb evaporates less than the other species.

The use of n^+ polysilicon gates for both NMOS and PMOS devices raises some interesting device issues. Since the Fermi level in the n^+ gate is

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 453

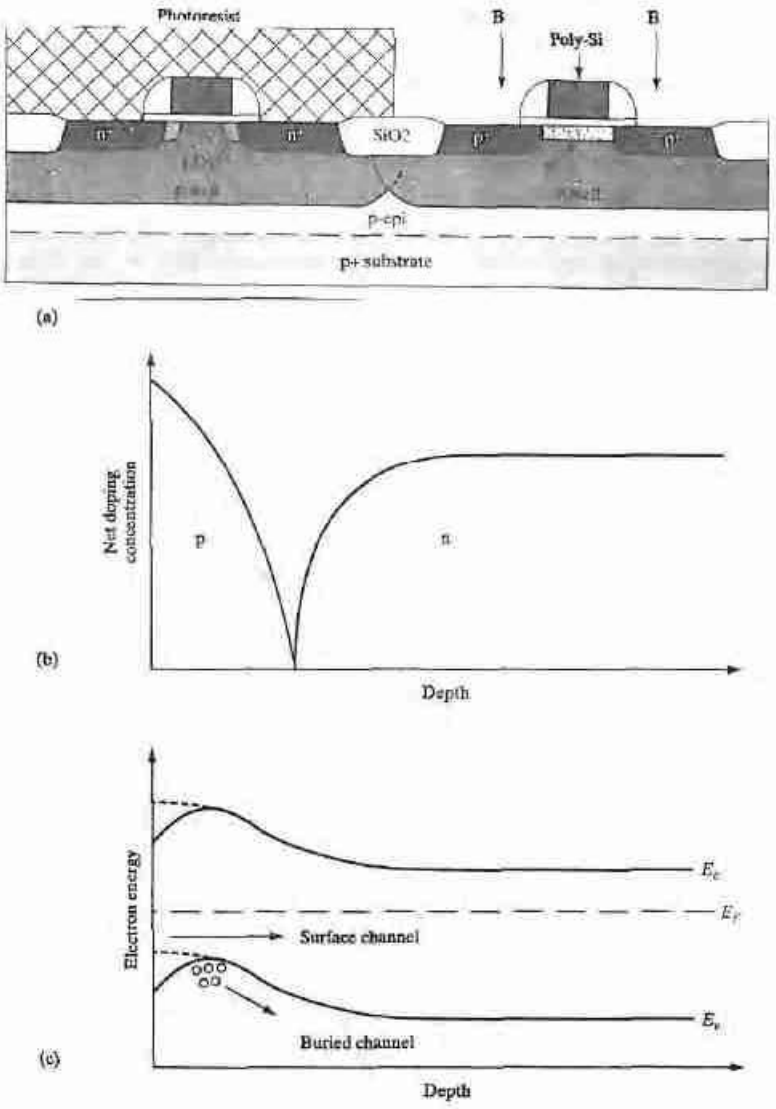


Figure 9–8 Buried channel PMOS: (a) self-aligned p^+ source–drain implant with no LDD using photoresist to protect NMOSFETs. A p-type V_T adjust implant is shown in color in the channel; (b) doping profile as a function of depth in the middle of the channel, showing the p-type V_T adjust implant near the surface; (c) electron potential energy as a function of depth in the middle of the channel, showing holes collecting in the "buried" channel. For higher gate bias, the PMOS operation changes to a surface channel, as indicated by the dashed line.

very close to the Si conduction band, its work function is well suited to achieving a low V_T for NMOS (Φ_{ms}~−1 V), but not for PMOS (Φ_{ms}~0 V). From the V_T expression [Eq. (6–38)], we notice that the second and third terms approach zero as thin-oxide technology evolves because C_i is getting larger. For high drive current we want V_T to be in the neighborhood of ~0.3 to ~0.7 V for NMOS (−0.3 to −0.7 V for PMOS). We find from Eq. (6–38) that the p-well doping can be optimized to achieve the correct V_T for the NMOS transistor, while at the same time being high enough to prevent punch-through breakdown between source and drain. For the PMOS transistor, on the other hand,

an n-well doping of the order of 10^{18} cm^{-3} prevents punch-through, but the Fermi potential, ϕ_F, is so large and negative that the V_T is too negative. That forces us to do a separate acceptor implant to adjust V_T for the PMOS devices (Fig. 9-8b). The acceptor dose is low enough that the p-layer is fully depleted at zero gate bias, leading to enhancement-mode, rather than depletion-mode, transistors. In CMOS we try to make the negative V_T of the PMOS device about the same value as the positive V_T of the NMOS.

Close examination of the band diagram in the channel of the PMOSFET along the vertical direction (perpendicular to the gate oxide) shows that the energy minimum for the holes in the inversion layer is slightly below (~100 nm) the oxide–silicon interface, leading to what is known as *buried channel operation* for PMOS (Fig. 9-8c). On the other hand, for NMOS, the electron energy minimum in the inversion layer occurs right at the oxide-silicon interface, leading to surface channel operation.[1] There are good and bad aspects of this buried channel behavior for PMOS. Since the holes in the inversion layer of the PMOSFET travel slightly away from the oxide-silicon interface, they do not suffer as much channel mobility degradation as the electrons in the NMOSFET due to surface roughness scattering. That is good, because hole mobilities in Si are generally lower than electron mobilities, which forces us to make PMOS devices wider than the NMOSFETs to get similar drive currents. However, buried channel devices have a greater propensity towards DIBL and punch-through breakdown. Hence, as the size of MOSFETs is reduced, the DIBL problem becomes worse, and there is a desire to have surface channel operation for both NMOS and PMOS. Thus, there is interest in so-called *dual gate* CMOS, where n^+ polysilicon gates are used for NMOSFETs and p^+ gates are used for PMOSFETs. Such dual work function gates can be achieved by depositing the polysilicon undoped and then using the source and drain implants themselves to also dope the gates appropriately. This approach exploits the high polysilicon grain boundary diffusivities to degenerately dope the gates, while at the same time having ultrashallow source and drain junctions to minimize DIBL.

As a historical footnote, it may be added that MOSFETs initially used Al gates which could not withstand high temperature processing. Hence, the source and drain regions had to be formed first, either by diffusion or by implant, and then the Al gate was deposited and patterned. Such non-self-aligned processes suffered from the Miller capacitance mentioned previously. What made self-aligned source and drain regions viable was the use of polysilicon, which has a sufficiently high melting point to withstand subsequent processing. Recent research on MOSFET technology is going back full circle to metal gates, but this time using refractory metals such as tungsten (W). These metals have better conductivity than heavily doped polysilicon, and a

[1]We can qualitatively understand why the acceptor implant in the channel leads to such buried channel behavior. Assume for a moment that the acceptor dose was high enough that the p-layer was not depleted at zero bias. In such a depletion-mode device, a positive gate bias to turn off the device would first deplete the surface region of holes, still leading to hole conduction deeper in the substrate away from the oxide–Si interface.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 455

work function that is better suited for CMOS. The Fermi level of W is near the middle of the Si band gap, which makes the flat band voltage and the threshold voltage more symmetric and better matched between NMOS and PMOS, and avoids the buried channel effect.

The next step is to form a metal–silicon alloy or silicide in the source/drain and gate regions of the MOSFETs in order to reduce the series resistance (and thereby the *RC* time constants), and increase the drive current. This involves depositing a thin layer of a refractory metal such as Ti over the entire wafer by sputtering, and reacting the Ti with Si wherever they come in *direct* contact, by doing a two-step heat treatment in an N ambient (Fig. 9–9a). A 600°C anneal results in the formation of Ti_2Si (the C49 phase according to metallurgists, which has fairly high resistivity), followed by an 800°C anneal which converts the Ti_2Si to the C54 phase, $TiSi_2$, which has an extremely low resistivity of ~17 $\mu\Omega$-cm, much lower than that of the most heavily doped Si. On the other hand, the Ti on top of the sidewall oxide spacers does not form a silicide and stays as unreacted Ti or forms TiN because the process is done in an N ambient. The Ti and TiN can be etched off selectively by using a wet hydrogen peroxide–based etch, which does not attack the titanium disilicide, thereby electrically isolating the gates from the source/drains. Note that this process results in a SALICIDE without a separate masking level only on the source/drains and the polysilicon gate, where the silicide is often termed a *polycide*. This results in very high-performance MOSFETs.

Finally, the MOSFETs have to be properly interconnected according to the circuit layout, using the metallization level. This involves LPCVD of an oxide dielectric layer doped with B and P, which is known as *borophospho-silicate glass* (BPSG) on the entire wafer, patterning it by means of the contact-level reticle and using RIE to open up the contact holes to the substrate (Fig. 9–9b). The B and P allow the oxide layer to soften and reflow more readily upon annealing, thereby helping planarize or smooth out the topography on the wafer. This shaping of the millions of very small contact holes is critical on a ULSI chip, because otherwise metal deposited on the surface into the contact holes may not reach completely into the holes, leading to a catastrophic open circuit. In fact, sometimes a CVD tungsten layer is selectively deposited in the contact holes to form a contact *plug* before one proceeds to the next step. Then a suitable metal layer such as Al (alloyed with ~1 percent Si and ~4 percent Cu) is sputter deposited over the wafer, patterned using the metal interconnect level, and subjected to metal RIE. The Si is added to the Al to solve the junction *spiking* problem, whereby pure Al can incorporate the solid solubility limit of Si from the shallow source/drain regions. This would allow the Al to "spike," or short, through the p-n junction. The Cu is added to enlarge the grain size in the Al interconnect films, which are polycrystalline, making it harder for the electrons moving during current flow to nudge the Al atoms along, thereby opening voids (producing open circuits) in the interconnect. This is an example of an *electromigration* phenomenon. A cross-sectional picture of an LDD MOSFET with SALICIDE is shown in Fig. 9–9c.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 456

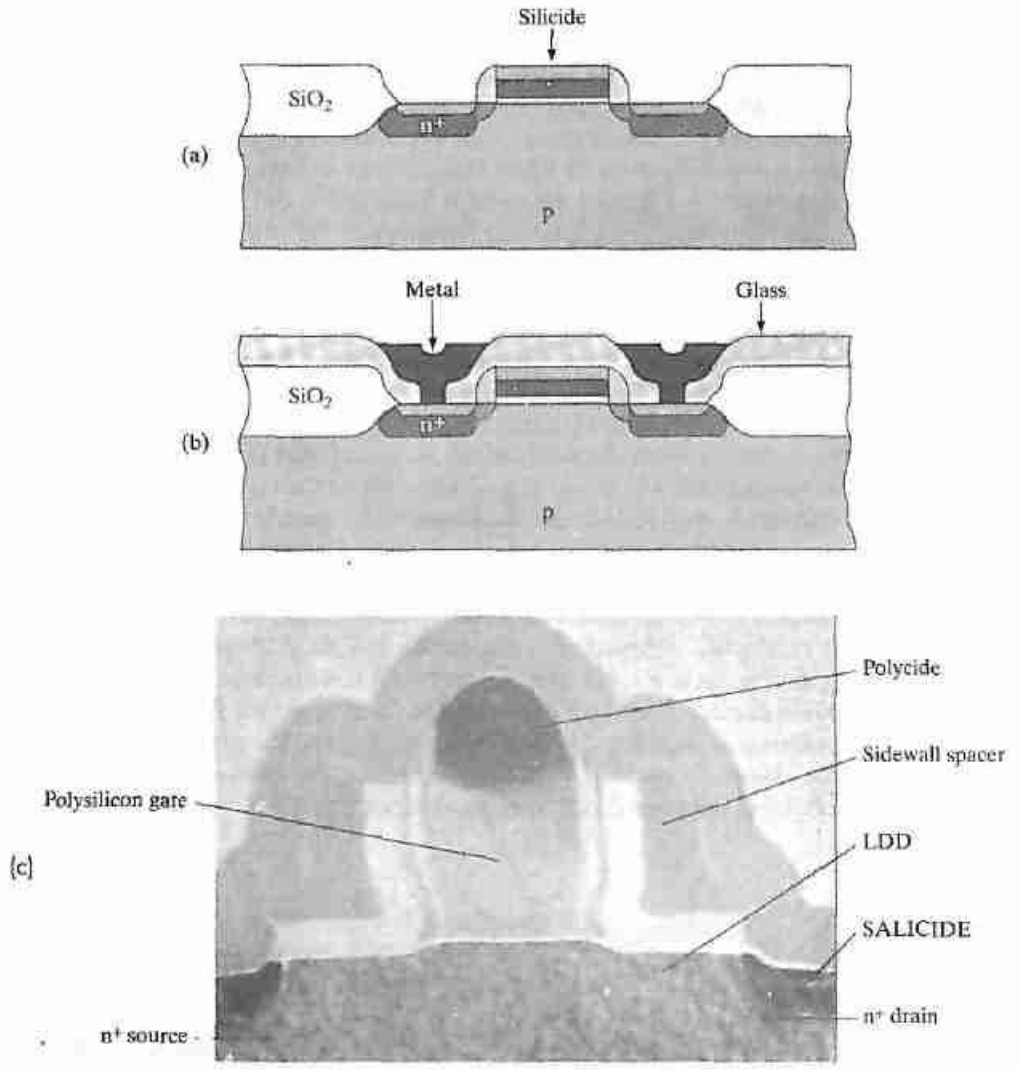


Figure 9–9
The formation of silicided source–drain and gate regions for low-resistance contacts: (a) a layer of refractory metal is deposited and reacted in regions of exposed Si to form a conducting silicide layer; (b) the unreacted metal is removed and a CVD glass is deposited and patterned for contact metallization; (c) cross-sectional micrograph of LDD MOSFET with gate length of 52 nm. (Photograph courtesy of *Texas Instruments.*)

In a modern ULSI chip, the complexity of the device layout generally demands that multiple levels of metallization be used for interconnecting the devices (Fig. 9–10). Hence, after the first metal is deposited, an intermetal dielectric isolation layer such as SiO_2 is deposited by low-temperature CVD.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 457

Low temperatures are very important in this *back-end* part of the processing because by now all the active devices are in place and one cannot allow the dopants to diffuse significantly. Also, the Al metallization cannot withstand temperatures higher than ~500°C. The dielectric isolation layer must be suitably planarized prior to the deposition of the next layer of metal, and this is generally done by CMP. Planarization is important because if metal is deposited on a surface with rough topography and subjected to RIE, there can be residual metal sidewall filaments or "stringers" at the steps for the same reason one gets sidewall oxide spacers on either side of the MOSFET gate in Fig. 9–7. These metal stringers can cause short circuits between adjacent metal lines. Planarization is also important in maintaining good depth of focus during photolithography. After planarization of the isolation layer, one uses photolithography to open up a new set of contact holes called *vias*, followed by deposition, patterning and RIE of the next layer of metal, and so on for multi-level metallization. As mentioned previously, W metal plugs are sometimes selectively deposited to fill up the via holes prior to the metal deposition, and reduce the likelihood of an open circuit.

Finally, a protective overcoat is deposited on the IC to prevent contamination and failure of the devices due to the ambient (Fig. 9–10). This generally involves plasma CVD of silicon nitride, which has the nice attribute that it blocks the diffusion of water vapor and Na through it. Sodium, as mentioned in Section 6.4.3, causes a mobile ion problem in the gate dielectric of MOS devices. Sometimes, the protective overcoat is a BPSG layer. After the overcoat is deposited, openings are etched for the metal bond pads. After the chips are tested in an automated tester, the *known good dies* are packaged and wire bonded, as discussed in Section 9.6.

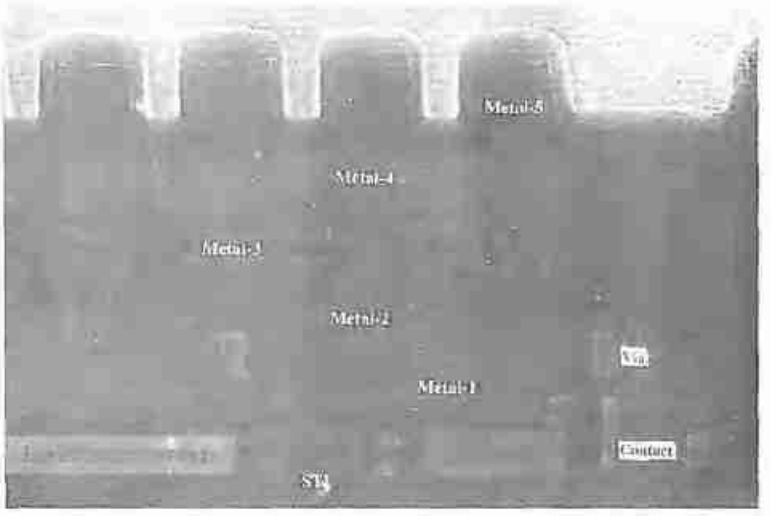


Figure 9–10
Multilevel interconnect: Cross section of IC showing five levels of Al interconnections with suitably planarized intermetal dielectrics. The transistors are at the very bottom and are electrically isolated by shallow-trench isolation (STI). (Photograph courtesy of Freescale Semiconductor.)

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 458

9.3.2 Silicon-on-Insulator (SOI)

An interesting and useful extension of the Si MOS process can be achieved by growing very thin films of single-crystal Si on insulating substrates (Fig. 9–11). Two such substrates which have the appropriate thermal expansion match to Si are sapphire and spinel ($MgO-Al_2O_3$). Epitaxial Si films can be grown on these substrates by chemical vapor deposition (e.g., the pyrolysis of silane), with typical film thickness of about 1 μm. The film can be etched by standard photolithographic techniques into islands for each transistor. Implantation of p^+ and n^+ areas into these islands for source and drain regions result in the MOS devices. Since the film is so thin, the source and drain regions can be made to extend entirely through the film to the insulating substrate. As a result, the junction capacitance is reduced to the very small capacitance associated with the sidewalls between the source/drain and the channel region. In addition, since interconnections between devices pass over the insulating substrate, the usual interconnection–Si substrate capacitance is eliminated (along with the possibility of parasitic induced channels in the field between devices). These capacitance reductions improve considerably the high-frequency operation of circuits using such devices.

Other insulators can be used for SOI devices, including SiO_2. Since oxide can easily be grown on Si substrates, it serves as an attractive insulator for subsequent growth of thin-film Si. Since polycrystalline Si can be deposited directly over SiO_2, devices can be made in thin poly-Si films. However, to avoid grain boundaries and other defects typical of polycrystalline material, a variety of techniques have been developed to grow single-crystal Si on oxide. For example, the oxide layer can be formed beneath the surface of a Si wafer by high-dose oxygen implantation. The thin Si layer remaining on the surface above the implanted oxide is usually about 0.1 μm thick, and can be used as the thin film for CMOS or other device fabrication. This process is called *separation by implantation of oxygen (SIMOX)*. In some cases a thicker Si film is grown epitaxially on the SIMOX wafer, using the thin Si crystalline layer over the oxide as a seed for the epitaxy.

Another approach for making SOI is to place two oxidized Si wafers face to face, and thermally bond the oxide layers by high-temperature

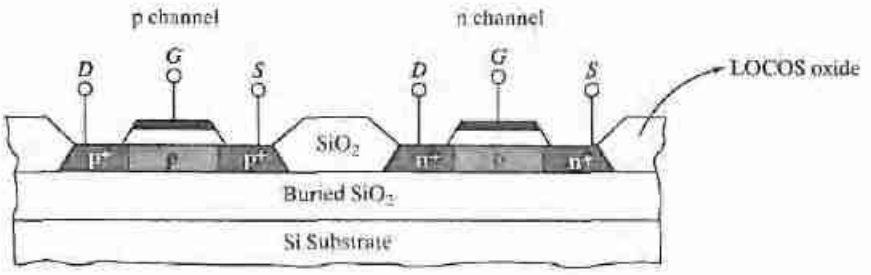


Figure 9–11
Silicon on insulator. Both n-channel and p-channel enhancement transistors are made in islands of Si film on the insulating substrate. These devices can be interconnected for CMOS applications.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 459

annealing in a furnace. One of the wafers is then chemically etched back almost completely, leaving about a micron of single-crystal Si material on top of the SiO_2. This approach is known as *Bond-and-Etch-back* SOI (BE-SOI). Since it is challenging to etch off about 600 μm of Si and stop controllably so as to leave about a micron of Si, an initial p^+ implanted layer is often used to act as an etch stop near the end of the chemical etch. The chemical recipe that is used for etching Si has a much lower etch rate in a p^+ layer than in lightly doped Si.

Another recent approach is a modification of BE-SOI using very high dose H implantation into one of the oxidized Si wafers, such that the peak of the H profile is about a micron below the Si surface. During the high-temperature annealing step, the H atoms coalesce to form tiny H bubbles which cause a thin layer of Si to cleave off from the implanted wafer, leaving the Si layer bonded to the oxidized Si surface. This way, one does not have to chemically etch off one Si wafer for every BE-SOI wafer. If properly done, the cleaved-off Si wafer has a smooth surface and can be reused.

Examining the two devices of Fig. 9-11 more closely, one of them is unlike the transistors we have considered thus far. The thin Si film is lightly doped p-type, and therefore the device labeled "p channel" appears *junctionless*. Such a device is able to operate in the enhancement mode (normally off) because of the equilibrium effects of the work function difference and the interface charge. With the usual Φ_{ms} and Q_i a depletion region is formed in the central p material of each device with zero gate voltage. In fact, for a Si film of about 0.1 μm or less, this depletion region can extend all the way through the Si to the insulator. Such a device is called *fully depleted* and no drain-to-source current flows. In the n-channel device a positive gate voltage greater than V_T induces an inverted region at the surface, as usual for an n-channel enhancement device. For the p-channel case a small negative voltage V_G removes the depletion and causes hole accumulation beneath the gate. The result is the formation of a conducting channel by a small negative gate voltage, as is the case for a conventional p-channel enhancement device. Although the fully depleted type of p-channel device operates by a somewhat different mechanism, its current–voltage characteristics are similar to the conventional device. Since both p-channel and n-channel transistors can be made on the same insulating surface, the silicon-on-insulator technique is quite compatible with CMOS circuit fabrication.

The SOI approach does not suffer from the latchup problem of bulk CMOS because there is no p-n-p-n thyristor from the power supply to ground. For circuits requiring high speeds, low standby power (due to the elimination of junction leakage to the substrate), and radiation tolerance (due to the elimination of the Si substrate), the extra expense of preparing sapphire substrates or growing crystalline Si films on oxide is compensated by increased performance.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 460

9.3.3 Integration of Other Circuit Elements

One of the most revolutionary developments of integrated circuit technology is the fact that integrated transistors are cheaper to make than are more mundane elements such as resistors and capacitors. There are, however, numerous applications calling for diodes, resistors, capacitors, and inductors in integrated form. In this section we discuss briefly how these circuit elements can be implemented on the chip. We will also discuss a very important circuit element – the interconnection pattern which ties all of the integrated devices together in a working system.

Diodes. It is simple to build p-n junction diodes in a monolithic circuit. It is also common practice to use transistors to perform diode functions. Since many transistors are included in a monolithic circuit, no special diffusion step is required to fabricate the diode element. There are a number of ways in which a transistor can be connected as a diode. Perhaps the most common method is to use the emitter junction as the diode, with the collector and base shorted. This configuration is essentially the narrow base diode structure, which has high switching speed with little charge storage (Prob. 5.35). Since all the transistors can be made simultaneously, the proper connections can be included in the metallization pattern to convert some of the transistors into diodes.

Resistors. Diffused or implanted resistors can be obtained in monolithic circuits by using the shallow junctions used in forming the transistor regions (Fig. 9-12a). For example, during the base implant, a resistor can be implanted which is made up of a thin p-type layer within one of the n-type islands. Alternatively a p region can be made during the base implant, and an n-type resistor channel can be included within the resulting p region during the emitter implant step. In either case, the resistance channel can be isolated from the rest of the circuit by proper biasing of the surrounding material. For example, if the resistor is a p-type channel obtained during the base implant, the surrounding n material can be connected to the most positive potential in the circuit to provide reverse-bias junction isolation. The resistance of the channel depends on its length, width, depth of the implant, and resistivity of the

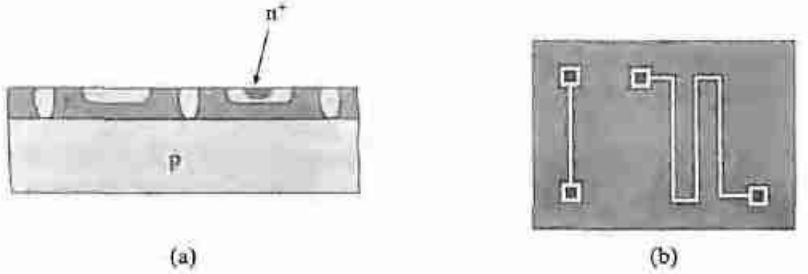


Figure 9–12
Monolithic resistors: (a) cross section showing use of base and emitter diffusions for resistors; (b) top view of two resistor patterns.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 461

implanted material. Since the depth and resistivity are determined by the requirements of the base or emitter implant, the variable parameters are the length and width. Two typical resistor geometries are shown in Fig. 9-12b. In each case the resistor is long compared with its width, and a provision is made on each end for making contact to the metallization pattern.

Design of diffused resistors begins with a quantity called the *sheet resistance* of the diffused layer. If the average resistivity of a diffused region is ρ, the resistance of a given length L is $R = \rho L/wt$, where w is the width and t is the thickness of the layer. Now if we consider one square of the material, such that $L = w$, we have the sheet resistance $R_s \equiv \rho/t$ in units of ohms per square. We notice that R_s measured for a given layer is numerically the same for any size square. This quantity is simple to measure for a thin diffused layer by a four-point probe technique.[2] Therefore, for a given diffusion, the sheet resistance is generally known with good accuracy. The resistance then can be calculated from the known value of R_s and the ratio L/w (the *aspect ratio*) for the resistor. We can make the width w as small as possible within the requirements of heat dissipation and photolithographic limitations and then calculate the required length from w and R_s. Design criteria for diffused resistors include geometrical factors, such as the presence of high current density at the inside corner of a sharp turn. In some cases it is necessary to round corners slightly in a folded or zigzag resistor (Fig. 9-12b) to reduce this problem.

To reduce the amount of space used for resistors or to obtain larger resistance values, it is often necessary to obtain surface layers having larger sheet resistance than is available during the standard base or emitter implants. We can use a different implant, such as the V_T adjust implant, to form shallow regions having very high sheet resistance ($\sim 10^5$ Ω/square). This procedure can provide a considerable saving of space on the chip. In integrated FET circuits it is common to replace load resistors with depletion-mode transistors, as mentioned in Section 6.5.5.

Capacitors. One of the most important elements of an integrated circuit is the capacitor. This is particularly true in the case of memory circuits, where charge is stored in a capacitor for each bit of information. Figure 9-13 illustrates a one-transistor DRAM cell, in which the n-channel MOS transistor provides access to the adjacent MOS capacitor. The top plate of the capacitor is polysilicon, and the bottom plate is an inversion charge contacted by an n^+ region of the transistor. The terms *bit line* and *word line* refer to the row and column organization of the memory (Section 9.5.2). One can also make use of the capacitance associated with p-n junctions, as discussed in Section 5.5.5.

Inductors. Inductors have not been incorporated into ICs in the past, because it is much harder to integrate inductors than the other circuit elements. Also,

[2]This is a very useful method in which current is introduced into a wafer at one probe and is collected at another probe, and the voltage is measured by two probes in between. Special formulas are required to calculate resistivity or sheet resistance from these measurements.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 462

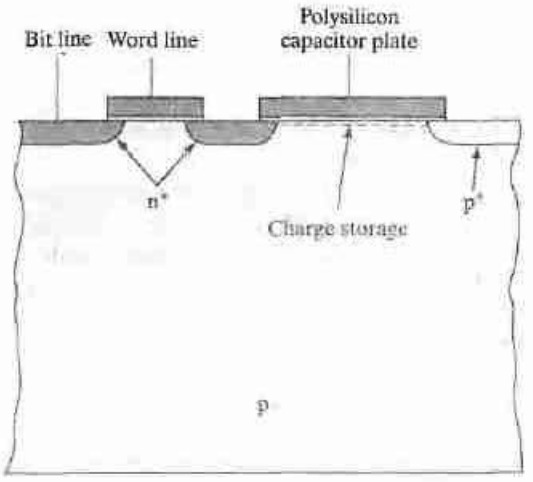


Figure 9–13
Integrated capacitor for DRAM cells. A one-transistor memory cell in which the transistor stores and accesses charge in an adjacent planar MOS capacitor.

there has not been a great need for integrating inductors. Recently, that has changed because of the growing need for rf analog ICs for portable communication electronics. Inductors are very important for such applications, and can be made with reasonable Q factors using spiral wound thin metal films on an IC. Such spiral patterns can be defined by photolithography and etching techniques compatible with IC processing, or they can be incorporated in a hybrid IC.

Contacts and Interconnections. During the metallization step, the various regions of each circuit element are contacted and proper interconnection of the circuit elements is made. Aluminum is commonly used for the top metallization, since it adheres well to Si and to SiO_2 if the temperature is raised briefly to about 550°C after deposition. Gold is used on GaAs devices, but the adherence properties of Au to Si and SiO_2 are poor. Gold also creates deep traps in Si.

As mentioned throughout this chapter, silicide contacts and doped polysilicon conductors are commonly used in integrated circuits. By opening windows through the oxide layers to these conductors, Al metallization can be used to contact them and connect them to other parts of the circuit. In cases where Al is used to contact the Si surface, it is usually necessary to use Al containing about 1 percent Si to prevent the metal from incorporating Si from the layer being contacted, thereby causing "spikes" in the surface. Thin diffusion barriers are also used between the Al and Si layers, to prevent migration between the two. The refractory silicides mentioned in Section 9.3.1 serve this purpose.

Increased complexity and packing density in integrated circuits inevitably leads to a need for multilayer metallization. Multiple levels of Cu metallization can be incorporated with interspersing dielectrics. In general, the metals may all be Al, Cu or they may be different conductors such as polysilicon or refractory metals (depending on the heat each is subjected to in subsequent processing). Also, the dielectrics may be deposited oxides, boro-phospho-silicate

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 463

glass for reflow planarization, nitrides, etc. The planarization of the surface is extremely important to prevent breaks in the metallization, which can occur in traversing a step on the surface. Various approaches using reflow glass, polyimide, and other materials to achieve planarization have been used, along with chemical mechanical polishing.

The most important challenge in designing interconnects is the *RC* time constant, which affects the speed and active power dissipation of the chip. A very simplistic model of two layers of interconnects with an inter-metal dielectric (Fig. 9-14b) shows that it can be regarded as a parallel plate capacitor. Regarding the interconnect as a rectangular resistor, its resistance is given by

$$R = \frac{\rho L}{tw} = R_s \frac{L}{w} \tag{9-1a}$$

where R_s is the sheet resistance, and the other symbols are defined in Fig. 9-14b. The capacitance is given by

$$C = \frac{\epsilon L w}{d} \tag{9-1b}$$

The *RC* time constant is then

$$\left(\frac{\rho L}{wt}\right)\left(\frac{\epsilon L w}{d}\right) = \frac{\rho \epsilon L^2}{td} = \frac{R_s \epsilon L^2}{d} \tag{9-2}$$

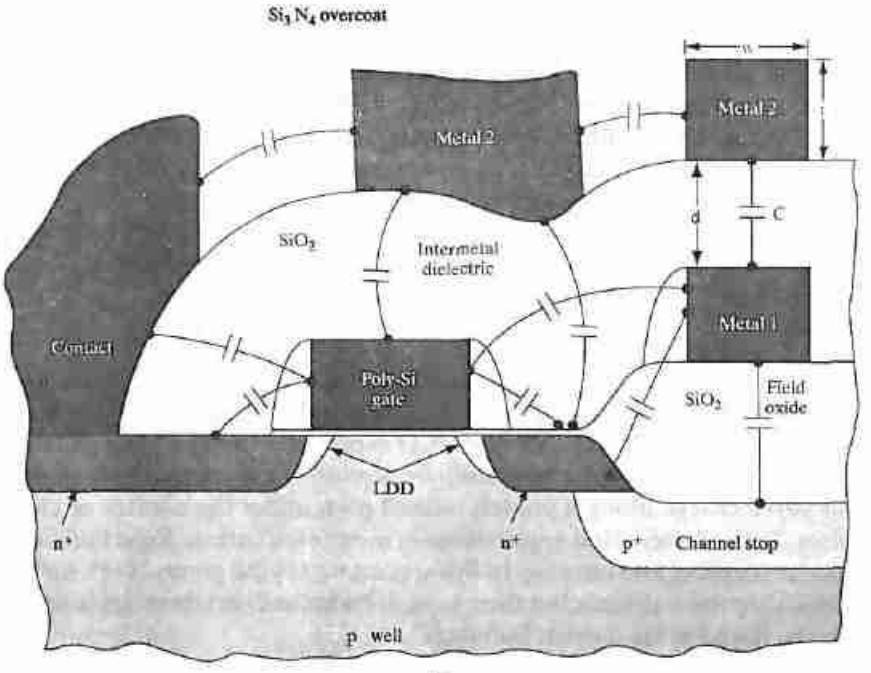
Interestingly, for this simple one-dimensional model, the width of the interconnect w cancels out. Therefore, it does not make sense to use wider conductors for high-speed operation. It is also impractical to do so in terms of packing density. Of course, in reality we must account for the fringing electric fields, and therefore account for width dependence. From Eq. (9-2), it is clear we need as thick a metal layer (within practical limits of deposition times and etching times) and as low a resistivity as possible. Low resistivities are also important in minimizing ohmic voltage drops in metal bus lines that carry power from one end of a chip to the other. Aluminum is very good in this regard, and thus was a mainstay for Si technology for many years. Aluminum also has other nice attributes such as good ohmic contacts to both n and p-type Si, and good adhesion to oxides.

Copper has even lower resistivity (1.7 $\mu\Omega$-cm) than Al (3 $\mu\Omega$-cm) and is about two orders of magnitude less susceptible to electromigration. Hence, it is an excellent alternative to Al for very high speed ICs (Fig. 9-14a). The process breakthroughs that have made Cu viable for metallization include new electrodeposition and electroplating techniques because CVD or sputter deposition is not very practical for Cu. It is also very difficult to use reactive ion etching (RIE) for Cu because the etch byproducts for Cu are not very volatile. Hence, instead of RIE, Cu patterning is based on the so-called *Damascene* process where grooves are first etched in a dielectric layer, Cu is deposited on it, and the metal layer is chemically-mechanically polished down, leaving inlaid metal lines in the oxide grooves. In this method, the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 464



(a)



(b)

Figure 9–14
Multi-level interconnects: (a) Micrograph of multi-level interconnects in an IC. The inter-level dielectrics have been etched off to reveal the copper interconnect lines; (b) equivalent circuit illustrating the various parasitic capacitive elements associated with a multi-level interconnect. On the top right hand corner of the figure, we focus on the parallel plate capacitor model referred to in Eqs. (9–1) and (9–2). (Photograph courtesy of IBM.)

metal does not have to be etched directly using RIE, which is always a difficult process. The name "damascene" is derived from a metallurgical technique in ancient Turkey where metal artwork was inlaid into swords and other artifacts using this type of process. Copper can create traps deep in the band gap of Si; hence, a suitable barrier layer such as Ti is needed between the Cu layer and the Si substrate.

Other parameters in Eq. (9-2) that can minimize the *RC* time constant are clearly the use of a thick inter-metal dielectric layer (once again within the limits of practicality in terms of deposition times), and as low a dielectric constant material as possible. Silicon dioxide has a relative dielectric constant of 3.9. There is active research in low dielectric constant materials (sometimes referred to as low-K materials). These include organic materials such as polyimides, or xerogels/aerogels which have air pockets or porosity purposely built in to minimize the dielectric constant.

In designing the layout of elements for a monolithic circuit, topological problems must be solved to provide efficient interconnection without *crossovers*—points at which one conductor crosses another conductor. If crossovers must be made on the Si surface, they can be accomplished easily at a resistor. Since the implanted or diffused resistor is covered by SiO_2, a conductor can be deposited crossing the insulated resistor. In cases requiring crossovers where no resistor is available, a low-value implanted resistor can be inserted in one of the conductor paths. For example, a short n^+ region can be implanted during the source/drain step and contacted at each end by one of the conductors. The other conductor can then cross over the oxide layer above the n^+ region. Usually, this can be accomplished without appreciable increase in resistance, since the n^+ region is heavily doped and its length can be made small.

During the metallization step, appropriate points in the circuit are connected to relatively large *pads* to provide for external contacts. These metal pads are visible in photographs of monolithic circuits as rectangular areas spaced around the periphery of the device. In the mounting and packaging process, these pads are contacted by small Au or Al wires or by special techniques such as those discussed in Section 9.6.

9.4 CHARGE TRANSFER DEVICES

One of the most interesting and broadly useful integrated devices is the *charge-coupled device (CCD)*. The CCD is part of a broader class of structures known generally as *charge transfer devices*. These are dynamic devices that move charge along a predetermined path under the control of clock pulses. These devices find applications in memories, various logic functions, signal processing, and imaging. In this section we lay the groundwork for understanding these devices, but their present forms and variety of applications must be found in the current literature.

9.4.1 Dynamic Effects in MOS Capacitors

The basis of the CCD is the dynamic storage and withdrawal of charge in a series of MOS capacitors. Thus we must begin by extending the MOS discussion of Chapter 6 to include the basics of dynamic effects. Figure 9-15 shows an MOS capacitor on a p-type substrate with a large positive gate pulse

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 466

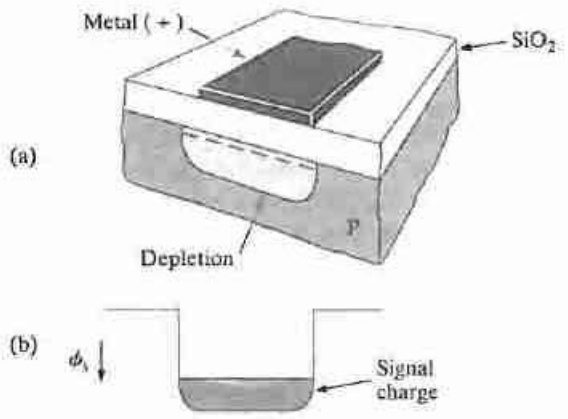


Figure 9–15
An MOS capacitor with a positive gate pulse: (a) depletion region and surface charge; (b) potential well at the interface, partially filled with electrons corresponding to the surface charge shown in (a).

applied. A depletion region exists under the gate, and the surface potential increases considerably under the gate electrode. In effect, the surface potential forms a *potential well*, which can be exploited for the storage of charge.

If the positive gate bias has been applied for a sufficiently long time, electrons accumulate at the surface and the steady state inversion condition is established. The source of these carriers is the thermal generation of electrons at or near the surface. In effect, the inversion charge tells us the capacity of the well for storage charge. The time required to fill the well thermally is called the *thermal relaxation time*, and it depends on the quality of the semiconductor material and its interface with the insulator. For good materials the thermal relaxation time can be much longer than the charge storage times involved in CCD operation.

If instead of a steady state bias we apply a large positive pulse to the MOS gate electrode, a deep potential well is first created. Before inversion has occurred by thermal generation, the depletion width is greater than it would be at equilibrium ($W > W_m$). This transient condition is sometimes called *deep depletion*. If we can inject electrons into this potential well electrically or optically, they will be stored there.[3] The storage is temporary, however, because we must move the electrons out to another storage location before thermal generation becomes appreciable.

What is needed is a simple method for allowing charge to flow from one potential well to an adjacent one quickly and without losing much charge in the process. If this is accomplished, we can inject, move, and collect packets of charge dynamically to do a variety of electronic functions.

[3]The potential well should not be confused with the depletion region, which extends into the bulk of the semiconductor. The "depth" of the well is measured in electrostatic potential, not distance. Electrons stored in the potential well are in fact located very near the semiconductor surface.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 467

9.4.2 The Basic CCD

The original CCD structure proposed in 1969 by Boyle and Smith of Bell Laboratories consisted of a series of metal electrodes forming an array of MOS capacitors as shown in Fig. 9-16. Voltage pulses are supplied in three lines (L_1, L_2, L_3), each connected to every third electrode in the row (G_1, G_2, G_3). These voltages are clocked to provide potential wells, which vary with time as in Fig. 9-16. At time t_1 a potential well exists under each G_1 electrode, and we assume this well contains a packet of electrons from a previous operation. At time t_2 a potential is applied also to the adjacent electrode G_2, and the charge equalizes across the common G_1-G_2 well. It is easy to visualize this process by thinking of the mobile charge in analogy with a fluid, which flows to equalize its level in the expanding container. This fluid model continues at t_3 when V_1 is reduced, thus decreasing the potential well under G_1. Now the charge flows into the G_2 well, and this process is completed at t_4 when V_1 is zero. By this process the packet of charge has been moved from under G_1 to G_2. As the procedure is continued, the charge is next passed to the G_3 position, and continues down the line as time proceeds. In this way charge can be injected using an input diode, transported down the line, and detected at the other end.

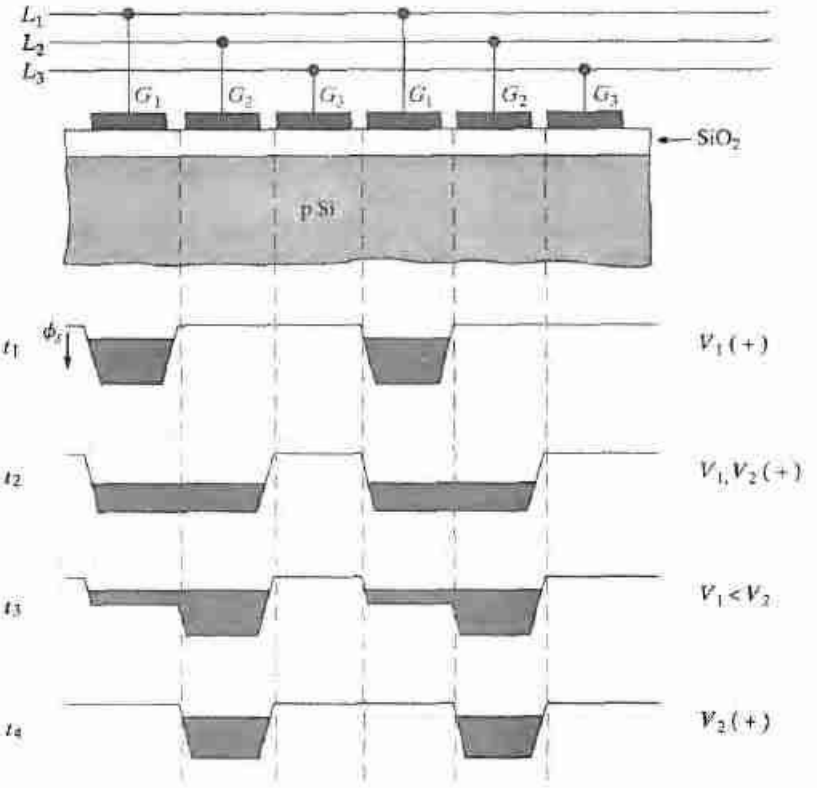


Figure 9–16
The basic CCD, composed of a linear array of MOS capacitors. At time t_1, the G_1 electrodes are positive, and the charge packet is stored in the G_1 potential well. At t_2 both G_1 and G_2 are positive, and the charge is distributed between the two wells. At t_3 the potential on G_1 is reduced, and the charge flows to the second well. At t_4 the transfer of charge to the G_2 well is completed.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 468

9.4.3 Improvements on the Basic Structure

Several problems arise in the implementation of the CCD structure of Fig. 9–16. For example, the separation between electrodes must be very small to allow coupling between the wells. An improvement can be made by using an overlapping gate structure such as that shown in Fig. 9–17. This can be done, for example, with poly-Si electrodes separated by SiO_2 or with alternating poly-Si and metal electrodes.

One of the problems inherent to the charge transfer process is that some charge is inevitably lost during the many transfers along the CCD. If the charges are stored at the Si-SiO_2 interface, surface states trap a certain amount of charge. Thus if the "0" logic condition is an empty well, the leading edge of a train of pulses is degraded by the loss of charge required in filling the traps which were empty in the "0" condition. One way of improving this situation is to provide enough bias in the "0" state to accommodate the interface and bulk traps. This procedure is colorfully referred to as using a *fat zero*. Even with the use of fat zeros, the signal is degraded after a number of transfers, by inherent inefficiencies in the transfer process.

Transfer efficiency can be improved by moving the charge transfer layer below the semiconductor–insulator interface. This can be accomplished by using ion implantation or epitaxial growth to create a layer of opposite type than the substrate. This shifts the maximum potential under each electrode into the semiconductor bulk, thus avoiding the semiconductor–insulator interface. This type of device is referred to as a *buried channel* CCD.

The three-phase CCD shown in Fig. 9–16 is only one example of a variety of CCD structures. Figure 9–18 illustrates one method for achieving a two-phase system, in which voltages are sequentially applied to alternating gate electrodes from two lines. A two-level poly-Si gate structure is used, in which the gate electrodes overlap, and a donor implant near the Si surface creates a built-in well under half of each electrode. When both gates are turned off (b), potential wells exist only under the implanted regions, and charge can be stored in any of these wells. With electrode G_2 pulsed positively, the charge packet shown in (b) is transferred to the deepest well under G_2, which is its implanted region as (c) indicates. Then with both gates off, the wells appear as in (b) again, except that the charge is now under the G_2 electrode. The next step in the transfer process is obviously to pulse G_1 positively, so that the charge moves to the implanted region under the G_1 electrode to the right.

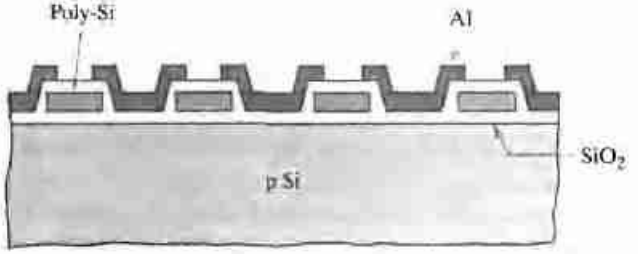


Figure 9–17
An overlapping gate CCD structure. One set of electrodes is polycrystalline Si, and the overlapping gates are Al in this case. SiO_2 separates the adjacent electrodes.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 469

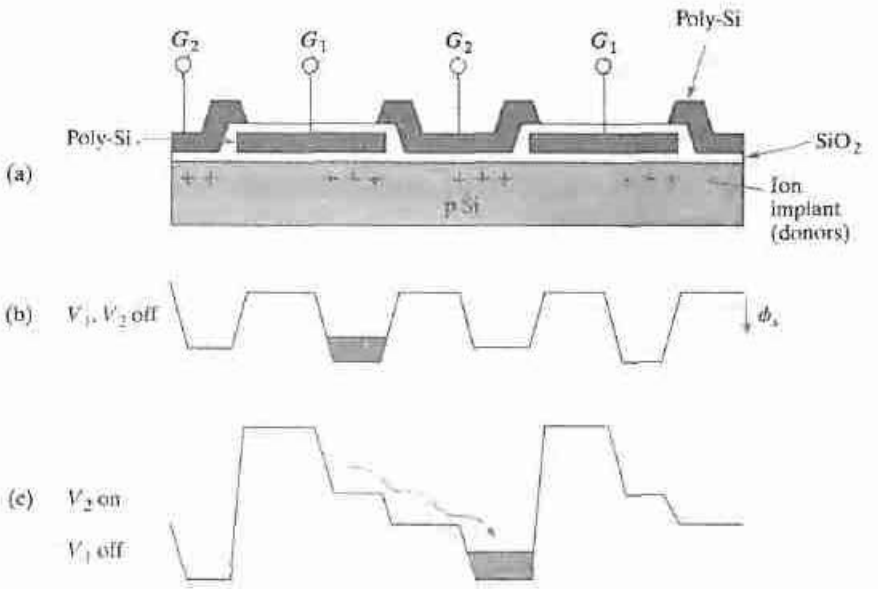


Figure 9–18
A two-phase CCD with an extra potential well built in under the right half of each electrode by donor implantation.

Other improvements to the basic structure are important in various applications. These include channel stops or other methods for achieving lateral confinement for the stored charge. Regeneration points must be included in the array to refresh the signal after it has been degraded.

9.4.4 Applications of CCDs

CCDs are used in a number of ways, including signal processing functions such as delay, filtering, and multiplexing several signals. Another interesting application of CCDs is in imaging for astronomy or in solid state TV cameras, in which an array of photosensors is used to form charge packets proportional to light intensity, and these packets are shifted to a detector point for readout. There are numerous ways of accomplishing this in CCDs, including the linear array line scanner, in which the second dimension is obtained by moving the scanner relative to the image. Alternatively, an area image sensor can be made which scans the image electronically in both dimensions. The latter device can be used as an alternative to the electron beam–addressed television imaging tube (Fig. 9-19).

9.5 ULTRA LARGE-SCALE INTEGRATION (ULSI)

In the early development of integrated circuits it was felt that the inevitable defects that occur in processing would prevent the fabrication of devices containing more than a few dozen logic gates. One approach to integration on a larger scale tried in the late 1960s involved fabricating many identical logic gates on a wafer, testing them, and interconnecting the good ones (a process

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 470

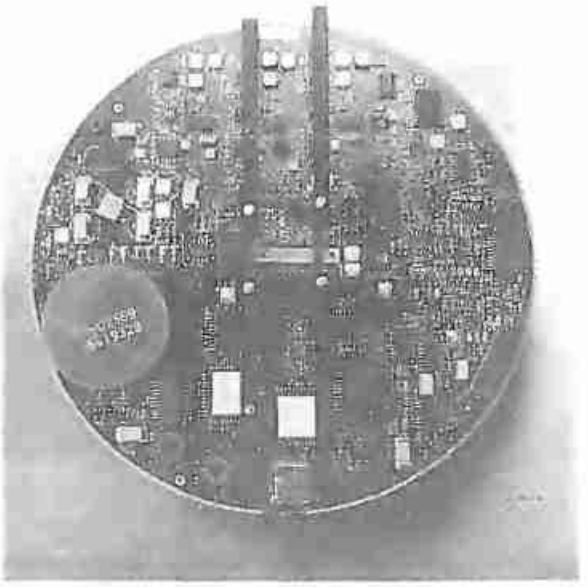


Figure 9–19
A charge-coupled device image sensor shown as large white rectangular areas, with peripheral signal processing circuitry. (Photograph courtesy of Texas Instruments.)

called *discretionary wiring*). While this approach was being developed, however, radical improvements were made in device processing which increased the yield of good chips on a wafer dramatically. By the early 1970s it was possible to build circuits with many hundreds of components per chip, with reasonable yield. These improvements made discretionary wiring obsolete almost as soon as it was developed. By reducing the number of processing defects, improving the packing density of components, and increasing the wafer size, it is now possible to place millions of device elements on a single chip of silicon and to obtain many perfect chips per wafer.

A major factor in the development of integrated circuits has been the continual reduction in size of the individual elements (transistors, capacitors) within each circuit. Through improved design and better lithography, there has been a dramatic shrinking of the minimum feature size (e.g., a transistor gate) used in these devices. The results of shrinking the elements in a 256-Mb-DRAM are shown in Fig. 9–20. By reducing the minimum feature size in successive steps from 0.13 to 0.11 μm, the die area was reduced from about 135 mm^2 in the first-generation design to less than 42 mm^2 in the fifth-generation device. Obviously, more of the smaller chips can be made by batch fabrication on the wafer, and the effort in shrinking the design is rewarded in a more profitable device.

Successive designs using reduced feature sizes have made dramatically increased circuit complexity possible. DRAM design has set the pace over the past two decades, in which successive 1-Mb, 4-Mb, and 16-Mb memories led to similar powers of two increase to the 128-Mb range. Figure 9–21 illustrates the size comparison of a 128-Mb memory with an equivalent amount

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 471

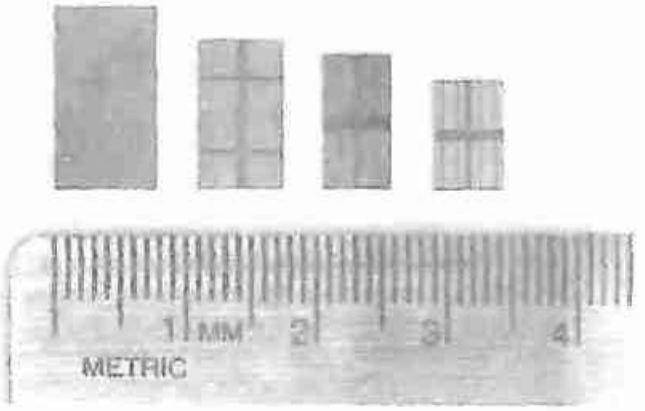


Figure 9–20
Size reduction of a 256-Mb DRAM die as the minimum feature size is reduced from 0.13 μm for the first-generation design (die on the left) to 0.11 μm for the die on the right. (© 2004 Micron Technology, Inc. All Rights Reserved. Used with permission.)

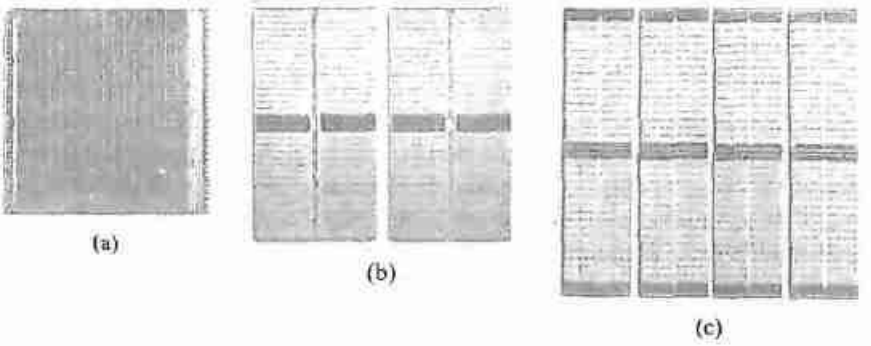


Figure 9–21
Three ways of obtaining 128 million bits of DRAM: (a) one 128-Mb chip; (b) two 64-Mb chips; (c) eight 16-Mb chips. The 128-Mb die measures almost 1 by 1 inch. (© 2004 Micron Technology, Inc. All Rights Reserved. Used with permission.)

of memory in the form of two 64-Mb and eight 16-Mb chips. These are examples of ultra large-scale integration (ULSI).

Although the achievement of many powers of two in memory is impressive and important, other ULSI chips are important for the integration of many different system functions. A *microprocessor* includes functions for a computer central processing unit (CPU), along with memory, control, timing, and interface circuits required to perform very complex computing functions. The complexity of such devices is shown in Fig. 9–22, which illustrates a microprocessor chip with various areas outlined by function.

Before leaving this section it might be useful to provide some calibration regarding the dimensions we have been discussing. Figure 9–23 compares the size of 64-Mb DRAM circuit interconnect elements with a human hair, on

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 472

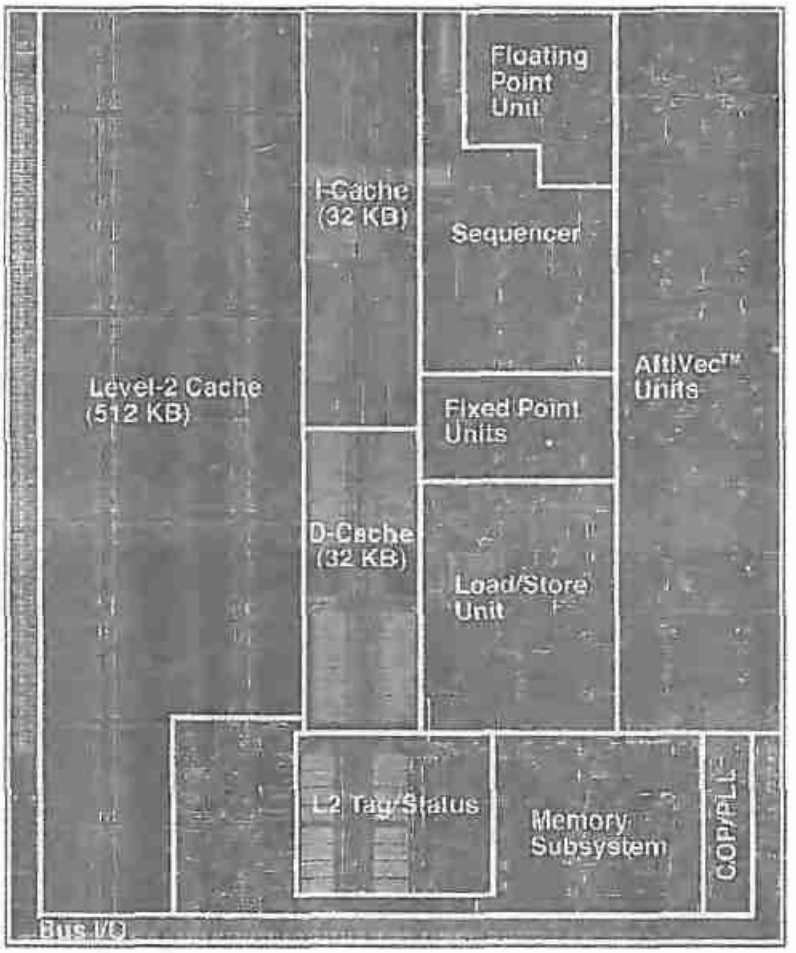
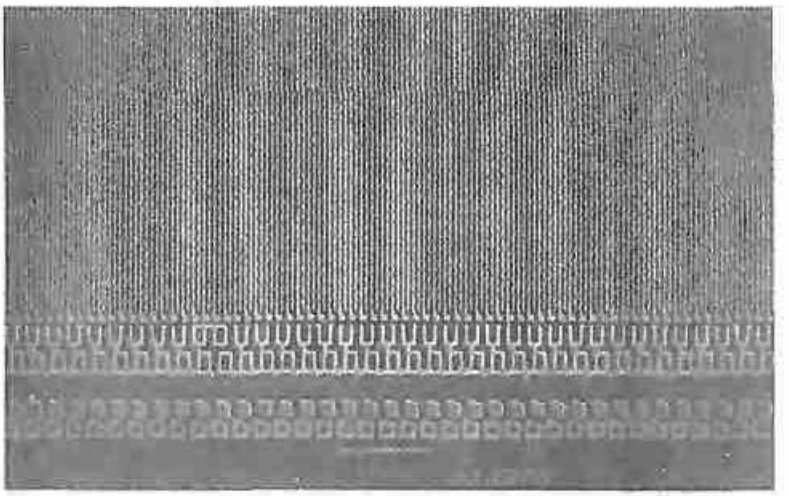


Figure 9–22
An example of ULSI, the PowerPC G4 MPC7447A microprocessor employing 90-nm CMOS technology, and running at 1.5 GHz. (Photograph courtesy of Freescale Semiconductor.)

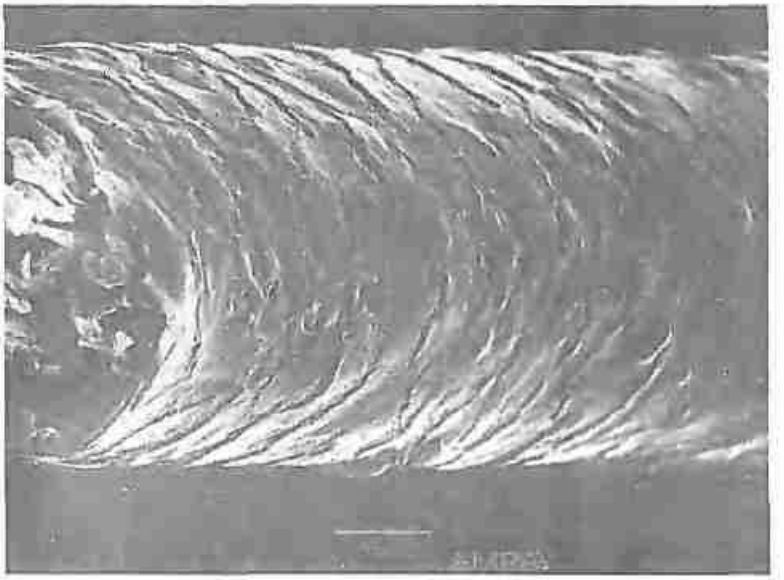
the same scale. We can see that the densely packed 0.18 μm lines on this ULSI memory chip are dwarfed by the scanning electron micrograph of a human hair which has the diameter of about 50 microns. This makes it dramatically clear why ULSI chips must be fabricated in ultra-clean environments.

Although the focus of this book is devices and not circuits, it is important to look at some typical applications of MOS capacitors and FETs in semiconductor logic and memory ULSI, which constitute about 90 percent of all ICs. This should give the reader a better feel for why we have studied the physics of MOS devices in Chapter 6. This is clearly not a comprehensive discussion, because the design and analysis of circuits is a large subject covered in other books and courses. We will first look at some digital logic applications, followed by some typical memory devices.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 473



(a)



(b)

Figure 9–23
Size comparison of ULSI circuit elements with human hair: (a) densely packed 0.18-μm interconnect lines in the array of a 64-Mb DRAM chip; (b) scanning electron micrograph of a human hair. Both micrographs are shown to the same scale. (© 2004 Micron Technology, Inc. All Rights Reserved. Used with permission.)

9.5.1 Logic Devices

A very simple and basic circuit element is the inverter, which serves to flip the logic state. When its input voltage is high (corresponding to logic "1"), its output voltage is low (logic "0"), and vice versa. Let us start the analysis with a resistor-loaded n-channel MOSFET inverter to illustrate the principles in the simplest possible manner (Fig. 9-24a). Then, we will extend the treatment to

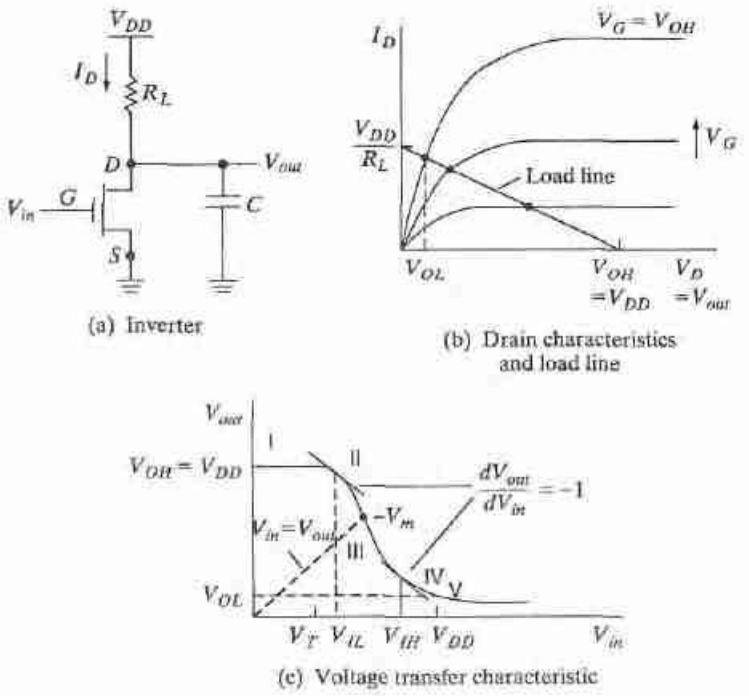


Figure 9–24
Resistor load inverter voltage transfer characteristics (VTC): (a) NMOSFET with load resistor R_L and load parasitic capacitance C; (b) determination of VTC by superimposing the load line (linear I–V ohmic characteristics of the resistor) on NMOSFET output characteristics; (c) VTC showing output voltage as a function of input voltage. The five key points on the VTC are logic high (V_{OH}), logic low (V_{OL}), unity gain points (V_{IL} and V_{IH}), and logic threshold where input equals output (V_m).

the slightly more complicated CMOS inverters which are much more useful and more common today.

A key concept for inverters is the *voltage transfer characteristic* (VTC), which is a plot of the output voltage as a function of the input bias (Fig. 9-24c). The VTC gives us information, for example, about how much noise the digital circuit can handle, and the speed of switching of the logic gates. There are five key operating points (marked I through V) on the VTC. They include V_{OH}, corresponding to the logic high or "1", V_{OL}, corresponding to the logic low or "0", and V_m, corresponding to the intersection of a line with unity slope (where $V_{out} = V_{in}$) with the VTC. V_m, known as the logic threshold (not to be confused with the V_T of the MOSFETs), is important when two inverters are cross-coupled in a flip-flop circuit because the output of one is fed to the input of the other, and vice versa. Two other key points are the unity gain points, V_{IL} and V_{IH}. The significance of these points is that if the input voltage is between them, the change of the input is amplified and we get a larger change of the output voltage. Outside of this operating range, the change of the input voltage is attenuated. Clearly, any noise voltage which puts the input voltage between V_{IL} and V_{IH} would be amplified, and lead to a potential problem with the circuit operation.

Let us see how to go about determining the VTC. From the circuit in Fig. 9-24a, we see that in the output loop from the power supply to ground, the current through the resistor load is the same as the drain current of the MOSFET. The power supply voltage is equal to the voltage drop across the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 475

resistor plus the drain-to-source voltage. To determine the VTC, we superimpose the load line of the load element (in this case a straight line for an ohmic resistor) on the output characteristics of the MOSFET (Fig. 9–24b). This is similar to our load line discussion in Section 6.1.1. The load line goes through V_{DD} on the voltage axis because when the current in the output loop is zero, there is no voltage drop across the resistor and all the voltage appears across the MOSFET. On the current axis, the load line goes through V_{DD}/R_L because when the voltage across the MOSFET is zero, the voltage across the resistor must be V_{DD}. As we change the input bias, V_{in}, we change the gate bias on the MOSFET, and thus in Fig. 9–24b, we go from one constant V_G curve to the next. At each input bias (and a corresponding constant V_G curve) the intersection of the load line with that curve tells us what the drain bias V_D is, which is the same as the output voltage. This is because at the point of intersection, we satisfy the condition that for the d-c case where the capacitor does not play any role, the current through the resistor is the same as the MOSFET current. (Later on, we shall see that in the a-c case when the logic gates are switched, we need to worry about the displacement current through the capacitor when it is charged or discharged.) It can be clearly seen from Fig. 9–24c that as the input voltage (or V_G) changes from low to high, the output voltage decreases from a high of V_{DD} to a low of V_{OL}. We can solve for any point on this VTC curve analytically simply by recognizing whether the MOSFET is in the linear region or in saturation, using the corresponding drain current expression [Eq. (6–49) or (6–53)] and setting it equal to the resistor current. As an illustration, suppose we want to determine the logic "0" level, V_{OL}. This occurs when the input V_G is high and the output V_D is low, putting the transistor in the linear region. Using Eq. (6–49), we can write

$$I_D = k\left[V_G - V_T - \frac{V_D}{2}\right]V_D = k\left[V_{DD} - V_T - \frac{V_{OL}}{2}\right]V_{OL} \qquad (9\text{–}3a)$$

Since in the d-c case the current through the MOSFET is the same as that through the resistor,

$$I_D = I_L = \frac{V_{DD} - V_{OL}}{R_L} \qquad (9\text{–}3b)$$

We can solve for V_{OL} if we know R_L and the MOSFET parameters. Alternatively, we can design the value of R_L to achieve a certain V_{OL}. What might dictate the choice of R_L? We shall see later in this section that for many applications we use two of these inverters in a cross-coupled manner to form a bistable flip-flop. The output of one flip-flop is fed back to the input of the other, and vice versa. Clearly, the V_{OL} must be designed to be significantly less than the V_T of the MOSFET. Otherwise, neither MOSFET will be fully turned off, and the flip-flop will not function properly. Similarly, all the other points on the VTC can be determined analytically by using the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 476

appropriate MOSFET drain current expression, and setting it equal to the current through the resistor.

We can make some general observations from this analysis. We want the transition region of the VTC (between V_{IL} and V_{IH}) to be as steep (i.e., high gain) as possible, and the transition should be around $V_{DD}/2$. High gain guarantees a high-speed transition from one logic state to the other. It is necessary to increase the load resistance to increase this gain in the transition region.

The transition around $V_{DD}/2$ guarantees high *noise immunity* or *margin* for both logic "1" and logic "0" levels. To appreciate the importance of noise immunity, we must recognize that in combinatorial or sequential digital circuits, the output of one inverter or logic gate is often fed into the input of the next stage. Noise immunity is a measure of how much noise voltage the circuit can tolerate at the input, and still have the digital outputs be at the correct logic level in the subsequent stages. For example in Fig. 9–24c, if the input is nominally at zero, the output should be high (logic "1"). If this is fed into another inverter stage, its output should be low, and so on. If a noise spike causes the input of the first stage to go above V_m, the output voltage decreases sufficiently to potentially create errors in the digital levels in subsequent stages. Having a symmetric transition of the VTC around $V_{DD}/2$ ensures that the noise margin is high for both logic levels.

One problem with the resistor load inverter is that the V_{OL} is low, but not zero. This, coupled with the fact that the load element is a passive resistor that cannot be turned off, causes high standby power dissipation in this circuit. These problems are addressed by the CMOS structure described next.

We can determine the VTC for the CMOS case exactly as for the resistor load, although the math is somewhat more messy (Fig. 9–25). As mentioned previously, for an input voltage V_{in}, the V_G of the NMOSFET is V_{in}, but that of the PMOSFET is $V_{in}-V_{DD}$. Similarly, if the output voltage is V_{out}, the V_D of the NMOSFET is V_{out}, but that of the PMOS is $V_{out}-V_{DD}$. The load element now is not a simple resistor with a linear current–voltage relationship, but instead is the PMOSFET device whose "load line" is a set of I_D–V_D output characteristics (Fig. 9–25b). The V_{out} can be determined as a function of the V_{in} by recognizing whether the NMOSFET and the PMOSFET are in the linear or saturation region of their characteristics, and using the appropriate current expressions. At each point, we would set the NMOSFET I_D equal to the PMOSFET I_D.

As in the case of the resistive load, there are five key points on the VTC (Fig. 9–25c). They are logic "1" equal to V_{DD}, logic "0" equal to 0, logic threshold V_m where $V_{in} = V_{out}$, and the two unity gain points, V_{IH} and V_{IL}. In region I in Fig. 9–25c, the NMOSFET is OFF, and $V_{out} = V_{DD}$. Similarly, in region V, the PMOSFET is OFF, and $V_{out} = 0$. We can illustrate the calculation in region II, where the NMOSFET is in saturation and the PMOSFET is in the linear region. In this case, we must use Eq. (6–53) for the saturation drain current of the NMOSFET.

$$I_{DN} = \frac{k_N}{2}(V_{in} - V_{TN})^2 \qquad (9\text{–}4a)$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 477

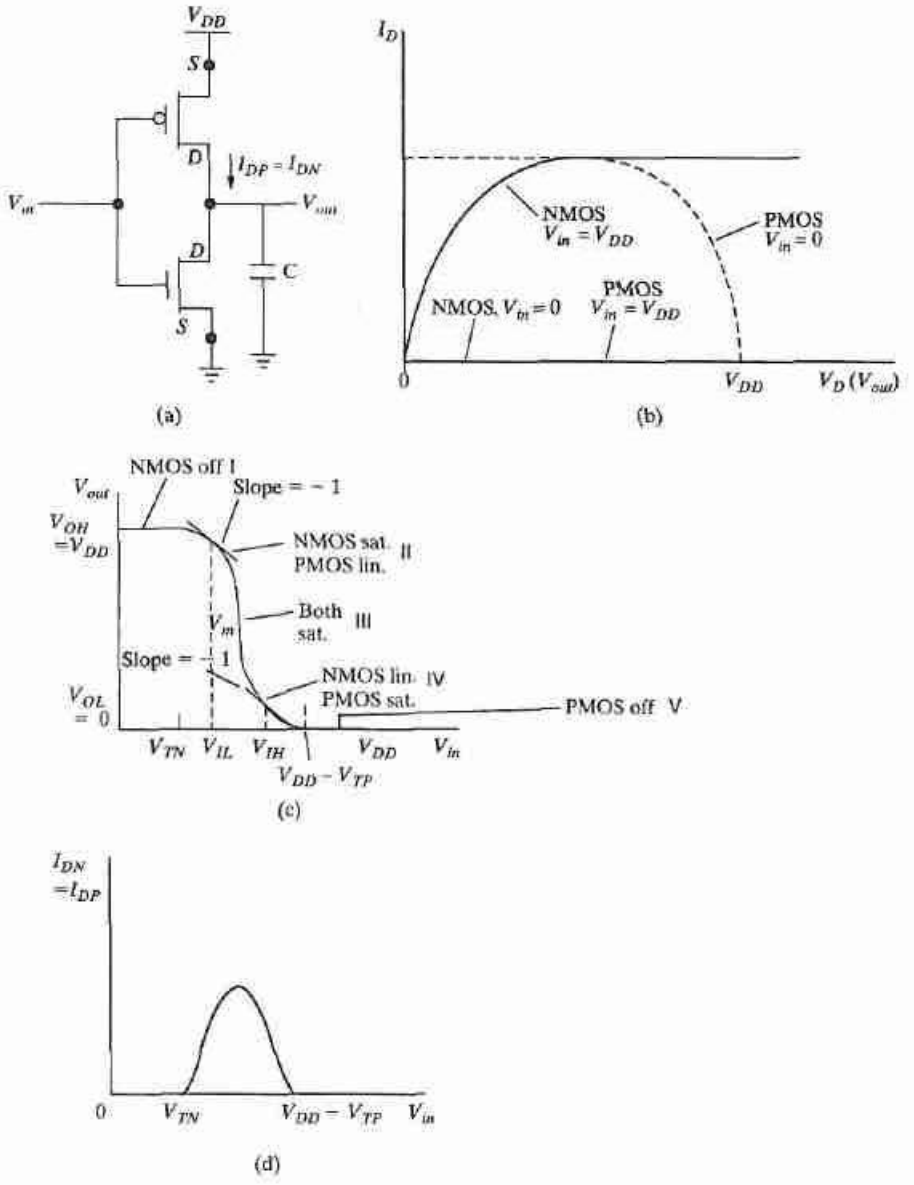


Figure 9–25
CMOS inverter voltage transfer characteristics: (a) NMOSFET with PMOSFET load and load parasitic capacitance, C; (b) determination of VTC by superimposing load line (output characterisitics of PMOSFET shown as dotted line) on NMOSFET output characteristics; (c) VTC showing output voltage as a function of input voltage. The five key points on the VTC are logic high (V_{OH}), logic low (V_{OL}), unity gain points (V_{IL} and V_{IH}), and logic threshold where input equals output (V_m); (d) switching current from V_{DD} to ground when the input voltage is in a range where both the NMOSFET and the PMOSFET are on.

On the other hand, we must use Eq. (6-49) for the PMOSFET in the linear region.

$$I_{DP} = k_P\left[(V_{DD} - V_{in}) + V_{TP} - \frac{(V_{DD} - V_{out})}{2}\right](V_{DD} - V_{out}) \quad (9\text{-}4b)$$

Here V_{TN} and V_{TP} are the n- and p-channel threshold voltages. In the d-c case, since the output load capacitor does not play a role, the drain current through the PMOSFET device must be equal in magnitude to that through the NMOSFET. (However, for the a-c case, we need to consider the displacement current through the capacitor.)

$$I_{DN} = I_{DP} \quad (9\text{-}5a)$$

Using Eq. (6-53) for the NMOSFET in saturation, and Eq. (6-49) for the PMOSFET in the linear region.

$$\frac{k_N}{2k_P}(V_{in} - V_{TN})^2 = \left[V_{DD} - V_{in} + V_{TP} - \frac{V_{DD} - V_{out}}{2}\right](V_{DD} - V_{out})$$

$$= \left[\frac{V_{DD}}{2} - V_{in} + V_{TP} + \frac{V_{out}}{2}\right](V_{DD} - V_{out}) \quad (9\text{-}5b)$$

From Eq. (9-5b), we can get an analytical relation between the input and output voltages valid in region II. We can get similar relationships in the other regions of the VTC.

Region IV is very similar to region II in Fig. 9-25c, except that now the NMOS is in the linear regime, while the PMOSFET is in saturation. In region III, both the NMOSFET and the PMOSFET are in saturation. Since the output impedance of a MOSFET is very high, this is tantamount to a semi-infinite load resistor, thereby resulting in a very steep transition region. That is why a CMOS inverter switches faster than the resistor load case. The CMOS inverter is also preferable because in either logic state (regions I or V), either the NMOSFET or the PMOSFET is OFF, and the standby power dissipation is very low. In fact, the current in either logic state corresponds to the (very low) source/drain diode leakage.

We want the transition region (region III) to be at $V_{DD}/2$ from the point of view of symmetry and noise immunity. Once again, by setting the NMOSFET I_D equal to that of the PMOSFET, it can be shown that the transition occurs at

$$V_{in} = (V_{DD} + \chi V_{TN} + V_{TP})/(1 + \chi) \quad (9\text{-}6a)$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 479

where

$$\chi = \left(\frac{k_N}{k_\mu}\right)^{\frac{1}{2}} = \frac{\left[\bar{\mu}_n C_i \left(\frac{Z}{L}\right)_N\right]^{\frac{1}{2}}}{\left[\bar{\mu}_p C_i \left(\frac{Z}{L}\right)_P\right]^{\frac{1}{2}}} \tag{9-6b}$$

We can design V_{in} to be at $V_{DD}/2$ by choosing $V_{TN} = -V_{TP}$ and $\chi = 1$. Since the effective electron mobility in the channel of a Si MOSFET is roughly twice that of the hole mobility, we must design CMOS circuits to have a $(Z/L)_P = 2(Z/L)_N$ to achieve the condition $\chi = 1$.

We can combine such CMOS inverters to form other logic gates for combinatorial circuits such as NOR gates and NAND gates (Fig. 9-26). The truth tables for these gates are shown in Fig. 9-27. By applying combinations

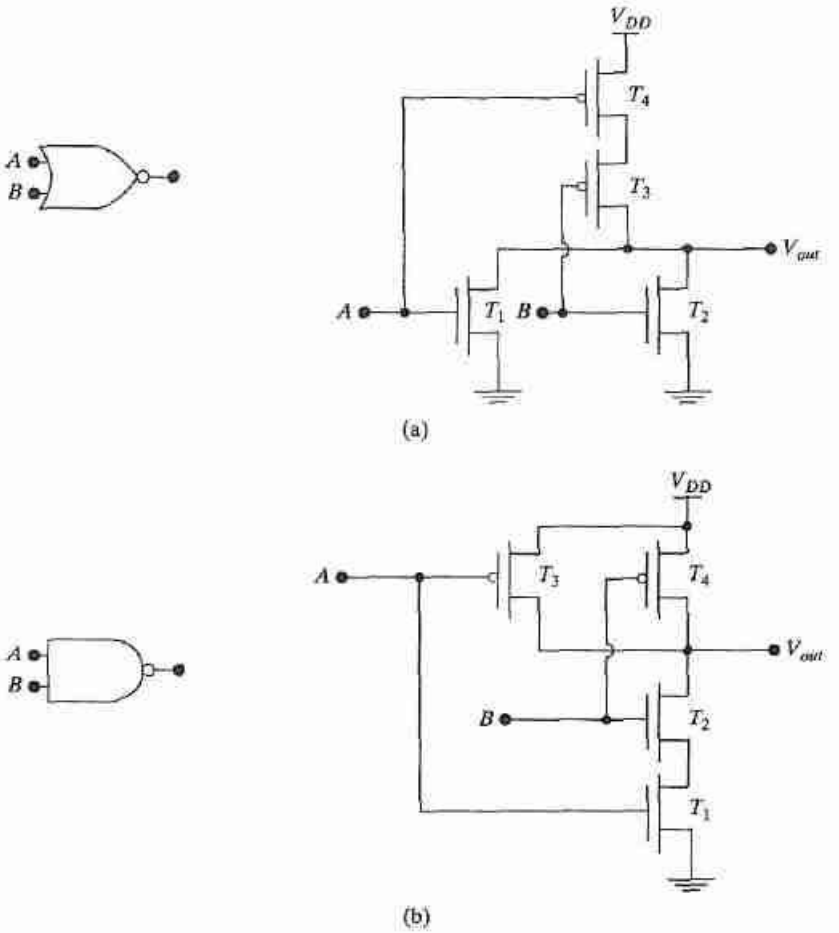
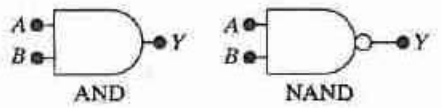


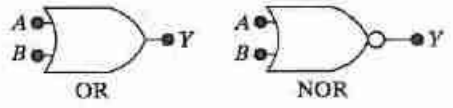
Figure 9–26 Logic gates and CMOS implementation of (a) NOR gate (b) NAND gate.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 480



Input		Output Y	
A	B	AND	NAND
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

(a)



Input		Output Y	
A	B	OR	NOR
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

(b)



Input		Output Y
A	B	XOR
0	0	0
0	1	1
1	0	1
1	1	0

(c)

Figure 9–27
(a) AND/NAND logic symbols and truth table; (b) OR/NOR logic symbols and truth table; (c) XOR logic symbol and truth table.

of logic "high" or logic "low" to inputs A and B, we get the output states corresponding to the truth tables. The synthesis of logic circuits corresponding to these truth tables can be done using Boolean algebra and De Morgan's laws. The upshot of these laws is that any logic circuit can be made using inverters in conjunction with either NAND gates or NOR gates. Which would be preferable from a device physics point of view? We see from Fig. 9–26, that in the NOR gate the PMOSFET devices T_3 and T_4 are in series, while for the NAND it is the NMOSFETs (T_1 and T_2). Since the electron channel mobilities are twice hole mobilities, we would obviously prefer NMOSFETs. Therefore, the preferred choice is NAND, along with inverters.

We can also estimate the power dissipation in the inverter circuit. We already know that the standby power dissipation is very small, being governed by the OFF state leakage current of either the NMOSFET or the PMOSFET, depending on the logic state. This leakage current depends on the source and drain diode leakage currents, or if the V_T is low, on the subthreshold leakage of the MOSFET that is turned OFF (see Section 6.5.7).

While the inverter is switching, there is also a transient current from the power supply to ground when both the transistors are ON (see Fig. 9–25d). This is known as the *switching current* or the *commutator current*. The magnitude of this current will clearly depend on the values of V_{TN} and V_{TP}. The higher the magnitudes of the thresholds, the less is the input voltage swing for which both

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 481

the PMOSFET and the NMOSFET will be ON while the input voltage is being changed. The commutator current is then less during switching, which is desirable from a reduced power dissipation point of view. However, this reduction of power dissipation by increasing threshold voltages is obtained at the expense of reduced drive current and, therefore, overall speed of the circuit.

The speed penalty due to reduction of drive current is because in a digital circuit, while switching between logic states, the MOSFET drive currents must charge and discharge the parasitic capacitors that are inevitably associated with the output node (Fig. 9-25a). There is also some power dissipation involved in charging and discharging load capacitors attached to the output of the inverter. This load capacitance depends mostly on the input gate oxide capacitance of the MOSFETs of the next inverter stage (or logic gate) that this inverter (or logic gate) may be driving, along with some small parasitic capacitances. The input load capacitance of a single inverter is given by gate oxide capacitance per unit area C_i times the device areas.

$$C_{inv} = C_i \{(ZL)_N + (ZL)_P\} \tag{9-7}$$

The total load capacitance is then multiplied by a factor that depends on the *fan-out* of the circuit, which is the number of gates that are being driven in parallel by the inverter (or logic gate). It is necessary to add up the load capacitances for all the inverters or logic gates that are being driven by this inverter stage. The energy expended in charging up the equivalent load capacitor, C, is the integral of the product of the time-dependent voltage times the time-dependent displacement current through the capacitor during the charging cycle.

$$\begin{aligned} E_C &= \int i_p(t)[V_{DD} - v(t)]dt \\ &= V_{DD}\int i_p(t)dt - \int i_p(t)v(t)dt \end{aligned} \tag{9-8a}$$

The energy stored in C is then obtained by considering the displacement current ($i_p(t) = C\,dv/dt$) through the capacitor:

$$E_c = V_{DD}\int C\frac{dv}{dt}\,dt - \int Cv\frac{dv}{dt}\,dt = CV_{DD}\int_0^{V_{DD}} dv - C\int_0^{V_{DD}} v\,dv = CV_{DD}^2 - \frac{1}{2}CV_{DD}^2 \tag{9-8b}$$

Similarly, during one discharging cycle we get

$$E_d = \int i_n(t)v(t)dt = -\int_{V_{DD}}^{0} Cv\,dv = \frac{1}{2}CV_{DD}^2 \tag{9-9}$$

If the inverter (or gate) is being charged and discharged at a frequency f, we get an active power disspation

$$P = CV_{DD}^2 f \tag{9-10}$$

In addition to power dissipation, we are also concerned with the speed of logic circuits. The speed of a gate, such as the one shown in Fig. 9-25, is determined

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 482

by the *propagation delay* time t_P. We define the time required for the output to go from the logic high V_{OH} to $V_{OH}/2$ as t_{PHL}. The converse (to go from logic low $V_{OL}(=0)$ to $V_{OH}/2$) is defined as t_{PLH}. We can write down approximate estimates for these times by recognizing that for the output to go from high to low (or logic "1" to "0"), the NMOSFET has to discharge the output node towards ground. During this period, the NMOSFET will be in saturation. Assuming a constant saturation current as an approximation, we obtain from Eq. (6-53)

$$t_{PHL} = \frac{\frac{1}{2}CV_{DD}}{I_{DN}} = \frac{\frac{1}{2}CV_{DD}}{\frac{k_N}{2}(V_{DD} - V_{TN})^2} \tag{9-11a}$$

This is the decrease of charge on the capacitor divided by the discharging current. Conversely,

$$t_{PLH} = \frac{\frac{1}{2}CV_{DD}}{I_{DP}} = \frac{\frac{1}{2}CV_{DD}}{\frac{k_P}{2}(V_{DD} + V_{TP})^2} \tag{9-11b}$$

Knowing these times helps us considerably in designing circuits that meet the speed requirements of a design. Of course, for accurate numerical estimates of these propagation time delays or of the power dissipation we need to use computers. A very popular program to do so is the Simulation Program with Integrated Circuit Emphasis (SPICE). This discussion illustrates that the device physics plays an important role in the design and analysis of such circuits.

9.5.2 Semiconductor Memories

In addition to logic devices such as microprocessors, integrated circuits depend on semiconductor memories. We can illustrate many key MOS device physics issues by looking at three of the most important types of semiconductor memory cells: the *static random-access memory (SRAM)*, the *dynamic random-access memory (DRAM)*, and the non-volatile *flash memory cell*. SRAMs and DRAMs are volatile in the sense that the information is lost if the power supply is removed. For flash memories, however, information is stored indefinitely. For SRAMs, the information is static, meaning that as long as the power supply is on, the information is retained. On the other hand, the information stored in the cells of a DRAM must periodically be refreshed because stored charge representing one of the logic states leaks away rapidly. The refresh time must be short compared with the time needed for stored charge to degrade.

The overall organization of all these types of memories is rather similar, and is shown in Fig. 9-28. We will not describe the memory organization in great detail here, but will instead focus on the device physics. We need to know the type of cell that is used at the intersection of the rows or wordlines, and the columns or bitlines. These memories are all random access in

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 483

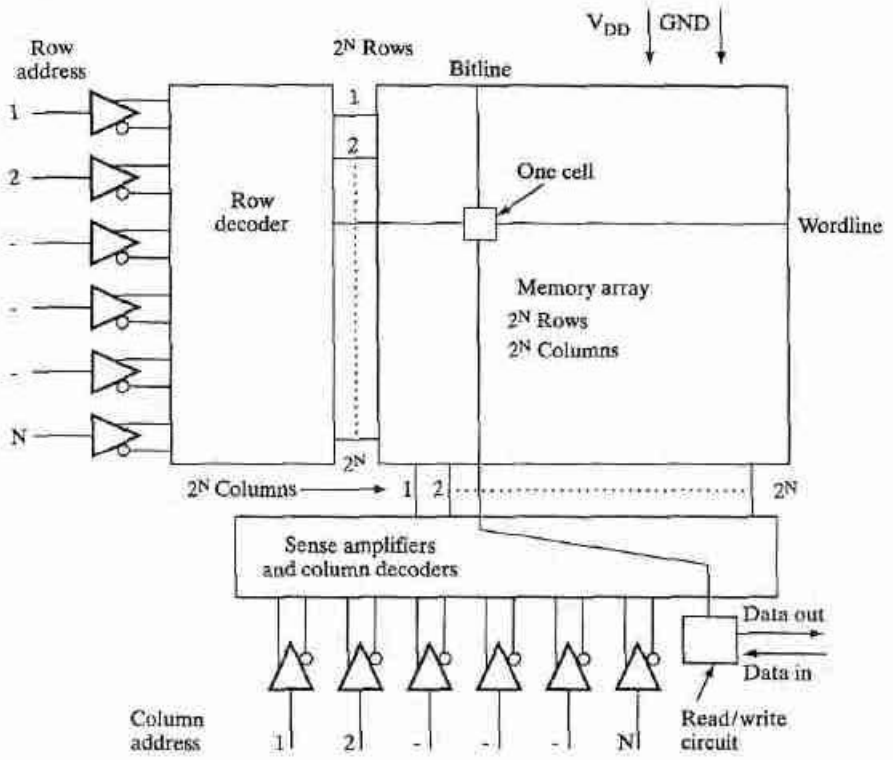


Figure 9–28
Organization of a random access memory (RAM): The memory array consists of memory cells arranged in an orthogonal array. There is one memory cell at the intersection of one row (wordline) and one column (bitline). To address a particular memory cell, the N row addresses are latched in from the N address pins, and decoded by the 2^N row decoders. All the memory cells on the selected row are read by the 2^N sense amplifiers. Of those, a cell (one bit) or group of cells (byte or word) is selected for transfer to the data output buffers depending on the column addresses that are decoded by the 2^N column decoders. Generally, to save pin count on the package, the N column addresses are provided in a multiplexed fashion to the same N address pins as the row addresses, *after* the row addresses have already been latched in.

the sense that the cells can be addressed for write or read operations in any order, depending on the row and column addresses provided to the address pins, unlike memories such as hard disks or floppy disks on a computer which can only be addressed sequentially. Generally, the same set of pins is used for both the row and the column addresses, in order to save pin count. This forces us to use what is known as address multiplexing. First, the row addresses are provided at the address pin, and decoded using row decoders. For N row addresses, we can have 2^N rows or wordlines. The row decoders then cause the selected wordline to go high, so that all the 2^N cells (corresponding to N column addresses) on this wordline are accessed for either read or write, through

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 484

sense amplifiers at the end of the 2^N columns or bitlines. After the appropriate row has been decoded, the appropriate column addresses are provided to the same address pins, and the column decoders are used to select the bit or group of bits (known as *byte* or *word*) out of all the 2^N bits on the selected wordline. We can either write into or read from the selected bit (or group of bits) using the sense amplifiers, which are basically flip-flops used as differential amplifiers.

SRAMs. A group of four 6-transistor CMOS SRAM cells is shown in Fig. 9-29. Each cell is found in this case at the intersection of a row or wordline, and a column or bitline (along with its logical complement known as bitline-bar). The cell is a flip-flop, consisting of two cross-coupled CMOS inverters. Clearly, it is bistable: if the output of one inverter is high (corresponding to the NMOSFET being OFF, and the PMOSFET ON), that high voltage is fed to the input of the other cross-coupled inverter, and the output of the other inverter will be low. This is one logic state (say "1") of the SRAM. Conversely, the other stable state of the flip-flop can be considered to be the other logic state (say "0"). Many of the device issues are identical to those described in Section 9.5.1 in connection with the VTC of inverters. We aim for a symmetric transition from V_{OH} to V_{OL} at $V_{DD}/2$ with a high gain in the transition region, to improve noise

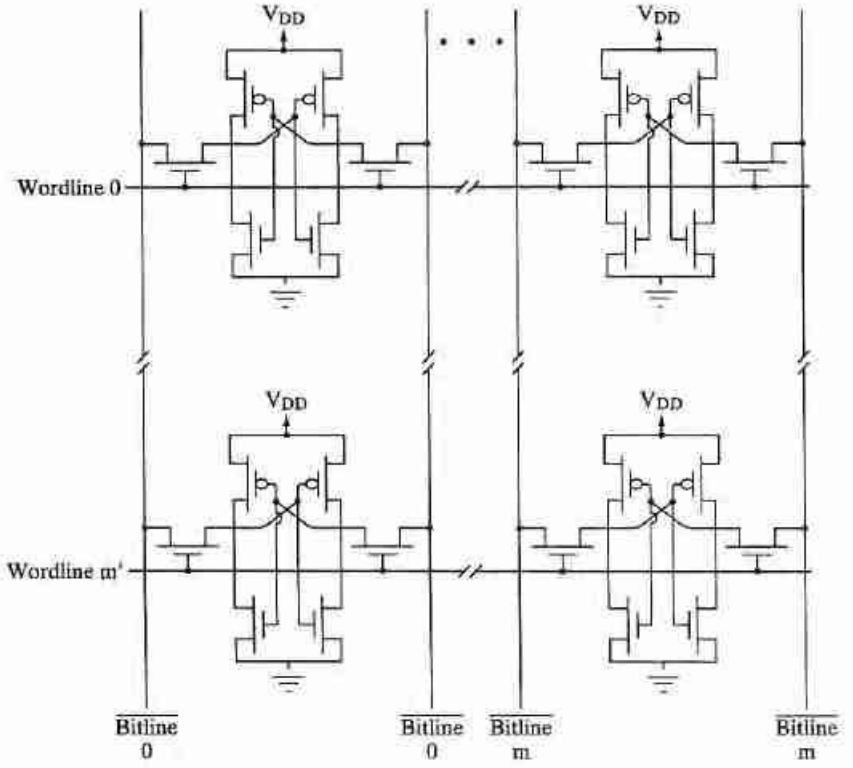


Figure 9-29
An array of 4 CMOS SRAM cells. The bitline and $\overline{\text{bitline}}$ (bitline-bar) are logical complements of each other.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 485

immunity and speed of convergence of the SRAM cell. The speed of convergence determines how fast the SRAM flip-flop latches into one stable logic state or the other. The cells are accessed through two access transistors whose gates are controlled by the wordline. That is why this is called a 6-transistor cell. Other SRAM cells use load resistors in the inverters, rather than PMOSFETs, leading to a 4-transistor, 2-resistor cell. As discussed in Section 9.5.1, the CMOS cell has superior performance, but at the expense of occupying more area.

Unless the row decoders cause a particular wordline to go high, the SRAM cells on that wordline are electrically isolated. By selecting a particular wordline, the access transistors on that row are turned ON and act as logic transmission gates between the output nodes of the SRAM cell and the bitline and its complement, the bitline-bar. During a read operation, the bitline and its complement are both precharged to the same voltage. Once the access transistors are turned ON, a small voltage differential develops between bitline and bitline-bar because the output nodes of the SRAM are at different voltages (0 and V_{DD}). The voltage differential that is established is due to a charge redistribution that occurs between the parasitic capacitance associated with the output nodes of the SRAM and the bitline capacitance. This voltage difference is amplified by the sense amplifiers. As mentioned previously, the sense amplifiers are differential amplifiers, very similar in configuration to the SRAM flip-flop cell itself. The bitline and bitline-bar (complement of the bitline) are fed to the two inputs of the sense amplifier, and the voltage differential is amplified until the voltage separation is V_{DD}.

DRAMs. The DRAM cell structure is shown in Fig. 9-30. The information is stored as charge on an MOS capacitor, which is connected to the bitline through a switch which is an MOS *pass transistor*, the gate of which is controlled by the wordline. There is one such cell at each intersection of the orthogonal array of wordlines and bitlines, exactly as for SRAMs. When the

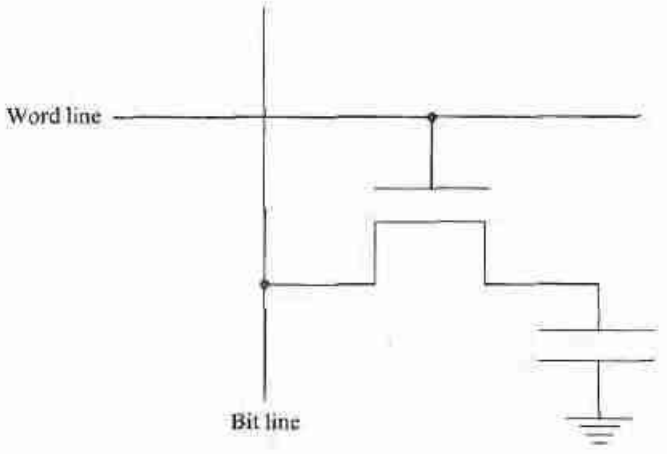


Figure 9–30
One transistor, one capacitor DRAM cell equivalent circuit: the storage MOS capacitor is connected to the bitline through the pass transistor (MOSFET switch) whose gate is controlled by the wordline.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 486

wordline voltage becomes higher than the V_T of the pass transistor (MOSFET between the bitline and the storage capacitor), the channel is turned ON, and connects the bitline to the MOS storage capacitor. The gate of this capacitor (or capacitor plate) is permanently connected to the power supply voltage V_{DD}, thereby creating a potential well under it which tends to be full of inversion electrons for a p-type substrate (Fig. 9-31a). We apply either 0 V to the bitline (generally corresponding to logic "0"), or V_{DD} (corresponding to logic "1"), and the appropriate voltage appears as the substrate potential of the MOS capacitor. For a stored "0" in the cell, the potential well that is created under the MOS capacitor by the plate voltage is full of inversion charge (Fig. 9-31b,c). When the wordline voltage is turned low such that the MOS pass transistor is turned off, the inversion charge under the storage capacitor stays the same; this is the stable state of the capacitor. On the other hand, if a positive voltage (V_{DD}) is applied to the bitline, it draws out the inversion electrons through the pass transistor (Fig. 9-31d,e). When the pass transistor is cut off, we end up with an empty potential well under the MOS capacitor plate. Over a period of time, the potential well tends to be filled up by minority carrier electrons that are constantly created by thermal generation-recombination in the substrate and are collected under the charged MOS capacitor plate. Hence, the logic "1" degrades towards the logic "0". That is why a DRAM is considered to be "dynamic" unlike an SRAM. It is necessary to periodically restore the logic levels or "refresh" the stored information.

There are interesting device physics issues regarding the pass transistor. This is like the access transistor in the SRAM, or a logic transmission gate. We see that in this MOSFET, neither the source nor the drain is permanently grounded. In fact, which side acts as the source and which as the drain depends on the circuit operation. When we are writing a logic "1" into the cell, the bitline voltage is held high ($=V_{DD}$). As this voltage is written into the cell, it is as if the source of the pass transistor gets charged up to V_{DD}. Another way of looking at this is that with respect to the source, the substrate bias of the pass transistor is $-V_{DD}$. The body effect of the MOSFET (Section 6.5.6) causes its V_T to increase. This is very important because for the pass transistor to operate as a transmission gate it is necessary that it be in the linear regime throughout, and not get into saturation (with a concomitant voltage drop across the pinch-off region). Hence, the gate or the wordline voltage must be held at V_{DD} (which is the final voltage of the source/drains) *plus* the V_T of the MOSFET, taking body effect into account. It is also important to make sure that the leakage of the pass transistor is low enough to satisfy refresh requirements of the DRAM. Not only must the source/drain diodes be low leakage, but the V_T and the subthreshold slope must be optimized such that subthreshold leakage for the grounded wordline case is low enough.

The stored charge difference between the two logic states can be determined by looking at the capacitance-voltage (C-V) characteristics of the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 487

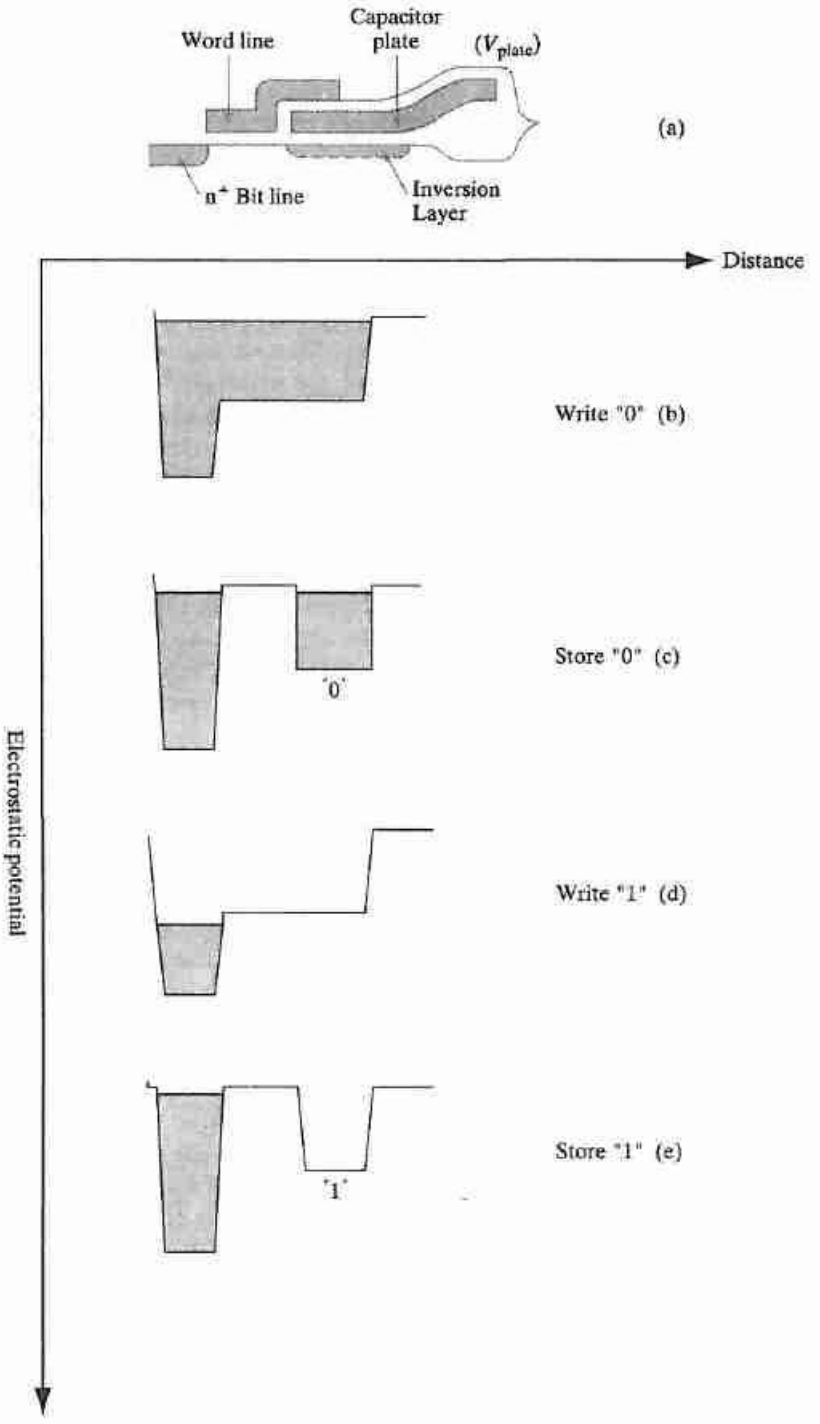


Figure 9–31 DRAM cell structure and cell operation: (a) cell structure corresponding to equivalent circuit of Fig. 9–30; (b)–(e) potentials under bitline, pass transistor channel and storage capacitor during write "0", store "0", write "1" and store "1" operations. It shows that the logic state "0" corresponds to a filled potential well (stable state), while the logic state "1" corresponds to an empty potential well (unstable state) that is filled up over time by minority carriers generated in the substrate and leakage through the pass transistor.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 488

MOS capacitor (Fig. 9-32). For a stored "1", essentially there is a substrate bias applied to the MOS capacitor, which raises its V_T due to the body effect (Section 6.5.6). Hence, the C–V characteristics shift to the right for a stored "1". Since the MOS capacitance is not a fixed capacitance, but is voltage dependent, we saw earlier that it must be defined in a differential form [Eq.(6-34a)]. Alternatively, we can write down the stored charge under the capacitor as

$$Q = \int C(V)dV \tag{9-12}$$

This is simply the area under the C–V curve. The charge differential that distinguishes the logic "1" and the logic "0" is the difference of areas under the capacitance-voltage curves in the two cases (Fig. 9-32).

When reading the cell, the pass transistor is turned on, and the MOS storage capacitor charge is dumped on the bitline capacitance C_B, precharged to V_B (typically = V_{DD}). The swing of the bitline voltage will clearly depend on the voltage V_C stored in the storage cell capacitance C_C. As in the case of the SRAM, the change of the bitline voltage depends on the capacitance ratio between the bitline and the cell. To do differential sensing in the case of DRAMs, we do not use two bitlines per cell as for SRAMs. Instead, we compare the bitline voltage for the selected cell with a reference bitline voltage to which is connected a dummy cell whose MOS capacitance, C_D, is roughly half that of the actual cell capacitance, C_C. Typical values of C_B, C_C,

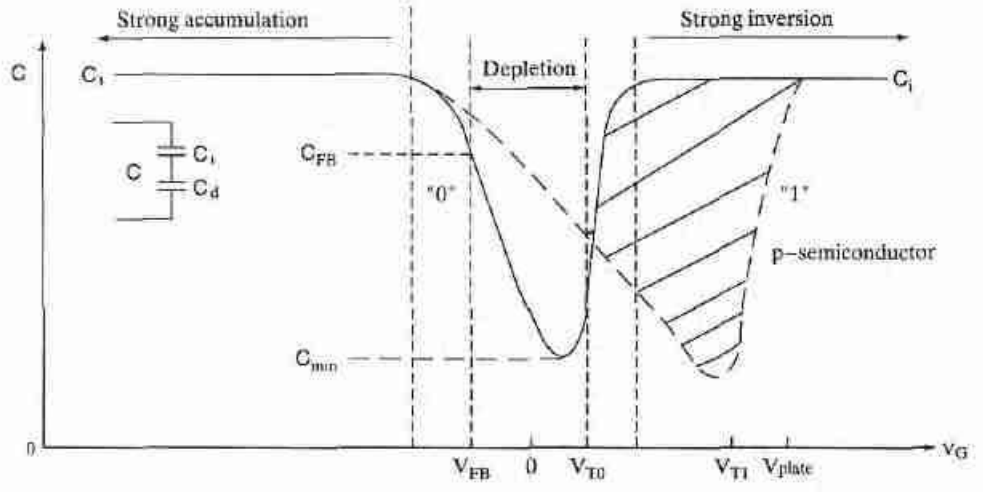


Figure 9–32
C–V characteristics of DRAM MOS capacitor in stored "0" and stored "1" states. The difference of area under the C–V curves shown by hatch-marked pattern reflects the charge differential between the two states.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 489

and C_D in a DRAM are 800 fF, 50 fF, and 20 fF, respectively. The voltage differential that is applied to the sense amplifier then becomes (Fig. 9-33)

$$\Delta V = \frac{C_C V_C + C_B V_B}{C_B + C_C} - \frac{C_C V_D + C_B V_B}{C_B + C_D}$$
$$= \frac{(V_B - V_D)C_B C_D - (V_B - V_C)C_B C_D - (V_C - V_D)C_C C_D}{(C_B + C_C)(C_B + C_D)} \quad (9\text{-}13a)$$

If V_D is set to zero, the expression simplifies to:

$$\Delta V = \frac{(V_B C_B + V_C C_C)C_D - (V_B - V_C)C_B C_C}{(C_B + C_C)(C_B + C_D)} \quad (9\text{-}13b)$$

Putting the cell voltage V_C equal to 0 V or 5 V, and typical, acceptable bitline-to-cell capacitance ratios C_B/C_C (= 15–20) in Eq. (9-13b), we get different polarities of the differential voltage of the order of ±100 mV for logic "1" and logic "0", respectively, which can be detected by sense amplifiers. From Eq. (9-13b), it can be seen that for much higher bitline-to-cell capacitance ratios, the swing of the bitline voltage will be negligible, regardless of the cell voltage. The minimum required cell capacitance C_C is about 50 fF, governed by so-called soft errors. DRAMs, like everything else on Earth, are constantly being bombarded by cosmic rays, and high-energy alpha particles can create electron-hole pairs in semiconductors. A typical collected charge due to one of these events is about 100 fC. This spurious charge can be neglected if the cell capacitance is 50 fF and 5 V is applied to the cell, for which the stored charge is roughly 250 fC. The DRAM cell then becomes immune to typical alpha particle hits.

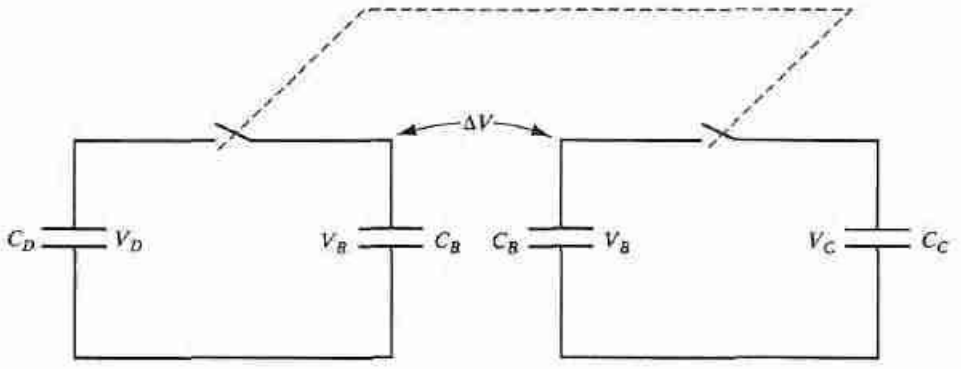


Figure 9-33
Equivalent circuit showing charge redistribution between cell capacitance (C_C) and bitline capacitance (C_B) on one side, versus dummy cell (C_D) and bitline capacitance (C_B) on other.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 490

Maintaining a cell capacitance of 50 fF as the cell dimensions are reduced from one generation of DRAM to the next is a tremendous technological challenge. One way to look at this problem is shown in Fig. 9-34. The challenge is to store more charge per unit area on the planar surface (A_s) of the Si substrate. Approximating the MOS capacitance as a fixed, voltage-independent capacitor, we can write the stored charge Q as

$$Q = CV = (\epsilon A_C/d)V \tag{9-14}$$

where ϵ is the permittivity of the dielectric, d is its thickness, and A_C is the capacitor area. As shown in Fig. 9-34, the historical way of achieving the desired capacitance has been to scale the dielectric thickness, d. But that runs into the problems discussed in Section 6.4.7. Another approach, which is being taken currently, is to use fabrication schemes to increase the area devoted to the MOS storage capacitor, A_c, even as we reduce the planar

Time	Past	Present	Future
Approaches	Scaled dielectric	Trench/ stacked capacitor	Alternate dielectric
Problems	Tunneling & wearout	Fabrication	Material properties

$$\frac{Q}{A_s} = \frac{CV}{A_s} = \frac{V}{d} \times \frac{A_c}{A_s} \times \epsilon$$

Figure 9-34
Various approaches (past, present, and future) of achieving higher DRAM cell capacitance and charge storage density without increasing cell size. A_s = Area on wafer taken by capacitor; A_c = Area of capacitor. For a planar capacitor $A_c = A_s$; however, for nonplanar structures $A_c > A_s$. $C = A_c \epsilon/d$ is the capacitance; and $Q = CV$ is the total stored charge in a fixed, voltage-independent capacitor.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 491

surface area on the wafer, A_s, used for making this storage capacitance. Obviously, this can be done by moving away from a purely planar structure, and exploiting the third dimension. We can go down into the Si by digging "trenches" in the substrate with RIE and forming a trench storage capacitor on the sidewalls of the trench (Fig. 9–35a). Alternatively, we can go up from the substrate by stacking multiple layers of capacitor electrodes to increase the "stacked" capacitor area (Fig. 9–35b). Other tricks that have been tried are to purposely create a rough polysilicon surface on the capacitor plates to increase the surface area. In the future, alternative materials may be used. For example, the ferroelectrics have much higher dielectric constant than SiO_2 and offer larger capacitance without increasing area or reducing thickness. Promising materials include barium strontium titanate and zirconium oxide.

Flash Memories. Another interesting MOS device is the flash memory, which is rapidly becoming the most important type of non-volatile memory. The memory cell structure is shown in Fig. 9–36. It is very simple and compact, and looks just like a MOSFET, except that it has two gate electrodes, one on top of the other. The top electrode is the one that we have direct electrical access to, and is known as the control gate. Below that we have a so-called

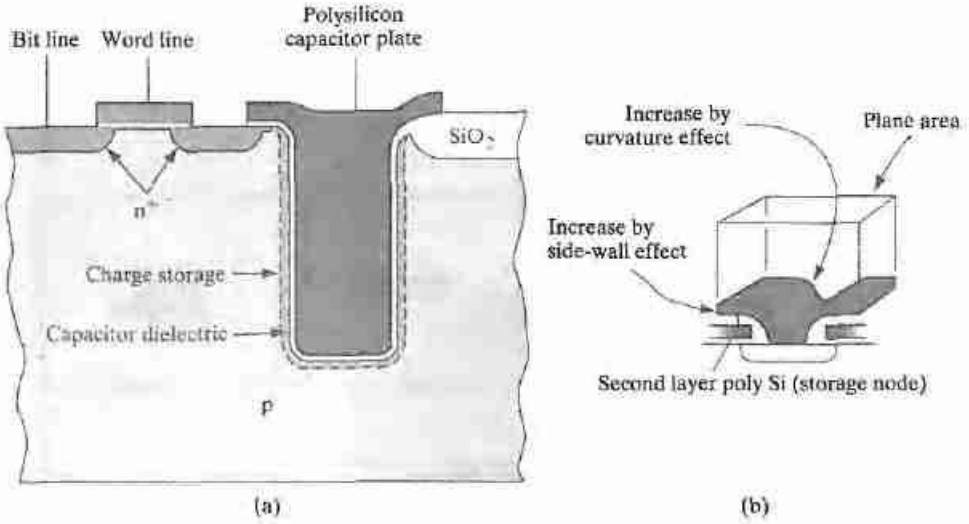


Figure 9–35
Increasing cell capacitance by exploiting the vertical dimension: (a) trench capacitors involve etching a trench in the substrate so that the larger area on the sidewalls can be used to increase capacitance; (b) stacked capacitors go "up" rather than "down" as in trenches, and increase capacitor area by using multiple polysilicon capacitor plates or "fins," as well as by exploiting the topography of the cell surface.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 492

"floating" gate that is capacitively coupled to the control gate and the underlying silicon.

The capacitive coupling of the floating gate to the various terminals is illustrated in Fig. 9-36 in terms of the various coupling capacitance components. The floating gate and the control gate are separated by a stacked oxide-nitride-oxide dielectric in typical flash devices. The capacitance between these two gates is called C_{ONO} because of the oxide-nitride-oxide makeup of the dielectric stack. The total capacitance C_{TOT} is the sum of all the parallel components shown in Fig. 9-36.

$$C_{TOT} = C_{ONO} + C_{TOX} + C_{FLD} + C_{SRC} + C_{DRN} \quad (9\text{-}15)$$

where C_{TOX} is the floating gate-to-channel capacitance through the tunnel oxide, C_{FLD} is the floating gate-to-substrate capacitance in the LOCOS field oxide region, and C_{SRC} and C_{DRN} are the gate-to-source/drain overlap capacitances.

Since it is isolated by the surrounding dielectrics, the charge on the floating gate Q_{FG} is not changed by (moderate) changes of the terminal biases.

$$Q_{FG} = 0 = C_{ONO}(V_{FG} - V_G) + C_{SRC}(V_{FG} - V_S) + C_{DRN}(V_{FG} - V_D) \quad (9\text{-}16)$$

We assume that the substrate bias is fixed, and hence ignore the contributions from C_{TOX} and C_{FLD}, which couple the floating gate to the substrate. The floating gate voltage can be indirectly determined by the various terminal voltages, in terms of the gate, drain, and source coupling ratios as defined in Eq. (9-17).

$$V_{FG} = V_G \cdot GCR + V_S \cdot SCR + V_D \cdot DCR \quad (9\text{-}17)$$

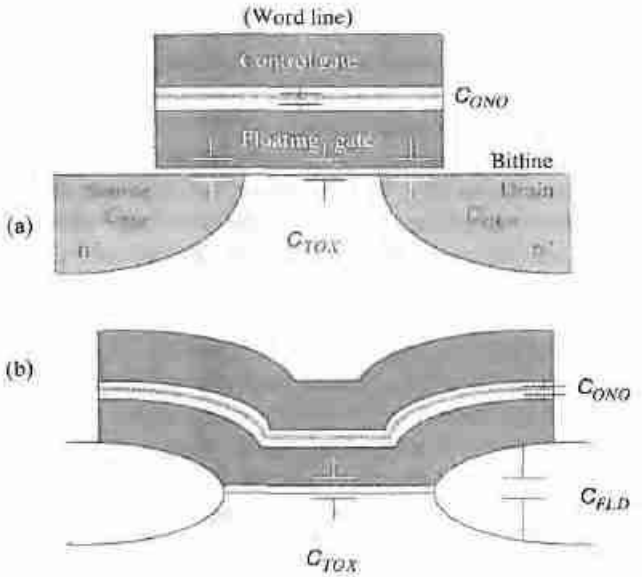


Figure 9-36
Flash memory cell structure: (a) cell structure shown along the channel length showing the control gate (wordline), floating gate below it, the source and the drain (bitline); (b) view of cell along the width of the MOSFET. The various coupling capacitors to the floating gate are shown.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 493

where

$$GCR = \frac{C_{ONO}}{C_{TOT}}$$

$$DCR = \frac{C_{DRN}}{C_{TOT}}$$

$$SCR = \frac{C_{SRC}}{C_{TOT}}$$

The basic cell operation involves putting charge on the floating gate or removing it, in order to program the MOSFET to have two different V_T's, corresponding to two logic levels. We can think of the stored charge on the floating gate to be like the fixed oxide charge in the V_T expression [Eq. (6–38)]. If many electrons are stored in the floating gate, the V_T of an NMOSFET is high; the cell is considered to have been "programmed" to exhibit the logic state "0". On the contrary, if electrons have been removed from the floating gate, the cell is considered to have been "erased" into a low V_T state or logic "1".

How do we go about transferring charges into and out of the floating gate? To program the cell, we can use channel hot carrier effects that we discussed in Section 6.5.9. We apply a high field to both the drain (bitline) and floating gate (wordline) such that the MOSFET is in saturation. It was discussed in Section 6.5.9 that the high longitudinal electric field in the pinch-off region accelerates electrons toward the drain and makes them energetic (hot). We maximize such hot carrier effects near the drain pinch-off region in a flash device by making the drain junction somewhat shallower than the source junction (Fig. 9-37a). This can be achieved by a separate higher energy source implant that is masked in the drain region. If the kinetic energy of electrons is high enough, a few can become hot enough to be scattered into the floating gate. They must surmount the 3.1-eV energy barrier that exists between the conduction band of Si and that of SiO_2, or hot electrons can tunnel through the oxide (Fig. 9–37b). Once they get into the floating gate, electrons become trapped in the 3.1-eV potential well between the floating polysilicon gate and the oxides on either side. This barrier is extremely high for a trapped (low kinetic energy) electron. Therefore the trapped electrons essentially stay in the floating gate forever, unless the cell is intentionally erased. That is why a flash memory is non-volatile.

To erase the cell, we use Fowler–Nordheim tunneling between the floating gate and the source in the overlap region (Fig. 9-38a). A high positive voltage (say −12 V) is applied to the source with the control gate grounded. The polarity of the field is such that electrons tunnel from the floating gate into the source region, through the oxide barrier (Section 6.4.7). The band diagram (along a vertical line in this overlap region) during the operation is shown in Fig. 9–38b. Interestingly, in a flash device we make use of two effects that are considered to be "problems" in regular MOS devices: hot carrier effects and Fowler–Nordheim tunneling.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 494

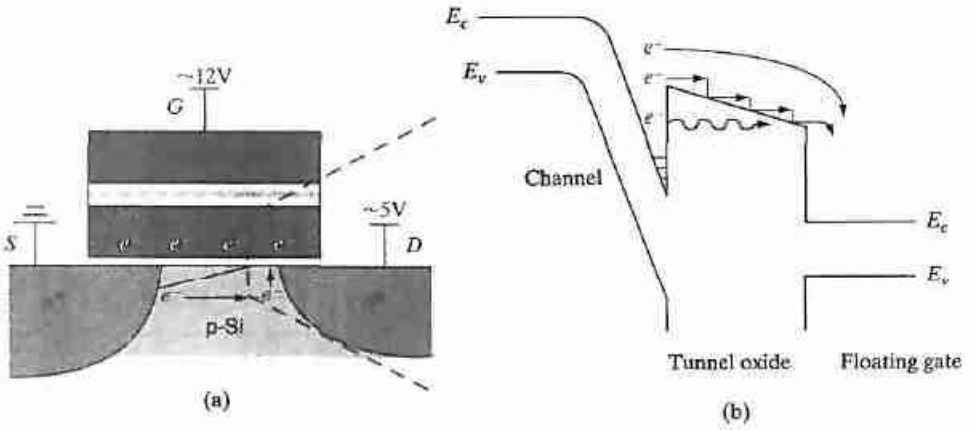


Figure 9–37
Hot carrier programming of the flash cell: (a) flash memory cell structure with typical biases required for writing into the cell. The channel of the MOSFET is pinched off in saturation; (b) band diagram along a vertical line in the middle of MOSFET channel showing hot electrons in the channel being injected across the gate oxide and getting trapped in the floating gate.

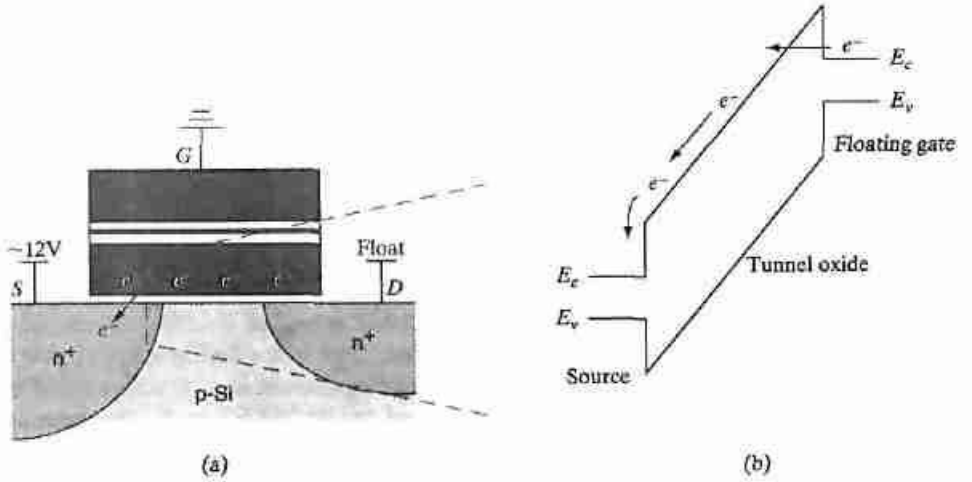


Figure 9–38
Fowler–Nordheim tunneling erasure: (a) flash memory cell structure with typical biases required for erasing the cell; (b) band diagram as a function of depth in the gate/source overlap region of the MOSFET showing quantum mechanical tunneling of carriers from the floating gate into the oxide, and subsequent drift to the source.

During the read operation, we apply a moderate voltage (~1 V) to the bitline (drain of the MOSFET), and a wordline (control gate) voltage V_{CG} that causes the capacitively coupled floating gate voltage to be between that of the high V_T and the low V_T state of the programmed flash memory cell

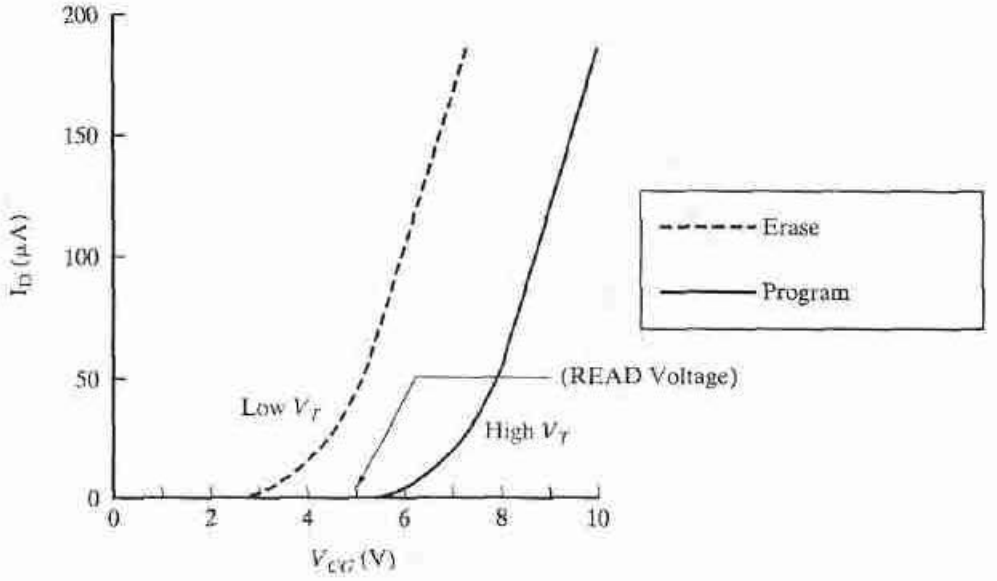


Figure 9–39
Drain (bitline) current versus control gate (wordline) voltage transfer characteristics of the MOSFET in a flash cell: if the cell is programmed to a high V_T (logic "0"), and a read voltage is applied to the wordline that is below this V_T, the MOSFET does not conduct, and there is negligible bitline current. On the other hand, if the cell had been erased to a low V_T state (logic "1") the MOSFET is turned ON, and there is significant bitline current.

(Fig. 9–39). There will be negligible drain current flow in the bit line (drain) for the high V_T case because the gate voltage is less than the threshold voltage. We will then interpret the selected cell as being in state "1". For the low V_T case, since the applied gate voltage is higher than the threshold voltage of the cell, there will be drain current flow in the bitline (drain), and this can be interpreted as state "0". The read operation can be understood by looking at the transfer characteristics of the MOSFET in the programmed and erased states (Fig. 9–39).

9.6 TESTING, BONDING, AND PACKAGING

After the preceding discussions of rather dramatic fabrication steps in monolithic circuit technology, the processes of attaching leads and packaging the devices could seem rather mundane. Such an impression would be far from accurate, however, since the techniques discussed in this section are crucial to the overall fabrication process. In fact, the handling and packaging of individual circuits can be the most critical steps of all from the viewpoints of cost and reliability. The individual IC chip must be connected properly to outside leads and packaged in a way that is convenient for use in a larger circuit or system. Since the devices are handled individually once they are separated

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 496

from the wafer, bonding and packaging are expensive processes. Considerable work has been done to reduce the steps required in bonding. We shall discuss the most straightforward technique first, which involves bonding individual leads from the contact pads on the circuit to terminals in the package. Then we shall consider two important methods for making all bonds simultaneously. Finally, we shall discuss a few typical packaging methods for ICs.

9.6.1 Testing

After the wafer of monolithic circuits has been processed and the final metallization pattern defined, it is placed in a holder under a microscope and is aligned for testing by a multiple-point probe (Fig. 9-40). The probe contacts the various pads on an individual circuit, and a series of tests are made of the electrical properties of the device. The various tests are programmed to be made automatically in a very short time. These tests may take only milliseconds for a simple circuit, to several seconds for a complex ULSI chip. The information from these tests is fed into a computer, which compares the results with information stored in its memory, and a decision is made regarding the acceptability of the circuit. If there is some defect so that the circuit falls below specifications, the computer remembers that chip must be discarded. The probe automatically steps the prescribed distance to the next circuit on the wafer and repeats the process. After all of the circuits have been tested and the substandard ones noted, the wafer is removed from the testing machine, sawed between the circuits, and broken apart (Fig. 9-41). Then each die that passed the test is picked up and placed in the package. In the testing process, information from tests on each die can be stored to facilitate analysis of the rejected circuits or to evaluate the fabrication process for possible changes.

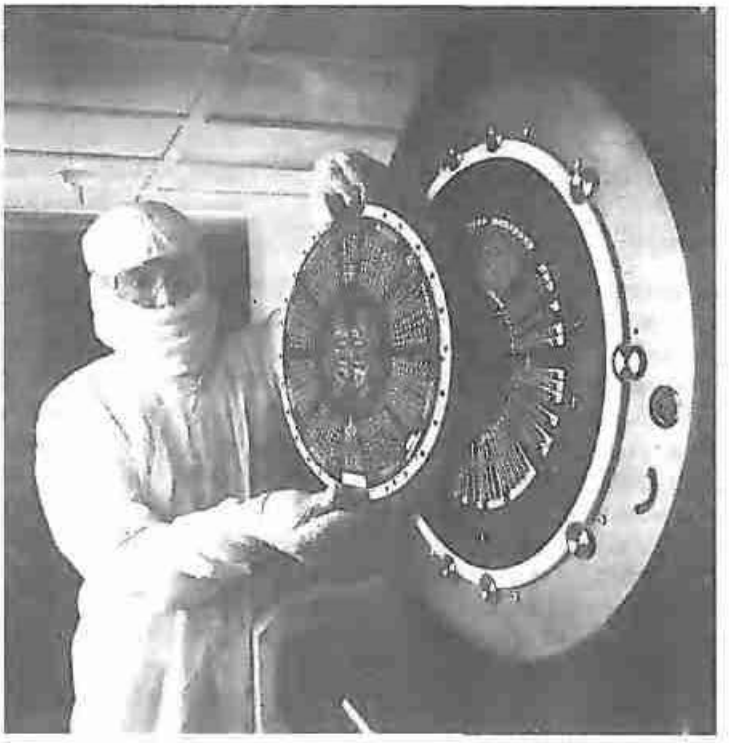
9.6.2 Wire Bonding

The earliest method used for making contacts from the monolithic chip to the package was the bonding of fine Au wires. Later techniques expanded wire bonding to include Al wires and several types of bonding processes. Here we shall outline only a few of the most important aspects of wire bonding.

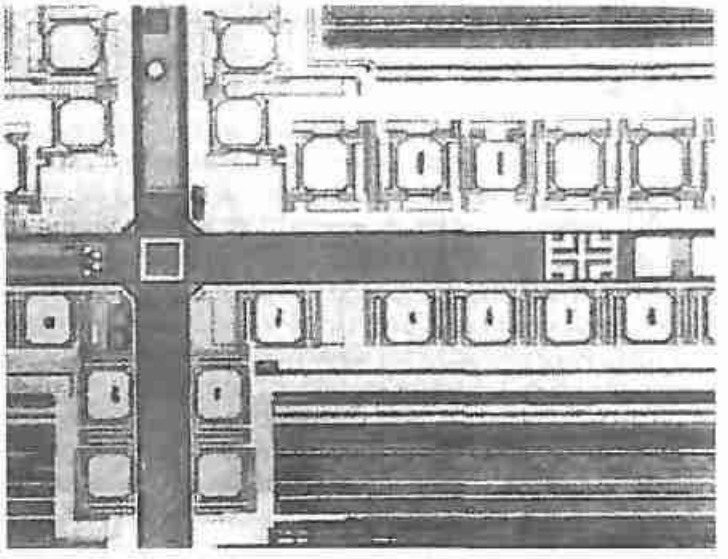
If the chip is to be wire bonded, it is first mounted solidly on a metal lead frame or on a metallized region in the package. In this process a thin layer of Au (perhaps combined with Ge or other elements to improve the metallurgy of the bond) is placed between the bottom of the chip and the substrate; heat and a slight scrubbing motion are applied, forming an alloyed bond which holds the chip firmly to the substrate. This process is called *die bonding*. Generally, die bonding is done by a robotic arm that picks up each die, orients it, and places it for bonding. Once the chip is mounted, the interconnecting wires are attached from the various contact pads to posts on the package (Fig. 9-42).

In Au wire bonding, a spool of fine Au wire (about 0.007–0.002-inch diameter) is mounted in a *lead bonder* apparatus, and the wire is fed through a glass or tungsten carbide *capillary* (Fig. 9-43a). A hydrogen gas

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 497



(a)



(b)

Figure 9–40
Automated probing of devices: (a) high-speed testing of ICs is done using probe cards having a rigidly fixed array of probes that corresponds to the bond pad pattern on the IC to be tested. Many electrical signals are provided or measured by the automated tester at the various pins. After one chip is tested, the tester mechanically moves the wafer to the next die location; (b) array of Al bond pads near the chip periphery, some showing probe marks. The space between the arrays of bond pads is the "scribe line" along which the wafer will be sawed into individual chips after testing. (© 2004 Micron Technology, Inc. All Rights Reserved. Used with permission.)

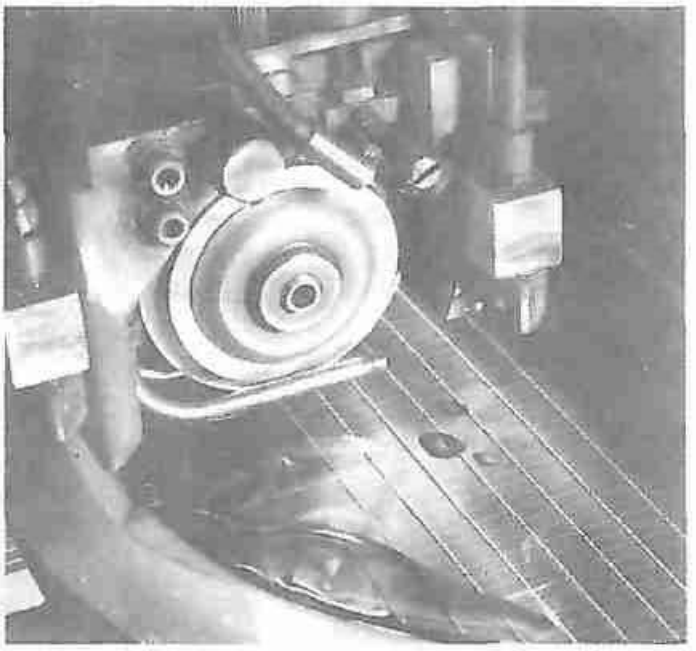


Figure 9–41
Sawing of a wafer along scribe lines: After the wafer is tested, the "known good dies" are "inked" or identified. The wafer is then sawed into individual dies for subsequent packaging. (© 2004 Micron Technology, Inc. All Rights Reserved. Used with permission.)

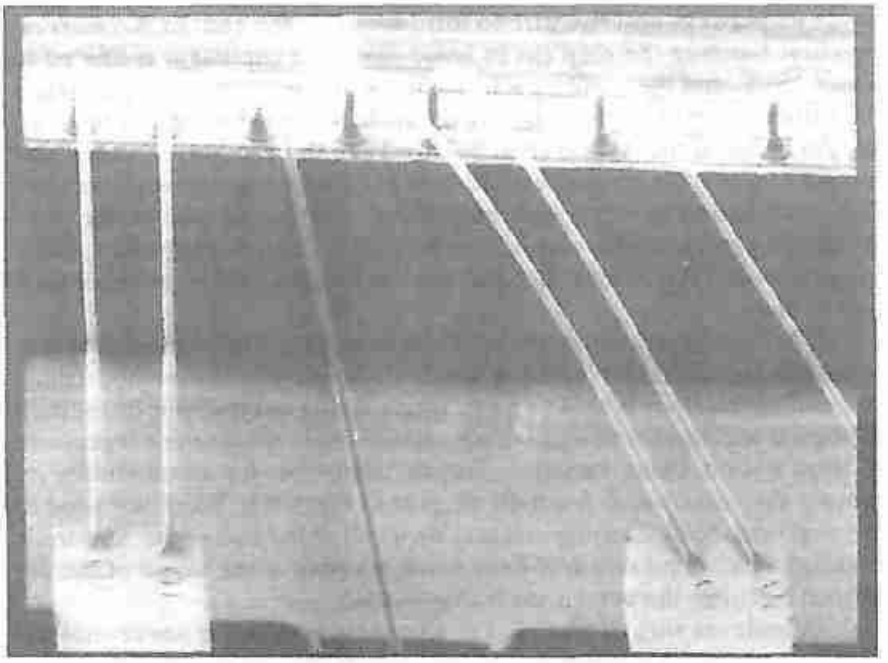


Figure 9–42
Attachment of leads from the Al pads on the periphery of the chip to posts on the package. (© 2004 Micron Technology, Inc. All Rights Reserved. Used with permission.)

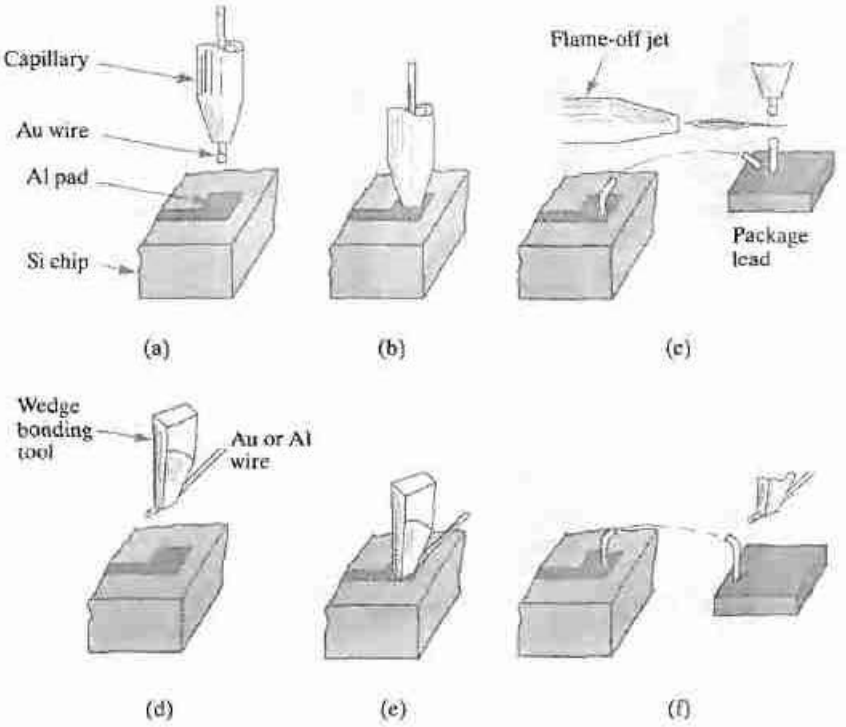


Figure 9–43 Wire bonding techniques: (a) capillary positioned over one of the contact pads for a ball (nail-head) bond; (b) pressure exerted to bond the wire to the pad; (c) post bond and flame-off; (d) wedge bonding tool; (e) pressure and ultrasonic energy applied; (f) post bond completed and wire broken or cut for next bond.

flame jet is swept past the wire to form a ball on the end. In *thermocompression bonding* the chip (or in some cases the capillary) is heated to about 360°C, and the capillary is brought down over the contact pad. When pressure is exerted by the capillary on the ball, a bond is formed between the Au ball and the Al pad (Fig. 9-43b). Then the capillary is raised and moved to a post on the package. The capillary is brought down again, and the combination of force and temperature bonds the wire to the post. After raising the capillary again, the hydrogen flame is swept past, forming a new ball (Fig. 9-43c); then the process is repeated for the other pads on the chip.

There are many variations in this basic method. For example, the substrate heating can be eliminated by *ultrasonic bonding*. In this method a tungsten carbide capillary is held by a tool connected to an ultrasonic transducer. When it is in contact with a pad or a post, the wire is vibrated under pressure to form a bond. Other variations include techniques for automatically removing the "tail," which is left on the post in Fig. 9-43c. When the bond to the chip is made by exerting pressure on a ball at the end of the Au wire, it is called a *ball bond* or a *nail-head bond*, because of the shape of the deformed ball after the bond is made (Fig. 9-44a).

Aluminum wire can be used in ultrasonic bonding; it has several advantages over Au, including the absence of possible metallurgical problems

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 500

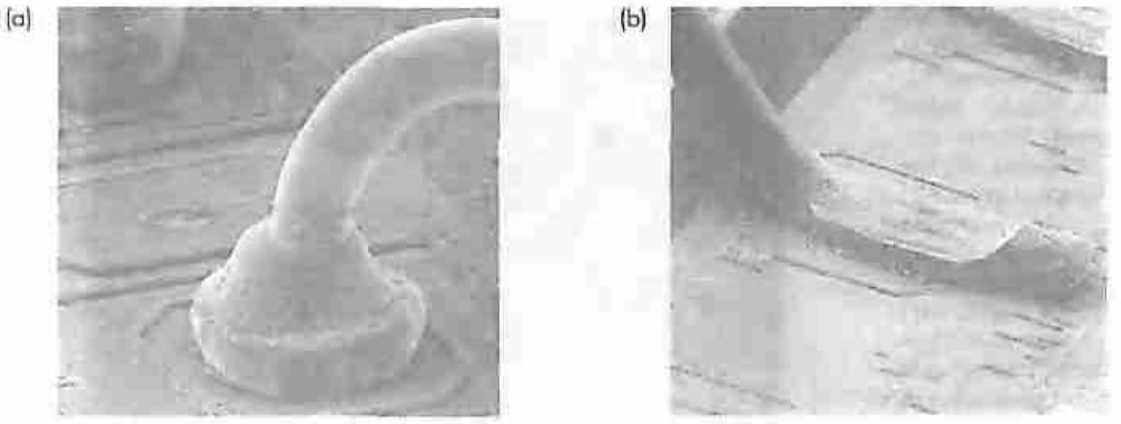


Figure 9–44
Scanning electron micrographs of a ball bond (a) and a wedge bond (b). (© 2004 Micron Technology, Inc. All Rights Reserved. Used with permission.)

in bonds between Au and Al pads. When Al wire is used, the flame-off step is replaced by cutting or breaking the wire at appropriate points in the process. In forming a bond, the wire is bent under the edge of a wedge-shaped bonding tool (Fig. 9-43d). The tool then applies pressure and ultrasonic vibration, forming the bond (Fig. 9-43e and f). The resulting flat bond, formed by the bent wire wedged between the tool and the bonding surface, is called a *wedge bond*. A closeup view of ball and wedge bonds is given in Fig. 9-44.

9.6.3 Flip-Chip Techniques

The time consumed in bonding wires individually to each pad on the chip can be overcome by several methods of simultaneous bonding. The *flip-chip* approach is typical of these methods. Relatively thick metal is deposited on the contact pads before the devices are separated from the wafer. After separation, the deposited metal is used to contact a matching metallized pattern on the package substrate.

In the flip-chip method, "bumps" of solder or special metal alloys are deposited on each contact pad. These metal bumps can be distributed over the die (Fig. 9-45). After separation from the wafer, each chip is turned upside down, and the bumps are properly aligned with the metallization pattern on the substrate. At this point, ultrasonic bonding or solder attaches each bump to its corresponding connector on the substrate. An obvious advantage of this method is that all connections are made simultaneously. Disadvantages include the fact that the bonds are made under the chip and therefore cannot be inspected visually. Furthermore, it is necessary to heat and/or exert pressure on the chip.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 501

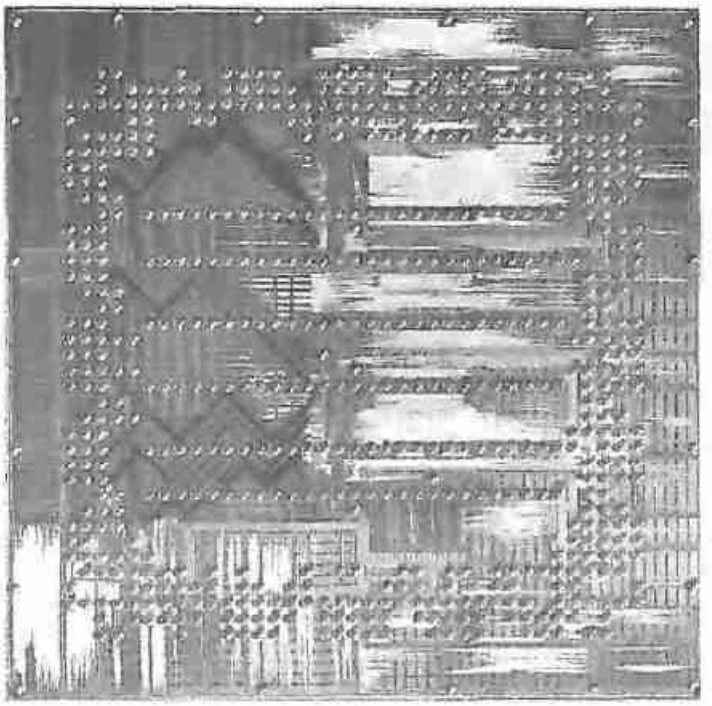


Figure 9–45
Flip-chip bonding. The Power PC chip has metallized "bumps" distributed over the surface instead of contact pads around the periphery. These bumps are aligned with the interconnection pattern on the package and bonded simultaneously. (Photograph courtesy of IBM Corp.)

9.6.4 Packaging

The final step in IC fabrication is packaging the device in a suitable medium that can protect it from the environment of its intended application. In most cases this means the surface of the device must be isolated from moisture and contaminants and the bonds and other elements must be protected from corrosion and mechanical shock. The problems of surface protection are greatly minimized by modern passivation techniques, but it is still necessary to provide some protection in the packaging. In every case, the choice of package type must be made within the requirements of the application and cost considerations. There are many techniques for encapsulating devices, and the various methods are constantly refined and changed. Here we shall consider just a few general methods for the purpose of illustration.

In the early days of IC technology, all devices were packaged in metal headers. In this method the device is alloyed to the surface of the header, wire bonds are made to the header posts, and a metal lid is welded over the device and wiring. Although this method has several drawbacks, it does provide complete sealing of the unit from the outside environment. This is often called a *hermetically sealed* device. After the chip is mounted on the header and bonds are made to the posts, the header cap can be welded shut in a

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 502

controlled environment (e.g., an inert gas), which maintains the device in a prescribed atmosphere.

Integrated circuits are now mounted in packages with many output leads (Fig. 9–46). In one version the chip is mounted on a stamped metal lead frame and wire bonding is done between the chip and the leads. The package

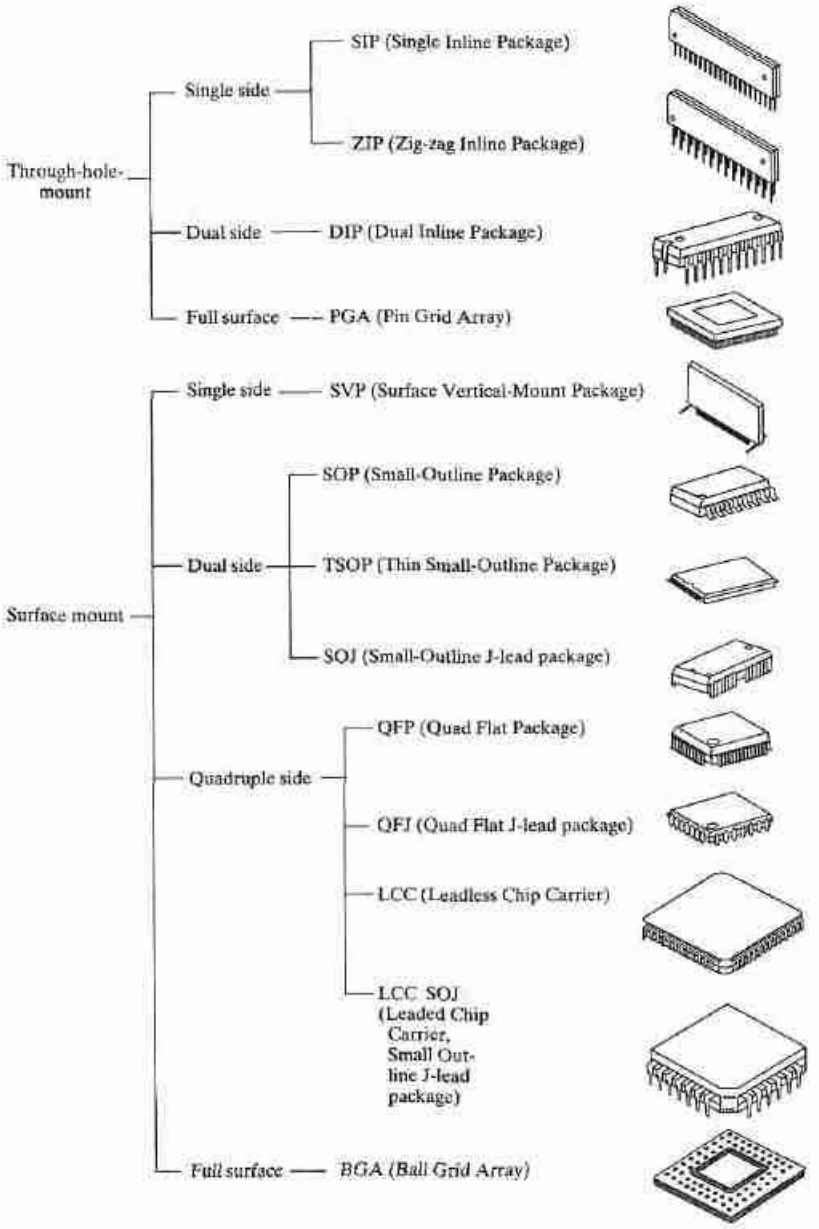


Figure 9–46
Various types of packaging for ICs: Packages can be through-hole-mount or surface-mount type, and be made out of plastic or ceramic. The pins can be on one side (SIP), two sides (DIP), or four sides (quad) of the package, or distributed over the surface of the package (PGAs or BGAs).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 503

is formed by applying a ceramic or plastic case and trimming away the unwanted parts of the lead frame.

Broadly speaking, packages can be through-hole-mount that involve inserting the package pins through holes on the printed circuit board (PCB) before soldering, or surface-mount type where the leads do not pass through holes in the PCB. Instead, surface-mounted package leads are aligned to electrical contacts on the PCB, and are connected simultaneously by solder reflow. Most packages can be made using ceramic or plastic (which is cheaper). The ICs are hermetically sealed for protection from the environment. The pins can be on one side (single inline or zig-zag pattern of leads), two sides (dual inline package or DIP) or four sides of the package (quad package) (Fig. 9-46). More advanced packages have leads distributed over a large portion of the surface of the package as in through-hole-mounted pin grid arrays (PGAs) (Fig. 9-47) or surface-mounted ball grid arrays (BGAs) (Fig. 9-48). By not restricting the leads to the edges of the package, the pin count can be increased dramatically, which is very attractive for advanced ULSI in which a large number of electrical leads must be accessed.

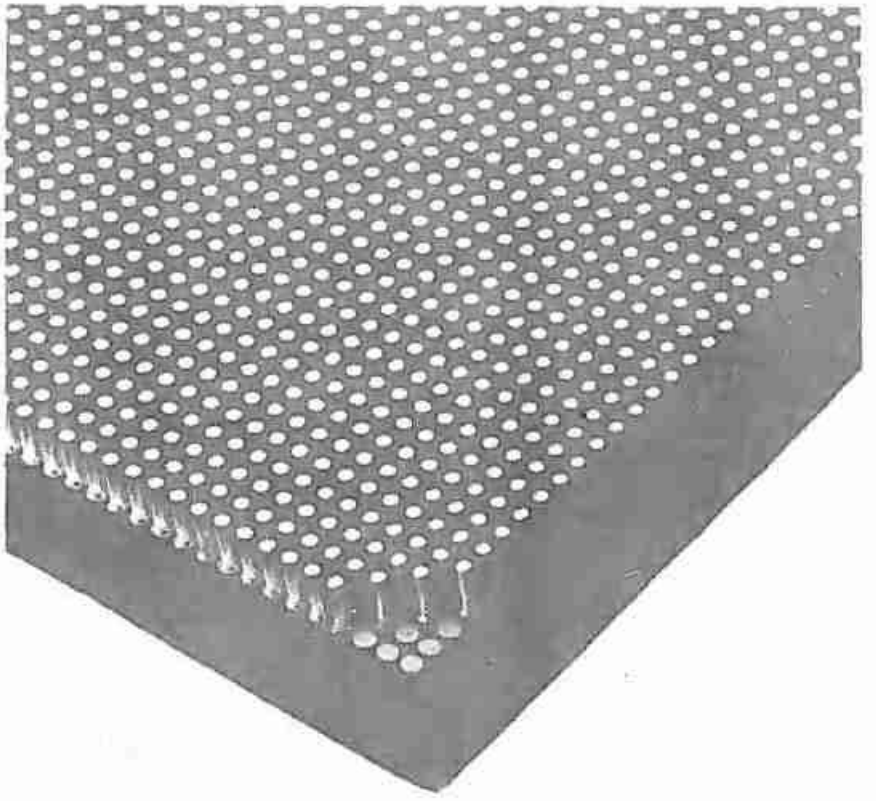


Figure 9-47
Ceramic column grid array (CCGA): This advanced ceramic package is a type of pin grid array made up of several hundred metal columns. Several ICs with metallized solder bumps on them as in Fig. 9-45 can be flip-chip bonded on the back of this package, making this a multi-chip module (MCM). (Photograph courtesy of IBM.)

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 504

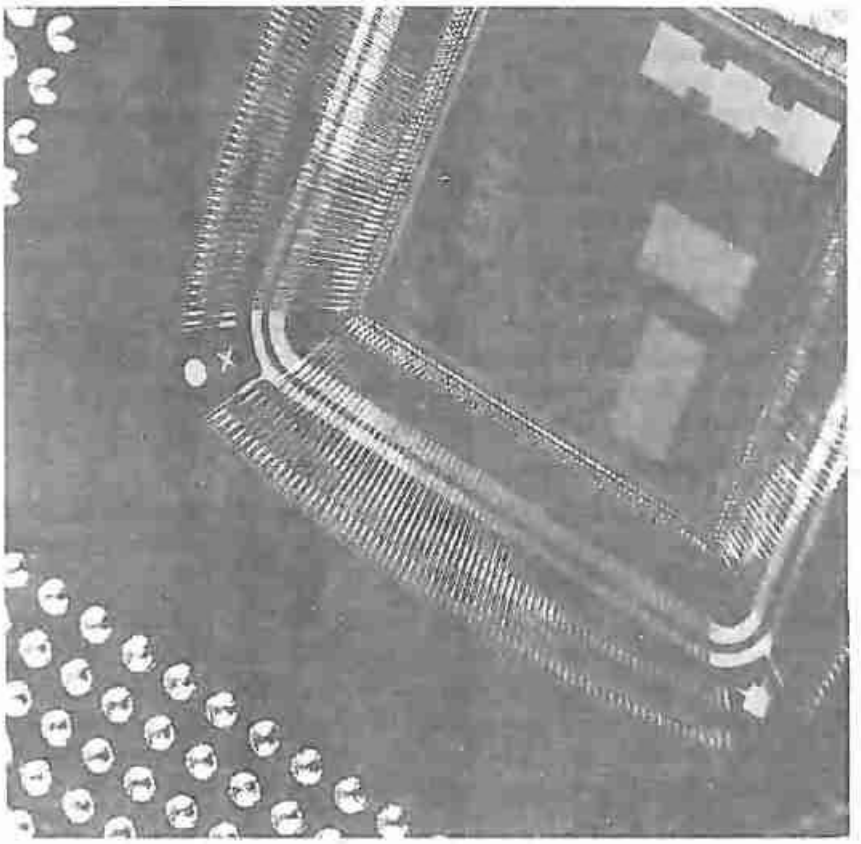


Figure 9–48
Ball grid array. In this package, the IC in the middle is wire bonded to electrical connections on the package. The package itself has an array of solder "balls" on the top, which can be properly aligned and surface-mount connected simultaneously to electrical sockets on a PCB using solder reflow. (Photograph courtesy of IBM.)

Since a sizable fraction of the cost of an IC is due to bonding and packaging, there have been a number of innovations for automating the process. These include the use of film reels that contain the metal contact pattern onto which the chips can be bonded. The film can then be fed into packaging equipment, where the position registration capabilities of a film reel can be used for automated handling. This process, called *tape-automated bonding (TAB)*, is particularly useful in mounting several chips on a large ceramic substrate having multilevel interconnection patterns (called a *multichip module*).

SUMMARY

9.1 The progress of ICs can be characterized by an exponential increase in the transistor count and a corresponding decrease in feature size with time (Moore's law). The economics are driven by device scaling and batch processing.

9.2 The dominant IC technology is digital CMOS, which is the basis of logic circuits (microprocessors), memories (DRAMs, SRAMs, and NVM), and application-specific ICs (ASICs).

PROBLEMS

9.1 Assume that boron is diffused into a uniform n-type Si sample, resulting in a net doping profile $N_a(x)$-N_d. Set up an expression relating the sheet resistance of the diffused layer to the acceptor profile $N_a(x)$ and the junction depth x_j. Assume that $N_a(x)$ is much greater than the background doping N_d over most of the diffused layer.

9.2 A typical sheet resistance of a base diffusion layer is 200 Ω/square.

(a) What should be the aspect ratio of a 10-k Ω resistor, using this diffusion?

(b) Draw a pattern for this resistor (see Fig. 9-12b) which uses little area for a width $w = 5$ μm.

9.3 A 3-μm n-type epitaxial layer ($N_d = 10^{16}$ cm^{-3}) is grown on a p-type Si substrate. Areas of the n layer are to be junction isolated (see Fig. 9-12a) by a boron diffusion at 1200°C ($D = 2.5 \times 10^{-12}$ cm^2/s). The surface boron concentration is held constant at 10^{20} cm^{-3} (see Prob. 5.2).

(a) What time is required for this isolation diffusion?

(b) How far does an Sb-doped buried layer ($D = 2 \times 10^{-13}$ cm^2/s) diffuse into the epitaxial layer during this time, assuming the concentration at the substrate-epitaxial boundary is constant at 10^{20} cm^{-3}?

9.4 A 500-μm-thick p-type Si wafer with a doping level of 1×10^{15} cm^{-3} has a certain region in which we do a constant source solid-solubility-limited P diffusion, resulting in a junction depth of 0.8 μm and a surface concentration of 6×10^{19} cm^{-3}. We do sheet resistance measurements on the two parts of the wafer. What is the measured sheet resistance of the p-type part? If we have a sheet resistance of 90 Ω/square in the n-type part, what is the *average* resistivity there?

READING LIST

Information about the evolution of integrated circuit technology is available at http://public.itrs.net/

Campbell, S. A. *The Science and Engineering of Microelectronic Fabrication*. New York: Oxford, 2001.

Chang, C. Y., and S. M. Sze. *ULSI Technology*. New York: McGraw-Hill, 1996.

Chang, C. Y., and S. M. Sze. *ULSI Devices*. New York: John Wiley, 2000.

Howe, R. T., and C. G. Sodini. *Microelectronics: An Integrated Approach*. Upper Saddle River, NJ: Prentice Hall, 1997.

Jaeger, R. C. *Modular Series on Solid State Devices: Vol. V. Introduction to Microelectronic Fabrication*. Reading, MA: Addison-Wesley, 1988.

Rabaey, J. M. *Digital Integrated Circuits*. Upper Saddle River, NJ: Prentice Hall, 1996.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 506

Seraphim, D. P., R. C. Lasky, and C. Y. Li, eds. *Principles of Electronic Packaging.* New York: McGraw-Hill, 1989.

Sharma, A. K. *Semiconductor Memories.* New York: IEEE Press, 1997.

Tummala, R. R. and E. J. Rymaszewski. *Microelectronics Packaging Handbook.* New York: Chapman and Hall, 1997.

Wolf, S., and R. N. Tauber. *Silicon Processing for the VLSI Era.* Sunset Beach, CA: Lattice Press, 2000.

SELF QUIZ

Question 1

Study the ITRS roadmap chapters on Process, Integration, Devices and Structures (PIDS) and on Front End Processes (FEP) available at http://public.itrs.net/. This has projections about next-generation CMOS devices.

- Plot some of the projected MOS device parameters from the various tables as a function of time. Do they obey Moore's law?
- Based on what you have learned in Chapters 6 and 9, do the required $I_{D(Sat)}$ numbers for NMOSFETs for various technology nodes in Table 47b make sense? How about some of the other MOSFET requirements in other tables that are color-coded red?

Question 2

Discuss consequences (one good, one bad) of quantum mechanical tunneling in MOSFETs.

Question 3

What is hot electron damage, and is it more or less severe than hot hole damage? Why? How can you minimize hot carrier damage?

Question 4

Why are MOSFETs manufactured with {100} planes parallel to the $Si–SiO_2$ interface?

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 507

CHAPTER 10

High-Frequency and High-Power Devices

OBJECTIVES

1. Understand how tunneling, transit time effects, and electron transfer can lead to NDR
2. Understand SCRs
3. Describe how IGFETs switch power

We have discussed a number of devices that are useful in microwave circuits, such as the varactor and specially designed high-frequency transistors, which can provide amplification and other functions at microwave frequencies up to 10^{11} Hz. However, transit time and other effects limit the application of transistors beyond the 10^{11}-Hz range. Therefore, other devices are required to perform electronic functions such as switching and d-c-to-microwave power conversion at higher frequencies.

Several important devices for high-frequency applications use the instabilities that occur in semiconductors. An important type of instability involves *negative conductance*. Here we shall concentrate on three of the most commonly used negative conductance devices: *Esaki* or *tunnel* diodes, which depend on quantum-mechanical tunneling; transit-time diodes, which depend on a combination of carrier injection and transit-time effects; and *Gunn* diodes, which depend on the transfer of electrons from a high-mobility state to a low-mobility state. Each is a two-terminal device that can be operated in a negative conductance mode to provide amplification or oscillation at microwave frequencies in a proper circuit.

10.1 TUNNEL DIODES

The tunnel diode is a p-n junction device that operates in certain regions of its I–V characteristic by the quantum mechanical tunneling of electrons through the potential barrier of the junction. (See Sections 2.4.4 and 5.4.1.) The tunneling process for reverse current is essentially the Zener effect, although negligible reverse bias is needed to initiate the process in tunnel diodes. As we shall see in this section, the tunnel diode (often called the Esaki diode after L. Esaki, who received the Nobel Prize in 1973 for his

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 508

work on the effect) exhibits the important feature of *negative resistance* over a portion of its I–V characteristic.

10.1.1 Degenerate Semiconductors

Thus far, we have discussed the properties of relatively pure semiconductors; any impurity doping represented a small fraction of the total atomic density of the material. Since the few impurity atoms were so widely spaced throughout the sample, we could be confident that no charge transport could take place within the donor or acceptor levels themselves. At high doping, the impurities are so close together that we can no longer consider the donor level as being composed of discrete, noninteracting energy states. Instead, the donor states form a band, which may overlap the bottom of the conduction band. If the conduction-band electron concentration n exceeds the effective density of states N_c, the Fermi level is no longer within the band gap, but lies within the conduction band. When this occurs, the material is called *degenerate* n-type. The analogous case of degenerate p-type material occurs when the acceptor concentration is very high and the Fermi level lies in the valence band. We recall that the energy states below E_F are mostly filled and states above E_F are empty, except for a small distribution dictated by the Fermi statistics. Thus, in a degenerate n-type sample, the region between E_c and E_F is for the most part filled with electrons, and in a degenerate p-type sample, the region between E_v and E_F is almost completely filled with holes.

A p-n junction between two degenerate semiconductors is illustrated in terms of energy bands in Fig. 10–1a. This is the equilibrium condition, for which the Fermi level is constant throughout the junction. We notice that E_{Fp} lies below the valence-band edge on the p side, and E_{Fn} is above the conduction-band edge on the n side. Thus, the bands must overlap on the energy scale in order for E_F to be constant. This overlapping of bands is very important; it means that, with a small forward or reverse bias, filled states and empty states appear opposite each other, separated by essentially the width of the depletion region. If the metallurgical junction is sharp, the depletion region will be very narrow for such high-doping concentrations, and the electric field at the junction will be quite large. Hence, the conditions for electron tunneling are met: filled and empty states separated by a narrow potential barrier of finite height. In Fig. 10–1, the bands are shown filled to the Fermi level for convenience of illustration, with the understanding that a distribution is implied.

Since the bands overlap under equilibrium conditions, a small reverse bias (Fig. 10–1b) allows electron tunneling from the filled valence-band states below E_{Fp} to the empty conduction-band states above E_{Fn}. This condition is similar to the Zener effect, except that no bias is required to create the condition of overlapping bands. As the reverse bias is increased, E_{Fn} continues to move down the energy scale with respect to E_{Fp}, placing more filled states on the p side opposite empty states on the n side. Thus, the tunneling of electrons from p to n increases with increasing reverse bias. The resulting conventional current is opposite to the electron flow—that is, from n to p. At

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 509

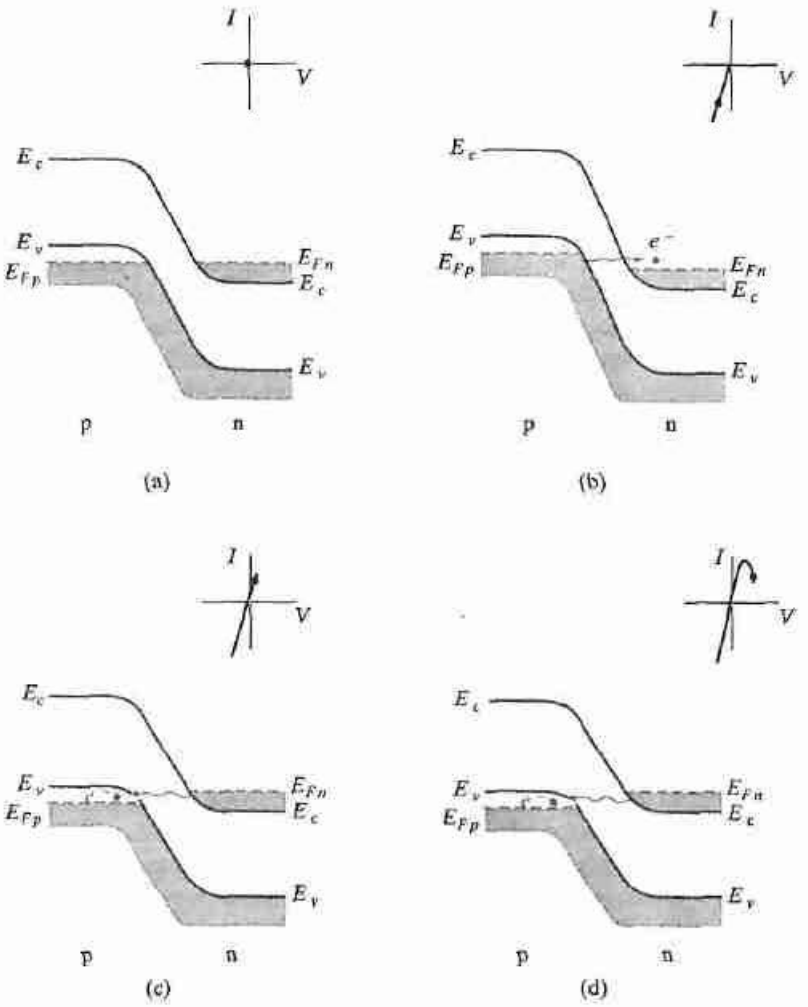


Figure 10–1 Tunnel diode band diagrams and *I–V* characteristics for various biasing conditions: (a) equilibrium (zero bias) condition, no net tunneling; (b) small reverse bias, electron tunneling from p to n; (c) small forward bias, electron tunneling from n to p; (d) increased forward bias, electron tunneling from n to p decreases as bands pass by each other

equilibrium (Fig. 10–1a), there is equal tunneling from n to p and from p to n, given a zero net current.

When a small forward bias is applied (Fig. 10–1c), E_{Fn} moves up in energy with respect to E_{Fp} by the amount qV. Thus, electrons below E_{Fn} on the n side are placed opposite empty states above E_{Fp} on the p side. Electron tunneling occurs from n to p as shown, with the resulting conventional current from p to n. This forward-tunneling current continues to increase with increased bias as more filled states are placed opposite empty states. However, as E_{Fn} continues to move up with respect to E_{Fp}, a point is reached at which the bands begin to pass by each other. When this occurs, the number of filled states opposite empty states decreases. The resulting decrease in tunneling current is illustrated in Fig. 10–1d. This region of the *I–V* characteristic is important in that the *decrease* in tunneling current with *increased* bias produces

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 510

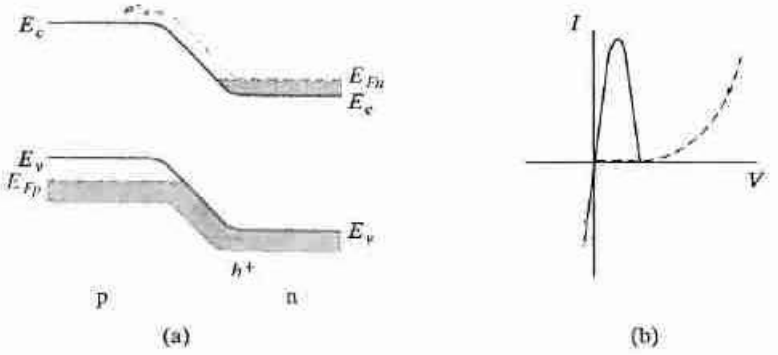


Figure 10–2
Band diagram (a) and *I–V* characteristic (b) for the tunnel diode beyond the tunnel current region. In (b), the tunneling component of current is shown by the solid curve and the diffusion current component is dashed.

a region of negative slope; that is, the *dynamic resistance dV/dI* is negative. This negative-resistance region is useful in oscillators.

If the forward bias is increased beyond the negative-resistance region, the current begins to increase again (Fig. 10–2). Once the bands have passed each other, the characteristic resembles that of a conventional diode. The forward current is now dominated by the diffusion current—electrons surmounting their potential barrier from n to p and holes surmounting their potential barrier from p to n. Of course, the diffusion current is present in the forward tunneling region, but it is negligible compared with the tunneling current.

The total tunnel diode characteristic (Fig. 10–3) has the general shape of an *N* (if a little imagination is applied); therefore, it is common to refer to this characteristic as exhibiting a *type-N negative resistance*. It is also called a *voltage-controlled negative resistance*, meaning that the current decreases rapidly at some critical voltage (in this case, the *peak voltage* V_p, taken at the point of maximum forward tunneling).

The values of *peak tunneling current* I_p and *valley current* I_v (Fig. 10–3) determine the magnitude of the negative-resistance slope for a diode of given material. For this reason, their ratio I_p/I_v is often used as a figure of merit for

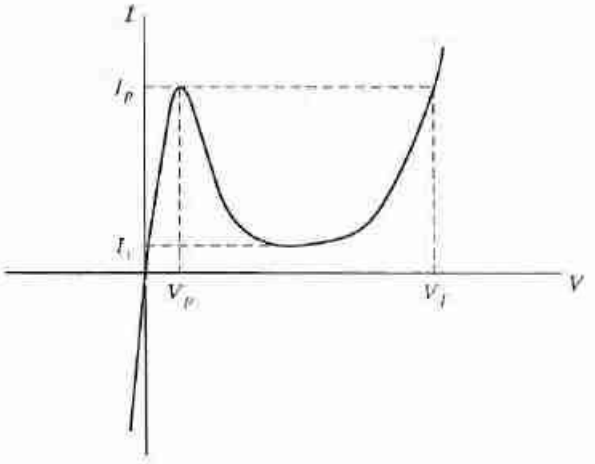


Figure 10–3
Total tunnel diode characteristic.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 511

the tunnel diode. Similarly, the ratio V_p/V_f is a measure of the voltage spread between the two positive-resistance regions.

The negative resistance of the tunnel diode can be used in a number of ways to achieve oscillation and other circuit functions. The fact that the tunneling process does not present the time delays of drift and diffusion makes the tunnel diode a natural choice for certain high-speed circuits. However, the tunnel diode has not achieved widespread application, because of its relatively low current operation and competition from other devices.

10.2 THE IMPATT DIODE

In this section, we describe a type of microwave negative-conductance device that operates by a combination of carrier injection and transit-time effects. Diodes with simple p-n junction structure, or with variations on that structure, are biased to achieve tunneling or avalanche breakdown, with an a-c voltage superimposed on the d-c bias. The carriers generated by the injection process are swept through a drift region to the terminals of the device. We shall see that the a-c component of the resulting current can be approximately 180° out of phase with the applied voltage under proper conditions of bias and device configuration, giving rise to negative conductance and oscillation in a resonant circuit. Transit-time devices can convert d-c to microwave a-c signals with high efficiency and are very useful in the generation of microwave power for many applications.

The original suggestion for a microwave device employing transit-time effects was made by W. T. Read and involved an n^+-p-i-p^+ structure such as that shown in Fig. 10-4. This device operates by injecting carriers into the drift region and is called an *impact avalanche transit-time (IMPATT)* diode. Although IMPATT operation can be obtained in simpler structures, the Read diode is best suited for illustration of the basic principles. The device consists essentially of two regions: (1) the n^+-p region, at which avalanche multiplication occurs, and (2) the i (essentially intrinsic) region, through which generated holes must drift in moving to the p^+ contact. Similar devices can be built in the p^+-n-i-n^+ configuration, in which electrons resulting from

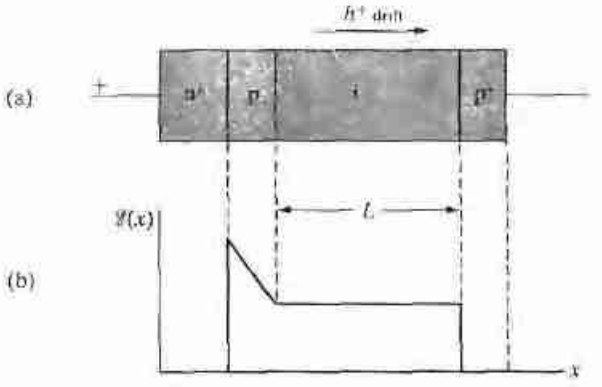


Figure 10-4 The Read diode: (a) basic device configuration; (b) electric field distribution in the device under reverse bias.

avalanche multiplication drift through the i region, taking advantage of the higher mobility of electrons compared with holes.

Although detailed calculations of IMPATT operation are complicated and generally require computer solutions, the basic physical mechanism is simple. Essentially, the device operates in a negative-conductance mode when the a-c component of current is negative over a portion of the cycle during which the a-c voltage is positive, and vice versa. The negative conductance occurs because of two processes, causing the current to lag behind the voltage in time: (1) a delay due to the avalanche process and (2) a further delay due to the transit time of the carriers across the drift region. If the sum of these delay times is approximately one-half cycle of the operating frequency, negative conductance occurs and the device can be used for oscillation and amplification.

From another point of view, the a-c conductance is negative if the a-c component of carrier flow drifts opposite to the influence of the a-c electric field. For example, with a d-c reverse bias on the device of Fig. 10-4, holes drift from left to right (in the direction of the field), as expected. Now, if we superimpose an a-c voltage such that $\mathscr{E}$ decreases during the negative half-cycle, we would normally expect the drift of holes to decrease also. However, in IMPATT operation, the drift of holes through the i region actually increases while the a-c field is decreasing. To see how this happens, let us consider the effects of avalanche and drift for various points in the cycle of applied voltage (Fig. 10-5).

To simplify the discussion, we shall assume that the p region is very narrow and that all the avalanche multiplication takes place in a thin region near the n^+-p junction. We shall approximate the field in the narrow p region by a uniform value. If the d-c bias is such that the critical field for avalanche E_a is just met in the n^+-p space charge region (Fig. 10-5a), avalanche multiplication begins at $t = 0$. Electrons generated in the avalanche move to the n^+ region, and holes enter the i drift region. We assume that the device is mounted in a resonant microwave circuit so that an a-c signal can be maintained at a given frequency. As the applied a-c voltage goes positive, more and more holes are generated in the avalanche region. In fact, the pulse of holes (dotted line) generated by the multiplication process continues to grow as long as the electric field is above E_a (Fig. 10-5b). It can be shown that the particle current due to avalanche increases exponentially with time while the field is above the critical value. The important result of this growth is that the hole pulse reaches its peak value, not at $\pi/2$, when the voltage is maximum, but at π (Fig. 10-5c). Therefore, there is a phase delay of $\pi/2$ inherent in the avalanche process itself. A further delay is provided by the drift region. Once the avalanche multiplication stops ($\omega t > \pi$), the pulse of holes simply drifts toward the p^+ contact (Fig. 10-5d). But during this period, the a-c terminal voltage is negative. Therefore, the dynamic conductance is negative, and energy is supplied to the a-c field.

If the length of the drift region is chosen properly, the pulse of holes is collected at the p^+ contact just as the voltage cycle is completed, and the cycle then repeats itself. The pulse will drift through the length L of the i

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 513

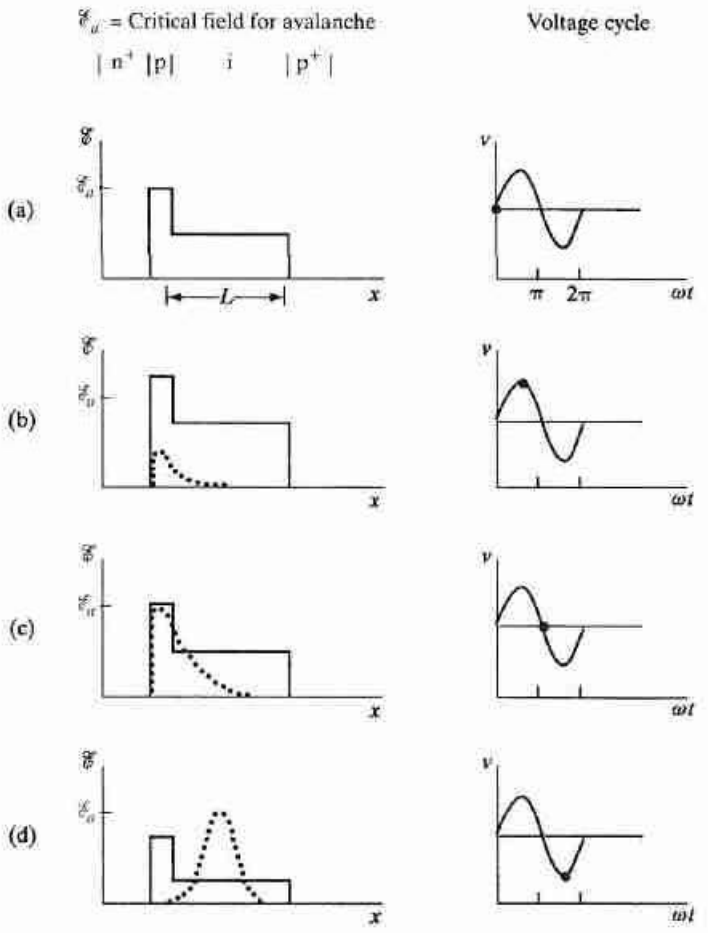


Figure 10–5
Time dependence of the growth and drift of holes during a cycle of applied voltage for the Read diode: (a) $\omega t = 0$; (b) $\omega t = \pi/2$; (c) $\omega t = \pi$; (d) $\omega t = 3\pi/2$. The hole pulse is sketched as a dotted line on the field diagram.

region during the negative half-cycle if we choose the transit time to be one-half the oscillation period; that is,

$$\frac{L}{v_d} = \frac{1}{2}\frac{1}{f}, f = \frac{v_d}{2L} \tag{10–1}$$

where f is the operating frequency and v_d is the drift velocity for holes.[1] Therefore, for a Read diode, the optimum frequency is one-half the inverse transit time v_d/L of holes across the drift region. In choosing an appropriate resonant circuit for this device, the parameter L is critical. For example, taking $v_d = 10^7$ cm/s for Si, the optimum operating frequency for a device with an i region length of 5 μm is $f = 10^7/2(5 \times 10^{-4}) = 10^{10}$ Hz. Negative

[1] In general, v_d is a function of the local electric field. However, these devices are normally operated with fields in the i region sufficiently large that holes drift at their scattering-limited velocity (Fig. 3–24). In this case, the drift velocity does not vary appreciably with the a-c variations in the field.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 514

resistance is exhibited by an IMPATT diode for frequencies somewhat above and below this optimum frequency for exact 180° phase delay. A careful analysis of the small-signal impedance shows that the minimum frequency for negative conductance varies as the square root of the d-c bias current for frequencies in the neighborhood of that described by Eq. (10–1).

Although the Read diode of Fig. 10–4 displays most directly the operation of IMPATT devices, simpler structures can be used, and in some cases they may be more efficient. Negative conductance can be obtained in simple p-n junctions or in p-i-n devices. In the case of the p-i-n, most of the applied voltage occurs across the i region, which serves as a uniform avalanche region and also as a drift region. Therefore, the two processes of delay due to avalanche and drift, which were separate in the case of the Read diode, are distributed within the i region of the p-i-n. This means that both electrons and holes participate in the avalanche and drift processes.

10.3 THE GUNN DIODE

Microwave devices that operate by the *transferred-electron* mechanism are often called *Gunn diodes*, after J. B. Gunn, who first demonstrated one of the forms of oscillation. In the transferred-electron mechanism, the conduction electrons of some semiconductors are shifted from a state of high mobility to a state of low mobility by the influence of a strong electric field. Negative conductance operation can be achieved in a diode[2] for which this mechanism applies, and the results are varied and useful in microwave circuits.

First, we shall describe the process of electron transfer and the resulting change in mobility. Then we shall consider some of the modes of operation for diodes using this mechanism.

10.3.1 The Transferred-Electron Mechanism

In Section 3.4.4, we discussed the nonlinearity of mobility at high electric fields. In most semiconductors, the carriers reach a scattering-limited velocity, and the plot of velocity vs. field saturates at high fields (Fig. 3–24). In some materials, however, the energy of electrons can be raised by an applied field to the point that they transfer from one region of the conduction band to another, higher-energy region. For some band structures, negative conductivity can result from this electron transfer. To visualize the process, let us recall the discussion of energy bands in Section 3.1.

A simplified $E(k)$ band structure for GaAs is shown in Fig. 10–6 for reference; some of the detail has been omitted in this diagram to isolate the essential features of electron transfer between bands. In n-type GaAs, the valence band is filled and the *central valley* (or *minimum*) of the conduction band at $\Gamma(\mathbf{k} = 0)$ normally contains the conduction electrons. There is a set of *subsidiary minima* at L (sometimes called *satellite valleys*) at higher

[2]These devices are called diodes, since they are two-terminal devices. No p-n junction is involved, however. Gunn effect and related devices utilize bulk instabilities, which do not require junctions.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 515

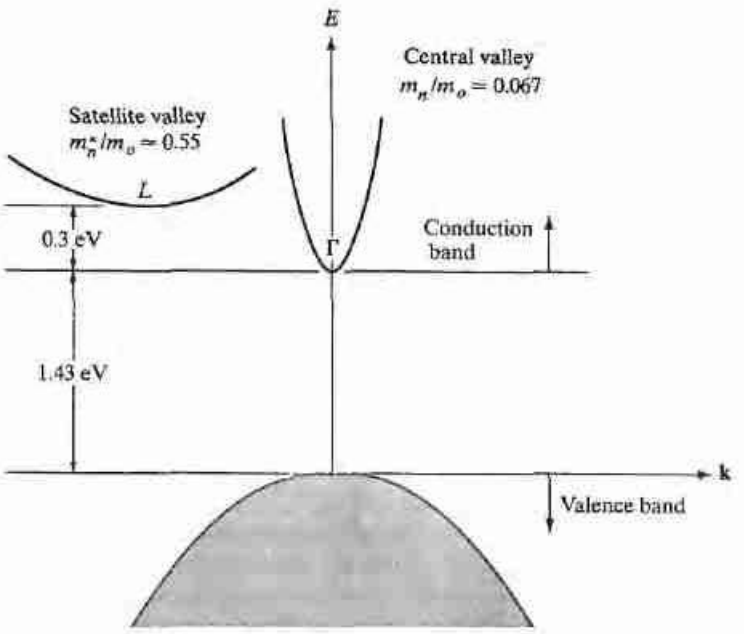


Figure 10–6 Simplified band diagram for GaAs, illustrating the lower (Γ) and upper (*L*) valleys in the conduction band.

energy,[3] but these minima are many kT above the central valley and are normally unoccupied. Therefore, the direct band gap at Γ and the energy bands centered at $\mathbf{k} = 0$ are generally used to describe the conduction processes in GaAs. This was true, for example, of GaAs lasers, discussed in Section 8.4. The presence of the satellite valleys at *L* is crucial to the Gunn effect, however. If the material is subjected to an electric field above some critical value (about 3000 V/cm), the electrons in the central Γ valley of Fig. 10–6 gain more energy than the 0.30 eV separating the valleys; therefore, there is considerable scattering of electrons into the higher-energy satellite valley at L.

Once the electrons have gained enough energy from the field to be transferred into the higher-energy valley, they remain there as long as the field is greater than the critical value. The explanation for this involves the fact that the combined effective density of states for the upper valleys is much greater than for the central valley (by a factor of about 24). Although we shall not prove it here, it seems reasonable that the probability of electron scattering between valleys should depend on the density of states available in each case and that scattering from a valley with many states into a valley with few states would be unlikely. As a result, once the field increases above the critical value, most conduction electrons in GaAs reside in the satellite valleys and exhibit properties typical of that region of the conduction band. In particular, the effective mass for electrons in the higher *L* valleys is almost eight times as great as in the central valley because of the smaller band curvature, and the electron mobility is

[3]We have shown only one satellite valley for convenience; there are other equivalent valleys for different directions in **k**-space. The effective mass ratio of 0.55 refers to the combined satellite valleys.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 516

much lower. This is an important result for the negative conductivity mechanism: As the electric field is increased, the electron velocity increases until a critical field is reached; then the electrons *slow down* with further increase in field. The electron transfer process allows electrons to gain energy at the expense of velocity over a range of values of the electric field. Taking the current density as $qv_d n$, we see clearly that current also drops in this range of increasing field, giving rise to a negative differential conductivity $dJ/d\mathscr{E}$.

A possible dependence of electron velocity vs. electric field for a material capable of electron transfer is shown in Fig. 10–7. For low values of field, the electrons reside in the lower (Γ) valley of the conduction band, and the mobility ($\mu_\Gamma = v_d/E$) is high and constant with field. For high values of field, electrons transfer to the satellite valleys, where their velocity is smaller and their mobility lower. Between these two states is a region of negative slope on the plot of v_d vs. $\mathscr{E}$, indicating a negative differential mobility $dv_d/dE = -\mu^*$.[4]

The actual dependence of electron drift velocity on electric field for GaAs and InP is shown in Fig. 10-8. The negative resistance due to electron transfer occurs at a higher field for InP, and the electrons achieve a higher peak velocity before transfer from Γ to L occurs.

The existence of a drop in mobility with increasing electric field and the resultant possibility of negative conductance were predicted by Ridley and Watkins and by Hilsum several years before Gunn demonstrated the effect in GaAs. The mechanism of electron transfer is therefore often called the Ridley–Watkins–Hilsum mechanism. This negative conductivity effect depends

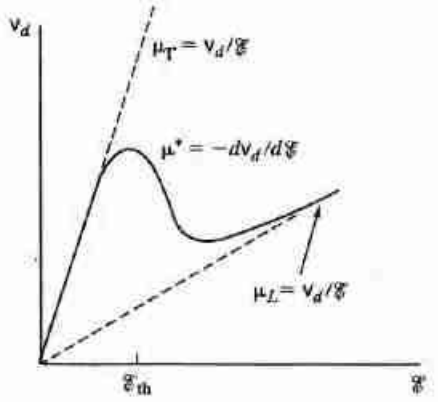


Figure 10–7
A possible characteristic of electron drift velocity vs. field for a semiconductor exhibiting the transferred electron mechanism.
μ_Γ = Mobility in central (Γ) valley
μ_L = Mobility in satellite (L) valley
μ^* = Average magnitude of negative differential mobility during transition

[4]This is a rather crude approximation, since μ^* is not a constant, but varies considerably with field; the negative dielectric relaxation time therefore changes with time as the domain grows.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 517

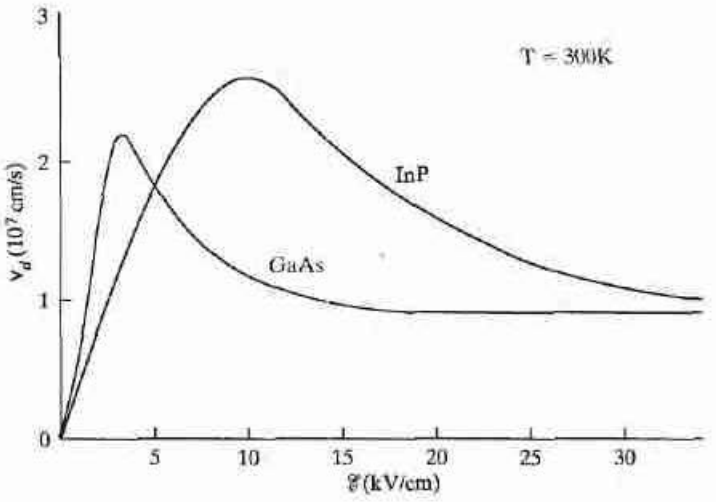


Figure 10–8 Electron drift velocity vs. field for GaAs and InP.

only on the bulk properties of the semiconductor and not on junction or surface effects. It is therefore called a *bulk negative differential conductivity (BNDC)* effect.

10.3.2 Formation and Drift of Space Charge Domains

If a sample of GaAs is biased such that the field falls in the negative conductivity region, space charge instabilities result, and the device cannot be maintained in a d-c stable condition. To understand the formation of these instabilities, let us consider first the dissipation of space charge in the usual semiconductor. It can be shown from treatment of the continuity equation that a localized space charge dies out exponentially with time in a homogeneous sample with positive resistance (Prob. 10.3). If the initial space charge is Q_0, the instantaneous charge is

$$Q(t) = Q_0 e^{-t/\tau_d} \tag{10-2}$$

where $\tau_d = \epsilon/\sigma$ is called the *dielectric relaxation time*. Because of this process, random fluctuations in carrier concentration are quickly neutralized and space charge neutrality is a good approximation for most semiconductors in the usual range of conductivities. For example, the dielectric relaxation time for a 1.0 Ω-cm Si or GaAs sample is approximately 10^{-12} s.

Equation (10-2) gives a rather remarkable result for cases in which the conductivity is negative. For these cases, τ_d is negative also and *space charge fluctuations build up* exponentially in time rather than dying out. This means that normal random fluctuations in the carrier distribution can grow into large space charge regions in the sample. Let us see how this occurs in a GaAs sample biased in the negative conductivity regime. The velocity–field diagram for n-type GaAs is illustrated in Fig. 10-9a. If we assume a small shift of electron concentration in some region of the device, a dipole layer can form as shown

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 518

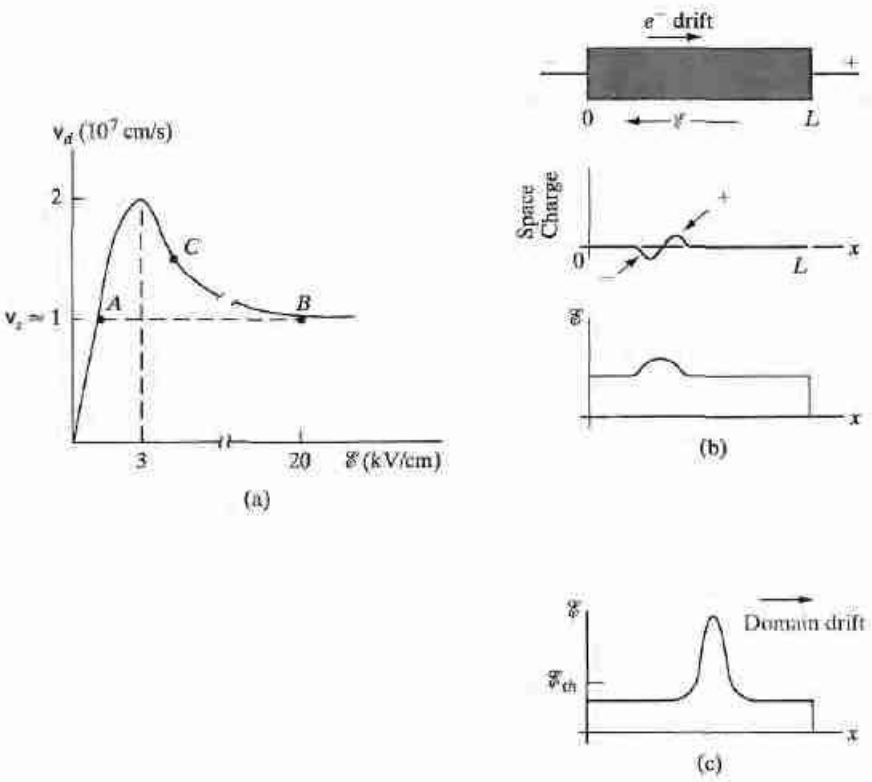


Figure 10–9
Buildup and drift of a space charge domain in GaAs: (a) velocity–field characteristic for n-type GaAs; (b) formation of a dipole; (c) growth and drift of a dipole for conditions of negative conductivity.

in Fig. 10–9b. Under normal conditions, this dipole would die out quickly. However, under conditions of negative conductivity, the charge within the dipole, and therefore the local electric field, builds up as shown in Fig. 10–9c. Of course, this buildup takes place in a stream of electrons drifting from the cathode to the anode, and the dipole (now called a *domain*) drifts along with the stream as it grows. Eventually, the drifting domain will reach the anode, where it gives up its energy as a pulse of current in the external circuit.

During the initial growth of the domain, an increasing fraction of the applied voltage appears across it, at the expense of electric field in the rest of the bar. As a result, it is unlikely that more than one domain will be present in the bar at a time; after the formation of one domain, the electric field in the rest of the bar quickly drops below the threshold value for negative conductivity. If the bias is d-c, the field outside the moving domain will stabilize at a positive conductivity point such as *A* in Fig. 10–9a, and the field inside the domain will stabilize at the high-field value *B*. A small dipole forms from a random noise fluctuation (or, more likely, at a permanent *nucleation site* such as a crystal defect, a doping inhomogeneity, or the cathode itself), and this dipole grows and drifts down the bar as a domain. The formation of stable domains is not the only mode of operation for transferred electron devices. Nor is it the most desirable mode for most applications, since the resulting short pulses of current are inefficient sources

of microwave power. Removal of heat is a very serious problem in these devices. The power dissipation may be 10^7 W/cm^3 or greater (Prob. 10.5), giving rise to considerable heating of the sample. If the application does not require continuous operation, peak powers of hundreds of watts can be achieved in pulses of microwave oscillation.

10.4 THE P-N-P-N DIODE

One of the most common applications of electronic devices is in switching, which requires the device to change from an "off" or *blocking* state to an "on" or *conducting* state. We have discussed the use of transistors in this application, in which base current drives the device from cutoff to saturation. Similarly, diodes and other devices can be used to serve as certain types of switches. There are a number of important switching applications that require a device remain in the blocking state under forward bias until switched to the conducting state by an external signal. Several devices that fulfill this requirement have been developed, including the *semiconductor-controlled rectifier (SCR)*.[5] These devices are typified by a high impedance (the "off" condition) under forward bias until a switching signal is applied; after switching, they exhibit low impedance (the "on" condition). The signal required for switching can be varied externally; therefore, the devices can be used to block or pass currents at predetermined levels.

The SCR is a four-layer (p-n-p-n) structure that effectively blocks current through two terminals until it is turned on by a small signal at a third terminal. We shall begin by investigating the current flow in a two-terminal p-n-p-n device and then extend the discussion to include triggering by a third terminal. We shall see that the p-n-p-n structure can be considered for many purposes as a combination of p-n-p and n-p-n transistors, and the analysis in Chapter 7 can be used as an aid in understanding its behavior.

10.4.1 Basic Structure

First we consider a four-layer diode structure with an *anode* terminal A at the outside p region and with a *cathode* terminal K at the outside n region (Fig. 10-10a). We shall refer to the junction nearest the anode as j_1, the center junction as j_2, and the junction nearest the cathode as j_3. When the anode is biased positively with respect to the cathode (v positive), the device is forward biased. However, as the I–V characteristic of Fig. 10-10b indicates, the forward-biased condition of this diode can be considered in two separate states: the high-impedance, or *forward-blocking*, state and the low-impedance, or *forward-conducting*, state. In the device illustrated here, the forward I–V characteristic switches from the blocking to the conducting state at a critical peak forward voltage V_β.

We can anticipate the discussion of conduction mechanisms that follows by noting that an initial positive voltage v places j_1 and j_3 under forward bias and the center junction j_2 under reverse bias. As v is increased, most of

[5]Since Si is the material commonly used for this device, it is often called a silicon-controlled rectifier.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 520

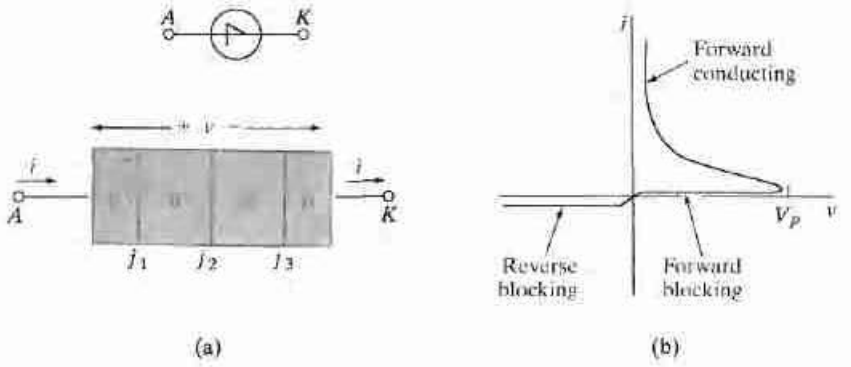


Figure 10–10
A two-terminal p-n-p-n device: (a) basic structure and common circuit symbol; (b) *I–V* characteristic.

the forward voltage in the blocking state must appear across the reverse-biased junction j_2. After switching to the conducting state, the voltage from A to K is very small (less than 1 V), and we conclude that, in this condition, all three junctions must be forward biased. The mechanism by which j_2 switches from reverse bias to forward bias is the subject of much of the discussion that follows.

In the *reverse-blocking* state (v negative), j_1 and j_3 are reverse biased and j_2 is forward biased. Since the supply of electrons and holes to j_2 is restricted by the reverse-biased junctions on either side, the device current is limited to a small saturation current arising from the thermal generation of EHPs near j_1 and j_3. The current remains small in the reverse-blocking condition until avalanche breakdown occurs at a large reverse bias. In a properly designed device, with guards against surface breakdown, the reverse breakdown voltage can be several thousand volts. We shall now consider the mechanism by which this device, often called a *Shockley diode*, switches from the forward-blocking state to the forward-conducting state.

10.4.2 The Two-Transistor Analogy

The four-layer configuration of Fig. 10–10a suggests that the p-n-p-n diode can be considered as two coupled transistors: j_1 and j_2 form the emitter and collector junctions, respectively, of a p-n-p transistor; similarly, j_2 and j_3 form the collector and emitter junctions of an n-p-n transistor. (Note that the emitter of the n-p-n is on the right, which is the reverse of what we usually draw.) In this analogy, the collector region of the n-p-n is in common with the base of the p-n-p, and the base of the n-p-n serves as the collector region of the p-n-p. The center junction j_2 serves as the collector junction for both transistors.

This two-transistor analogy is illustrated in Fig. 10–11. The collector current i_{C1} of the p-n-p transistor drives the base of the n-p-n, and the base current i_{B1} of the p-n-p is dictated by the collector current i_{C2} of the n-p-n. If we associate an emitter-to-collector current transfer ratio α with each transistor, we can use the analysis in Chapter 7 to solve for the current i.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 521

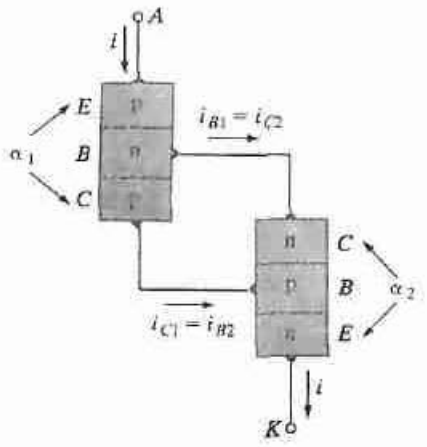


Figure 10–11
Two-transistor analogy of the p-n-p-n diode.

Using Eq. (7-37b) with $\alpha_1 = \alpha_N$ for the p-n-p, $\alpha_2 = \alpha_N$ for the n-p-n, and with I_{CO1} and I_{CO2} for the respective collector saturation currents, we have

$$i_{C1} = \alpha_1 i + I_{C01} = i_{B2} \tag{10-3a}$$
$$i_{C2} = \alpha_2 i + I_{C02} = i_{B1} \tag{10-3b}$$

But the sum of i_{C1} and i_{C2} is the total current through the device:

$$i_{C1} + i_{C2} = i \tag{10-4}$$

Taking this sum in Eq. (10-3) we have

$$i(\alpha_1 + \alpha_2) + I_{C01} + I_{C02} = i$$
$$i = \frac{I_{C01} + I_{C02}}{1 - (\alpha_1 + \alpha_2)} \tag{10-5}$$

As Eq. (10-5) indicates, the current i through the devices is small (approximately the combined collector saturation currents of the two equivalent transistors) as long as the sum $\alpha_1 + \alpha_2$ is small compared with unity. As the sum of the alphas approaches unity, the current i increases rapidly. The current does not increase without limit as Eq. (10-5) implies, however, because the derivation is no longer valid as $\alpha_1 + \alpha_2$ approaches unity. Since j_2 becomes forward biased in the forward-conducting state, both transistors become saturated after switching. The two transistors remain in saturation in the forward-conducting state.

10.4.3 Variation of α with Injection

Since the two-transistor analogy implies that switching involves an increase in the alphas to the point that $\alpha_1 + \alpha_2$ approaches unity, it may be helpful to review how alpha varies with injection for a transistor. The emitter-to-collector

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 522

current transfer ratio α is given in Section 7.2 as the product of the emitter injection efficiency γ and the base transport factor B. An increase in α with injection can be caused by increases in either of these factors or both. At very low currents (such as in the forward-blocking state of p-n-p-n diodes), γ is usually dominated by recombination in the transition region of the emitter junction (Section 7.7.4). As the current is increased, injection across the junction begins to dominate over recombination within the transition region (Section 5.6.2), and γ increases. There are several mechanisms by which the base transport factor B increases with injection, including the saturation of recombination centers as the excess carrier concentration becomes large. Whichever mechanism dominates, the increase in $\alpha_1 + \alpha_2$ required for switching of the p-n-p-n diode is accomplished automatically. In general, no special design is required to maintain $\alpha_1 + \alpha_2$ smaller than unity during the forward-blocking state; this requirement is usually met at low currents by the dominance of recombination within the transition regions of j_1 and j_3.

10.4.4 Forward-Blocking State

When the device is biased in the forward-blocking state (Fig. 10–12a), the applied voltage v appears primarily across the reverse-biased junction j_2. Although j_1 and j_3 are forward biased, the current is small. The reason for this becomes clear if we consider the supply of electrons available to n_1 and holes to p_2. Focusing attention first upon j_1, let us assume that a hole is injected from p_1 into n_1. If the hole recombines with an electron in n_1 (or in the j_1 transition region), that electron must be resupplied to the n_1 region to maintain space charge neutrality. The supply of electrons in this case is severely restricted, however, by the fact that n_1 is terminated in j_2, a reverse-biased

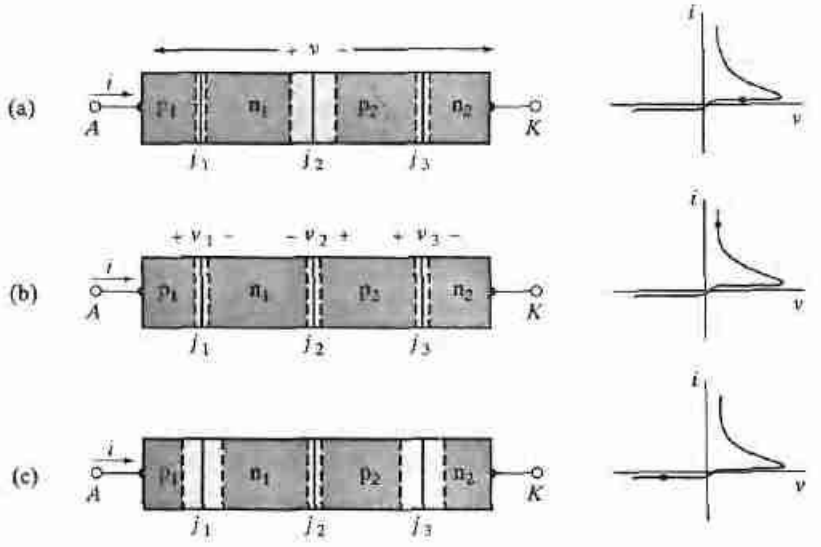


Figure 10–12
Three bias states of the p-n-p-n diode: (a) the forward-blocking state; (b) the forward-conducting state; (c) the reverse-blocking state.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 523

junction. In a normal p-n diode, the n region is terminated in an ohmic contact, so that the supply of electrons required to match recombination (and injection into p) is unlimited. In this case, however, the electron supply is restricted essentially to those electrons generated thermally within a diffusion length of j_2. As a result, the current passing through the j_1 junction is approximately the same as the reverse saturation current of j_2. A similar argument holds for the current through j_3; holes required for injection into n_2 and to feed recombination in p_2 must originate in the saturation current of the center junction j_2. The applied voltage v divides appropriately among the three junctions to accommodate this small current throughout the device.

In this discussion, we have tacitly assumed that the current crossing j_2 is strictly the thermally generated saturation current. This implies that electrons injected by the forward-biased junction j_3 do not diffuse across p_2 in any substantial numbers, to be swept across the reverse-biased junction into n_1 by transistor action. This is another way of saying that α_2 (for the "n-p-n transistor") is small. Similarly, the supply of holes to p_2 is primarily thermally generated, since few holes injected at j_1 reach j_2 without recombination (i.e., α_1 is small for the "p-n-p"). Now we can see physically why Eq. (10–5) implies a small current while $\alpha_1 + \alpha_2$ is small: Without the transport of charge provided by transistor action, the thermal generation of carriers is the only significant source of electrons to n_1 and holes to p_2.

10.4.5 Conducting State

The charge transport mechanism changes dramatically when transistor action begins. As $\alpha_1 + \alpha_2$ approaches unity by one of the mechanisms described, many holes injected at j_1 survive to be swept across j_2 into p_2. This helps to feed the recombination in p_2 and to support the injection of holes into n_2. Similarly, the transistor action of electrons injected at j_3 and collected at j_2 supplies electrons for n_1. Obviously, the current through the device can be much larger once this mechanism begins. The transfer of injected carriers across j_2 is regenerative, in that a greater supply of electrons to n_1 allows greater injection of holes at j_1 while maintaining space charge neutrality; this greater injection of holes further feeds p_2 by transistor action, and the process continues to repeat itself.

If $\alpha_1 + \alpha_2$ is large enough, so that many electrons are collected in n_1 and many holes are collected in p_2, the depletion region at j_2 begins to decrease. Finally, the reverse bias disappears across j_2 and is replaced by a forward bias, in analogy with a transistor that is biased deep in saturation. When this occurs, the three small forward-bias voltages appear as shown in Fig. 10-12b. Two of these voltages essentially cancel in the overall v, so that the forward voltage drop of the device from anode to cathode in the conducting state is not much greater than that of a single p-n junction. For Si, this forward drop is less than 1 V, until ohmic losses become important at high current levels.

We have discussed the current transport mechanisms in the forward-blocking and forward-conducting states, but we have not indicated how

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 524

switching is initiated from one state to the other. Basically, the requirement is that the carrier injection at j_1 and j_2 must somehow be increased so that significant transport of injected carriers across j_2 occurs. Once this transport begins, the regenerative nature of the process takes over and switching is completed.

10.4.6 Triggering Mechanisms

The most common method of triggering a two-terminal p-n-p-n is simply to raise the bias voltage to the peak value V_p. This type of *voltage triggering* results in a breakdown (or significant leakage) of the reverse-biased junction j_2; the accompanying increase in current provides the injection at j_1 and j_3, as well as the transport required for switching to the conducting state. The breakdown mechanism commonly occurs by the combination of *base-width narrowing* and *avalanche multiplication*.

When carrier multiplication occurs in j_2, many electrons are swept into n_1 and holes into p_2. This process provides to these regions the majority carriers needed for increased injection by the emitter junctions. Because of transistor action, the full breakdown voltage of j_2 need not be reached. As we showed in Eq. (7-52), breakdown occurs in the collector junction of a transistor with $i_B = 0$ when $M\alpha = 1$. In the coupled-transistor case of the p-n-p-n diode, breakdown occurs at j_2 when

$$M_p\alpha_1 + M_n\alpha_2 = 1 \tag{10-6}$$

where M_p is the hole multiplication factor and M_n is the multiplication factor for electrons.

As the bias v increases in the forward-blocking state, the depletion region about j_2 spreads to accommodate the increased reverse bias on the center junction. This spreading means that the neutral base regions on either side (n_1 and p_2) become thinner. Since α_1 and α_2 increase as these base widths decrease, triggering can occur by the effect of base-width narrowing. A true punch-through of the base regions is seldom required, since moderate narrowing of these regions can increase the alphas enough to cause switching. Furthermore, switching may be the result of a combination of avalanche multiplication and base-width narrowing, along with possible leakage current through j_2 at high voltage. From Eq. (10-6), it is clear that, with avalanche multiplication present, the sum $\alpha_1 + \alpha_2$ need not approach unity to initiate breakdown of j_2. Once breakdown begins, the increased numbers of carriers in n_1 and p_2 drive the device to the forward-conducting state by the regenerative process of coupled transistor action. As switching proceeds, the reverse bias is lost across j_2 and the junction breakdown mechanisms are no longer active. Therefore, base narrowing and avalanche multiplication serve only to start the switching process.

If a forward-bias voltage is applied rapidly to the device, switching can occur by a mechanism commonly called *dv/dt triggering*. Basically, this type of triggering occurs as the depletion region of j_2 adjusts to accommodate the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 525

increasing voltage. As the depletion width of j_2 increases, electrons are removed from the n_1 side, and holes are removed from the p_2 side, of the junction. For a slow increase in voltage, the resulting flow of electrons toward j_1 and holes toward j_3 does not constitute a significant current. If dv/dt is large, however, the rate of charge removal from each side of j_2 can cause the current to increase significantly. In terms of the junction capacitance (C_{j2}) of the reverse-biased junction, the transient current is given by

$$i(t) = \frac{dC_{j2}v_{j2}}{dt} = C_{j2}\frac{dv_{j2}}{dt} + v_{j2}\frac{dC_{j2}}{dt} \tag{10-7}$$

where v_{j2} is the instantaneous voltage across j_2. This type of current flow is often called *displacement current*. The rate of change of C_{j2} must be included in calculating current, since the capacitance varies with time as the depletion width changes.

The increase in current due to a rapid rise in voltage can cause switching well below the steady state triggering voltage V_P. Therefore, a dv/dt rating is usually specified along with V_P for p-n-p-n diodes. Obviously, dv/dt triggering can be a disadvantage in circuits subjected to unpredictable voltage transients.

The various triggering mechanisms discussed in this section apply to the two-terminal p-n-p-n diode. As we shall see in the next section, the semiconductor-controlled rectifier is triggered by an external signal applied to a third terminal.

10.5 THE SEMICONDUCTOR-CONTROLLED RECTIFIER

The semiconductor-controlled rectifier (SCR) is useful in many applications, such as power switching and various control circuits. This device can handle currents from a few milliamperes to hundreds of amperes. Since it can be turned on externally, the SCR can be used to regulate the amount of power delivered to a load simply by passing current only during selected portions of the line cycle. A common example of this application is the light-dimmer switch used in many homes. At a given setting of this switch, an SCR is turned on and off repetitively, such that all or only part of each power cycle is delivered to the lights. As a result, the light intensity can be varied continuously from full to dark. The same control principle can be applied to motors, heaters, and many other systems. We shall discuss this type of application next, after first establishing the fundamentals of device operation.

The most important four-layer device in power circuit applications is the three-terminal SCR[6] (Fig. 10–13). This device is similar to the p-n-p-n diode, except that a third lead (*gate*) is attached to one of the base regions. When the SCR is biased in the forward-blocking state, a small current supplied to

[6]This device is often called a *thyristor* to indicate its function as a solid state analogue of the gas thyratron, a gas-filled tube that passes current when an arc discharge occurs at a critical firing voltage. In a manner analogous to the gate current control of the SCR, this firing voltage can be varied by a voltage applied to a third electrode.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 526

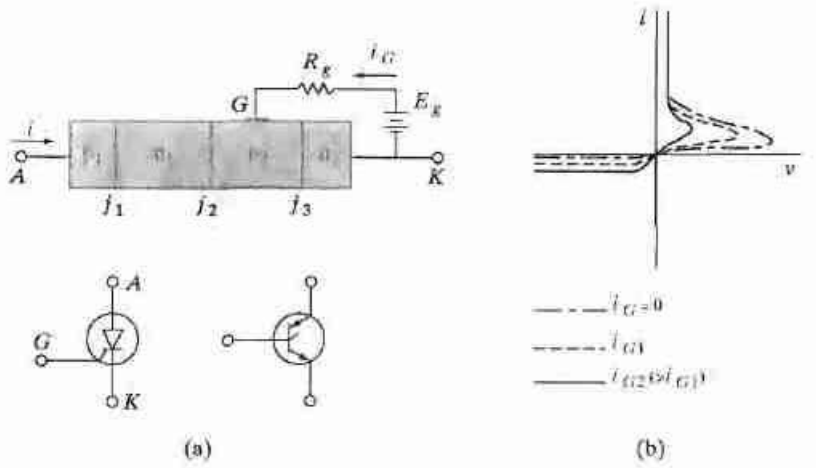


Figure 10–13
A semiconductor-controlled rectifier: (a) four-layer geometry and common circuit symbols; (b) *I–V* characteristics.

the gate can initiate switching to the conducting state. As a result, the anode switching voltage V_P decreases as the current i_G applied to the gate is increased (Fig. 10–13b). This type of turn-on control makes the SCR a useful and versatile device in switching and control circuits.

To visualize the *gate-triggering* mechanism, let us assume that the device is in the forward-blocking state, with a small saturation current flowing from anode to cathode. A positive gate current causes holes to flow from the gate into p_2, the base of the n-p-n transistor. This added supply of holes and the accompanying injection of electrons from n_2 into p_2 initiates transistor action in the n-p-n. After a transit time τ_{t2}, the electrons injected by j_3 arrive at the center junction and are swept into n_1, the base of the p-n-p. This causes an increase in hole injection of j_1, and these holes diffuse across the base n_1 in a transit time τ_{t1}. Thus, after a delay time of approximately $\tau_{t1} + \tau_{t2}$, transistor action is established across the entire p-n-p-n, and the device is driven into the forward-conducting state. In most SCRs, the delay time is less than a few microseconds and the gate current required for turn-on is only a few milliamperes. Therefore, the SCR can be turned on by a very small amount of power in the gate circuit. On the other hand, the device current i can be many amperes, and the power controlled by the device may be very large.

It is not necessary to maintain the gate current once the SCR switches to the conducting state; in fact, the gate essentially loses control of the device after regenerative transistor action is initiated. For most devices, a gate current pulse lasting a few microseconds is sufficient to ensure switching. Ratings of minimum gate pulse height and duration are generally provided for particular SCR devices.

10.5.1 Turning Off the SCR

Turning off the SCR—changing it from the conducting state to the blocking state—can be accomplished by reducing the current i below a critical value

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 527

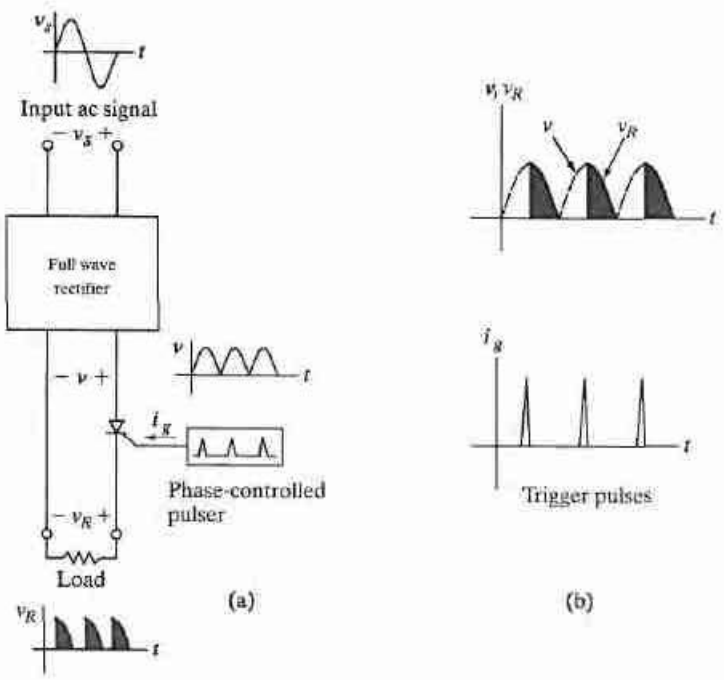


Figure 10–14 Example of the use of an SCR to control the power delivered to a load: (a) schematic diagram of the circuit; (b) waveforms of the delivered signal and the phase-variable trigger pulse.

(called the *holding current*) required to maintain the $\alpha_1 + \alpha_2 = 1$ condition. In some SCR devices, gate turnoff can be used to reduce the alpha sum below unity. For example, if the gate voltage is reversed in Fig. 10–14, holes are extracted from the p_2 base region. If the rate of hole extraction by the gate is sufficient to remove the n-p-n transistor from saturation, the device turns off. However, there are often problems involving the lateral flow of current in p_2 to the gate; nonuniform biasing of j_3 can result from the fact that the bias on this emitter junction varies with position when a lateral current flows. Therefore, SCR devices must be specifically designed for turn-off control; at best, this turn-off capability can be utilized only over a limited range for a given device.

10.6 INSULATED-GATE BIPOLAR TRANSISTOR

We saw in Section 10.5 that the SCR has difficulty in efficiently turning off the device using the gate. We need to use additional circuitry to reduce the anode-to-cathode current below the holding current to change the SCR from the conducting state to the blocking state. This is, of course, clumsy and expensive.

Hence, the *insulated-gate bipolar transistor (IGBT)* was invented by Baliga in 1979 to address this issue. The aforesaid variation on the SCR can easily be turned off from the conducting to the blocking state by the action of the gate. This device is also known by several other names, such as conductivity-modulated FET (COMFET), insulated-gate transistor (IGT), insulated-gate rectifier (IGR), gain-enhanced MOSFET (GEMFET), and bipolar FET (BiFET).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 528

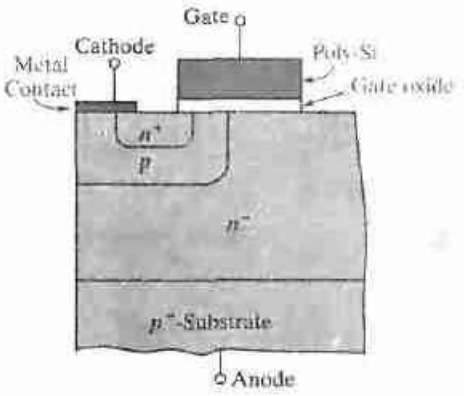


Figure 10–15
Structure of an insulated-gate bipolar transistor.

The basic structure is shown for an n-channel device in Fig. 10–15. It combines an SCR with a MOSFET that is able to connect or disconnect the n^+ cathode to the n^- base region, depending on the gate bias of the MOSFET. The MOSFET channel length is determined by the p region, which is formed by diffusion of the acceptors implanted in the same region as the n^+ cathode. In other words, the channel length is determined not by the lithography of the gate, as in a conventional MOSFET but rather by the diffusion of the acceptors. Such a MOSFET structure is known as a *double-diffused MOSFET* (DMOS) The DMOS device is essentially an NMOSFET.

The main part of the IGBT is the n^- region, which acts as the drain of the DMOS device. This is generally a thick (~50 μm) epitaxial region with a low doping (~10^{14} cm^{-3}) grown on a heavily p^+-doped substrate that forms the anode. The n^- region can therefore support a large blocking voltage in the "off" state. In the "on" state, the conductivity of this lightly doped region is modulated (increased) by the electrons injected from the n^+ cathode and the holes injected from the p^+ anode—hence the alternative name *conductivity-modulated FET* (COMFET). The increased conductivity allows the voltage drop across the device to be minimal in the "on" state.

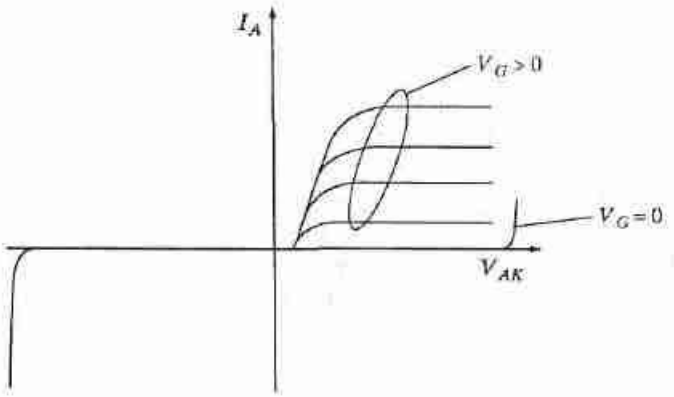


Figure 10–16
Output current–voltage characteristics of an insulated-gate bipolar transistor (n channel).

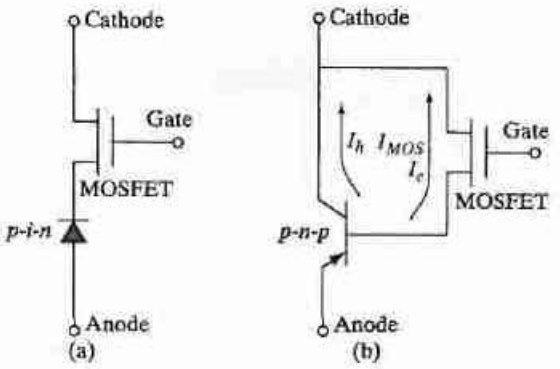


Figure 10–17
IGBT equivalent circuit: (a) below the offset voltage, for low V_{AK}; (b) above the offset voltage, for high V_{AK}.

The current–voltage characteristics are shown in Fig. 10–16. If the DMOS gate voltage is zero (or below the threshold voltage), an n-type inversion region is not formed in the p-type channel region and the n^+ cathode is not shorted to the n^- base. The structure then looks exactly like a conventional SCR that allows minimum current flow in either polarity until breakdown is reached. For positive anode-to-cathode bias V_{AK}, avalanche breakdown occurs at the n^--p junction, while for negative V_{AK} avalanche occurs at the n^--p^+ junction.

When a gate bias is applied to the DMOS gate, we see that, for positive V_{AK}, there is significant current flow (Fig. 10–17). The characteristics look like that of a MOSFET, with one difference: Instead of the current starting to increase from the origin, there is an offset or cut-in voltage of ~0.7 V, as with a diode. The reason for this can be understood by looking at the equivalent circuit in Fig. 10–17a. For small V_{AK} up to the offset voltage, the structure looks like a DMOS in series with a p-i-n diode made up of the p^+ substrate (the anode), the n^- blocking region (the base), which is essentially like an intrinsic region, and the n^+ cathode. In this regime, there is negligible voltage drop across the DMOS device, and the p-i-n device is in forward bias. The injected carriers from the anode and the cathode recombine in the n^- region. As we saw in Chapter 5, for a diode dominated by recombination in the depletion region, the current–voltage characteristics show an exponential behavior, with a diode ideality factor of $\mathbf{n} = 2$. Therefore, in this region, we get

$$I_A \propto \exp(qV_{AK}/2kT) \tag{10-8}$$

On the other hand, when V_{AK} is larger than the offset voltage (~0.7 V), the characteristics look like that of a MOSFET, multiplied by a p-n-p bipolar junction transistor gain term. The equivalent circuit in this region is shown in Fig. 10–17b. In this regime, not all the injected carriers recombine in the near-intrinsic n^- region. The current they supply, which is essentially the DMOSFET current I_{MOS}, acts as the base current of the vertical p-n-p BJT formed between the p^+ substrate (anode), the n^- base, and the p^- channel of the DMOS device. Hence, the current is now given by

$$I_A = (1 + \beta_{pnp})I_{MOS} \tag{10-9}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 530

The shape of the characteristics looks like that of the DMOS device. This is the preferred mode of operation of the IGBT.

Finally, if the current levels are too high, the IGBT latches into a low impedance state like that of a conventional SCR in the "on" state. This is undesirable, because it means that the gate of the DMOS device has now lost control.

The IGBT clearly incorporates some of the best features of MOSFETs and BJTs. Like a MOSFET, it has high input impedance and low input capacitance. On the other hand, in the "on" state, it has low resistance and high current-handling capability, like a BJT or an SCR. Because of these factors, and because it can turn off more easily than an SCR, the IGBT is gradually becoming the power device of choice, in place of the more traditional SCR.

SUMMARY

10.1 High frequencies in the microwave regime can be obtained by negative differential resistance achievable by tunneling (Esaki diodes), transferred electron effects (Gunn diodes) or transit-time effects (IMPATT diodes).

10.2 High power switching is possible with p-n-p-n diodes, thyristors, or SCRs. These devices can be understood in terms of two coupled BJTs. Similar effects in CMOS cause latchup.

10.3 Modern power devices are based on the IGBT, which is a combination of an SCR and a MOSFET.

PROBLEMS

10.1 Sketch the band diagram for an abrupt junction in which the doping on the p side is degenerate and the Fermi level on the n side is aligned with the bottom of the conduction band. Draw the forward- and reverse-bias band diagrams and sketch the *I–V* characteristic. This diode is often called a *backward diode. Can you explain why?*

10.2 What determines the peak tunneling voltage V_p of a tunnel diode? Explain.

If a large density of trapping centers is present in a tunnel diode (Fig. P10–2), tunneling can occur from the n-side conduction band to the trapping level (A–B).

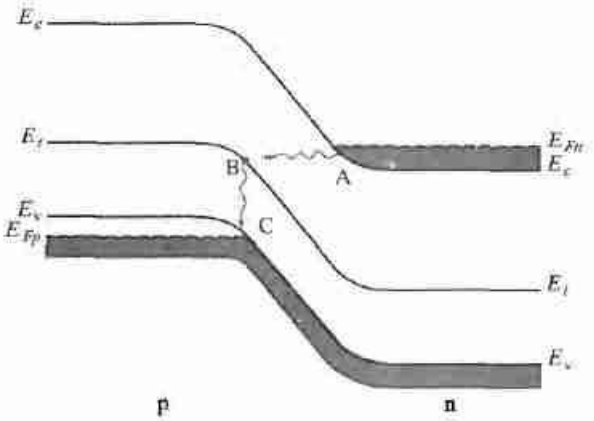


Figure P10–2

Then the electrons may drop to the valence band on the p side (B–C), thereby completing a two-step process of charge transport across the junction. In fact, if the density of trapping centers is large, it is possible to observe an increase in current as the states below E_{Fn} pass by the trapping level with increased bias. In Fig. P10–2, the trapping level E_t is located 0.3 eV above the valence band. Assume that $E_g = 1$ eV, $E_{Fn} - E_c$ on the n side equals $E_v - E_{Fp}$ on the p side, and both $E_{Fn} - E_c$ and $E_v - E_{Fp}$ equal 0.1 eV.

(a) Calculate the minimum forward bias at which tunneling through E_t occurs.

(b) Calculate the maximum forward bias for tunneling via E_t.

(c) Sketch the I–V curve for this tunnel diode. Assume that the maximum tunneling current via E_t is about one-third of the peak band-to-band tunneling current.

10.3 (a) Use Poisson's equation, the continuity equation, and the definition of current density in terms of the gradient of electrostatic potential to relate the variation of space charge density ρ with time to the conductivity σ and the permittivity ϵ of a material, neglecting recombination.

(b) Assuming a space charge density ρ_0 at $t = 0$, show that $\rho(t)$ decays exponentially with a time constant equal to the dielectric relaxation time τ_d.

(c) Given a sample of thickness L and area A, calculate the inherent RC time constant if the conductivity is σ and the permittivity is ϵ.

10.4 Assuming that n_Γ electrons/cm^3 are in the lower (central) valley of the GaAs conduction band at time t and n_L are in the satellite (L) valleys, show that the criterion for negative differential conductivity ($dJ/dE < 0$) is

$$\frac{\mathscr{E}(\mu_\Gamma - \mu_L)\dfrac{dn_\Gamma}{dE} + \mathscr{E}\left(n_\Gamma \dfrac{d\mu_\Gamma}{d\mathscr{E}} + n_L \dfrac{d\mu_L}{d\mathscr{E}}\right)}{n_\Gamma \mu_\Gamma + n_L \mu_L} < -1$$

where μ_Γ and μ_L are the electron mobilities in the Γ and L valleys, respectively. (*Note:* $n_0 = n_\Gamma + n_L$.) Discuss the conditions for negative differential conductivity, assuming that the mobilities are approximately proportional to $\mathscr{E}^{-1}$.

10.5 We wish to estimate the d-c power dissipated in a GaAs Gunn diode. Assume that the diode is 5 μm long and operates in the stable domain mode.

(a) What is the minimum electron concentration n_0? What is the time between current pulses?

(b) Using data from Fig. 10–9a, calculate the power dissipated in the sample per unit volume when it is biased just below threshold if n_0 is chosen from the calculation of part (a). In general, does operation at a higher frequency result in greater power dissipation?

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 532

10.6 (a) Calculate the ratio N_L/N_Γ of the effective density of states in the upper (L) valleys to the effective density of states in the lower (Γ) valley of the GaAs conduction band (Fig. 10–6).

(b) Assuming a Boltzmann distribution $n_L/n_\Gamma = (N_L/N_\Gamma)\exp(-\Delta E/kT)$, calculate the ratio of the concentration of conduction-band electrons in the upper valley to the concentration in the central valley in equilibrium at 300 K.

(c) As a rough calculation, assume that an electron at the bottom of the central valley has kinetic energy kT. After it is promoted to the satellite (I.) valley, what is its approximate equivalent temperature?

10.7 Explain why two separate transistors cannot be connected as in Fig. 10–11 to achieve the p-n-p-n switching action of Fig. 10–10.

10.8 In the p-n-p-n diode (Fig. 10–12a), the junction j_3 is forward biased during the forward-blocking state. Why, then, does the forward bias provided by the gate-to-cathode voltage in Fig. 10–13 cause switching?

10.9 (a) Sketch the energy-band diagrams for the p-n-p-n diode in equilibrium, in the forward-blocking state, and in the forward-conducting state.

(b) Sketch the excess minority carrier distributions in regions n_1 and p_2 when the p-n-p-n diode is in the forward-conducting state.

10.10 Use schematic techniques such as those illustrated in Fig. 7–3 to describe the hole flow and electron flow in a p-n-p-n diode for the forward-blocking state and for the forward-conducting state. Explain the diagrams and be careful to define any new symbols (e.g., those representing EHP generation and recombination).

10.11 Using the coupled transistor model, rewrite Eqs. (10–3) to include avalanche multiplication in j_2, and show that Eq. (10–6) is valid for the p-n-p-n diode.

READING LIST

Baliga, B. J. *Power Semiconductor Devices.* Boston: PWS, 1996.

Esaki, L. "Discovery of the Tunnel Diode." *IEEE Trans. Elec. Dev.*, ED-23 (1976).

Gentry, F. E., F. W. Gutzwiller, N. Holonyak, Jr., and E. E. Von Zastrow. *Semiconductor Controlled Rectifiers: Principles and Application of p-n-p-n Devices.* Englewood Cliffs, NJ: Prentice-Hall, 1964.

Gunn, J. B. "Microwave Oscillations of Current in III-V Semiconductors." *Solid State Comm.*, 1 (1963).

Moll, J. L., M. Tanenbaum, J. M. Goldey, and N. Holonyak. "P-N-P-N Transistor Switches." *Proc. IRE*, 44 (1956).

Read, W. T. "A Proposed High Frequency, Negative Resistance Diode." *Bell Syst. Tech. J.*, 37 (1958).

Ridley, B. K., and T. B. Watkins. "The Possibility of Negative Resistance Effects in Semiconductors." *Proc. Phys. Soc. Lond.* 78 (1961).

Shockley, W. "Negative Resistance Arising From Transit Time in Semiconductor Diodes." *Bell Syst. Tech. J.*, 33 (1954).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 533

SELF QUIZ

Question 1

Study the ITRS roadmap chapter on Emerging Research Devices available at http://public.itrs.net/. This has projections about next-generation CMOS devices as well as novel devices using nanotechnology.

- Write a report on which of these devices you think will be used in products in the next 5 years, 10 years, and 20 years.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 534

Appendix I

Definitions of Commonly Used Symbols[1]

a	Chapter 1: unit cell dimension (Å); Chapter 6: metallurgical channel half-width for an FET (cm)
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	basis vectors
A	area (cm^2)
$\mathscr{B}$	magnetic flux density (Wb/cm^2)
B	base transport factor for a BJT
B, E, C	base, emitter, collector of a BJT
c	speed of light (cm/s)
C	capacitance/area in MOS (F/cm^2)
C_i, C_d, C_{it}	insulator, depletion, interface-state MOS capacitance/area (F/cm^2)
C_j	junction capacitance (F)
C_s	charge storage capacitance (F)
D, D_n, D_p	diffusion coefficient for dopants, electrons, holes (cm^2/s)
D, G, S	drain, gate, source of an FET
e	*Napierian base*
e^-	electron
$\mathscr{E}$	electric field strength (V/cm)
E	energy[2] (J, eV); battery voltage (V)
E_a, E_d	acceptor, donor energy level (J, eV)
E_c, E_v	conduction band, valence band edge (J, eV)
E_F	equilibrium Fermi level (J, eV)
E_g	band gap energy (J, eV)
E_i	intrinsic level (J, eV)
E_r, E_t	recombination, trapping energy level (J, eV)
$f(E)$	Fermi–Dirac distribution function
F_n, F_p	quasi-Fermi level for electrons, holes (J, eV)
g, g_{op}	EHP generation rate, optical generation rate (cm^{-3}-s^{-1})

[1]This list does not include some symbols that are used only in the section where they are defined. Units are given in common semiconductor usage, involving cm where appropriate; it is important to note, however, that calculations should be made in the MKS system in some formulas.

[2]In the Boltzmann factor $\exp\{-\Delta E/kT\}$, ΔE can be expressed in J or eV if k is expressed in J/K or eV/K, respectively.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 535

g_m	mutual transconductance (Ω^{-1}, S)
h	Planck's constant (J-s, eV-s); Chapter 6: FET channel half-width (cm)
$\hbar$	Planck's constant divided by 2π (J-s, eV-s)
$h\nu$	photon energy (J, eV)
h, k, l	Miller indices
h^+	hole
i, I	current[3] (A)
I (subscript)	inverted mode of a BJT
i_B, i_C, i_E	base, collector, emitter current in a BJT (A)
I_{CO}, I_{EO}	magnitude of the collector, emitter saturation current with the emitter, collector open (A)
I_{CS}, I_{ES}	magnitude of the collector, emitter saturation current with the emitter, collector shorted (A)
I_D	channel current in an FET, directed from drain to source (A)
I_0	reverse saturation current in a p-n junction (A)
j	$\sqrt{-1}$
J	current density (A/cm^2)
k	Boltzmann's constant (J/K, eV/K)
k_N, k_P	transconductance of NMOSFET, PMOSFET divided by W_D (A/V^2)
$\mathbf{k}$	wave vector (cm^{-1})
k_d	distribution coefficient
K	scaling factor
K	$4\pi\epsilon_0$ (F/cm)
l, L	length (cm)
L_D	Debye length (cm)
$\bar{l}$	mean free path for carriers in random motion (cm)
m, m^*	mass, effective mass (kg)
m_n^*, m_p^*	effective mass for electrons, holes (kg)
m_l, m_t	longitudinal, transverse electron effective mass (kg)
m_{lh}, m_{hh}	light, heavy hole effective mass (kg)
m_0	rest mass of the electron (kg)
M	avalanche multiplication factor
m, n	integers; exponents
n	concentration of electrons in the conduction band (cm^{-3})
n	n-type semiconductor material
n^+	heavily doped n-type material
n_i	intrinsic concentration of electrons (cm^{-3})
n_n, n_p	equilibrium concentration of electrons in n-type, p-type material (cm^{-3})
n_0	equilibrium concentration of electrons (cm^{-3})
N (subscript)	normal mode of a BJT
N_a, N_d	concentration of acceptors, donors (cm^{-3})
N_a^-, N_d^+	concentration of ionized acceptors, donors (cm^{-3})
N_c, N_v	effective density of states at the edge of the conduction band, valence band (cm^{-3})
p	concentration of holes in the valence band (cm^{-3})
p	p-type semiconductor material

[3]See note at the end of this list.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 536

p^+	heavily doped p-type material
p	momentum (kg-m/s)
p_i	intrinsic hole concentration (cm^{-3}) = n_i
p_n, p_p	equilibrium concentration of holes in n-type, p-type material (cm^{-3})
p_0	equilibrium hole concentration (cm^{-3})
q	magnitude of the electronic charge (C)
Q_+, Q_-	total positive, negative charge (C)
Q_d	depletion region charge/area (C/cm^2)
Q_f	oxide fixed charge/area (C/cm^2)
Q_i	effective MOS interface charge/area (C/cm^2)
Q_{it}	interface trap charge/area (C/cm^2)
Q_m	mobile ionic charge/area (C/cm^2)
Q_n, Q_p	charge stored in an electron, hole distribution (C)
Q_n	mobile charge/area in FET channel (C/cm^2)
Q_{ot}	oxide-trapped charge/area (C/cm^2)
$R_p, \Delta R_p$	projected range, straggle (cm)
r, R	resistance (Ω)
R_H	Hall coefficient (cm^3/C)
S	subthreshold slope (mV/decade)
t	time (s)
t	sample thickness (cm)
$\bar{t}$	mean free time between scattering collisions (s)
t_{sd}	storage delay time (s)
T	temperature (K)
v, V	voltage[4] (V)
V	potential energy (J)
$\mathcal{V}$	electrostatic potential (V)
V_{CB}, V_{EB}	voltage from collector to base, emitter to base in a BJT (V)
V_D, V_G	voltage from drain to source, gate to source in an FET (V)
$\mathcal{V}_n, \mathcal{V}_p$	electrostatic potential in the neutral n, p material (V)
V_0	contact potential (V)
V_P	Chapter 6: pinch-off voltage for an FET; Chapter 11: forward breakover voltage for an SCR (V)
V_T, V_{FB}	MOS threshold voltage, flat-band voltage (V)
v, v_d	velocity, drift velocity (cm/s)
w	sample width (cm)
W	depletion region width (cm)
W_b	base width in a BJT, measured between the edges of the emitter and collector junction depletion regions (cm)
x	distance (cm), alloy composition
x_n, x_p	distance in the neutral n region, p region of a junction, measured from the edge of the transition region (cm)
x_{n0}, x_{p0}	penetration of the transition region into the n region, p region, measured from the metallurgical junction (cm)
Z	atomic number; dimension in z-direction (cm)
α	emitter-to-collector current transfer ratio in a BJT

[4]See note at the end of this list.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 537

α	optical absorption coefficient (cm^{-1})
α_r	recombination coefficient (cm^3/s)
β	base-to-collector current amplification factor in a BJT
γ	emitter injection efficiency; in a p-n-p, the fraction of i_E due to the hole current i_{E_p}
δ, Δ	incremental change
$\delta n, \delta p$	excess electron, hole concentration (cm^{-3})
$\Delta n_p, \Delta p_n$	excess electron, hole concentration at the edge of the transition region on the p side, n side (cm^{-3})
$\Delta p_C, \Delta p_E$	excess hole concentration in the base of a BJT, evaluated at the edge of the transition region of the collector, emitter junction (cm^{-3})
$\epsilon, \epsilon_r, \epsilon_0$	permittivity, relative dielectric constant, permittivity of free space (F/cm); $\epsilon = \epsilon_r \epsilon_0$
λ	wavelength of light (μm, Å)
μ	mobility ($cm^2/V\text{-}s$)
ν	frequency of light (s^{-1})
ρ	resistivity (Ω-cm); charge density (C/cm^3)
σ	conductivity $(\Omega\text{-cm})^{-1}$
τ_d	dielectric relaxation time (s); in a BJT, delay time (s)
τ_n, τ_p	recombination lifetime for electrons, holes (s)
τ_t	transit time (s)
ϕ	flux density $(cm^2\text{-}s)^{-1}$; potential (V), dose (cm^{-2})
ϕ_F	$(E_i - E_F)/q$ (V)
ϕ_s	surface potential (V)
Φ	work function potential (V)
Φ_B	metal-semiconductor barrier height (V)
Φ_{ms}	metal-semiconductor work function potential difference (V)
ψ, Ψ	*time-independent, time-dependent wave function*
ω	angular frequency (s^{-1})
$\langle\ \rangle$	average of the enclosed quantity

Note: For d-c voltage and current, capital symbols with capital subscripts are used; lowercase symbols with lowercase subscripts represent a-c quantities; lowercase symbols with capital subscripts represent total (a-c + d-c) quantities. For voltage symbols with double subscripts, V is positive when the potential at the point referred to by the first subscript is higher than that of the second point. For example, V_{GD} is the potential difference $V_G - V_D$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 538

Appendix II
Physical Constants and Conversion Factors[1]

Avogadro's number	$N_A = 6.02 \times 10^{23}$ molecules/mole		
Boltzmann's constant	$k = 1.38 \times 10^{-23}$ J/K		
	$= 8.62 \times 10^{-5}$ eV/K		
Electronic charge (magnitude)	$q = 1.60 \times 10^{-19}$ C		
Electronic rest mass	$m_0 = 9.11 \times 10^{-31}$ kg		
Permittivity of free space	$\epsilon_0 = 8.85 \times 10^{-14}$ F/cm		
	$= 8.85 \times 10^{-12}$ F/m		
Planck's constant	$h = 6.63 \times 10^{-34}$ J-s		
	$= 4.14 \times 10^{-15}$ eV-s		
Room temperature value of kT	$kT = 0.0259$ eV		
Speed of light	$c = 2.998 \times 10^{10}$ cm/s		
	Prefixes:		
1 Å (angstrom) $= 10^{-8}$ cm	milli-,	m-	$= 10^{-3}$
1 μm (micron) $= 10^{-4}$ cm	micro-,	μ-	$= 10^{-6}$
1 nm $= 10$ Å $= 10^{-7}$ cm	nano-,	n-	$= 10^{-9}$
2.54 cm = 1 in.	pico-,	p-	$= 10^{-12}$
1 eV $= 1.6 \times 10^{-19}$ J	kilo-,	k-	$= 10^{3}$
	mega-,	M-	$= 10^{6}$
	giga-,	G-	$= 10^{9}$

A wavelength λ of 1 μm corresponds to a photon energy of 1.24 eV.

[1]Since cm is used as the unit of length for many semiconductor quantities, caution must be exercised to avoid unit errors in calculations. When using quantities involving length in formulas which contain quantities measured in MKS units, it is usually best to use all MKS quantities. Conversion to standard semiconductor usage involving cm can be accomplished as a last step. Similar caution is recommended in using J and eV as energy units.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 539

Appendix III

Properties of Semiconductor Materials

		E_g (eV)	μ_n (cm^2/V-s)	μ_p (cm^2/V-s)	m^*_n/m_e (m_l, m_t)	m^*_p/m_e (m_{lh}, m_{hh})	a (Å)	ϵ_r	Density (g/cm^3)	Melting point (°C)
Si	(i/D)	1.11	1350	480	0.98, 0.19	0.16, 0.49	5.43	11.8	2.33	1415
Ge	(i/D)	0.67	3900	1900	1.64, 0.082	0.04, 0.28	5.65	16	5.32	936
SiC (α)	(i/W)	2.86	500	—	0.6	1.0	3.08	10.2	3.21	2830
AlP	(i/Z)	2.45	80	—	—	0.2, 0.63	5.46	9.8	2.40	2000
AlAs	(i/Z)	2.16	1200	420	2.0	0.15, 0.76	5.66	10.9	3.60	1740
AlSb	(i/Z)	1.6	200	300	0.12	0.98	6.14	11	4.26	1080
GaP	(i/Z)	2.26	300	150	1.12, 0.22	0.14, 0.79	5.45	11.1	4.13	1467
GaAs	(d/Z)	1.43	8500	400	0.067	0.074, 0.50	5.65	13.2	5.31	1238
GaN	(d/Z, W)	3.4	380	—	0.19	0.60	4.5	12.2	6.1	2530
GaSb	(d/Z)	0.7	5000	1000	0.042	0.06, 0.23	6.09	15.7	5.61	712
InP	(d/Z)	1.35	4000	100	0.077	0.089, 0.85	5.87	12.4	4.79	1070
InAs	(d/Z)	0.36	22600	200	0.023	0.025, 0.41	6.06	14.6	5.67	943
InSb	(d/Z)	0.18	10^5	1700	0.014	0.015, 0.40	6.48	17.7	5.78	525
ZnS	(d/Z, W)	3.6	180	10	0.28	—	5.409	8.9	4.09	1650*
ZnSe	(d/Z)	2.7	600	28	0.14	0.60	5.671	9.2	5.65	1100*
ZnTe	(d/Z)	2.25	530	100	0.18	0.65	6.101	10.4	5.51	1238*
CdS	(d/W, Z)	2.42	250	15	0.21	0.80	4.137	8.9	4.82	1475
CdSe	(d/W)	1.73	800	—	0.13	0.45	4.30	10.2	5.81	1258
CdTe	(d/Z)	1.58	1050	100	0.10	0.37	6.482	10.2	6.20	1098
PbS	(i/H)	0.37	575	200	0.22	0.29	5.936	17.0	7.6	1119
PbSe	(i/H)	0.27	1500	1500	—	—	6.147	23.6	8.73	1081
PbTe	(i/H)	0.29	6000	4000	0.17	0.20	6.452	30	8.16	925

All values at 300 K.

*Vaporizes

The first column lists the semiconductor, the second indicates band structure type and crystal structure. Definitions of symbols: *i* is indirect; *d* is direct; *D* is diamond; *Z* is zinc blende; *W* is wurtzite; *H* is halite (NaCl). Values of mobility are for material of high purity.

Crystals in the wurtzite structure are not described completely by the single lattice constant given here, since the unit cell is not cubic. Several II–VI compounds can be grown in either the zinc blende or wurtzite structures.

Many values quoted here are approximate or uncertain, particularly for the II–VI and IV–VI compounds. The gaps indicate that the values are unknown.

For electrons, the first set of band curvature effective masses is the longitudinal mass, the second set the transverse. For holes, the first set is for light holes, the second for heavy holes.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 540

Appendix IV

Derivation of the Density of States in the Conduction Band

In this derivation we shall consider the conduction band electrons to be essentially free. Constraints of the particular lattice can be included in the effective mass of the electron at the end of the derivation. For a free electron, the three-dimensional Schrödinger wave equation becomes

$$-\frac{\hbar^2}{2m}\nabla^2\psi = E\psi \tag{IV-1}$$

where ψ is the wave function of the electron and E is its energy. The form of the solution to Eq. (IV-1) is

$$\psi = (\text{const.})e^{j\mathbf{k}\cdot\mathbf{r}} \tag{IV-2}$$

We must describe the electron in terms of a set of boundary conditions within the lattice. A common approach is to use periodic boundary conditions, in which we quantize the electron energies in a cube of material of side L. This can be accomplished by requiring that

$$\psi(x + L, y, z) = \psi(x, y, z) \tag{IV-3}$$

and similarly for the y- and z-directions. Thus our wave function can be written as

$$\psi_n = A \exp\left[j\frac{2\pi}{L}(\mathbf{n}_x x + \mathbf{n}_y y + \mathbf{n}_z z)\right] \tag{IV-4}$$

where the $2\pi\mathbf{n}/L$ factor in each direction guarantees the condition described by Eq. (IV-3), and A is a normalizing factor. Substituting ψ_n into the Schrödinger equation (IV-1), we obtain

$$-\frac{\hbar^2}{2m}A\nabla^2 \exp\left[j\frac{2\pi}{L}(\mathbf{n}_x x + \mathbf{n}_y y + \mathbf{n}_z z)\right] = EA \exp\left[j\frac{2\pi}{L}(\mathbf{n}_x x + \mathbf{n}_y y + \mathbf{n}_z z)\right] \tag{IV-5}$$

Let us determine the number of allowed states per unit volume as a function of energy [the density of states, $N(E)$] in various cases such as one, two, or three dimensions. We first count states in $\mathbf{k}$-space, then we can use the band structure, $E(\mathbf{k})$, to convert to $N(E)$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 541

For the the 3-D case in Eq. (IV-5), the components of the **k**-vector are $\mathbf{k}_x = 2\pi\,\mathbf{n}_x/L$, $\mathbf{k}_y = 2\pi\,\mathbf{n}_y/L$, and $\mathbf{k}_z = 2\pi\,\mathbf{n}_z/L$. Since there is one **k**-state for every distinct choice of integer quantum numbers, $(\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_z)$, the volume per **k**-state is $(2\pi)^3/L^3 = (2\pi)^3/V$, where $V = L^3$ is the three-dimensional volume. Hence, the number of states for 3-D in a **k**-space of $\Delta\mathbf{k}$, taking into account the factor of 2 spin degeneracy, is

$$\left\{\frac{L^3}{(2\pi)^3}\Delta\mathbf{k}\right\} \times (2)\,spin \tag{IV-6a}$$

The number of states per unit volume for 3-D is

$$\frac{2}{(2\pi)^3}(\Delta\mathbf{k}) \tag{IV-6b}$$

In general, for p-dimensions we can generalize this expression as

$$\text{Number of states per unit volume} = \frac{2}{(2\pi)^p}(\Delta\mathbf{k}) \tag{IV-7a}$$

We can then transform from **k**-space to E-space using the $E(\mathbf{k})$ band structure relationship by setting

$$N(E)\,\Delta E = \frac{2}{(2\pi)^p}(\Delta\mathbf{k}) \tag{IV-7b}$$

As described in Section 3.2.2, the simplest band structure is parabolic:

$$E(\mathbf{k}) = \frac{\hbar^2 k^2}{2m^*} \tag{IV-8a}$$

This is often a good approximation, particularly near the bottom of the conduction band or top of the valence band. Using this, we get the relation between **k** and E as follows:

$$k = \sqrt{\frac{2m^*E}{\hbar^2}} \tag{IV-8b}$$

$$dk = \left\{\sqrt{\frac{m^*}{2}}\,\frac{1}{\hbar}\right\}\frac{1}{\sqrt{E}}\,dE \tag{IV-8c}$$

For $p = 3$ we have the 3-D case, which is typical of bulk semiconductors. The volume in **k**-space between two constant-k spherical surfaces at k and $k + dk$ is (Fig. IV-1a):

$$\Delta\mathbf{k} = 4\pi k^2 dk \tag{IV-9a}$$

neglecting terms with dk multiplied by itself.

The density-of-states then becomes:

$$N(E)dE = \frac{2}{(2\pi)^3}4\pi k^2 dk = \frac{\sqrt{2}}{\pi^2}\left(\frac{m^*}{\hbar^2}\right)^{3/2}E^{1/2}dE \tag{IV-9b}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 542

We see that if we plot $N(E)$ versus E, we get a parabolic density-of-states function in 3-D for a parabolic band structure relationship (Fig. IV-2a).

For $p = 2$, we get a so-called 2-D electron gas (2-DEG) or hole gas. This can arise, for example, in a quantum well (Section 3.2.5) or in the inversion layer of a MOSFET.

In this case, the "volume" in **k**-space is the annular region between two circles, k and $k+dk$, as shown in Fig. IV-1b, where

$$\Delta \mathbf{k} = (2\pi k)dk \tag{IV-10a}$$

again neglecting dk^2.

Using Eq. (IV-7a), this leads to a density of states (per unit area)

$$N(E)dE = \frac{2}{(2\pi)^2}(2\pi k)dk = \frac{m^*}{\pi\hbar^2}dE \tag{IV-10b}$$

We see that for 2-D, the density of states is a constant in energy, unlike the parabolic density of states for 3-D (Fig. IV-2b). Actually, for the 2-DEG

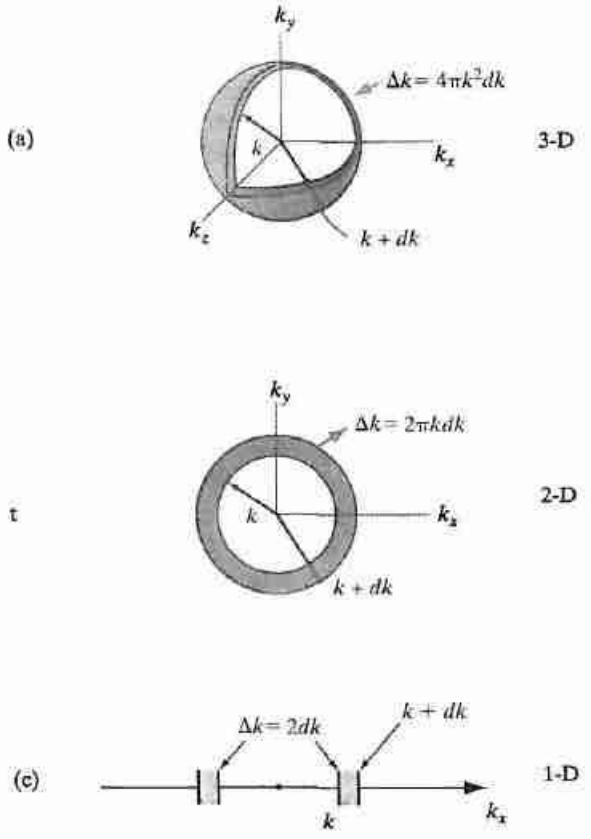


Figure IV-1 Volume in **k**-space: (a) 3-D systems; (b) 2-D systems; (c) 1-D systems.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 543

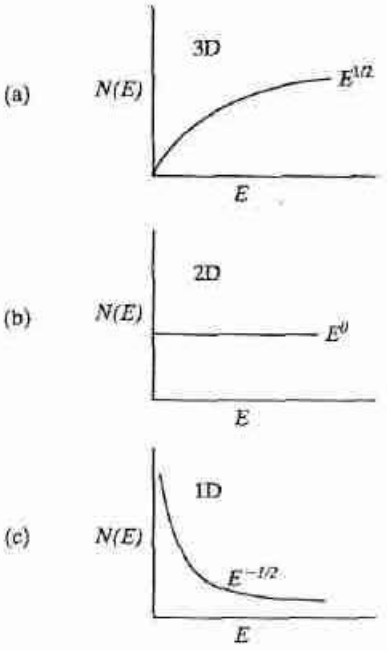


Figure IV–2
Density of states: (a) in 3-D or bulk; (b) in 2-D electron or hole gases; (c) in 1-D quantum "wires."

in a quantum well or inversion layer (see Chapter 6) we must add the various constant 2-D densities-of-states for the different "particle-in-a-box" levels that were discussed in Sections 2.4.3 and 3.2.5, leading to a so-called "staircase" density of states.

For $p = 1$, we get 1-D quantum "wires." These more esoteric structures can be grown, for example, by MBE or MOCVD. In this case, the "volume" in **k**-space in the region between k and $k+dk$ in 1-D is (Fig. IV-1c):

$$\Delta \mathbf{k} = 2(dk) \tag{IV-11a}$$

Using Eq. (IV-7a), this leads to a density of states

$$N(E)dE = \frac{2}{(2\pi)^1}(2dk) = \frac{\sqrt{2m^*}}{\pi\hbar\sqrt{E}}dE \tag{IV-11b}$$

By examining the density of states in 3-, 2-, and 1-D [Eqs. (IV-9b), (IV-10b), and (IV-11b), respectively] we notice a very interesting trend. Every time we go to a lower dimensionality system, the dependence of density of states on energy changes by $1/\sqrt{E}$. In fact, one finds that for 0-D quantum "dots" the density of states is indeed proportional to $1/E$. In the 1- and 0-D cases, we see that the density of states has singularities in energy, which has very important implications for semiconductor devices. Unfortunately, those discussions are beyond the scope of this book.

To include the probability of occupation of any energy level E, we use the Fermi-Dirac distribution function:

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 544

$$f(E) = \frac{1}{e^{(E-E_F)/kT} + 1} \tag{IV-12}$$

The concentration of electrons in the range dE is given by the product of the density of allowed states in that range and the probability of occupation. Thus the density of occupied electron states N_e in dE is

$$N_e dE = N(E)f(E)dE \tag{IV-13}$$

For the 3-D case we may calculate the concentration of electrons in the conduction band at a given temperature by integrating Eq. (IV-13) across the band:

$$n = \int_0^\infty N(E)f(E)dE = \frac{1}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} e^{E_F/kT} \int_0^\infty E^{1/2} e^{-E/kT} dE \tag{IV-14}$$

In this integration we have referred the energies in the conduction band to the band edge (E_c taken as $E = 0$). Furthermore, we have taken the function $f(E)$ to be

$$f(E) = e^{(E_F - E)/kT} \tag{IV-15}$$

for energies such that $(E - E_F) \gg kT$.

The integral in Eq. (IV-14) is of the standard form:

$$\int_0^\infty x^{1/2} e^{-ax} dx = \frac{\sqrt{\pi}}{2a\sqrt{a}} \tag{IV-16}$$

Thus Eq. (IV-14) gives

$$n = 2\left(\frac{2\pi mkT}{h^2}\right)^{3/2} e^{E_F/kT} \tag{IV-17}$$

If we refer to the bottom of the conduction band as E_c instead of $E = 0$, the expression for the electron concentration is

$$n = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{3/2} e^{(E_F - E_C)/kT} \tag{IV-18}$$

which corresponds to Eq. (3–15). We have included constraints of the lattice through the effective mass of the electron in the crystal, m_n^*.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 545

Appendix V
Derivation of Fermi–Dirac Statistics

In this section, we will give a simplified derivation of Fermi–Dirac statistics. We will not go through all the details, but will instead point out the physical assumptions involved. The distribution function is determined by calculating the number of distinct ways (W_k) we can put n_k indistinguishable electrons in g_k states at an energy level E_k, subject to the Pauli exclusion principle.

The assumptions are:

1. Each allowed state has a maximum of one electron (Pauli principle).
2. The probability of occupancy of each allowed (degenerate) quantum state is the same.
3. All electrons are indistinguishable.

The number of distinct ways we can put the electrons in a particular level is

$$W_k = \frac{(g_k)(g_k - 1)(g_k - \overline{n_k - 1})}{n_k!} = \frac{g_k!}{(g_k - n_k)!n_k!} \tag{V-1}$$

For N levels in a band, the number of distinct ways we can put in the various electrons gives us the so-called "multiplicity function,"

$$W_b = \prod_k W_k = \prod_k \frac{g_k!}{(g_k - n_k)!n_k!} \tag{V-2}$$

If we ask, "What is the most probable distribution of the n_k electrons in the various E_k levels (degeneracy of g_k in level E_k)?" the statistical mechanical answer is:

> In thermal equilibrium, the distribution which is most disordered (i.e., has the maximum entropy, or which can occur in the largest number of ways) is the most probable.

We therefore have to maximize W_b with respect to n_k.

We assume here that the total number of electrons in the band is fixed.

$$\sum_k n_k = n = \text{constant} \Rightarrow \sum_k dn_k = 0 \tag{V-3}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 546

We also assume that the total energy in the band is constant.

$$E_{tot} = \sum_k E_k n_k = \text{constant, implying } \sum_k E_k dn_k = 0 \qquad \text{(V-4)}$$

To maximize or minimize some function $f(x_i)$ of q variables $x_i (i = 1, \ldots, q)$ subject to the constraints that $g(x_i)$ and $h(x_i)$ are constant, we use the method of Lagrange undetermined multipliers.

We have

$$df = 0 \quad \text{(for extremal value of } f\text{)} \qquad \text{(V-5)}$$

$$dg = 0, dh = 0 \quad \text{(because } g \text{ and } h \text{ are constant)} \qquad \text{(V-6)}$$

Introducing two Lagrange undetermined multipliers α and β, we get

$$\sum_i \frac{\partial}{\partial x_i}[f(x_i) + \alpha g(x_i) + \beta h(x_i)]dx_i = 0$$

$$\frac{\partial}{\partial x_i}[f(x_i) + \alpha g(x_i) + \beta h(x_i)] = 0 \qquad \text{(V-7)}$$

for $i = 1, \ldots q$

$$g(x_i) = \text{const.} \quad h(x_i) = \text{const.} \qquad \text{(V-8)}$$

We thus get $(q + 2)$ equations in $(q + 2)$ unknowns of (x_i, α, β).

We apply this technique to our problem at hand. Instead of maximizing W_b, we maximize $\ln W_b$, because it makes the mathematics simpler. Since the log function increases monotonically with the argument, maximizing one is the same as maximizing the other.

$$\ln W_b = \sum_k [\ln(g_k)! - \ln(g_k - n_k)! - \ln(n_k)!] \qquad \text{(V-9)}$$

To simplify these terms, we use Stirling's approximation for factorials of large numbers. $\ln x! = x \ln x - x$ for large x.

$$\ln W_b = \sum_k [g_k \ln(g_k) - g_k - (g_k - n_k)\ln(g_k - n_k) + (g_k - n_k) - n_k \ln(n_k) + n_k]$$

$$= \sum_k [g_k \ln(g_k) - (g_k - n_k)\ln(g_k - n_k) - n_k \ln(n_k)] \qquad \text{(V-10)}$$

Now $dg_k = 0$ because these are system constraints. We then get

$$d(\ln W_b) = \sum_k \frac{\partial[\ln W_b]}{\partial n_k} dn_k = \sum_k \ln\left(\frac{g_k}{n_k} - 1\right) dn_k = 0 \qquad \text{(V-11)}$$

Also, from the two constraints we get

$$\sum_k dn_k = 0 \text{ and } \sum_k E_k dn_k = 0 \qquad \text{(V-12)}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 547

Then,

$$\sum_k \left[\ln\left(\frac{g_k}{n_k} - 1\right) - \alpha - \beta E_k \right] dn_k = 0 \tag{V-13}$$

$$\ln\left(\frac{g_k}{n_k} - 1\right) - \alpha - \beta E_k = 0 \tag{V-14}$$

From this,

$$\frac{n_k}{g_k} = f(E_k) = \frac{1}{1 + e^{\alpha + \beta E_k}} \tag{V-15}$$

From basic thermodynamics, it can be shown that

$$\alpha = -\frac{E_F}{kT}, \quad \beta = \frac{1}{kT} \tag{V-16}$$

to get the Fermi-Dirac distribution function,

$$f(E_k) = \frac{1}{\exp\left[\dfrac{E_k - E_F}{kT}\right] + 1} \tag{V-17}$$

For the limit of high energies,

$$E \gg E_F, \quad f(E) \simeq \exp \frac{E_F - E}{kT}. \tag{V-18}$$

This is the classical Maxwell-Boltzmann limit of the Fermi-Dirac distribution function. Once we have the probabilities of electron occupancy, the probability of hole occupancy becomes

$$1 - f(E) = \frac{1}{\exp \dfrac{E_F - E}{kT} + 1} \tag{V-19}$$

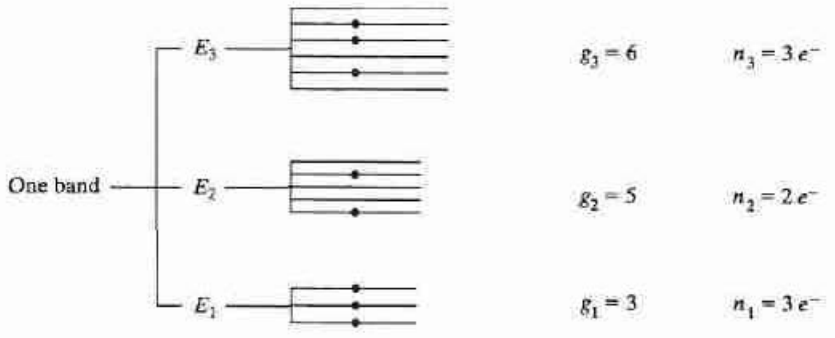


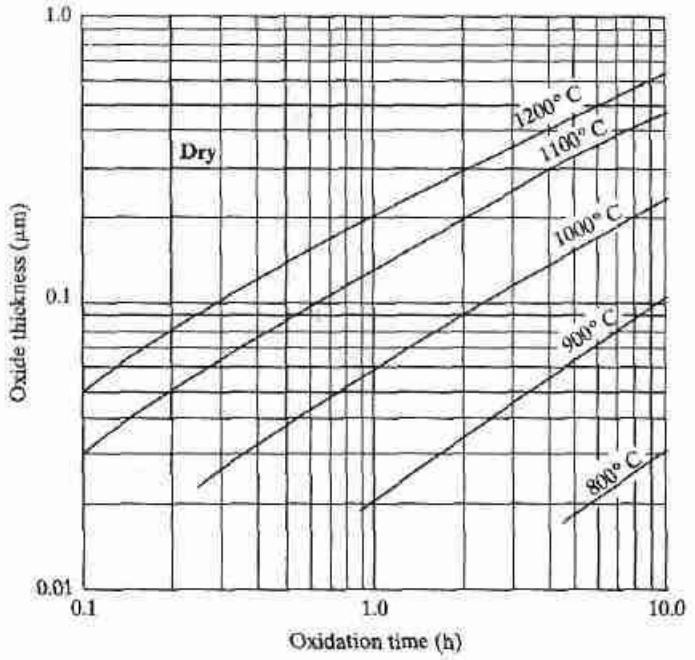
Figure V–1
Example showing three energy levels in a band, having different degeneracies, g, and electron occupancies, n, as shown.

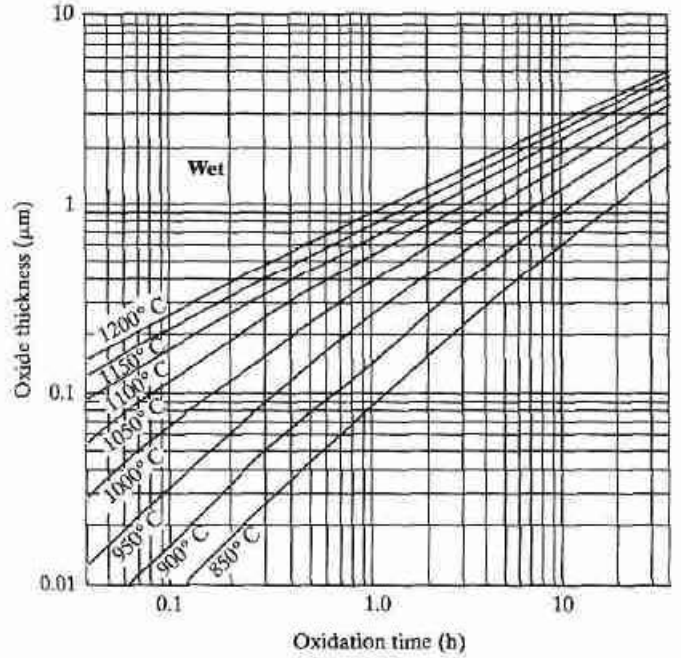
Appendix VI

Dry and Wet Thermal Oxide Thickness Grown on Si (100) as a Function of Time and Temperature[1]

[1]From **B. Deal.** "The Oxidation of Silicon in Dry Oxygen, Wet Oxygen and Steam." *J. Electrochem. Soc.* 110 (1963): 527.

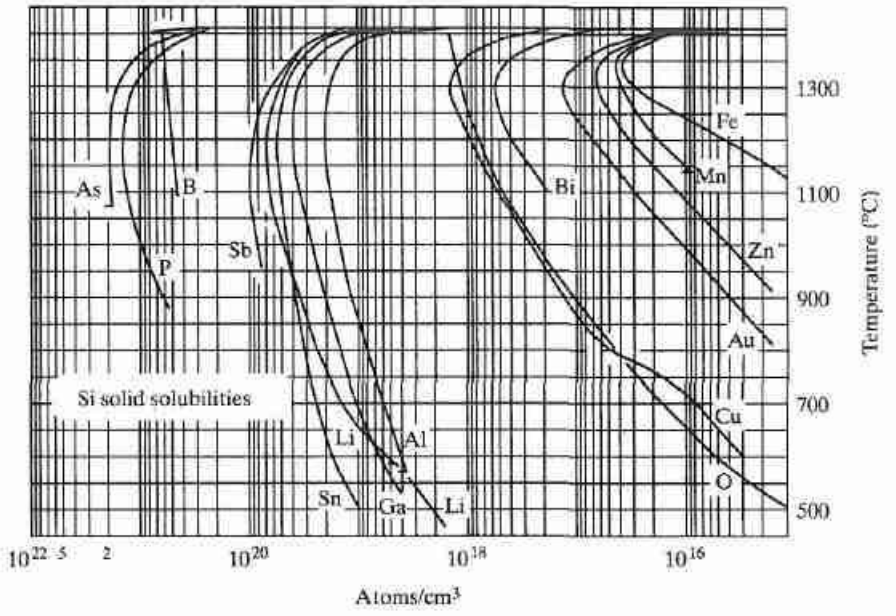
Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 550


Dry
1200° C
1100° C
1000° C
900° C
800° C
Oxide thickness (μm)
1.0
0.1
0.01
0.1
1.0
10.0
Oxidation time (h)


Wet
1200° C
1150° C
1100° C
1050° C
1000° C
950° C
900° C
850° C
Oxide thickness (μm)
10
1
0.1
0.01
0.1
1.0
10
Oxidation time (h)

Appendix VII

Solid Solubilities of Impurities in Si[1]



[1]From **F. A. Trumbore.** "Solid Solubilities of Impurity Elements in Si and Ge," *Bell System Technical Journal* 39, no. 1, pp. 205–233 (January 1960) copyright 1960, The American Telephone and Telegraph Co., reprinted by permission. Alterations have been made to include later data.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 552

Appendix VII

Diffusivities of Dopants in Si and SiO$_2$

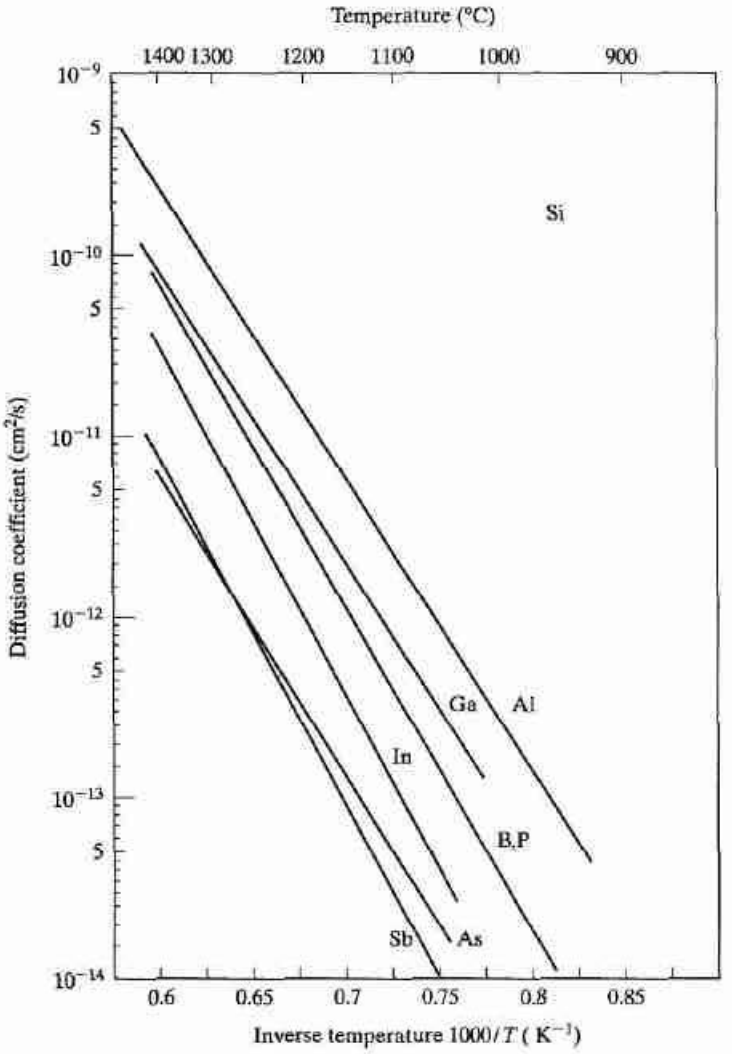
Appendix VIII

Diffusivities of Dopants in Si and SiO_2[1]

[1]Silicon diffusivity data from **C. S. Fuller and J. A. Ditzenberger.** "Diffusion of Donor and Acceptor Elements in Silicon." *J. Appl. Physics*, 27 (1956), 544.

SiO_2 diffusivity data from **M. Ghezzo and D. M. Brown.** "Diffusivity Summary of B, Ga, P, As and Sb in SiO_2," *J. Electrochem. Soc.* 120 (1973), 146.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 554



Diffusivity of various impurities in SiO_2

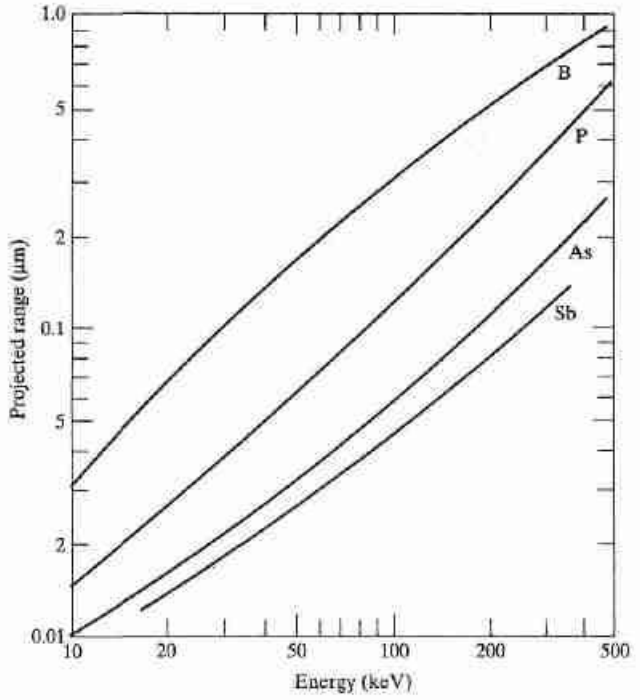
Element	$D_o(cm^2/sec)$	$E_A(eV)$
Boron	3×10^{-4}	3.53
Phosphorus	0.19	4.03
Arsenic	250	4.90
Antimony	1.31×10^{16}	8.75

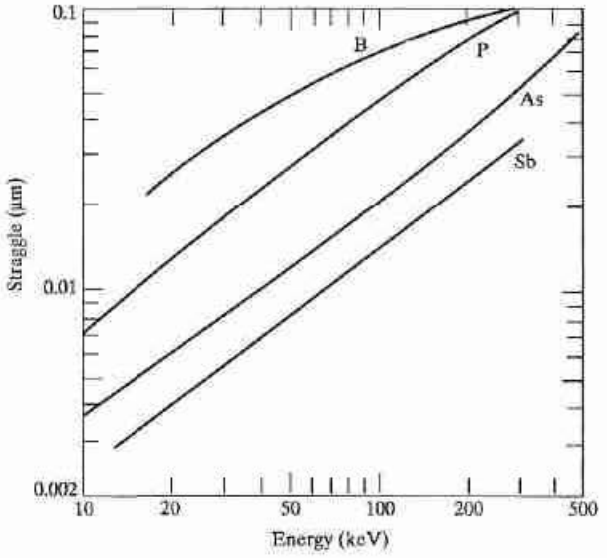
Appendix IX

Projected Range and Straggle as Function of Implant Energy in Si[1]

[1]From **J. F. Gibbons, W. S. Johnson, and S. W. Mylroie.** *Projected Range Statistics: Semiconductors and Related Materials.* Stroudsburg: Dowden, Hutchison and Ross, 1975.
The projected ranges in SiO_2 are very close to those in Si.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 556


1.0
5
2
0.1
5
2
0.01
Projected range (μm)
B
P
As
Sb
10
20
50
100
200
500
Energy (keV)


0.1
0.01
0.002
Straggle (μm)
B
P
As
Sb
10
20
50
100
200
500
Energy (keV)

Answers to Selected Self Quiz Questions

CHAPTER 1

1. (a) (111); (100); (110)
2. (a) 1 and 2
   (b) {110}
   (c) B
3. (a) C
   (b) (1) zinc blende; (2) diamond
4. 0-D: vacancies, interstitials; 1-D: dislocations; 2-D: grain boundaries, surfaces
6. InP

CHAPTER 2

1. (a) not allowed
   (b) not allowed
   (c) allowed
2. (a) No
   (b) Yes
3. (a) No
   (b) Yes
   (c) Yes
4. Not enough information
5. p_0; unknown; unknown
6. (a) Energy increased; unchanged
   (b) Energy unchanged; decreased

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 558

CHAPTER 3

1. (a) Semiconductor, Metal, Insulator
   (b) Material 3
2. (a) X valley
   (b) μ valley
3. (a) GaAs
   (b) GaAs
   (c) 1.4 eV; 0.87 μm; infrared
   (d) 6, 1
4. (a) GaAs
   (b) equal
5. (a) p-type
   (b) equal
   (c) very high acceptor doping
6. (a) 1.08 eV
   (b) $n_0 = 10^{16}\ \text{cm}^{-3}$; $p_0 = 10^4\ \text{cm}^{-3}$
   (c) $p_0 = 10^{16}\ \text{cm}^{-3}$; $n_0 = 10^4\ \text{cm}^{-3}$
   (d) 0.36 eV

CHAPTER 4

1. (a) $10^{17}/\text{cm}^3\text{-s}$
   (b) $2 \times 10^{17}\ \text{eV/cm}^3\text{-s}$
   (c) $10^{17}\ \text{eV/cm}^3\text{-s}$
2. (a) Deep more harmful; Au
   (b) Si; indirect
   (c) increase
3. (a) E_F flat
   (b) Constant electric field points to right
4. (a) left, right, right, left
   (b) left, right, left, right
5. (b) 4
6. (a) right, right, left, right
   (b) right, right, right, left

CHAPTER 5

3. Short
4. 1.5 V
7. (b) ohmic

CHAPTER 6

1. (a) depletion
   (b) weak inversion
   (c) accumulation
   (d) strong inversion
   (e) flat band
   (f) threshold
2. (a) power gain
   (b) d.c. power supply
3. (b) 1.2 $\mu F/cm^2$
   (c) same curve
4. (a) n-channel
   (b) long
   (c) −0.5 V
   (d) depletion mode
5. (a) decrease
   (b) decrease
6. (a) increase
   (b) increase
7. (a) unchanged
   (b) decrease
   (c) decrease
8. (a) 1 mmhos
   (b) −0.2 V
   (c) short channel
   (d) *p*-channel
   (e) enhancement mode
9. (a) normally ON
   (b) long channel
   (c) 25 mV/decade
   (d) S cannot be less than 60 mV/decade

CHAPTER 7

1. (a) 50
   (c) 6 V
   (d) 21 V
3. (a) increase
   (b) increase

(c) increase
(d) decrease

5. (a) increase
(b) increase
(c) decrease

CHAPTER 8

1. (a) C
(b) 0.7 eV
(c) 7 mW
(d) 0.4 eV
(e) 14 mW
(f) 50%
2. 28 mW
3. 0.5 μm; No; Yes
4. Renewable and green; cost

CHAPTER 9

1. Yes; No
4. Lowest Q_{ox}

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 561

Index

A

Absorption coefficient, 3, 120, 121, 123, 145, 150, 152, 414
Absorption rate, stimulated emission rate vs., 420
a-c conductance, 216, 384, 513
a-c transconductance, 384
Acceleration tube, 159
Acceptor level, 82–84
Active regions, 270, 427–428, 449
Address multiplexing, 484
Alkali metals, 47, 63
Aluminum wire, and ultrasonic bonding, 500–501
Amorphous solids, 3
Amphoteric impurities, 84
Amplification, with bipolar junction transistors (BJTs), 340–346
Anderson affinity rule, 234
Anisotropic, use of term, 166
Annealing, 160–161
Anode terminal, 520
Anti-bonding orbital, 64
Anti-punchthrough implant, 319–320
Application-specific ICs (ASICs), 444, 446
Aspect ratio, 462
Atalla, N. H., 252
Atomic models, 47
Atomic spectra, 35–36
Atomic structure:
  hydrogen atom, 47–50
  and the periodic table, 47
Atoms, 31–61
Avalanche breakdown, 197, 198–200
Avalanche multiplication, 525
Avalanche photodiodes (APDs), 406, 409

B

Back-end processing, 168
Backward diode, 531
Baliga, 528
Ball bond, 500–501
Ball grid arrays (BGAs), 504, 505
Band diagrams, 70, 75, 88
Band gap, 202
Bandbending, 275, 289
Bands, 61, 66–67
Bandstructure, 79
Bardeen, John, 251–252, 343
Base Gummel number, 378
Base narrowing, 370–371
Base spreading resistance, 374
Base transport factor, 340, 391
Base-to-collector current amplification factor, 341
Base-wide modulation, 370
Base-width narrowing, 265, 525
Basis, 3
Basis vectors, 5, 8
Batch fabrication, 438–439

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 563

Bell Telephone Laboratories, 144, 251, 468
Beveling, 203
Bias-temperature stress test, 289
BiCMOS, 446
Binary compounds, 2
Binding energy, 83
Bipolar FET (BiFET), 528
Bipolar junction transistors (BJTs), 251, 336–397
  amplification with, 340–346
  avalanche breakdown, 372–373
  base narrowing, 370–371
  base resistance, 374–376
  drift in the base region, 369–370
  emitter crowding, 374–376
  fabrication of, 343–346
  frequency limitations of transistors, 382–388
    capacitance and charging times, 382–385
    high-frequency transistors, 386–388
    transit time effects, 385–386
    Webster effect, 386
  fundamentals of operation, 336–340
Bipolar junction transistors generalized
  biasing, 355–363
    charge control analysis, 361–363
    coupled-diode model, 356
  Gummel-Poon model, 376–380
  heterojunction, 388–390
  history of, 252
  injection level, 373–374
  Kirk effect, 380–382
  minority carrier distributions and terminal currents, 346–355
    approximations of the terminal currents, 352–354
    current transfer ratio, 354–355
    evaluation of the terminal currents, 349–352
    solution of the diffusion equation in the base region, 347–349
  switching, 363–368
    saturation, 365–366
    specifications for switching transistors, 368
    switching cycle, 366–368
  thermal effects, 373–374
Bipolar transistor, 251
Bird's beak, 449
Bit line, 462
BJTs, *See* Bipolar junction transistors (BJTs)
Blackbody radiation, 33, 56
Blocking state, p-n-p-n diodes, 520, 523–524
Body effect, 310–311, 326
Body-centered cubic (BCC) lattice, 5
Bohr, Niels, 36, 39, 56
Bohr orbit, 50
Bohr theory, 39
Boltzmann factor, 86, 230
Boltzmann's constant, 87
Bond-and-Etch-back SOI (BE-SOI), 460
Bonding, integrated circuits:
  ball bond, 500–501
  die bonding, 497
  flip-chip techniques, 501–502
  lead bond, 497
  nail-head bond, 500
  thermocompression bonding, 500
  ultrasonic bonding, 500–501
  wedge bond, 500–501
  wire bonding, 497–501
Bonding forces:
  energy bands, 64–67
  and energy bands in solids, 61–71
  in solids, 62–64
Bonding orbital, 64
Borophosphosilicate glass (BPSG), 456
Bose–Einstein statistics for photons, 86
Bowing parameter, 27
Boyle, W. S., 468
Brattain, Walter H., 251–252, 343
Bravais lattices, 13
Breakdown diodes, 197, 204–205
  defined, 204
Buffer layer, 261
Built-in fields, 139–140
Built-in potential barrier, 172, 239
Bulk crystal growth, 12–18
  doping, 16–18
  single crystal ingots, growth of, 13–15
  starting materials, 13
  wafers, 15–16

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 564

Bulk negative differential conductivity (BNDC) effect, 517
Buried channel CCD, 469
Buried channel operation for PMOS, 454–455
Buried collector, 345
Bytes, 485

C

Capacitance and charging times, bipolar junction transistors (BJTs), 382–385
Capacitance transient, 291
Capillary, 497, 500
Carrier concentrations, 86–108
  compensation and space charge neutrality, 96–98
  electron and hole concentrations at equilibrium, 89–94
  Fermi level, 86–89
  temperature dependence of, 94–96
Carrier confinement, heterojunction lasers, 427
Carrier diffusion, 134–148
  diffusion processes, 134–137
  and drift of carriers, 137–140
  Haynes-Shockley experiment, 144–147
  and recombination, 140–142
  steady state carrier injection, 142–144
Carrier drift in electric/magnetic fields:
  conductivity and mobility, 98–102
  doping, effect on mobility, 103–105
  drift and resistance, 102–103
  Hall effect, 106–110
  high-field effects, 105–106
  temperature, effect on mobility, 103–105
Carrier injection, 184–193
Carrier lifetime, and photoconductivity, 124–134
Carrier multiplication, 198
Cathode terminal, 520
Cathodoluminescence, 121
Cell, 4
Centered rectangular lattice, 5
Central processing unit (CPU), 472
Ceramic column grid array (CCGA), 504
Channel, 254
Channel length modulation parameter, 322
Channel stop, 270, 449
Channel stop implant, 308
Charge carriers in semiconductors, 71–86
  carrier concentrations, 86–108
  effective mass, 76–80
  electrons and holes, 73–76
  extrinsic material, 81–85
  intrinsic material, 80–81
Charge control analysis, of BJTs, 361–363, 391
Charge control approximation, 190
Charge sharing, 322, 326
Charge storage capacitance, 212
Charge transfer devices, 466–470
  applications of CCDs, 470
  basic CCD, 468
  basic structure improvements, 469–470
  charge-coupled devices (CCDs), 466
  dynamic effects in MOS capacitors, 466–467
Charge-coupled devices (CCDs), 466
  applications of, 468
  basic, 468
Chemical beam epitaxy, 25
Chemical mechanical polishing (CMP), 166, 449
Chemical potential, 148
Chemical vapor deposition (CVD), 18, 162
Chip, 163
Chromatic dispersion, 416
CMOS process integration, 444–458
Coherent radiation, 418–419
Collector, 338
Collector junction, 338
Common base configuration, 338
Common-base current gain, 341
Common-emitter circuit, 342
Common-emitter current gain, 341
Commutator current, 481–482
Compact discs (CDs), 431
Compensation, 97
  and space charge neutrality, 96–98
Complementary error function, 157

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 565

Complementary MOS (CMOS) circuits, 313, 444–445
  dual gate CMOS, 455
  structure, 445
  twin-well CMOS, 446–447
Compound semiconductors, 1–2, 64
Conducting state, p-n-p-n diodes, 520, 524–525
Conduction band, 67
Conductivity effective mass, 80
Conductivity modulation, 203–204, 224
Conductivity-modulated FET (COMFET), 528–529
Conformal LPCVD oxide, 453
Constant Fermi level, 148
Contact potential, 170–174
Continuity equation, 140–142
  for holes, 141
Corner effect, 449
Corpuscular view of light, 33
Coulomb potential, 47
Coulombic potential energy, 64
Coupled-diode property, 358
Coupled-diode model, 356–361
Covalent bonding, 63, 83
Crossovers, 466
Crystal, 3–4
Crystal lattices, 3–12
  cubic lattices, 5–7
  diamond lattice, 10–12
  periodic structures, 3–5
  planes and directions, 7–10
Crystals, arrangement of atoms in, 12
Cubic lattices, 5–7
Current transfer ratio, 340
Curve tracer, 252
Cutoff frequency, 385
Cutoff regime, 363
Czochralski growth, 15, 17

D

Damascene process, 464
Davisson, C. S., 34
D-c-to-microwave power conversion, 508
De Broglie, Louis, 34, 56
De Broglie relation, 166
De Broglie wave concept, 37
Debye screening length, 275–281
Deep depletion, 290–291, 467
Degenerate n-type, 509
Degenerate semiconductors, 509–512
Delay time, switching transistors, 368
Dennard, Robert H., 315
Density-of-states effective mass, 80
Depletion approximation, 175
Depletion capacitance of a junction, 214, 240
Depletion region, 172, 325
Depletion-mode condition, 308
Depletion-mode transistor, 267
Depth-of-focus (DOF), 165
Destriau effect, 124
Diamond lattice, 10–12
  unit cell, 12
Diamond structure, 10
Die, 163
Die bonding, 497
Dielectric relaxation time, 518
Diffraction-limited geometry, 166
Diffused junctions, approximations to, 226
Diffusion, 156–158, *See also* Carrier diffusion
  capacitance, 216, 240
  defined, 134
  equation, 141–142
  flux, 149
  length, 142
  potential, 172
Digital circuits, 440
Digital versatile discs (DVDs), 431
Diode equation, 187
Diodes, 515
  lateral nature of, 202
  long, 216
  short, 216
Direct semiconductors, 68–70
Direct tunneling, 292–293
Direction in a lattice, expression of, 8
Discrete energy, 85
Discretionary wiring, 471

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 566

Dispersion:
chromatic, 416
pulse, 416
Dispersion relationship, 35
Displacement current, 526
Distribution coefficient, 17
Domains, 519–520
Donor, 82
impurities, 82
Doped polysilicon layer, 269
Doping, 16–18, 81
concentration, 202
effect on mobility, 103–105
importance of, 84
Double polysilicon, self-aligned n-p-n Si BJT, fabrication of, 343–344
Double-diffused MOSFET (DMOS), 529
Double-heterojunction laser structure, 427–428
Double-polysilicon process, 346
Drain, 254
Drain-induced barrier lowering (DIBL), 298, 320–322, 326
Drift current, 182–183
Drift tube, 159
Dual gate CMOS, 455
Dual inline package (DIP), 504
dv/dt triggering, 525
DX center, 264
Dynamic equilibrium, 239
Dynamic random-access memories (DRAMs), 441, 444, 483, 486–492
Dynamic resistance dV/dI, 511

E

Early effect, 265, 370, 378, 391
Early, J. M., 370
Early voltage, 371
Ebers, J. J., 358
Ebers–Moll equations, 358–359, 365, 391
Effective channel mobility, 326
Effective density of states N_c, 90
Effective mass, 76–80
Effective number of lattice points, 4
EHPs, *See* Electron–hole pairs (EHPs)
Eigenenergy, 56
Eigenfunctions, 56
Eigenstates, 56
Einstein, Albert, 33, 56
Einstein coefficients, 420
Einstein relation, 139, 149, 377
Electric field, 180
Electrochemical potentials, 148
Electroluminescence, 121, 124
injection, 124, 411
Electromigration phenomenon, 456
Electron affinity, 228
Electron diffusion coefficient, 136
Electron energy diagram, 272
Electron flux density, 136
Electron mobility, 101–102
Electron projection lithography (EPL), 166
Electron–hole pairs (EHPs), 73–76
created by impact ionization, 199
generation of, visualizing in a qualitative way, 80
recombination of, 81, 122
thermal excitation of, 202
Electrons, 31–61
direct recombination of, 125–127
Electrostatic potential barrier, 180
Elemental semiconductors, 1–2
Emission from recombination centers, 222
Emission spectra for p-n junction lasers, 424–425
Emitter, 338
Emitter crowding, 375, 375–376
bipolar junction transistors (BJTs), 374–376
Emitter Gummel number, 378
Emitter injection efficiency, 340, 391
Emitter junction, 338
Energy bands, 2, 64–67
with alloy composition, variation of, 71
separation of, 182
Energy gap, 67
Enhancement mode, 267
Enhancement-mode NMOSFET, 325

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 567

Enhancement-type n-channel MOSFET, 266
Epitaxial growth (epitaxy), 18–25
  lattice matching in, 18–21
  molecular beam epitaxy (MBE), 23–25
  vapor-phase epitaxy (VPE), 21–23
Equilibrium conditions, 168–179
  contact potential, 170–174
  equilibrium Fermi levels, 174–175
  space charge at a junction, 175–179
Equilibrium, defined, 130
Equilibrium Fermi levels, 108–110, 174–175
Equivalent circuit for the MOSFET, 313–315
Esaki diodes, *See* Tunnel diodes
Esaki, Leo, 508–509
Etching, 166–167
Excess carriers, 118–153
  defined, 119
  luminescence, 121–124
  optical absorption, 118–121
  quasi-Fermi levels, 130–132
  steady state carrier generation, 130–132
Expectation values, 40, 56
Experimental observations, 33–36
  atomic spectra, 35–36
  photoelectric effect, 33–35
Extreme ultraviolet (EUV) sources, 165
Extrinsic detectors, 407

F

Fabrication:
  of bipolar junction transistors (BJTs), 343–346
  of metal-insulator-semiconductor FET (MISFET), 266–270
  of p-n junctions, 154–168
Fabry–Perot interferometer, 421
Face-centered cubic (FCC) unit cell, 5
Fairchild Corporation, 441
Fall time, switching transistors, 368
Fan-out, 482
Fast interface state density, 288–289
Fat zeroes, 469
Fermi level, 87, 87–88
Fermi–Dirac distribution function, 86–88, 109
  applied to semiconductors, 88
Fermi–Dirac statistics, 86
Fiber optic communications, 414–418
Fiber-optic transmission system, 416–417
Field, 304
Field ionization, 198
Field region, 270
Field-effect transistors (FETs), 251–335
  amplification, 252, 254
  curve tracer, 252
  defined, 251
  forms of, 251
  history of, 252
  junction FETs (JFETs), 251, 254–265
    current-voltage characteristics, 257–259
    gate control, 257–259
    gradual channel approximation, 259
    mutual transconductance, 260
    pinchoff, 254–256
    saturation, 254–256
  load line, 252–254
  metal-insulator-semiconductor (MISFET):
    basic operation and fabrication, 266–270
    current voltage characteristics of MOS gate oxides, 291–294
    ideal MOS capacitor, 270–281
    MOS capacitance-voltage analysis, 286–290
    real surfaces, effects of, 281–284
    threshold voltage, 284–286
    time-dependent capacitance measurements, 290–291
  metal-insulator-semiconductor (MISFET), 265–294
  metal-semiconductor FET (MESFET), 251, 261–265
    GaAs MESFET, 261–262
    high electron mobility transistor (HEMT), 262–264
    short channel effects, 264–265
  MOS field-effect transistor (MOSFET), 265, 294–325
    Drain-Induced Barrier Lowering (DIBL), 320–322
    equivalent circuit for the MOSFET, 313–315
    gate-induced drain leakage, 323–325

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 568

mobility models, 300–302
output characteristics, 295–297
scaling and hot electron effects, 315–320
short channel and narrow width effect, 322–323
short channel MOSFET I-V characteristics, 303
substrate bias effects, 309–311
subthreshold characteristics, 311–313
threshold voltage, control of, 304–309
transfer characteristics, 297–300
operation, 252–254
switching, 252, 254
Flash memories, 492–496
Flash memory cell, 483
drain current vs. control gate voltage characteristics of, 496
hot carrier programming of, 494–495
structure, 493
Flat band, 273
Flat-band conditions, 282, 325
Flip-chip techniques, 501–502
Fluorescence, 122
Forbidden band, 67
Forward resistance, 202
Forward-biased junctions, 180–196
Forward-blocking state, 520
p-n-p-n diodes, 523–524
Forward-conducting state, 520
Fowler–Nordheim tunneling, 291–293, 494
Fractional distillation, 13
Fully depleted device, 460

G

GaAs, band structure of, 69
GaAs MESFET, 261–262
GaAsP, 2
Gain–bandwidth product, 407
Gain-enhanced MOSFET (GEMFET), 528
Gas-source MBE, 25
Gate control, 257–259
Gates, 251, 254
Gate-triggering mechanism:
semiconductor-controlled rectifier, 527
three-terminal SCR, 526
Gate-tunneling leakage currents, 326
Gaussian distribution, 145
Generalized biasing:
bipolar junction transistors (BJTs), 355–363
charge control analysis, 361–363
couple-diode model, 356–361
Generation current, 183
Generation–recombination (G–R) of electron–hole pairs (EHPs), 118, 121–131, 133, 140–148, 240
Germer, L. H., 34
Gigascale integration (GLSI), 441
Graded index separate confinement heterostructure (GRINSCH), 428
Graded junctions, 169, 224–227, 240
transition region, properties of, 226
Gradient, 134
Gradient of the carrier concentration, 190–191
Gradual channel approximation, 259
Gross world product (GWP), 443
Ground state, 50
Guard ring, 203
Gummel plot, 379
Gummel-Poon model, 376–380, 391
Gunn diodes, 508, 515–520
space charge domains, formation/drift of, 518–520
transferred-electron mechanism, 515–517
Gunn, J. B., 515

H

Half-rectified sine wave, 201
Hall coefficient, 107–108
Hall effect, 106–110
defined, 107
Hall voltage, 107–108
Halo implants, 319–320
Haynes, J. R., 144
Haynes-Shockley experiment, 144–147
HBTs, *See* Heterojunction bipolar transistors (HBTs)
Heavy hole band, 79
Heisenberg, 39, 56

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 569

Heisenberg uncertainty principle, 40
Heisenberg's matrix mechanics, 47
HEMT, *See* High electron mobility transistor (HEMT)
Hess, K., 215
Heteroepitaxy, 18
Heterojunction bipolar transistors (HBTs), 388–390, 391
Heterojunction lasers, 426–430, 427
Heterojunctions, 86, 234–239, 241
High electron mobility transistor (HEMT), 262–264
  advantages of, 263
  pseudomorphic HEMTs, 264
  two-dimensional electron gas FET (2-DEG FET/TEGFET), 264
High-field effects, 105–106
High-frequency transistors, bipolar junction transistors (BJTs), 386–388
High-frequency/high-power devices, 508–534
  Gunn diodes, 515–520
    space charge domains, formation/drift of, 518–520
    transferred-electron mechanism, 515–517
  impact avalanche transit-time (IMPATT) diode, 512–515
    defined, 512
    Read diode, 514–515
  insulated-gate bipolar transistor (IGBT), 528–531
    basic structure, 528–529
    current–voltage characteristics, 529–530
    input impedance/input capacitance, 531
    n^- region, 529
    resistance/current-handling capability, 531
  p-n-p-n diodes, 520–526
    basic structure, 520–521
    conducting state, 524–525
    forward-blocking state, 523–524
    triggering mechanism, 525–526
    two-transistor analogy, 521–522
    variation of α with injection, 522–523
  semiconductor-controlled rectifier, 526–528
    gate-triggering mechanism, 527
    turning off, 527–528
  tunnel diodes, 508–512
    defined, 508–509
    degenerate semiconductors, 509–512
    negative resistance of, 512
    total tunnel diode characteristic, 511
High-level carrier injection, 240
High-low junction, 245
Hilsum, C., 517
Holding current, 527
Hole(s), 73–76
  accumulation, 272
  direct recombination of, 125–127
  gas, *See* Two-dimensional electron gas (2DEG)
  injection device, 337
Homojunction lasers, 427
Homojunctions, 234, 388
Hot carrier effect, 106
Hot electron effects, 315, 326
Huygens, Christiaan, 33
Hybrid circuits, 440–441
Hybrip-pi model, 384
Hydrogen atom, 47–50
Hyperabrupt junctions, 217

I

Ideal diode, 200
Ideal MOS capacitor, 270–281
Ideal Shockley diode, 240
Ideality factor, 221
IGBT, *See* Insulated-gate bipolar transistor (IGBT)
IGFET, *See* Insulated-gate field effect transistor (IGFET)
III—V compounds, 2
Impact avalanche transit-time (IMPATT) diode, 512–515
  defined, 512
  Read diode, 514–515
Impact ionization, 198, 204, 240
IMPATT diode, *See* Impact avalanche transit-time (IMPATT) diode

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 570

Impurities:
amphoteric, 84
donor, 82
Impurity scattering, 103
IMREF, 132
Index of refraction, 414
Indirect semiconductors, 68–70
Injection electroluminescence, 124, 411
Input impedance, 252, 325
Insulated-gate bipolar transistor (IGBT), 528–531
basic structure, 528–529
current–voltage characteristics, 529–530
input impedance/input capacitance, 531
n^- region, 529–530
resistance/current-handling capability, 531
Insulated-gate field effect transistor (IGFET), 265, 315
Insulated-gate rectifier (IGR), 528
Insulated-gate transistor (IGT), 528
Insulators, 67–68
Integrated circuits, 437–507
advantages of, in miniaturization, 438–440
applications of, 438–439
application-specific ICs (ASICs), 444, 446
background, 438–441
batch fabrication, 438–439
bonding:
ball bond, 500–501
die bonding, 497
flip-chip techniques, 501–502
lead bond, 497
nail-head bond, 500
thermocompression bonding, 500
ultrasonic bonding, 500
wedge bond, 500–501
wire bonding, 497–501
charge transfer devices, 466–470
applications of CCDs, 470
basic CCD, 468
basic structure improvements, 469–470
charge-coupled device (CCD), 466
dynamic effects in MOS capacitors, 466–467
complementary MOS (CMOS), 444–445
design/lithography, 471–472
digital circuits, 440
evolution of, 441–444
gigascale integration (GLSI), 441
hybrid circuits, 440–441
large-scale integration (LSI), 439, 441
linear IC, 440
medium-scale integration (MSI), 441
monolithic circuits, 440–441
monolithic device elements, 444–466
CMOS process integration, 444–458
integration of other circuit elements, 461–466
silicon-on-insulator (SOI), 459–460
Moore's law, 441–442
packaging, 502–505
hermetically sealed device, 502–505
output leads, 503–504
surface-mount packaging, 503
through-hold mount, 503–504
small-scale integration (SSI), 441
testing, 497
types of, 440–441
ultra large-scale integration (ULSI), 441, 466–470, 472
logic devices, 474–483
semiconductor memories, 483–496
very large-scale integration (LSI), 441
wafers, 438–440
Integration, 437
advantages of, 438–440
large-scale, 439
Interdigitated geometry, 376
Interface charge density, 283–284
Interface states, 283
International Technology Roadmap for Semiconductors (ITRS), 442
Intrinsic detectors, 406–407
Intrinsic semiconductor, 80
Inversion, 273–274
population, 420
at a junction, 422–424
strong, 273–274
Inversion region, 422
Inverted mode, 356
Inverters, voltage transfer characteristic (VTC), 475

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 571

Ion implantation, 159–162
Ionic bonding, 62–63, 110
Ionization region, 96
Isolation region, 270
Isotropic, use of term, 166

J

JFET, *See* Junction FET (JFET)
Johnson noise, 409
Junction capacitance, 212–213
Junction FET (JFET), 251, 254–265
  current-voltage characteristics, 257–259
  gate control, 257–259
  gradual channel approximation, 259
  mutual transconductance, 260
  pinchoff, 254–256
  saturation, 254–256
Junction spiking problem, 456
Junctionless device, 460
Junctions, 148, 154–250
  carrier injection, 184–193
  current flow at a junction, qualitative description of, 180–184
  depletion capacitance of, 214
  deviations from the simple theory, 218–227
    carrier injection, effects of contact potential on, 219–220
    graded junctions, 224–227
    ohmic losses, 223–224
    recombination and generation in the transition region, 221–223
  graded, 224–227, 240
  heterojunctions, 234–239
  hyperabrupt, 217
  metal–semiconductor junctions, 224–234
    ohmic contacts, 231–232
    rectifying contacts, 229–231
    Schottky barriers, 227–229, 233–234
  reverse bias, 193–196
  transient and a-c conditions, 205–218
    capacitance of p-n junctions, 212–217
    reverse recovery transient, 208–211
    switching diodes, 211–212
    time variation of stored charge, 205–208
    varactor diode, 217–218

K

Kahng, D., 252
Kilby, Jack, 441
Kirk effect, 380–382, 391
Known good dies, 458
Kooi effect, 449–450

L

Large-scale integration (LSI), 439, 441
Lasers, 416, 418–422
  coherent radiation, 418–419
  defined, 418
  Einstein coefficients, 420
  monochromatic radiation, 418
  negative temperature, 420
  optical resonant cavity, 420
  population inversion, 420
  semiconductor lasers, 422–432
    basic, 425–426
    emission spectra for p-n junction lasers, 424–425
    heterojunction lasers, 426–430
    materials for, 430–432
    population inversion at a junction, 422–424
  stimulated emission rate, 420
  thermal equilibrium, 419–421
Latchup, 446
Lateral moat encroachment, 449
Lattice constant, 6
Lattice matching, in epitaxial growth, 18–21
Lattice scattering, 103
Lattices:
  body-centered cubic (BCC), 5
  Bravais, 13
  centered rectangular, 5
  crystal, 3–12
  cubic, 5–7
  defined, 3
  diamond, 10–12
  direction in, 8
  rhombic, 4
  strained-layer superlattice (SLS), 20
  sublattices, 10
  Wurtzite, 11

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 572

Laux, S., 215
Lead bonder, 497
LEDs, *See* Light-emitting diodes (LEDs)
Light, 31, 33–35, 39, 56
Light hole band, 79
Light-emitting diodes (LEDs), 2–3, 398, 411–418
  chromatic dispersion, 416
  fiber optic communications, 414–418
  injection electroluminescence, 411
  light-emitting materials, 411–414
  optical fiber, 414
  optoelectronic isolator, 413
  optoelectronic pair, 413
  pulse dispersion, 416
  Rayleigh scattering, 415–416
  semiconductor laser, 411
  step-index fiber, 414
Light-emitting materials, 411–414
Lightly doped drain (LDD) implant, 451–453
  structure, 452
Lightly doped drains (LDDs), 316, 326
Lilienfeld, Julius Edgar, 252
Linear combinations of atomic orbitals (LCAO), 55, 64–65
Linear integrated circuits, 440
Liquid-encapsulated Czochralski (LEC) growth, 15
Liquid-phase epitaxy (LPE), 18
Load line, 252–254
Local oxidation of silicon (LOCOS), 270, 314, 449
Long diodes, 216
Low pressure chemical vapor deposition (LPCVD), 162, 345, 447–450
Luminescence, 121–124
  cathodoluminescence, 121
  defined, 121
  electroluminescence, 121, 124
  photoluminescence, 121, 122–124
Lyman, Balmer, and Paschen series, 35, 38

M

Majority carriers, 85, 230–231
  mobility, 145
Mask, 163
Mask aligners, 165
Mass separator, 159
Matrix mechanics, 39
Matter, 33, 34, 56
Maxwell–Boltzmann statistics for classical particles, 86
MBE, 429, 430
Mean free time, 99
Medium-scale integration (MSI), 441
MESFET, *See* Metal-semiconductor FET (MESFET)
Metal-insulator-semiconductor FET (MISFET), 251
  basic operation and fabrication, 266–270
  current voltage characteristics of MOS gate oxides, 291–294
  ideal MOS capacitor, 270–281
  MOS capacitance-voltage analysis, 286–290
  real surfaces, effects of, 281–284
  threshold voltage, 284–286
  time-dependent capacitance measurements, 290–291
Metallic bonding, 63
Metallization, 167–168, 270
Metal-organic vapor-phase epitaxy (MOVPE), 23
Metals, 67–68
Metal-semiconductor FET (MESFET), 251, 261–265
  GaAs MESFET, 261–262
  high electron mobility transistor (HEMT), 262–264
  short channel effects, 264–265
Metal–semiconductor junctions, 224–234
  ohmic contacts, 231–232
  rectifying contacts, 229–231
  Schottky barriers, 227–229, 233–234
mhos, 260
Microprocessor, 441, 472
Miller indices, 8–10
Miller overlap capacitance, 313–314, 451
Miniaturization, advantages of, 439
Minimum, conduction band, 515–516

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 573

Minority carrier distributions/terminal currents:
bipolar junction transistors (BJTs), 346–355
approximations of the terminal currents, 352–354
current transfer ratio, 354–355
evaluation of the terminal currents, 349–352
solution of the diffusion equation in the base region, 347–349
Minority carriers, 85, 231
extraction, 185, 193
lifetime, 126
mobility, 145
MISFET, *See* Metal-insulator-semiconductor (MISFET)
Misfit dislocations, 20
Mobile ion determination, 290
Mobility:
of electrons, 98, 101–102
of holes, 102
majority carriers, 145
minority carriers, 145
units of, 101
Mobility degradation parameter, 302
Mobility models, 300–302
MOCVD, 431
Modified work function, 270
Modulation doped field-effect transistor (MODFET), 262
Modulation doping, 262–263
Molecular beam epitaxy (MBE), 18, 23–25
Moll, J. L., 358
Monochromatic radiation, 418
Monolithic circuits, 440–441
Monolithic device elements, 444–466
CMOS process integration, 444–458
integration of other circuit elements, 461–466
capacitors, 462
contacts and interconnections, 463–464
diodes, 461
inductors, 462–463
resistors, 461–462
silicon-on-insulator (SOI), 459–460
Moore, Gordon, 441
Moore's law, 441–442
MOS field-effect transistor (MOSFET), 265, 294–325
cross section of, 305
drain-induced barrier lowering (DIBL), 320–322
equivalent circuit for, 313–315
gate-induced drain leakage, 323–325
hot carrier degradation in, 318–319
mobility models, 300–302
narrow width effect in, 323–324
output characteristics, 295–297
scaling and hot electron effects, 315–320
scaling down, 315
short channel and narrow width effect, 322–323
short channel effects in, 317
short channel MOSFET I-V characteristics, 303
substrate bias effects, 309–311
substrate current in, 319–320
subthreshold characteristics, 311–313
threshold voltage, 304–309
threshold voltage, control of, 304–309
transfer characteristics, 297–300
MOS gate oxides, current voltage characteristics of, 291–294
MOS pass transistor, 486
MOVPE, 430
Multi-chip module (MCM), 504–505
Multiplexing:
address, 484
ultradense-wavelength-division multiplexing (UDWDM), 484
Mutual transconductance, 260

N

N+ doped LPCVD polysilicon refractory (high melting point) gates, 304
Nail-head bond, 500
Narrow base diode, 208, 212, 246
Narrow width effect, 323
MOS field-effect transistor, 322–323
n-channel MOSFET, 267–269

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 574

Negative conductance, 508
Negative effective mass, 77
Negative resist, 163
Negative temperature, 420
Net drift, 100
Newton, Isaac, 34
Nichia Corporation (Japan), 430
Noise immunity, 477
Noise-equivalent power (NEP), 409
Normal active mode, 356
Noyce, Robert, 441
n-p-n transistors, 336–339, 387–388
Nucleation site, 519

O

Offset voltage, 200
Ohmic contacts, 241
  metal–semiconductor junctions, 231–232
Ohmic metal–semiconductor contact, 231
Ohm's law, 100
  modified, 148
OMVPE, 429
"On" state/"off" state, 363
Optical absorption, 118–122
  absorption coefficient, 120
Optical fiber, 414
Optical resonant cavity, 420
Optoelectronic devices, 398–436
  lasers, 416, 418–422
    coherent radiation, 418–419
    Einstein coefficients, 420
    monochromatic radiation, 418
    negative temperature, 420
    optical resonant cavity, 420
    population inversion, 420
    semiconductor lasers, 422–432
    stimulated emission rate, 420
    thermal equilibrium, 419–421
  light-emitting diodes (LEDs), 411–418
    chromatic dispersion, 416
    fiber optic communications, 414–418
    injection electroluminescence, 411
    light-emitting materials, 411–414
    optical fiber, 414
    optoelectronic isolator, 413
    optoelectronic pair, 413
    pulse dispersion, 416
    Rayleigh scattering, 415–416
    semiconductor laser, 411
    step-index fiber, 414
  photodetectors:
    multimode fibers, 416
    single-mode fibers, 416
  photodiodes, 398–411
    avalanche photodiodes (APDs), 406, 409
    current and voltage in an illuminated junction, 399–402
    equilibrium thermal generation—recombination rate, 400
    extrinsic detectors, 407
    fill factor, 404
    intrinsic detectors, 406–407
    optical energy, 402
    photodetectors, 405–407
    photodiode detector structures, 398
    photovoltaic effect, 400
    p-i-n photodetector, 406
    solar cells, 398, 402–405
  semiconductor lasers, 411, 422–432
    basic, 425–426
    emission spectra for p-n junction lasers, 424–425
    heterojunction lasers, 426–430
    materials for, 430–432
    population inversion at a junction, 422–424
    UV/blue semiconductor lasers, 431–432
Optoelectronic isolator, 413
Optoelectronic pair, 413
Optoelectronics, defined, 398
Orbital:
  anti-bonding, 64
  bonding, 64
  linear combinations of atomic orbitals (LCAO), 55, 64–65
Organometallic vapor-phase epitaxy (OMVPE), 23
Oversaturation, 366
Oxide-isolated gate (MOS capacitor), 325

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 575

P

p⁺-n junction, 189
Packaging:
  integrated circuits, 502–505
    hermetically sealed device, 502–505
    output leads, 503–504
    surface-mount packaging, 503
    through-hole mount, 503–504
Pankove, J. I., 431
Parasitic bipolar structures, 446
"Particle in a box" problem, 45, 53
Particles, 32, 34, 40, 43, 47, 56
Pauli exclusion principle, 50, 54, 56, 63, 65–66
p-channel MOS transistors, 269
Peak tunneling current, 511–512
Periodic structures, 3–5
Periodic table, 50–53, 64
  and the atomic structure, 47–56
Phonon scattering, 103
Phonons, 103, 239
Phosphorescence, 122
Phosphors, 122
Phosphorus dopant source, 450
Photoconductive cell, time response of, 133–134
Photoconductive decay, 129
Photoconductivity:
  and carrier lifetime, 124–134
  defined, 124
  devices, 132–134
  direct recombination of electrons and holes, 125–127
  indirect recombination, 127–130
  minority carrier lifetime, 126
  photoconductive decay, 129
  recombination lifetime, 126
  trapping, 127–130
Photoconductors:
  optical sensitivity of, 133
    evaluating, 133
Photodetectors:
  avalanche photodiodes (APDs), 406, 409
  extrinsic detectors, 407
  gain, bandwidth, and signal-to-noise ratio of, 407–411
  intrinsic detectors, 406–407
  multimode fibers, 416
  p-i-n photodetector, 406
  shot noise, 409
  single-mode fibers, 416
Photodiodes, 398–411
  current and voltage in an illuminated junction, 399–402
  equilibrium thermal generation–recombination rate, 400
  fill factor, 404
  optical energy, 402
  photodetectors, 405–407
  photodiode detector structures, 398
  photovoltaic effect, 400
  solar cells, 398, 402–405
  waveguide photodiode, 410
Photoelectric effect, 33–35
  defined, 33
Photolithography, 162–166
Photoluminescence, 121, 122–124
Photon energy, 35
Photons, 34
Photoresist, 163
Photovoltaic technology, cost/scalability of, and terrestrial applications, 405
Physical models, 32–33
Piecewise-linear equivalents, 201
Pilot waves, 37
Pin grid arrays (PGAs), 504
p-i-n photodetector, 406
Pinchoff, 254–256
Pinch-off, 265
Planarization, 458
Planck, Max Karl Ernst Ludwig, 33–35, 56
Planck's constant, 37, 40
Plasma-enhanced CVD (PECVD), 162
p-n junction lasers, emission spectra for, 424–425
p-n junction photodiode, 183
p-n junctions:
  capacitance of, 212–217
  chemical vapor deposition (CVD), 162
  diffusion, 156–158
  etching, 166–167
  fabrication of, 154–168
  ion implantation, 159–162
  metallization, 167–168

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 576

photolithography, 162–166
rapid thermal processing, 158–159
thermal oxidation, 155–156
p-n-p bipolar junction transistors, 336–339, 387–388
p-n-p-n diodes, 520–526
basic structure, 520–521
conducting state, 524–525
forward-blocking state, 523–524
triggering mechanism, 525–526
two-transistor analogy, 521–522
variation of α with injection, 522–523
Pocket implants, 319–320
Point contact transistor, 343
Poisson's equation, 176
Poly depletion effect, 451
Polycide, 456
Polycrystalline solids, 3
Polysilicon emitter (polyemitter), 346
Population inversion, 420
Positive resist, 163
Potential well, 467
compared to depletion region, 467
Potential well problem, 44–46, 85–86
Power gain, 325
Primitive cell, 4
Primitive vectors, 4
Principal quantum number, 50
Principle of indeterminacy, 40
Printed circuit board (PCB), 504
Probability density function, 40–41
Projected range, 159
Propagation delay, 483
Pseudomorphic HEMTs, 264
Pseudomorphic layer, 19–20
Pulse dispersion, 416
Punch-through, 202

Q

Quanta, 33
Quantization, 33
Quantum mechanics, 32, 39–47, 56
potential well problem, 44–46
probability and the uncertainty principle, 40–41
Schrödinger wave equation, 41–43
tunneling, 46–47
Quantum number, 45
Quantum state, 45, 50–51
Quasi-Fermi levels, 130–132, 148, 423
gradients in, 147–148
Quasi-steady state approximation, 207
Quaternary compounds, 2

R

Radiation:
blackbody, 33, 56
coherent radiation, 418–419
monochromatic, 418
stimulated emission of, 418
Radicals, 167
Range, 61
Rapid thermal anneal (RTA), 453
Rapid thermal processing (RTP), 158–159
Rayleigh scattering, 415–416
Reaction chamber, 23
Reactive ion etching (RIE), 166, 448, 464–465
Reactor, 23
Read diode, 514–515
Real surfaces, effects, 281–284
Recombination kinetics, 342
Recombination levels, 127
Recombination lifetime, 126
Recombination of EHPs, 81
Rectification, 201
Rectifiers, 180, 200–204
Rectifying contacts, metal–semiconductor junctions, 229–231
Red, yellow and green LEDs, 431
Reference diode, 205
Refractive index, 414
Resonant tunneling diode, 47
Reverse bias, 193–196
Reverse saturation current, 184
Reverse short channel effect (RSCE), 323
Reverse-bias breakdown, 196–205
avalanche breakdown, 198–200
breakdown diode, 204–205
rectifiers, 200–204
Zener breakdown, 197–198

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 577

Reverse-biased junctions, 180–196
Reverse-blocking state, p-n-p-n diodes, 521
Rhombic lattice, 4
Ridley, B. K., 517
Ridley–Watkins–Hilsum mechanism, 517
Rise time, switching transistors, 368
Ritz combination principle, 36
Roll-off, 323
Row decoders, 484
Rydberg constant, 36, 39

S

SACM APD, 407
Satellite valleys, 515–516
Saturation, 256
Saturation regime, 363
Saturation region, 269
Scattering limited velocity, 106
Schottky barrier diodes, 227, 230–231
Schottky barrier junction, 233
Schottky barriers, 227–229
  typical, 233–234
Schottky diode, 241, 325
Schottky effect, 227
Schrödinger equation, 41–43, 47, 61–62
Schrödinger, Erwin, 39, 56
Selectivity, 166
Self-aligned gate, 313
Self-aligned implant mask, 448
Self-aligned process, 269–270
Self-aligned silicides (SALICIDE), 326, 446
Self-alignment, defined, 346
Semiconductor lasers, 411, 422–432
  basic, 425–426
  emission spectra for p-n junction lasers, 424–425
  heterojunction lasers, 426–430
  materials for, 430–432
  population inversion at a junction, 422–424
  UV/blue semiconductor lasers, 431–432
Semiconductor memories, 483–496
  dynamic random access memory (DRAM), 486–492
  flash memories, 492–496
  static random access memory (SRAM), 485–486
Semiconductor-controlled rectifier (SCR), 520, 526–528
  gate-triggering mechanism, 527
  three-terminal SCR, 526
  turning off, 527–528
Semiconductors, 1–3, 67–68
  charge carriers in, 71–86
    carrier concentrations, 86–108
    effective mass, 76–80
    electrons and holes, 73–76
    extrinsic material, 81–85
    intrinsic material, 80–81
  common materials, 2
  compound, 1, 1–2, 64
  defined, 1
  degenerate, 509–512
  device development, 416
  direct, 68–70
  electronic and optical properties of, 3
  elemental, 1
  excess carriers in, 118–153
  indirect, 68–70
  intrinsic, 80
  metal–semiconductor junctions, 224–234
Semi-insulating GaAs, 261
Separate confinement, 428
Separately doped FET (SEDFET), 264
Separation by implantation of oxygen (SIMOX), 459
Separation of the absorption and multiplication (SAM) APDs, 407
Series resistance, 240
Shallow trench isolation (STI), 449, 458
Sheath regions, 167
Sheet resistance, 462
Shell, 50
Shockley diode, 224, 521
Shockley ideal diode approximation, 187
Shockley, W., 142–144, 251–252
Short channel effect (SCE), 264–265, 322
  MOS field-effect transistor, 322–323
Short channel MOSFET I-V characteristics, 303

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 578

Short diodes, 216
Shot noise, 409
Sidewall oxide spacers, 453
Siemens (S), 260
Signal-to-noise ratio, photodetectors, 409
Silicon band structure, 67
Silicon dioxide (SiO_2), 15
Silicon-on-insulator (SOI), 459–460
Simple cubic structure, 5
Simulation Program with Integrated Circuit Emphasis (SPICE), 483
Single crystal ingots, growth, 13–15
Six-transistor cell, 486
Small-scale integration (SSI), 441
Smith, G. E., 468
Soft errors, 490
Solar cells, 398, 402–405
  applications of, 404
Solids:
  amorphous, 3
  bonding forces in, 62–64
  energy bands in, 61–71
  polycrystalline, 3
Source, 254
Source/drain extension or tip, 453
Source-injection regime, 326
Space charge at a junction, 175–179
Space charge fluctuation build-up, 518
Space charge neutrality, 97
Space charge region, 172
Spacers, oxide, 345
Spherical coordinate system, 47–48
Split-off band, 79
Spontaneous emission, 418
Stable domains, 519–520
"Staircase" constant density of states, 277
Static random access memory (SRAM), 483, 485–486
Steady state carrier generation, 130–132
Steady state carrier injection, 142–144
Steady state, defined, 130
Steady-state conditions, 180–196
Step function, 169
Step-index fiber, 414
Stepper, 164
Stimulated emission of radiation, 418
Stimulated emission rate, 420
Storage delay time, 210
Straggle, 159
Strained-layer superlattice (SLS), 20
Strong inversion, 273–274
Sub-collector, 345
Sublattices, 10
Subthreshold conduction, 311
Subthreshold slope, 311
Surface electron mobility, 295
Surface field-effect transistor, *See* MOS field-effect transistor (MOSFET)
Surface roughness scattering, 326
Surface states, 233
Surface-mount packaging, 503
Surface-mounted ball grid arrays (BGAs), 504
Switching, 508
  bipolar junction transistors (BJTs), 363, 363–368
    saturation, 365–366
    specifications for switching transistors, 368
    switching cycle, 366–368
  field-effect transistors (FETs), 252, 254
Switching current, 481–482
Symmetrical transistor, 357

T

Tanks, 446
Tape-automated bonding (TAB), 505
Target chamber, 159
Temperature dependence of carrier concentrations, 94–96
Temperature, effect on mobility, 103–105
Temporary trapping, 128
Ternary compounds, 2
Tetraethyl-ortho-silicate (TEOS), 453
Texas Instruments, 441
Thermal budget, 156
Thermal equilibrium, 419–421
Thermal oxidation, 155–156
Thermal relaxation time, 467
Thermal runaway, 374
Thermionic emission, 234

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 579

Thermocompression bonding, 500
Three-terminal active devices, 325
Threshold current, 427
Threshold voltage, 279, 325
  MOSFET, 267, 304–309
Through-hole mount, 503–504
Through-hole–mounted pin grid arrays (PGAs), 504
Thyristor, 446
Time-dependent capacitance measurements, 290–291
Time-dependent dielectric breakdown (TDDB), 293–294
Total current, 183
Transconductance:
  a-c, 260
  mutual, 384
Transit time, 342
Transit time effects, bipolar junction transistors (BJTs), 385–386
Transition region, 172
Transition region width, 181
Trapping center, 128
Trapping levels, 128–129
Traps, 148
Tubs, 446
Tunnel diodes, 508–512
  defined, 508–509
  degenerate semiconductors, 509–512
  negative resistance of, 512
  total tunnel diode characteristic, 511
Tunneling, 46–47, 240
Twin-well CMOS, 446–447
Two-dimensional electron gas (2-DEG), 238, 263, 277
Two-dimensional electron gas FET (2-DEG FET/TEGFET), 264
Two-transistor analogy, 521–522
Type-N negative resistance, 511

U

Ultra large-scale integration (ULSI), 441, 441–443, 466–470, 472
  economic payoff of, 443–444
  logic devices, 474–483
  semiconductor memories, 483–496
Ultradense-wavelength-division multiplexing (UDWDM), 416
Ultrasonic bonding, 500
Unipolar transistor, 251
Unit cell, importance of, 5
UV/blue semiconductor lasers, 431–432

V

Valence band, 67
Valley current, 511–512
Vapor-phase epitaxy (VPE), 21–23
Varactor, 215, 217–218
Varactor diode, 217–218
Vegard's law, 18
Velocity-saturated region, 326
Vertical cavity surface-emitting lasers (VCSELs), 428–430
Very large-scale integration (LSI), 441
Vias, 458
Voltage regulators, 205
Voltage transfer characteristic (VTC), 475
Voltage triggering, 525
Voltage-controlled negative resistance, 511
Voltage-independent reverse current, 240
Voltage-variable capacitance, 215

W

Wafers, 15–16
Watkins, T. B., 517
Wave mechanics, 39
Wave nature of light, 39, 56
Wave vector, 69
Waveguide photodiode, 410
Wave-particle duality, 34, 56
Waveshaping, 202
Webster effect, bipolar junction transistors (BJTs), 386
Wedge bond, 500–501
Wells, 446
White ribbon effect, 449–450
Word line, 462

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 580

Wordlines, 484–485
Words, 485
Work function, 34
Wurtzite lattice, 11

X

Xenon, 130
X-ray crystallography, and identification of crystal planes, 15

Y

Yield, 440

Z

Zener breakdown, 197–198
Zener diodes, 204
Zener effect (tunneling), 197–198, 204, 224, 508
Zincblende structure, 10–11

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 581

SEMICONDUCTOR PHYSICS

Electron Momentum: $\mathrm{p} = m^*\mathrm{v} = \hbar\mathbf{k} = \dfrac{h}{\lambda}$ Planck: $E = h\nu = \hbar\omega$

Kinetic: $E = \dfrac{1}{2}m^*\mathrm{v}^2 = \dfrac{1}{2}\dfrac{\mathrm{p}^2}{m^*} = \dfrac{\hbar^2}{2m^*}\mathbf{k}^2$ (3-4) Effective mass: $m^* = \dfrac{\hbar^2}{d^2E/d\mathbf{k}^2}$ (3-3)

Total electron energy = $P.E. + K.E. = E_c + E(\mathbf{k})$

Fermi-Dirac e^- distribution: $f(E) = \dfrac{1}{e^{(E-E_F)/kT}+1} \cong e^{(E_F-E)/kT}$ for $E \gg E_F$ (3-10)

Equilibrium: $n_0 = \displaystyle\int_{E_c}^{\infty} f(E)N(E)dE = N_c f(E_c) = N_c e^{-(E_c-E_F)/kT}$ (3-15)

$N_c = 2\left(\dfrac{2\pi m_n^* kT}{h^2}\right)^{3/2}$ $N_v = 2\left(\dfrac{2\pi m_p^* kT}{h^2}\right)^{3/2}$ (3-16), (3-20)

$p_0 = N_v[1 - f(E_v)] = N_v e^{-(E_F-E_v)/kT}$ (3-19)

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

$n_i = \sqrt{N_c N_v}\, e^{-E_g/2kT} = 2\left(\dfrac{2\pi kT}{h^2}\right)^{3/2}(m_n^* m_p^*)^{3/4} e^{-E_g/2kT}$ (3-23), (3-26)

Equilibrium: $n_0 = n_i e^{(E_F-E_i)/kT}$, $p_0 = n_i e^{(E_i-E_F)/kT}$ (3-25) $n_0 p_0 = n_i^2$ (3-24)

Steady state: $n = N_c e^{-(E_c-F_n)/kT} = n_i e^{(F_n-E_i)/kT}$, $p = N_v e^{-(F_p-E_v)/kT} = n_i e^{(E_i-F_p)/kT}$ (4-15) $np = n_i^2 e^{(F_n-F_p)/kT}$ (5-38)

$\mathscr{E}(x) = -\dfrac{d^{\circ}V(x)}{dx} = \dfrac{1}{q}\dfrac{dE_i}{dx}$ (4-26)

Poisson: $\dfrac{d\mathscr{E}(x)}{dx} = -\dfrac{d^2{}^{\circ}V(x)}{dx^2} = \dfrac{\rho(x)}{\epsilon} = \dfrac{q}{\epsilon}(p - n + N_d^+ - N_a^-)$ (5-14)

$\mu \equiv \dfrac{q\bar{t}}{m^*}$ (3-40a) Drift: $\mathrm{v}_d \cong \dfrac{\mu\mathscr{E}}{1 + \mu\mathscr{E}/\mathrm{v}_s}\begin{cases} = \mu\mathscr{E} \text{ (low fields, ohmic)} \\ = \mathrm{v}_s \text{ (high fields, saturated vel.)} \end{cases}$ (Fig. 6-9)

Drift current density: $\dfrac{I_x}{A} = J_x = q(n\mu_n + p\mu_p)\mathscr{E}_x = \sigma\mathscr{E}_x$ (3-43)

$$J_n(x) = q\mu_n n(x)\mathscr{E}(x) + qD_n \frac{dn(x)}{dx}$$

Conduction Current: drift diffusion (4–23)

$$J_p(x) = q\mu_p p(x)\mathscr{E}(x) - qD_p \frac{dp(x)}{dx}$$

$$J_{\text{total}} = J_{\text{conduction}} + J_{\text{displacement}} = J_n + J_p + C\frac{dV}{dt}$$

Continuity: $\frac{\partial p(x,t)}{\partial t} = \frac{\partial \delta p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} - \frac{\delta p}{\tau_p}$ $\frac{\partial \delta n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - \frac{\delta n}{\tau_n}$ (4–31)

For steady state diffusion: $\frac{d^2\delta n}{dx^2} = \frac{\delta n}{D_n\tau_n} \equiv \frac{\delta n}{L_n^2}$ $\frac{d^2\delta p}{dx^2} = \frac{\delta p}{L_p^2}$ (4–34)

Diffusion length: $L \equiv \sqrt{D\tau}$ Einstein relation: $\frac{D}{\mu} = \frac{kT}{q}$ (4–29)

p-n JUNCTIONS

Equilibrium: $V_0 = \frac{kT}{q}\ln\frac{p_p}{p_n} = \frac{\text{kT}}{\text{q}}\ln\frac{N_a}{n_i^2/N_d} = \frac{\text{kT}}{\text{q}}\ln\frac{N_aN_d}{n_i^2}$ (5–8)

$\frac{p_p}{p_n} = \frac{n_n}{n_p} = e^{qV_0/kT}$ (5–10) $W = \left[\frac{2\epsilon(V_0 - V)}{q}\left(\frac{N_a + N_d}{N_aN_d}\right)\right]^{1/2}$ (5–57)

One-sided abrupt p^+-n: $x_{n0} = \frac{WN_a}{N_a + N_d} \simeq W$ (5–23) $V_0 = \frac{qN_dW^2}{2\epsilon}$

$\Delta p_n = p(x_{n0}) - p_n = p_n(e^{qV/kT} - 1)$ (5–29)

$\delta p(x_n) = \Delta p_n e^{-x_n/L_p} = p_n(e^{qV/kT} - 1)e^{-x_n/L_p}$ (5–31b)

Ideal diode: $I = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)(e^{qV/kT} - 1) = I_0(e^{qV/kT} - 1)$ (5–36)

Non-ideal: $I = I_0'(e^{qV/\mathbf{n}kT} - 1)$ ($\mathbf{n}$ = 1 to 2) (5–74)

With light: $I_{\text{op}} = qAg_{\text{op}}(L_p + L_n + W)$ (8–1)

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 583

Capacitance: $C = \left|\frac{dQ}{dV}\right|$ (5-55)

Junction Depletion: $C_j = \epsilon A\left[\frac{q}{2\epsilon(V_0 - V)}\frac{N_d N_a}{N_d + N_a}\right]^{1/2} = \frac{\epsilon A}{W}$ (5-62)

Stored charge exp. hole dist.: $Q_p = qA\int_0^\infty \delta p(x_n)dx_n = qA\Delta p_n\int_0^\infty e^{-x_n/L_p}dx_n = qAL_p\Delta p_n$ (5-39)

$I_p(x_n = 0) = \frac{Q_p}{\tau_p} = qA\frac{L_p}{\tau_p}\Delta p_n = qA\frac{D_p}{L_p}p_n(e^{qV/kT} - 1)$ (5-40)

$G_s = \frac{dI}{dV} = \frac{qAL_p p_n}{\tau_p}\frac{d}{dV}(e^{qV/kT}) = \frac{q}{kT}I$ (5-67c)

Long p$^+$-n: $i(t) = \frac{Q_p(t)}{\tau_p} + \frac{dQ_p(t)}{dt}$ (5-47)

MOS-n CHANNEL

Oxide: $C_i = \frac{\epsilon_i}{d}$ Depletion: $C_d = \frac{\epsilon_s}{W}$ MOS: $C = \frac{C_i C_d}{C_i + C_d}$ (6-36)

Threshold: $V_T = \underbrace{\Phi_{ms} - \frac{Q_i}{C_i}}_{\text{Flat band}} - \frac{Q_d}{C_i} + 2\phi_F$ (6-38)

Inversion: $\phi_s\,(\text{inv.}) = 2\phi_F = 2\frac{kT}{q}\ln\frac{N_a}{n_i}$ (6-15) $W = \left[\frac{2\epsilon_s\phi_s}{qN_a}\right]^{1/2}$ (6-30)

$Q_d = -qN_aW_m = -2(\epsilon_s qN_a\phi_F)^{1/2}$ (6-32) At V_{FB}: $C_{FB} = \frac{C_i C_{debye}}{C_i + C_{debye}}$

Debye screening length: $L_D = \sqrt{\frac{\epsilon_s kT}{q^2 p_0}}$ (6-25) $C_{debye} = \frac{\epsilon_s}{L_D}$ (6-40)

Substrate bias: $\Delta V_T = \frac{\sqrt{2\epsilon_s qN_a}}{C_i}(-V_B)^{1/2}$ (n channel) (6-63)

of f download.ir

$$I_D = \frac{\bar{\mu}_n Z C_i}{L}\left\{(V_G - V_{FB} - 2\phi_F - \tfrac{1}{2}V_D)V_D - \frac{2}{3}\frac{\sqrt{2\epsilon_s q N_a}}{C_i}[(V_D + 2\phi_F)^{3/2} - (2\phi_F)^{3/2}]\right\} \quad (6\text{-}50)$$

$$I_D \simeq \frac{\bar{\mu}_n Z C_i}{L}[(V_G - V_T)V_D - \tfrac{1}{2}V_D^2] \quad (6\text{-}49)$$

Saturation: $I_D(\text{sat.}) \simeq \frac{1}{2}\bar{\mu}_n C_i \frac{Z}{L}(V_G - V_T)^2 = \frac{Z}{2L}\bar{\mu}_n C_i V_D^2(\text{sat.})$ (6-53)

$$g_m = \frac{\partial I_D}{\partial V_G} \quad : \quad g_m(\text{sat.}) = \frac{\partial I_D(\text{sat.})}{\partial V_G} \simeq \frac{Z}{L}\bar{\mu}_n C_i (V_G - V_T) \quad (6\text{-}54)$$

For short L: $I_D \simeq Z C_i (V_G - V_T) v_s$ (6-60)

Subthreshold slope: $S = \dfrac{dV_G}{d(\log I_D)} = \dfrac{kT}{q}\ln 10\left[1 + \dfrac{C_d + C_{it}}{C_i}\right]$ (6-66)

ВЛТ-p-n-p

$$I_{Ep} = qA\frac{D_p}{L_p}\left(\Delta p_E \operatorname{ctnh}\frac{W_b}{L_p} - \Delta p_C \operatorname{csch}\frac{W_b}{L_p}\right) \quad (7\text{-}18) \qquad \begin{aligned}\Delta p_E &= p_n(e^{qV_{EB}/kT} - 1)\\ \Delta p_C &= p_n(e^{qV_{CB}/kT} - 1)\end{aligned} \quad (7\text{-}8)$$

$$I_C = qA\frac{D_p}{L_p}\left(\Delta p_E \operatorname{csch}\frac{W_b}{L_p} - \Delta p_C \operatorname{ctnh}\frac{W_b}{L_p}\right)$$

$$I_B = qA\frac{D_p}{L_p}\left[(\Delta p_E + \Delta p_C)\tanh\frac{W_b}{2L_p}\right] \quad (7\text{-}19)$$

$$B = \frac{I_C}{I_{Ep}} \simeq \frac{\operatorname{csch} W_b/L_p}{\operatorname{ctnh} W_b/L_p} = \operatorname{sech}\frac{W_b}{L_p} \simeq 1 - \left(\frac{W_b^2}{2L_p^2}\right) \quad (7\text{-}26)$$

(Base transport factor)

$$\gamma = \frac{I_{Ep}}{I_{En} + I_{Ep}} = \left[1 + \frac{L_p^n n_n \mu_n^p}{L_n^p p_p \mu_p^n}\tanh\frac{W_b}{L_p^n}\right]^{-1} \simeq \left[1 + \frac{W_b n_n \mu_n^p}{L_n^p p_p \mu_p^n}\right]^{-1} \quad (7\text{-}25)$$

(Emitter injection efficiency)

$$\frac{i_C}{i_E} = B\gamma \equiv \alpha \quad (7\text{-}3) \qquad \frac{i_C}{i_B} = \frac{B\gamma}{1 - B\gamma} = \frac{\alpha}{1 - \alpha} \equiv \beta \quad (7\text{-}6) \qquad \frac{i_C}{i_B} = \beta = \frac{\tau_p}{\tau_t} \quad (7\text{-}7)$$

(Common base gain) (Common emitter gain) (For $\gamma = 1$)

SOLID STATE ELECTRONIC DEVICES

SIXTH EDITION



BEN G. STREETMAN
SANJAY KUMAR BANERJEE
both of University of Texas at Austin

This classic introductory textbook in its sixth edition retains the two basic goals: (1) develop the basic semiconductor physics concepts to understand current and future devices and (2) provides a sound understanding of current semiconductor devices and technology so that their applications to electronic and optoelectronic circuits and systems can be appreciated.

The book is useful for advanced undergraduate and postgraduate students of electrical and electronics engineering and should also prove useful for practicing engineers and scientists, interested in updating their understanding of modern electronics.

Key Features

- The basics of semiconductor materials and conduction processes in solids. Incorporated to help students understand p-n junctions, bipolar and metal oxide semiconductor transistors, optoelectronic and other devices.
- Extensive discussion of circuit and other application examples to provide students with feedback about the practical relevance of the theory.
- Key equations for semiconductor devices are listed on the inside covers of the book, arranged thematically to help students digest concepts.
- Provides a summary of key concepts at end of each chapter, and chapter objectives at the beginning
- Approximately 200 problems along with current references that extend concepts in the text.

New to this Edition

- Discussion of advanced MOS and bipolar processes and structures such as high-k gate dielectrics, polyemitter BJTs and graded SiGe base HBTs.
- Consolidated treatment of optoelectronic devices, discusses sources such as white LEDs for *illumination, detectors such as SACM and waveguide photodiodes*, and advances in solar cells.
- Updated discussion of device fabrication processes and CMOS integrated circuit technology provides a useful understanding of how semiconductor devices are made.
- Summary of key concepts at end of each chapter.
- Answers to Selected Problems in the back of the book.

ISBN:978-81-203-3020-7
9 788120 330207

Rs. 295.00

www.phindia.com

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. Back cover