Chapter 4

Excess Carriers in Semiconductors

OBJECTIVES

1. Understand how photons interact with direct and indirect band gap semiconductors
2. Understand generation–recombination of excess carriers, possibly through trap sites
3. Introduce quasi-Fermi levels in non-equilibrium
4. Calculate diffusion currents from carrier concentration gradients and diffusivity
5. Use the continuity equation to study time dependence of carrier concentrations

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

4.1 OPTICAL ABSORPTION[1]

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

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

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 118

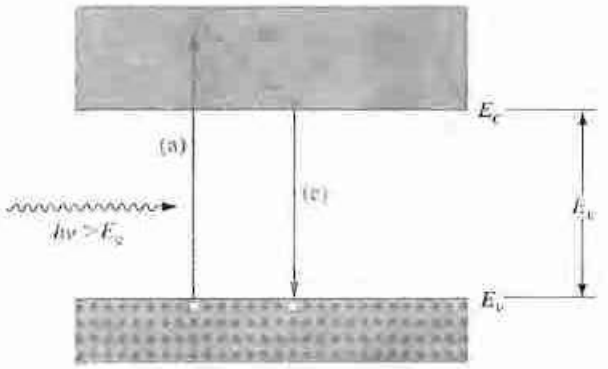


Figure 4–1
Optical absorption of a photon with $h\nu > E_g$: (a) An EHP is created during photon absorption; (b) the excited electron gives up energy to the lattice by scattering events; (c) the electron recombines with a hole in the valence band.

energies less than the band gap are transmitted, this experiment gives an accurate measure of the band gap energy.

It is apparent that a photon with energy $h\nu \geq E_g$ can be absorbed in a semiconductor (Fig. 4–1). Since the valence band contains many electrons and the conduction band has many empty states into which the electrons may be excited, the probability of photon absorption is high. As Fig. 4–1 indicates, an electron excited to the conduction band by optical absorption may initially have more energy than is common for conduction band electrons (almost all electrons are near E_c unless the sample is very heavily doped). Thus the excited electron loses energy to the lattice in scattering events until its velocity reaches the thermal equilibrium velocity of other conduction band electrons. The electron and hole created by this absorption process are *excess carriers;* since they are out of balance with their environment, they must eventually recombine. While the excess carriers exist in their respective bands, however, they are free to contribute to the conductivity of the material.

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

If a beam of photons with $h\nu > E_g$ falls on a semiconductor, there will be some predictable amount of absorption, determined by the properties of the material. We would expect the ratio of transmitted to incident light intensity to depend on the photon wavelength and the thickness of the sample. To calculate this dependence, let us assume that a photon beam of intensity $\mathbf{I}_0$ (photons/cm^2-s) is directed at a sample of thickness l (Fig. 4–2). The beam contains only photons of wavelength λ, selected by a monochromator. As the beam passes through the sample, its intensity at a distance x from the surface can be calculated by considering the probability of absorption within

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 119

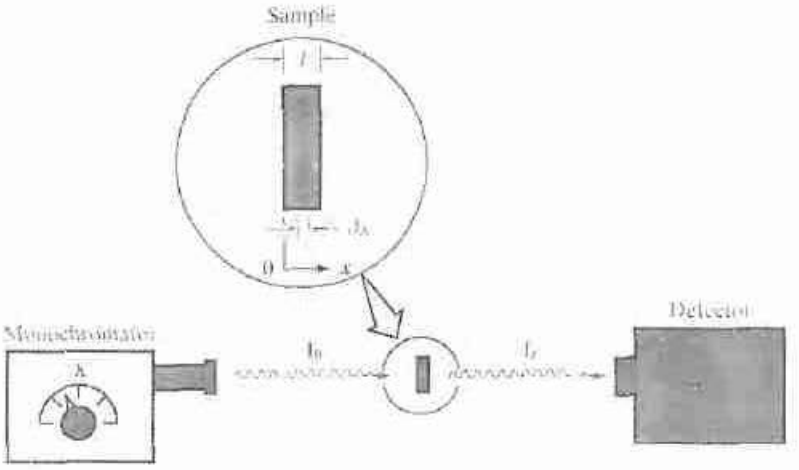


Figure 4–2
Optical absorption experiment.

any increment dx. Since a photon which has survived to x without absorption has no memory of how far it has traveled, its probability of absorption in any dx is constant. Thus the degradation of the intensity $-d\mathbf{I}(x)/dx$ is proportional to the intensity remaining at x:

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

The solution to this equation is

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

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

$$\mathbf{I}_t = \mathbf{I}_0 e^{-\boldsymbol{\alpha}l} \tag{4-3}$$

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

Figure 4-4 indicates the band gap energies of some of the common semiconductors, relative to the visible, infrared, and ultraviolet portions of the spectrum. We observe that GaAs, Si, Ge, and InSb lie outside the visible region, in the infrared. Other semiconductors, such as GaP and CdS, have band gaps wide enough to pass photons in the visible range. It is important to note here that a semiconductor absorbs photons with energies equal to the band gap, *or larger*. Thus Si absorbs not only band gap light (~1 μm) but also shorter wavelengths, including those in the visible part of the spectrum.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 120

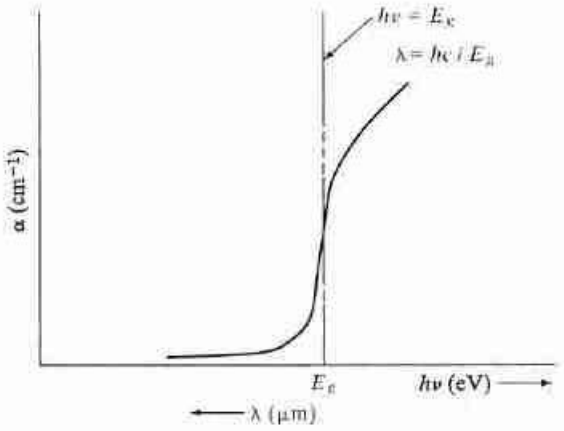


Figure 4–3
Dependence of optical absorption coefficient α for a semiconductor on the wavelength of incident light.

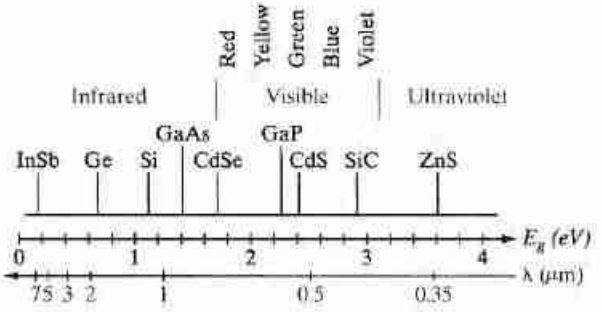


Figure 4–4
Band gaps of some common semiconductors relative to the optical spectrum.

4.2 LUMINESCENCE

When electron-hole pairs are generated in a semiconductor, or when carriers are excited into higher impurity levels from which they fall to their equilibrium states, light can be given off by the material. Many of the semiconductors are well suited for light emission, particularly the compound semiconductors with direct band gaps. The general property of light emission is called *luminescence.*[2] This overall category can be subdivided according to the excitation mechanism: If carriers are excited by photon absorption, the radiation resulting from the recombination of the excited carriers is called *photoluminescence;* if the excited carriers are created by high-energy electron bombardment of the material, the mechanism is called *cathodoluminescence;* if the excitation occurs by the introduction of current into the sample, the resulting luminescence is called *electroluminescence.* Other types of excitation are possible, but these three are the most important for device applications.

[2]The emission processes considered here should not be confused with radiation due to incandescence which occurs in heated materials. The various luminescent mechanisms can be considered "cold" processes as compared to the "hot" process of incandescence, which increases with temperature. In fact, most luminescent processes become more efficient as the temperature is lowered.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 121

4.2.1 Photoluminescence

The simplest example of light emission from a semiconductor occurs for direct excitation and recombination of an EHP, as depicted in Fig. 3–5a. If the recombination occurs directly rather than via a defect level, band gap light is given off in the process. For steady state excitation, the recombination of EHPs occurs at the same rate as the generation, and one photon is emitted for each photon absorbed. Direct recombination is a fast process; the mean lifetime of the EHP is usually on the order of 10^{-8} s or less. Thus the emission of photons stops within approximately 10^{-8} s after the excitation is turned off. Such fast luminescent processes are often referred to as *fluorescence*. In some materials, however, emission continues for periods up to seconds or minutes after the excitation is removed. These slow processes are called *phosphorescence*, and the materials are called *phosphors*. An example of a slow process is shown in Fig. 4–5. This material contains a defect level (perhaps due to an impurity) in the band gap which has a strong tendency to temporarily capture (*trap*) electrons from the conduction band. The events depicted in the figure are as follows: (a) An incoming photon with $h\nu_1 > E_g$ is absorbed, creating an EHP; (b) the excited electron gives up energy to the lattice by scattering until it nears the bottom of the conduction band; (c) the electron is *trapped* by the impurity level E_t and remains trapped until it can be thermally reexcited to the conduction band (d); (e) finally direct recombination occurs as the electron falls to an empty state in the valence band, giving off a photon ($h\nu_2$) of approximately the band gap energy. The delay time between excitation and recombination can be relatively long if the probability of thermal reexcitation from the trap (d) is small. Even longer delay times result if the electron is retrapped several times before recombination. If the trapping probability is greater than the probability of recombination, an electron may make several trips between the trap and the conduction band before recombination finally occurs. In such material the emission of phosphorescent light persists for a relatively long time after the excitation is removed.

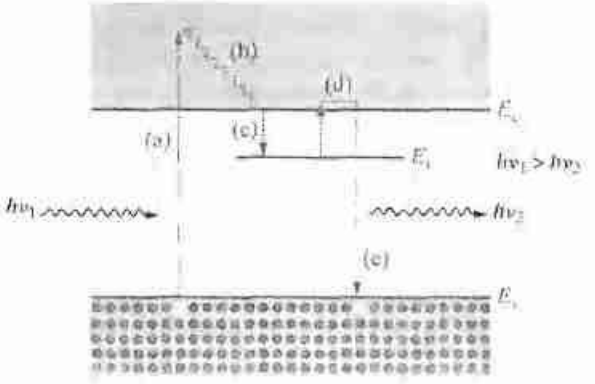


Figure 4–5 Excitation and recombination mechanisms in photoluminescence with a trapping level for electrons.

The color of light emitted by a phosphor such as ZnS depends primarily on the impurities present, since many radiative transitions involve impurity levels within the band gap. This selection of colors is particularly useful in the fabrication of a color television screen.

One of the most common examples of photoluminescence is the fluorescent lamp. Typically such a lamp is composed of a glass tube filled with gas (e.g., a mixture of argon and mercury), with a fluorescent coating on the inside of the tube. When an electric discharge is induced between electrodes in the tube, the excited atoms of the gas emit photons, largely in the visible and ultraviolet regions of the spectrum. This light is absorbed by the luminescent coating, and the visible photons are emitted. The efficiency of such a lamp is considerably better than that of an incandescent bulb, and the wavelength mixture in light given off can be adjusted by proper selection of the fluorescent material.

EXAMPLE 4–1

A 0.46-μm-thick sample of GaAs is illuminated with monochromatic light of $h\nu = 2$ eV. The absorption coefficient α is 5×10^4 cm^{-1}. The power incident on the sample is 10 mW.

(a) Find the total energy absorbed by the sample per second (J/s).

(b) Find the rate of excess thermal energy given up by the electrons to the lattice before recombination (J/s).

(c) Find the number of photons per second given off from recombination events, assuming perfect quantum efficiency.

SOLUTION

(a) From Eq. (4–3),

$$\mathbf{I}_t = \mathbf{I}_0 e^{-\alpha l} = 10^{-2} \exp(-5 \times 10^4 \times 0.46 \times 10^{-4})$$
$$= 10^{-2} e^{-2.3} = 10^{-3} \text{ W}$$

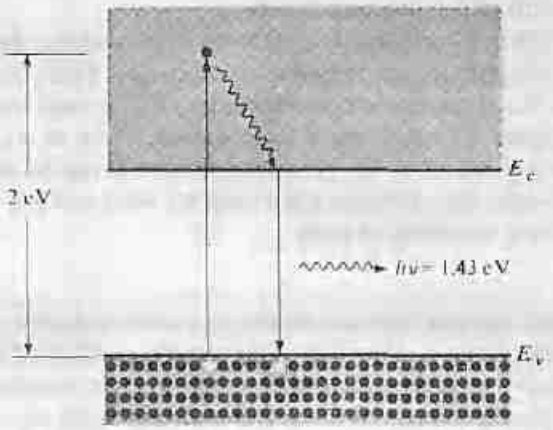


Figure 4–6
Excitation and band-to-band recombination leading to photoluminescence.

Thus the absorbed power is

$$10 - 1 = 9\,\text{mW} = 9 \times 10^{-3}\,\text{J/s}$$

(b) The fraction of each photon energy unit which is converted to heat is

$$\frac{2 - 1.43}{2} = 0.285$$

Thus the amount of energy converted to heat per second is

$$0.285 \times 9 \times 10^{-3} = 2.57 \times 10^{-3}\,\text{J/s}$$

(c) Assuming one emitted photon for each photon absorbed (perfect quantum efficiency), we have

$$\frac{9 \times 10^{-3}\,\text{J/s}}{1.6 \times 10^{-19}\,\text{J/eV} \times 2\,\text{eV/photon}} = 2.81 \times 10^{16}\,\text{photons/s}$$

Alternative solution: Recombination radiation accounts for 9 - 2.57 = 6.43 mW at 1.43 eV/photon.

$$\frac{6.43 \times 10^{-3}}{1.6 \times 10^{-19} \times 1.43} = 2.81 \times 10^{16}\,\text{photons/s}$$

4.2.2 Electroluminescence

There are many ways by which electrical energy can be used to generate photon emission in a solid. In LEDs an electric current causes the injection of minority carriers into regions of the crystal where they can recombine with majority carriers, resulting in the emission of recombination radiation. This important effect (*injection electroluminescence*) will be discussed in Chapter 8 in terms of p-n junction theory.

The first electroluminescent effect to be observed was the emission of photons by certain phosphors in an alternating electric field (the Destriau effect). In this device, a phosphor powder such as ZnS is held in a binder material (often a plastic) of a high dielectric constant. When an a-c electric field is applied, light is given off by the phosphor. Such cells can be useful as lighting panels, although their efficiency has thus far been too low for most applications and their reliability is poor.

4.3 CARRIER LIFETIME AND PHOTOCONDUCTIVITY

When excess electrons and holes are created in a semiconductor, there is a corresponding increase in the conductivity of the sample as indicated by Eq. (3-43). If the excess carriers arise from optical luminescence, the resulting increase in conductivity is called *photoconductivity*. This is an important effect, with useful applications in the analysis of semiconductor materials and in the operation of

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 124

several types of devices. In this section we shall examine the mechanisms by which excess electrons and holes recombine and apply the recombination kinetics to the analysis of photoconductive devices. However, the importance of recombination is not limited to cases in which the excess carriers are created optically. In fact, virtually every semiconductor device depends in some way on the recombination of excess electrons and holes. Therefore, the concepts developed in this section will be used extensively in the analyses of diodes, transistors, lasers, and other devices in later chapters.

4.3.1 Direct Recombination of Electrons and Holes

It was pointed out in Section 3.1.4 that electrons in the conduction band of a semiconductor may make transitions to the valence band (i.e., recombine with holes in the valence band) either directly or indirectly. In direct recombination, an excess population of electrons and holes decays by electrons falling from the conduction band to empty states (holes) in the valence band. Energy lost by an electron in making the transition is given up as a photon. Direct recombination occurs *spontaneously;* that is, the probability that an electron and a hole will recombine is constant in time. As in the case of carrier scattering, this constant probability leads us to expect an exponential solution for the decay of the excess carriers. In this case the rate of decay of electrons at any time t is proportional to the number of electrons remaining at t and the number of holes, with some constant of proportionality for recombination, α_r. The *net* rate of change in the conduction band electron concentration is the thermal generation rate $\alpha_r n_i^2$ from Eq. (3-7) minus the recombination rate

$$\frac{dn(t)}{dt} = \alpha_r n_i^2 - \alpha_r n(t)p(t) \tag{4-4}$$

Let us assume the excess electron-hole population is created at $t = 0$, for example by a short flash of light, and the initial excess electron and hole concentrations Δn and Δp are equal.[3] Then as the electrons and holes recombine in pairs, the instantaneous concentrations of excess carriers $\delta n(t)$ and $\delta p(t)$ are also equal. Thus we can write the total concentrations of Eq. (4-4) in terms of the equilibrium values n_0 and p_0 and the excess carrier concentrations $\delta n(t) = \delta p(t)$. Using Eq. (3-24) we have

$$\begin{aligned}\frac{d\delta n(t)}{dt} &= \alpha_r n_i^2 - \alpha_r[n_0 + \delta n(t)][p_0 + \delta p(t)] \\ &= -\alpha_r[(n_0 + p_0)\delta n(t) + \delta n^2(t)]\end{aligned} \tag{4-5}$$

[3]We will use $\delta n(t)$ and $\delta p(t)$ to mean instantaneous excess carrier concentrations, and Δn, Δp for their values at $t = 0$. Later we will use similar symbolism for spatial distributions, such as $\delta n(x)$ and $\Delta n(x = 0)$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 125

This nonlinear equation would be difficult to solve in its present form. Fortunately, it can be simplified for the case of low-level injection. If the excess carrier concentrations are small, we can neglect the δn^2 term. Furthermore, if the material is extrinsic, we can usually neglect the term representing the equilibrium minority carriers. For example, if the material is p-type ($p_0 \gg n_0$), Eq. (4–5) becomes

$$\frac{d\delta n(t)}{dt} = -\alpha_r p_0 \delta n(t) \tag{4–6}$$

The solution to this equation is an exponential decay from the original excess carrier concentration Δn:

$$\delta n(t) = \Delta n e^{-\alpha_r p_0 t} = \Delta n e^{-t/\tau_n} \tag{4–7}$$

Excess electrons in a p-type semiconductor recombine with a decay constant $\tau_n = (\alpha_r p_0)^{-1}$, called the *recombination lifetime*. Since the calculation is made in terms of the minority carriers, τ_n is often called the *minority carrier lifetime*. The decay of excess holes in n-type material occurs with $\tau_p = (\alpha_r n_0)^{-1}$. In the case of direct recombination, the excess majority carriers decay at exactly the same rate as the minority carriers.

There is a large percentage change in the minority carrier electron concentration in Example 4–2 and a small percentage change in the majority hole concentration. Basically, the approximations of extrinsic material and low-level injection allow us to represent $n(t)$ in Eq. (4–4) by the excess concentration $\delta n(t)$ and $p(t)$ by the equilibrium value p_0. Figure 4–7 indicates that this is a good approximation for the example. A more general expression for the carrier lifetime is

$$\tau_n = \frac{1}{\alpha_r(n_0 + p_0)} \tag{4–8}$$

This expression is valid for n- or p-type material if the injection level is low.

EXAMPLE 4–2 A numerical example may be helpful in visualizing the approximations made in the analysis of direct recombination. Let us assume a sample of GaAs is doped with 10^{15} acceptors/cm^3. The intrinsic carrier concentration of GaAs is approximately 10^6 cm^{-3}; thus the minority electron concentration is $n_0 = n_i^2/p_0 = 10^{-3}$ cm^{-3}. Certainly the approximation of $p_0 \gg n_0$ is valid in this case. Now if 10^{14} EHP/cm^3 are created at $t = 0$, we can calculate the decay of these carriers in time. The approximation of $\delta n \ll p_0$ is reasonable, as Fig. 4–7 indicates. This figure shows the decay in time of the excess populations for a carrier recombination lifetime of $\tau_n = \tau_p = 10^{-8}$ s.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 126

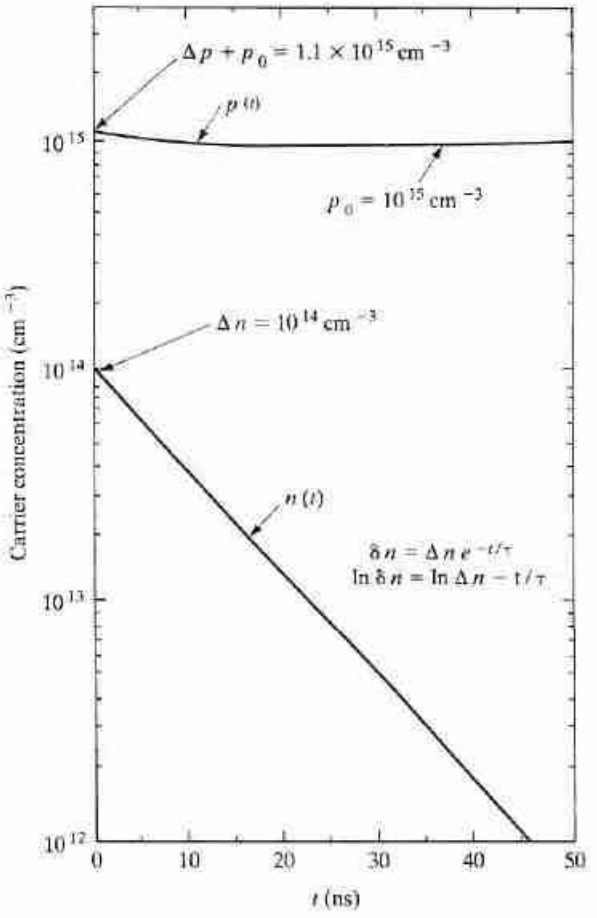


Figure 4–7
Decay of excess electrons and holes by recombination, for $\Delta n = \Delta p = 0.1 p_0$, with n_0 negligible, and $\tau = 10$ ns (Example 4–2). The exponential decay of $\delta n(t)$ is linear on this semilogarithmic graph.

4.3.2 Indirect Recombination; Trapping

In column IV semiconductors and in certain compounds, the probability of direct electron–hole recombination is very small (Appendix III). There is some band gap light given off by materials such as Si and Ge during recombination, but this radiation is very weak and may be detected only by sensitive equipment. The vast majority of the recombination events in indirect materials occur via *recombination levels* within the band gap, and the resulting energy loss by recombining electrons is usually given up to the lattice as heat rather than by the emission of photons. Any impurity or lattice defect can serve as a recombination center if it is capable of receiving a carrier of one type and subsequently capturing the opposite type of carrier, thereby annihilating the pair. For example, Fig. 4–8 illustrates a recombination level E_r which is below E_F at equilibrium and therefore is substantially filled with electrons. When excess electrons and holes are created in this material, each EHP recombines at E_r in two steps: (a) hole capture and (b) electron capture.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 127

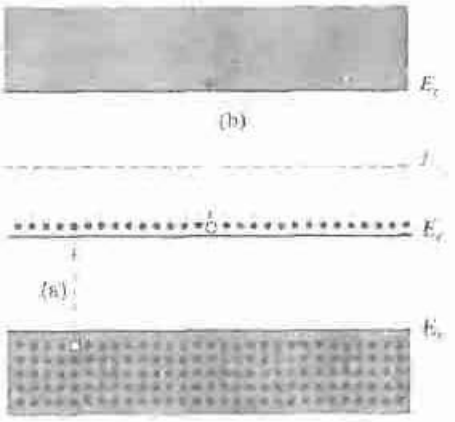


Figure 4–8
Capture processes at a recombination level: (a) hole capture at a filled recombination center; (b) electron capture at an empty center.

Since the recombination centers in Fig. 4–8 are filled at equilibrium, the first event in the recombination process is hole capture. It is important to note that this event is equivalent to an electron at E_t falling to the valence band, leaving behind an empty state in the recombination level. Thus in hole capture, energy is *given up* as heat to the lattice. Similarly, energy is given up when a conduction band electron subsequently falls to the empty state in E_t. When both of these events have occurred, the recombination center is back to its original state (filled with an electron), but an EHP is missing. Thus one EHP recombination has taken place, and the center is ready to participate in another recombination event by capturing a hole.

The carrier lifetime resulting from indirect recombination is somewhat more complicated than is the case for direct recombination, since it is necessary to account for unequal times required for capturing each type of carrier. In particular, recombination is often delayed by the tendency for a captured carrier to be thermally reexcited to its original band before capture of the opposite type of carrier can occur (Section 4.2.1). For example, if electron capture (b) does not follow immediately after hole capture (a) in Fig. 4–8, the hole may be thermally reexcited to the valence band. Energy is required for this process, which is equivalent to a valence band electron being raised to the empty state in the recombination level. This process delays the recombination, since the hole must be captured again before recombination can be completed.

When a carrier is trapped temporarily at a center and then is reexcited without recombination taking place, the process is often called *temporary trapping*. Although the nomenclature varies somewhat, it is common to refer to an impurity or defect center as a *trapping center* (or simply *trap*) if, after capture of one type of carrier, the most probable next event is reexcitation. If the most probable next event is capture of the opposite type of carrier, the center is predominately a recombination center. The recombination can be slow or fast, depending on the average time the first carrier is held before the second carrier is captured. In general, trapping levels located deep in the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 128

band gap are slower in releasing trapped carriers than are the levels located near one of the bands. This results from the fact that more energy is required, for example, to reexcite a trapped electron from a center near the middle of the gap to the conduction band than is required to reexcite an electron from a level closer to the conduction band.

As an example of impurity levels in semiconductors, Fig. 4–9[4] shows the energy level positions of various impurities in Si. In this diagram a superscript indicates whether the impurity is positive (donor) or negative (acceptor) when ionized. Some impurities introduce multiple levels in the band gap; for example, Zn introduces a level (Zn^-) located 0.31 eV above the valence band and a second level ($Zn^{=}$) near the middle of the gap. Each Zn impurity atom is capable of accepting two electrons from the semiconductor, one in the lower level and then one in the upper level.

The effects of recombination and trapping can be measured by a *photoconductive decay* experiment. As Fig. 4–7 shows, a population of excess electrons and holes disappears with a decay constant characteristic of the particular recombination process. The conductivity of the sample during the decay is

$$\sigma(t) = q[n(t)\mu_n + p(t)\mu_p] \tag{4-9}$$

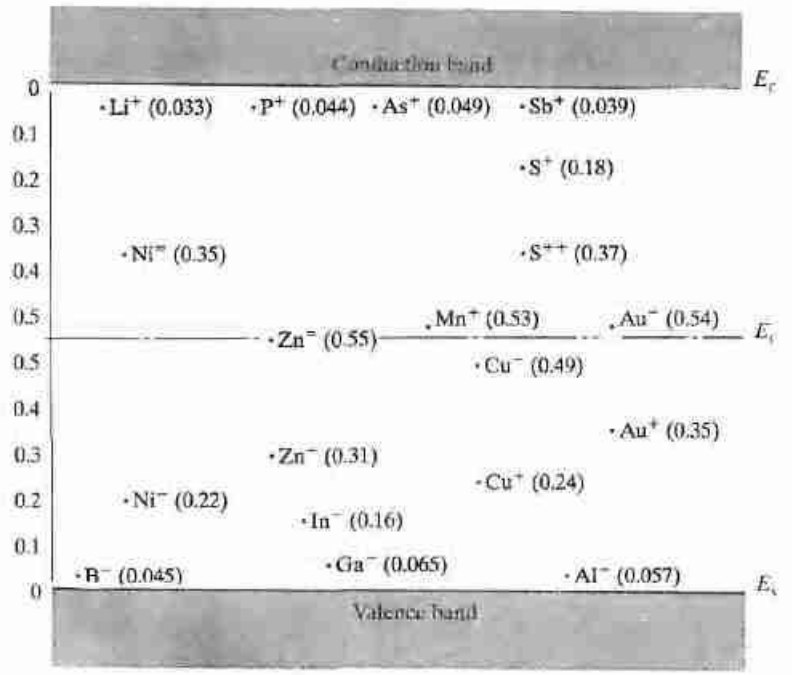


Figure 4–9
Energy levels of impurities in Si. The energies are measured from the nearest band edge (E_v or E_c); donor levels are designated by a plus sign and acceptors by a minus sign.

[4]References: S. M. Sze and J. C. Irvin, "Resistivity, Mobility, and Impurity Levels in GaAs, Ge and Si at 300 K," *Solid State Electronics*, vol. 11, pp. 599–602 (June 1968); E. Schibli and A. G. Milnes, "Deep Impurities in Silicon," *Materials Science and Engineering*, vol. 2, pp. 173–180 (1967).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 129

Therefore, the time dependence of the carrier concentrations can be monitored by recording the sample resistance as a function of time. A typical experimental arrangement is shown schematically in Fig. 4–10. A source of short pulses of light is required, along with an oscilloscope for displaying the sample voltage as the resistance varies. Microsecond light pulses can be obtained by periodically discharging a capacitor through a flash tube containing a gas such as xenon. For shorter pulses, special techniques such as the use of a pulsed laser must be used.

4.3.3 Steady State Carrier Generation; Quasi-Fermi Levels

In the previous discussion we emphasized the transient decay of an excess EHP population. However, the various recombination mechanisms are also important in a sample at thermal equilibrium or with a steady state EHP generation–recombination balance.[5] For example, a semiconductor at equilibrium experiences thermal generation of EHPs at a rate $g(T) = g_i$ described by Eq. (3-7). This generation is balanced by the recombination rate so that the equilibrium concentrations of carriers n_0 and p_0 are maintained:

$$g(T) = \alpha_r n_i^2 = \alpha_r n_0 p_0 \tag{4-10}$$

This equilibrium rate balance can include generation from defect centers as well as band-to-band generation.

If a steady light is shone on the sample, an optical generation rate g_{op} will be added to the thermal generation, and the carrier concentrations n and p will increase to new steady state values. We can write the balance between generation and recombination in terms of the equilibrium carrier concentrations and the departures from equilibrium δn and δp:

$$g(T) + g_{op} = \alpha_r np = \alpha_r(n_0 + \delta n)(p_0 + \delta p) \tag{4-11}$$

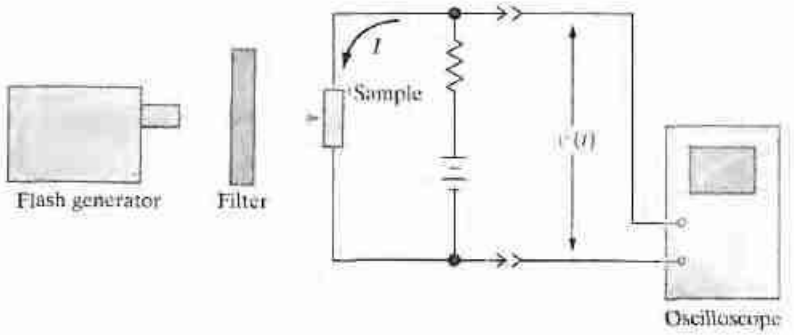


Figure 4–10 Experimental arrangement for photoconductive decay measurements, and a typical oscilloscope trace.

[5]The term *equilibrium* refers to a condition of no external excitation except for temperature, and no net motion of charge (e.g., a sample at a constant temperature, in the dark, with no fields applied). *Steady state* refers to a nonequilibrium condition in which all processes are constant and are balanced by opposing processes (e.g., a sample with a constant current or a constant optical generation of EHPs just balanced by recombination).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 130

For steady state recombination and no trapping, $\delta n = \delta p$; thus Eq. (4–11) becomes

$$g(T) + g_{op} = \alpha_r n_0 p_0 + \alpha_r[(n_0 + p_0)\delta n + \delta n^2] \tag{4–12}$$

The term $\alpha_r n_0 p_0$ is just equal to the thermal generation rate $g(T)$. Thus, neglecting the δn^2 term for low-level excitation, we can rewrite Eq. (4–12) as

$$g_{op} = \alpha_r(n_0 + p_0)\delta n = \frac{\delta n}{\tau_n} \tag{4–13}$$

The excess carrier concentration can be written as

$$\delta n = \delta p = g_{op}\tau_n \tag{4–14}$$

More general expressions are given in Eq. (4–16), which allow for the case $\tau_p \neq \tau_n$, when trapping is present.

EXAMPLE 4–3

As a numerical example, let us assume that 10^{13} EHP/cm^3 are created optically every microsecond in a Si sample with $n_0 = 10^{14}$ cm^{-3} and $\tau_n = \tau_p = 2$ μsec. The steady state excess electron (or hole) concentration is then 2×10^{13} cm^{-3} from Eq. (4–14). While the percentage change in the majority electron concentration is small, the minority carrier concentration changes from

$$p_0 = n_i^2/n_0 = (2.25 \times 10^{20})/10^{14} = 2.25 \times 10^6 \text{ cm}^{-3} \qquad (\textit{equilibrium})$$

to

$$p = 2 \times 10^{13} \text{ cm}^{-3} \qquad (\textit{steady state})$$

Note that the equilibrium equation $n_0 p_0 = n_i^2$ cannot be used with the subscripts removed; that is, $np \neq n_i^2$ when excess carriers are present.

It is often desirable to refer to the steady state electron and hole concentrations in terms of Fermi levels, which can be included in band diagrams for various devices. The Fermi level E_F used in Eq. (3–25) is meaningful only when no excess carriers are present. However, we can write expressions for the steady state concentrations in the same *form* as the equilibrium expressions by defining separate *quasi-Fermi levels* F_n and F_p for electrons and holes. The resulting carrier concentration equations

$$\boxed{\begin{aligned} n &= n_i e^{(F_n - E_i)/kT} \\ p &= n_i e^{(E_i - F_p)/kT} \end{aligned}} \tag{4–15}$$

can be considered as defining relations for the quasi-Fermi levels.[6]

[6] In some texts the quasi-Fermi level is called *IMREF*, which is Fermi spelled backward.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 131

EXAMPLE 4–4 In Example 4–3, the steady state electron concentration is

$$n = n_0 + \delta n = 1.2 \times 10^{14} = (1.5 \times 10^{10})e^{(F_n - E_i)/0.0259}$$

where $kT \simeq 0.0259$ eV at room temperature. Thus the electron quasi-Fermi level position $F_n - E_i$ is found from

$$F_n - E_i = 0.0259 \ln(8 \times 10^3) = 0.233 \text{ eV}$$

and F_n lies 0.233 eV above the intrinsic level. By a similar calculation, the hole quasi-Fermi level lies 0.186 eV below E_i (Fig. 4–11). In this example, the equilibrium Fermi level is $0.0259 \ln(6.67 \times 10^3) = 0.228$ eV above the intrinsic level.

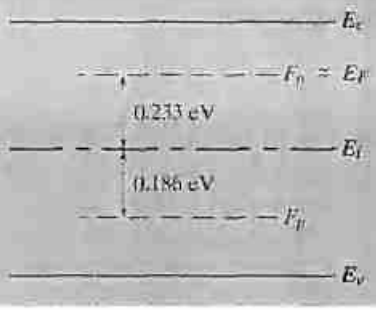


Figure 4–11 Quasi-Fermi levels F_n and F_p for a Si sample with $n_0 = 10^{14}$ cm^{-3}, $\tau_p = 2$ μs, and $g_{op} = 10^{19}$ EHP/cm^3-s (Example 4–4).

The quasi-Fermi levels of Fig. 4–11 illustrate dramatically the deviation from equilibrium caused by the optical excitation; the steady state F_n is only slightly above the equilibrium E_F, whereas F_p is greatly displaced below E_F. From the figure it is obvious that the excitation causes a large percentage change in minority carrier hole concentration and a relatively small change in the electron concentration.

In summary, the quasi-Fermi levels F_n and F_p are the steady state analogues of the equilibrium Fermi level E_F. When excess carriers are present, the deviations of F_n and F_p from E_F indicate how far the electron and hole populations are from the equilibrium values n_0 and p_0. A given concentration of excess EHPs causes a large shift in the minority carrier quasi-Fermi level compared with that for the majority carriers. The separation of the quasi-Fermi levels $F_n - F_p$ is a direct measure of the deviation from equilibrium (at equilibrium $F_n = F_p = E_F$). The concept of quasi-Fermi levels is very useful in visualizing minority and majority carrier concentrations in devices where these quantities vary with position.

4.3.4 Photoconductive Devices

There are a number of applications for devices which change their resistance when exposed to light. For example, such light detectors can be used in the

home to control automatic night lights which turn on at dusk and turn off at dawn. They can also be used to measure illumination levels, as in exposure meters for cameras. Many systems include a light beam aimed at the photoconductor, which signals the presence of an object between the source and detector. Such systems are useful in moving-object counters, burglar alarms, and many other applications. Detectors are used in optical signaling systems in which information is transmitted by a light beam and is received at a photoconductive cell.

Considerations in choosing a photoconductor for a given application include the sensitive wavelength range, time response, and optical sensitivity of the material. In general, semiconductors are most sensitive to photons with energies equal to the band gap or slightly more energetic than band gap. Less energetic photons are not absorbed, and photons with $h\nu \gg E_g$ are absorbed at the surface and contribute little to the bulk conductivity. Therefore, the table of band gaps (Appendix III) indicates the photon energies to which most semiconductor photodetectors respond. For example, CdS ($E_g = 2.42$ eV) is commonly used as a photoconductor in the visible range, and narrow-gap materials such as Ge (0.67 eV) and InSb (0.18 eV) are useful in the infrared portion of the spectrum. Some photoconductors respond to excitations of carriers from impurity levels within the band gap and therefore are sensitive to photons of less than band gap energy.

The optical sensitivity of a photoconductor can be evaluated by examining the steady state excess carrier concentrations generated by an optical generation rate g_{op}. If the mean time each carrier spends in its respective band before capture is τ_n and τ_p, we have

$$\delta n = \tau_n g_{op} \quad \text{and} \quad \delta p = \tau_p g_{op} \tag{4-16}$$

and the photoconductivity change is

$$\Delta\sigma = q g_{op}(\tau_n \mu_n + \tau_p \mu_p) \tag{4-17}$$

For simple recombination, τ_n and τ_p will be equal. If trapping is present, however, one of the carriers may spend little time in its band before being trapped. From Eq. (4-17) it is obvious that for maximum photoconductive response, we want high mobilities and long lifetimes. Some semiconductors are especially good candidates for photoconductive devices because of their high mobility; for example, InSb has an electron mobility of about 10^5 cm^2/V-s and therefore is used as a sensitive infrared detector in many applications.

The time response of a photoconductive cell is limited by the recombination times, the degree of carrier trapping, and the time required for carriers to drift through the device in an electric field. Often these properties can be adjusted by proper choice of material and device geometry, but in some cases improvements in response time are made at the expense of sensitivity. For example, the drift time can be reduced by making the device short, but this

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 133

substantially reduces the responsive area of the device. In addition, it is often desirable that the device have a large dark resistance, and for this reason, shortening the length may not be practical. There is usually a compromise between sensitivity, response time, dark resistance, and other requirements in choosing a device for a particular application.

4.4 DIFFUSION OF CARRIERS

When excess carriers are created nonuniformly in a semiconductor, the electron and hole concentrations vary with position in the sample. Any such spatial variation (*gradient*) in n and p calls for a net motion of the carriers from regions of high carrier concentration to regions of low carrier concentration. This type of motion is called *diffusion* and represents an important charge transport process in semiconductors. The two basic processes of current conduction are diffusion due to a carrier gradient and drift in an electric field.

4.4.1 Diffusion Processes

When a bottle of perfume is opened in one corner of a closed room, the scent is soon detected throughout the room. If there is no convection or other net motion of air, the scent spreads by diffusion. The diffusion is the natural result of the *random motion* of the individual molecules. Consider, for example, a volume of arbitrary shape with scented air molecules inside and unscented molecules outside the volume. All the molecules undergo random thermal motion and collisions with other molecules. Thus each molecule moves in an arbitrary direction until it collides with another air molecule, after which it moves in a new direction. If the motion is truly random, a molecule at the edge of the volume has equal probabilities of moving into or out of the volume on its next step (assuming the curvature of the surface is negligible on the molecular scale). Therefore, after a mean free time $\bar{t}$, half the molecules at the edge will have moved into the volume and half will have moved out of the volume. The net effect is that the volume containing scented molecules has increased. This process will continue until the molecules are uniformly distributed in the room. Only then will a given volume gain as many molecules as it loses in a given time. In other words, net diffusion will continue as long as gradients exist in the distribution of scented molecules.

Carriers in a semiconductor diffuse in a carrier gradient by random thermal motion and scattering from the lattice and impurities. For example, a pulse of excess electrons injected at $x = 0$ at time $t = 0$ will spread out in time as shown in Fig. 4-12. Initially, the excess electrons are concentrated at $x = 0$; as time passes, however, electrons diffuse to regions of low electron concentration until finally $n(x)$ is constant.

We can calculate the rate at which the electrons diffuse in a one-dimensional problem by considering an arbitrary distribution $n(x)$ such as Fig. 4-13a. Since the mean free path $\bar{l}$ between collisions is a small

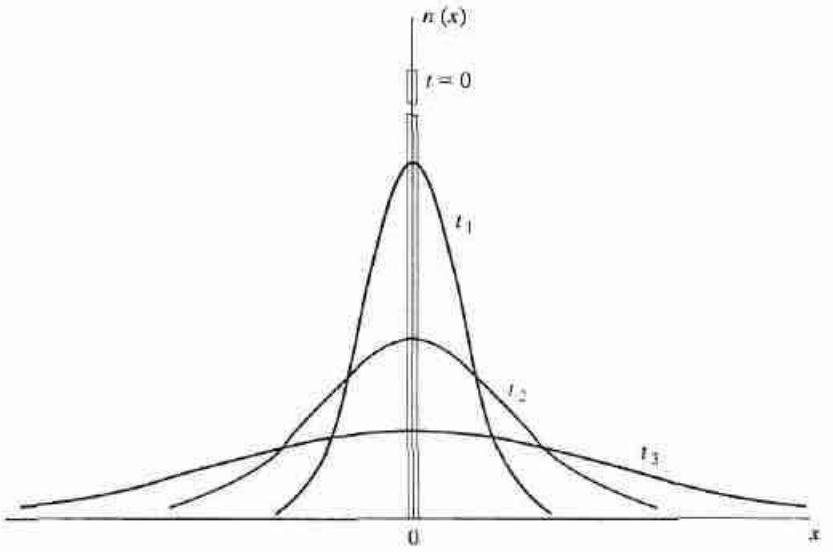


Figure 4–12
Spreading of a pulse of electrons by diffusion.

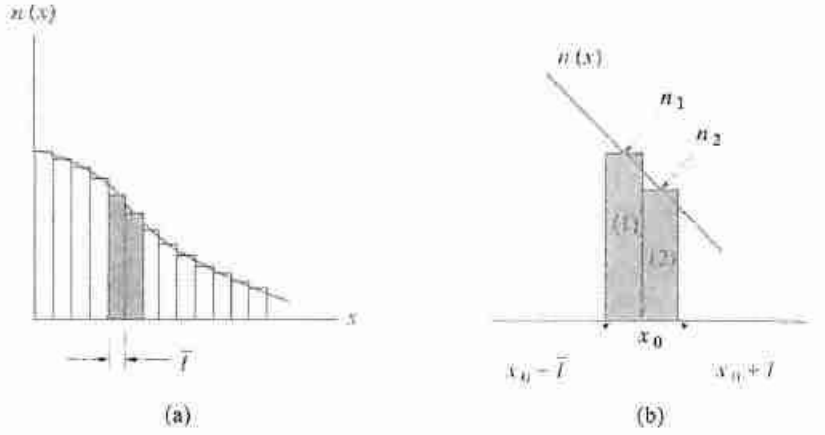


Figure 4–13
An arbitrary electron concentration gradient in one dimension: (a) division of $n(x)$ into segments of length equal to a mean free path for the electrons; (b) expanded view of two of the segments centered at x_0.

incremental distance, we can divide x into segments $\bar{l}$ wide, with $n(x)$ evaluated at the center of each segment (Fig. 4–13b).

The electrons in segment (1) to the left of x_0 in Fig. 4–13b have equal chances of moving left or right, and in a mean free time $\bar{\tau}$ one-half of them will move into segment (2). The same is true of electrons within one mean free path of x_0 to the right; one-half of these electrons will move through x_0 from right to left in a mean free time. Therefore, the *net* number of electrons passing x_0 from left to right in one mean free time is $\frac{1}{2}(n_1\bar{l}A) - \frac{1}{2}(n_2\bar{l}A)$, where the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 135

area perpendicular to x is A. The rate of electron flow in the $+x$-direction per unit area (the electron flux density ϕ_n) is given by

$$\phi_n(x_0) = \frac{\bar{l}}{2\bar{t}}(n_1 - n_2) \tag{4-18}$$

Since the mean free path $\bar{l}$ is a small differential length, the difference in electron concentration $(n_1 - n_2)$ can be written as

$$n_1 - n_2 = \frac{n(x) - n(x + \Delta x)}{\Delta x}\bar{l} \tag{4-19}$$

where x is taken at the center of segment (1) and $\Delta x = \bar{l}$. In the limit of small Δx (i.e., small mean free path $\bar{l}$ between scattering collisions), Eq. (4-18) can be written in terms of the carrier gradient $dn(x)/dx$:

$$\phi_n(x) = \frac{\bar{l}^2}{2\bar{t}} \lim_{\Delta x \to 0} \frac{n(x) - n(x + \Delta x)}{\Delta x} = \frac{-\bar{l}^2}{2\bar{t}} \frac{dn(x)}{dx} \tag{4-20}$$

The quantity $\bar{l}^2/2\bar{t}$ is called the *electron diffusion coefficient*[7] D_n, with units cm^2/s. The minus sign in Eq. (4-20) arises from the definition of the derivative; it simply indicates that the net motion of electrons due to diffusion is in the direction of *decreasing* electron concentration. This is the result we expect, since net diffusion occurs from regions of high particle concentration to regions of low particle concentration. By identical arguments, we can show that holes in a hole concentration gradient move with a diffusion coefficient D_p. Thus

$$\phi_n(x) = -D_n \frac{dn(x)}{dx} \tag{4-21a}$$

$$\phi_p(x) = -D_p \frac{dp(x)}{dx} \tag{4-21b}$$

The diffusion current crossing a unit area (the current density) is the particle flux density multiplied by the charge of the carrier:

$$J_n(\text{diff.}) = -(-q)D_n \frac{dn(x)}{dx} = +qD_n \frac{dn(x)}{dx} \tag{4-22a}$$

$$J_p(\text{diff.}) = -(+q)D_p \frac{dp(x)}{dx} = -qD_p \frac{dp(x)}{dx} \tag{4-22b}$$

[7] If motion in three dimensions were included, the diffusion would be smaller in the x-direction. Actually, the diffusion coefficient should be calculated from the true energy distributions and scattering mechanisms. Diffusion coefficients are usually determined experimentally for a particular material, as described in Section 4.4.5.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 136

It is important to note that electrons and holes move together in a carrier gradient [Eqs. (4–21)], but the resulting currents are in opposite directions [Eqs. (4–22)] because of the opposite charge of electrons and holes.

4.4.2 Diffusion and Drift of Carriers; Built-in Fields

If an electric field is present in addition to the carrier gradient, the current densities will each have a drift component and a diffusion component

$$J_n(x) = q\mu_n n(x)\mathscr{E}(x) + qD_n \frac{dn(x)}{dx} \qquad (4\text{–}23a)$$

drift diffusion

$$J_p(x) = q\mu_p p(x)\mathscr{E}(x) - qD_p \frac{dp(x)}{dx} \qquad (4\text{–}23b)$$

and the total current density is the sum of the contributions due to electrons and holes:

$$J(x) = J_n(x) + J_p(x) \qquad (4\text{–}24)$$

We can best visualize the relation between the particle flow and the current of Eqs. (4–23) by considering a diagram such as shown in Fig. 4–14. In this figure an electric field is assumed to be in the x-direction, along with carrier distributions $n(x)$ and $p(x)$ which decrease with increasing x. Thus the derivatives in Eqs. (4–21) are negative, and diffusion takes place in the $+x$-direction. The resulting electron and hole diffusion currents [J_n (diff.) and J_p (diff.)] are in opposite directions, according to Eqs. (4–22). Holes drift in the direction of the electric field [ϕ_p (drift)], whereas electrons drift in the opposite direction because of their negative charge. The resulting drift current is in the $+x$-direction in each case. Note that the drift and diffusion components of the current are additive for holes when the field is in the direction of decreasing hole concentration, whereas the two components are subtractive for electrons under similar conditions. The total current may be due primarily to the flow of electrons or holes, depending on the relative concentrations and the relative magnitudes and directions of electric field and carrier gradients.

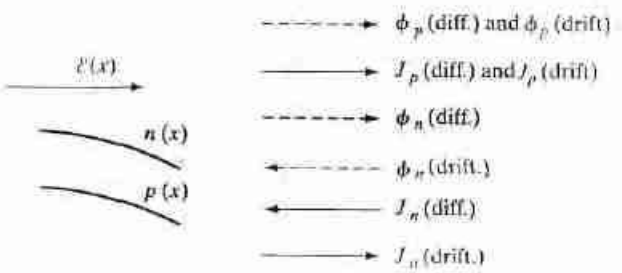


Figure 4–14
Drift and diffusion directions for electrons and holes in a carrier gradient and an electric field. Particle flow directions are indicated by dashed arrows, and the resulting currents are indicated by solid arrows.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 137

An important result of Eqs. (4–23) is that minority carriers can contribute significantly to the current through diffusion. Since the drift terms are proportional to carrier concentration, minority carriers seldom provide much drift current. On the other hand, diffusion current is proportional to the *gradient* of concentration. For example, in n-type material the minority hole concentration p may be many orders of magnitude smaller than the electron concentration n, but the gradient dp/dx may be significant. As a result, minority carrier currents through diffusion can sometimes be as large as majority carrier currents.

In discussing the motion of carriers in an electric field, we should indicate the influence of the field on the energies of electrons in the band diagrams. Assuming an electric field $\mathscr{E}(x)$ in the x-direction, we can draw the energy bands as in Fig. 4–15, to include the change in potential energy of electrons in the field. Since electrons drift in a direction opposite to the field, we expect the potential energy for electrons to increase in the direction of the field, as in Fig. 4–15. The electrostatic potential $\mathscr{V}(x)$ varies in the opposite direction, since it is defined in terms of positive charges and is therefore related to the electron potential energy $E(x)$ displayed in the figure by $\mathscr{V}(x) = E(x)/(-q)$.

From the definition of electric field,

$$\mathscr{E}(x) = -\frac{d\mathscr{V}(x)}{dx} \tag{4-25}$$

we can relate $\mathscr{E}(x)$ to the electron potential energy in the band diagram by choosing some reference in the band for the electrostatic potential. We are interested only in the spatial variation $\mathscr{V}(x)$ for Eq. (4–25). Choosing E_i as a convenient reference, we can relate the electric field to this reference by

$$\boxed{\mathscr{E}(x) = -\frac{d\mathscr{V}(x)}{dx} = -\frac{d}{dx}\left[\frac{E_i}{(-q)}\right] = \frac{1}{q}\frac{dE_i}{dx}} \tag{4-26}$$

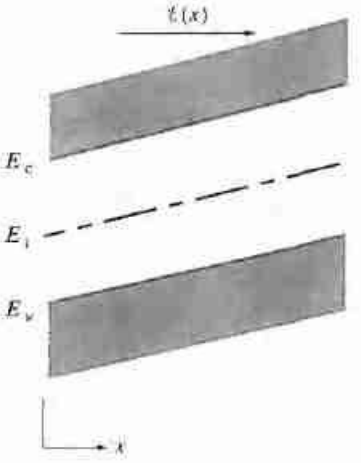


Figure 4–15 Energy band diagram of a semiconductor in an electric field $\mathscr{E}(x)$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 138

Therefore, the variation of band energies with $\mathscr{E}(x)$ as drawn in Fig. 4–15 is correct. The direction of the slope in the bands relative to $\mathscr{E}$ is simple to remember: Since the diagram indicates electron energies, we know the slope in the bands must be such that electrons drift "downhill" in the field. Therefore, $\mathscr{E}$ points "uphill" in the band diagram.

At equilibrium, no net current flows in a semiconductor. Thus any fluctuation which would begin a diffusion current also sets up an electric field which redistributes carriers by drift. An examination of the requirements for equilibrium indicates that the diffusion coefficient and mobility must be related. Setting Eq. (4–23b) equal to zero for equilibrium, we have

$$\mathscr{E}(x) = \frac{D_p}{\mu_p} \frac{1}{p(x)} \frac{dp(x)}{dx} \tag{4-27}$$

Using Eq. (3–25b) for $p(x)$,

$$\mathscr{E}(x) = \frac{D_p}{\mu_p} \frac{1}{kT} \left(\frac{dE_i}{dx} - \frac{dE_F}{dx} \right) \tag{4-28}$$

The equilibrium Fermi level does not vary with x, and the derivative of E_i is given by Eq. (4–26). Thus Eq. (4–28) reduces to

$$\boxed{\frac{D}{\mu} = \frac{kT}{q}} \tag{4-29}$$

This result is obtained for either carrier type. This important equation is called the *Einstein relation.* It allows us to calculate either D or μ from a measurement of the other. Table 4–1 lists typical values of D and μ for several semiconductors at room temperature. It is clear from these values that $D/\mu = 0.026$ V.

An important result of the balance of drift and diffusion at equilibrium is that *built-in* fields accompany gradients in E_i [see Eq. (4–26)]. Such gradients in the bands at equilibrium (E_F constant) can arise when the band gap varies due to changes in alloy composition. More commonly, built-in fields result from doping gradients. For example, a donor distribution $N_d(x)$ causes a gradient in $n_0(x)$, which must be balanced by a built-in electric field $\mathscr{E}(x)$.

Table 4–1 Diffusion coefficient and mobility of electrons and holes for intrinsic semiconductors at 300 K. *Note:* Use Fig. 3–23 for doped semiconductors.

	D_n (cm²/s)	D_p (cm²/s)	μ_n (cm²/V-s)	μ_p (cm²/V-s)
Ge	100	50	3900	1900
Si	35	12.5	1350	480
GaAs	220	10	8500	400

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 139

4.4.3 Diffusion and Recombination; The Continuity Equation

In the discussion of diffusion of excess carriers, we have thus far neglected the important effects of recombination. These effects must be included in a description of conduction processes, however, since recombination can cause a variation in the carrier distribution. For example, consider a differential length Δx of a semiconductor sample with area A in the yz-plane (Fig. 4–16). The hole current density leaving the volume, $J_p(x + \Delta x)$, can be larger or smaller than the current density entering, $J_p(x)$, depending on the generation and recombination of carriers taking place within the volume. The net increase in hole concentration per unit time, $\partial p/\partial t$, is the difference between the hole flux per unit volume entering and leaving, minus the recombination rate. We can convert hole current density to hole particle flux density by dividing J_p by q. The current densities are already expressed per unit area; thus dividing $J_p(x)/q$ by Δx gives the number of carriers per unit volume entering $\Delta x A$ per unit time, and $(1/q)J_p(x + \Delta x)/\Delta x$ is the number leaving per unit volume and time:

$$\left.\frac{\partial p}{\partial t}\right|_{x \to x + \Delta x} = \frac{1}{q}\frac{J_p(x) - J_p(x + \Delta x)}{\Delta x} - \frac{\delta p}{\tau_p} \tag{4-30}$$

Rate of hole buildup = increase of hole concentration in $\delta x A$ per unit time − recombination rate

As Δx approaches zero, we can write the current change in derivative form:

$$\frac{\partial p(x, t)}{\partial t} = \frac{\partial \delta p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} - \frac{\delta p}{\tau_p} \tag{4-31a}$$

The expression (4–31a) is called the *continuity equation* for holes. For electrons we can write

$$\frac{\partial \delta n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - \frac{\delta n}{\tau_n} \tag{4-31b}$$

since the electronic charge is negative.

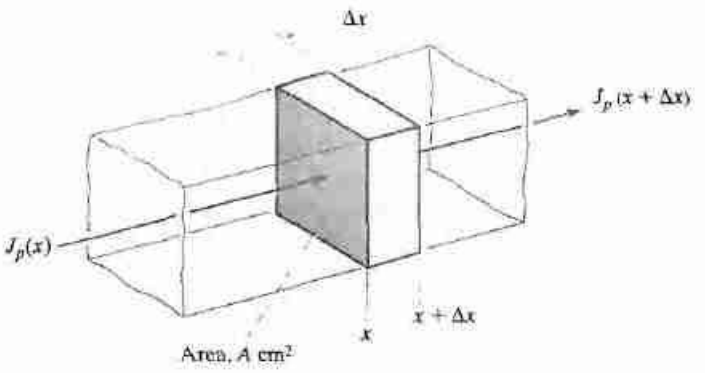


Figure 4–16
Current entering and leaving a volume $\Delta x A$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 140

When the current is carried strictly by diffusion (negligible drift), we can replace the currents in Eqs. (4-31) by the expressions for diffusion current; for example, for electron diffusion we have

$$J_n(\text{diff.}) = qD_n \frac{\partial \delta n}{\partial x} \tag{4-32}$$

Substituting this into Eq. (4-31b) we obtain the *diffusion equation* for electrons,

$$\boxed{\frac{\partial \delta n}{\partial t} = D_n \frac{\partial^2 \delta n}{\partial x^2} - \frac{\delta n}{\tau_n}} \tag{4-33a}$$

and similarly for holes,

$$\boxed{\frac{\partial \delta p}{\partial t} = D_p \frac{\partial^2 \delta p}{\partial x^2} - \frac{\delta p}{\tau_n}} \tag{4-33b}$$

These equations are useful in solving transient problems of diffusion with recombination. For example, a pulse of electrons in a semiconductor (Fig. 4-12) spreads out by diffusion and disappears by recombination. To solve for the electron distribution in time, $n(x, t)$, we would begin with the diffusion equation, Eq. (4-33a).

4.4.4 Steady State Carrier Injection; Diffusion Length

In many problems a steady state distribution of excess carriers is maintained, such that the time derivatives in Eqs. (4-33) are zero. In the steady state case the diffusion equations become

$$\boxed{\frac{d^2 \delta n}{dx^2} = \frac{\delta n}{D_n \tau_n} \equiv \frac{\delta n}{L_n^2}} \tag{4-34a}$$

$$\boxed{\frac{d^2 \delta p}{dx^2} = \frac{\delta p}{D_p \tau_p} \equiv \frac{\delta p}{L_p^2}} \tag{4-34b}$$

(steady state)

where $L_n \equiv \sqrt{D_n \tau_n}$ is called the electron *diffusion length* and L_p is the diffusion length for holes. We no longer need partial derivatives, since the time variation is zero for steady state.

The physical significance of the diffusion length can be understood best by an example. Let us assume that excess holes are somehow injected into a semi-infinite semiconductor bar at $x = 0$, and the steady state hole injection maintains a constant excess hole concentration at the injection point $\delta p(x = 0) = \Delta p$. The injected holes diffuse along the bar, recombining with

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 141

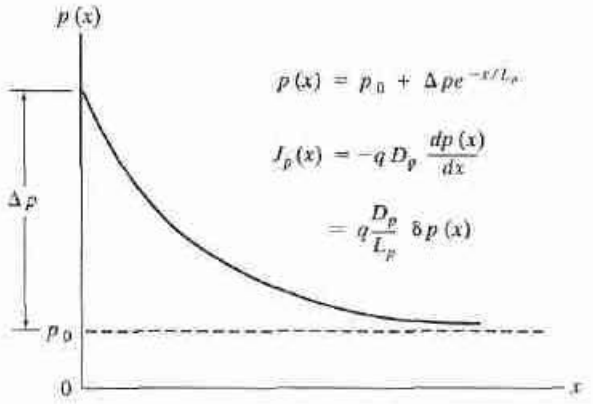


Figure 4–17
Injection of holes at $x = 0$, giving a steady state hole distribution $p(x)$ and a resulting diffusion current density $J_p(x)$.

a characteristic lifetime τ_p. In steady state we expect the distribution of excess holes to decay to zero for large values of x, because of the recombination (Fig. 4–17). For this problem we use the steady state diffusion equation for holes, Eq. (4–34b). The solution to this equation has the form

$$\delta p(x) = C_1 e^{x/L_p} + C_2 e^{-x/L_p} \tag{4–35}$$

We can evaluate C_1 and C_2 from the boundary conditions. Since recombination must reduce $\delta p(x)$ to zero for large values of x, $\delta p = 0$ at $x = \infty$ and therefore $C_1 = 0$. Similarly, the condition $\delta p = \Delta p$ at $x = 0$ gives $C_2 = \Delta p$, and the solution is

$$\boxed{\delta p(x) = \Delta p e^{-x/L_p}} \tag{4–36}$$

The injected excess hole concentration dies out exponentially in x due to recombination, and the diffusion length L_p represents the distance at which the excess hole distribution is reduced to $1/e$ of its value at the point of injection. We can show that *L_p is the average distance a hole diffuses before recombining*. To calculate an average diffusion length, we must obtain an expression for the probability that an injected hole recombines in a particular interval dx. The probability that a hole injected at $x = 0$ *survives* to x without recombination is $\delta p(x)/\Delta p = \exp(-x/L_p)$, the ratio of the steady state concentrations at x and 0. On the other hand, the probability that a hole at x will *recombine* in the subsequent interval dx is

$$\frac{\delta p(x) - \delta p(x + dx)}{\delta p(x)} = \frac{-(d\delta p(x)/dx)dx}{\delta p(x)} = \frac{1}{L_p} dx \tag{4–37}$$

Thus the total probability that a hole injected at $x = 0$ will recombine in a given dx is the product of the two probabilities:

$$(e^{-x/L_p})\left(\frac{1}{L_p} dx\right) = \frac{1}{L_p} e^{-x/L_p} dx \tag{4–38}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 142

Then, using the usual averaging techniques described by Eq. (2-21), the average distance a hole diffuses before recombining is

$$\langle x \rangle = \int_0^\infty x \frac{e^{-x/L_p}}{L_p} dx = L_p \tag{4-39}$$

The steady state distribution of excess holes causes diffusion, and therefore a hole current, in the direction of decreasing concentration. From Eqs. (4-22b) and (4-36) we have

$$J_p(x) = -qD_p \frac{dp}{dx} = -qD_p \frac{d\delta p}{dx} = q \frac{D_p}{L_p} \Delta p e^{-x/L_p} = q \frac{D_p}{L_p} \delta p(x) \tag{4-40}$$

Since $p(x) = p_0 + \delta p(x)$, the space derivative involves only the excess concentration. We notice that since $\delta p(x)$ is proportional to its derivative for an exponential distribution, the diffusion current at any x is just proportional to the excess concentration δp at that position.

Although this example seems rather restricted, its usefulness will become apparent in Chapter 5 in the discussion of p-n junctions. The injection of minority carriers across a junction often leads to exponential distributions as in Eq. (4-36), with the resulting diffusion current of Eq. (4-40).

EXAMPLE 4-5 In a very long p-type Si bar with cross-sectional area = 0.5 cm^2 and $N_a = 10^{17}$ cm^{-3}, we inject holes such that the steady state *excess* hole concentration is 5×10^{16} cm^{-3} at $x = 0$. What is the steady state separation between F_p and E_c at $x = 1000$ Å? What is the hole current there? How much is the excess stored hole charge? Assume that $\mu_p = 500$ cm^2/V-s and $\tau_p = 10^{-10}$ s.

SOLUTION

$$D_p = \frac{kT}{q} \mu_p = 0.0259 \times 500 = 12.95 \text{ cm/s}$$

$$L_p = \sqrt{D_p \tau_p} = \sqrt{12.95 \times 10^{-10}} = 3.6 \times 10^{-5} \text{ cm}$$

$$p = p_0 + \Delta p e^{-\frac{x}{L_p}} = 10^{17} + 5 \times 10^{16} e^{-\frac{10^{-5}}{3.6 \times 10^{-5}}}$$

$$= 1.379 \times 10^{17} = n_i e^{(E_i - F_p)/kT} = (1.5 \times 10^{10} \text{ cm}^{-3}) e^{(E_i - F_p)/kT}$$

$$E_i - F_p = \left(\ln \frac{1.379 \times 10^{17}}{1.5 \times 10^{10}} \right) \cdot 0.0259 = 0.415 \text{ eV}$$

$$E_c - F_p = 1.1/2 \text{ eV} + 0.415 \text{ eV} = \mathbf{0.965 \text{ eV}}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 143

We can calculate the hole current from Eq. (4–40)

$$I_p = -qAD_p\frac{dp}{dx} = qA\frac{D_p}{L_p}(\Delta p)e^{-\frac{x}{L_p}}$$

$$= 1.6 \times 10^{-19} \times 0.5 \times \frac{12.95}{3.6 \times 10^{-5}} \times 5 \times 10^{16}\, e^{-\frac{10^{-3}}{3.6\times10^{-5}}}$$

$$= \mathbf{1.09 \times 10^{3}\ A}$$

$$Q_p = qA(\Delta p)L_p$$

$$= 1.6 \times 10^{-19}(0.5)(5 \times 10^{16})(3.6 \times 10^{-5})$$

$$= \mathbf{1.44 \times 10^{-7}\ C}$$

4.4.5 The Haynes–Shockley Experiment

One of the classic semiconductor experiments is the demonstration of drift and diffusion of minority carriers, first performed by J. R. Haynes and W. Shockley in 1951 at the Bell Telephone Laboratories. The experiment allows independent measurement of the minority carrier mobility μ and diffusion coefficient D. The basic principles of the Haynes–Shockley experiment are as follows: A pulse of holes is created in an n-type bar (for example) that contains an electric field (Fig. 4–18); as the pulse drifts in the field and spreads out by diffusion, the excess hole concentration is monitored at some point down the bar; the time required for the holes to drift a given distance in the field gives a measure of the mobility; and the spreading of the pulse during a given time is used to calculate the diffusion coefficient.

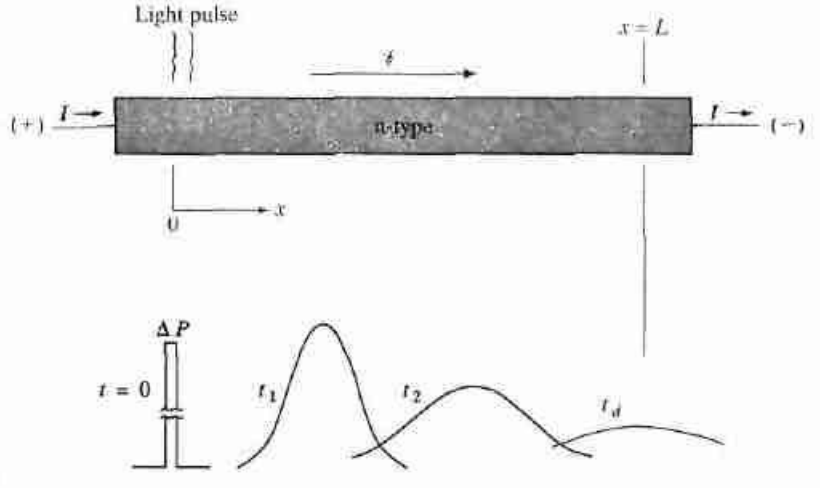


Figure 4–18 Drift and diffusion of a hole pulse in an n-type bar: (a) sample geometry; (b) position and shape of the pulse for several times during its drift down the bar.

In Fig. 4–18 a pulse of excess carriers is created by a light flash at some point $x = 0$ in an n-type semiconductor ($n_0 \gg p_0$). We assume that the excess carriers have a negligible effect on the electron concentration but change the hole concentration significantly. The excess holes drift in the direction of the electric field and eventually reach the point $x = L$, where they are monitored. By measuring the drift time t_d, we can calculate the drift velocity v_d and, therefore, the hole mobility:

$$v_d = \frac{L}{t_d} \tag{4-41}$$

$$\mu_p = \frac{v_d}{\mathscr{E}} \tag{4-42}$$

Thus the hole mobility can be calculated directly from a measurement of the drift time for the pulse as it moves down the bar. In contrast with the Hall effect (Section 3.4.5), which can be used with resistivity to obtain the *majority* carrier mobility, the Haynes–Shockley experiment is used to measure the *minority* carrier mobility.

As the pulse drifts in the $\mathscr{E}$ field it also spreads out by diffusion. By measuring the spread in the pulse, we can calculate D_p. To predict the distribution of holes in the pulse as a function of time, let us first reexamine the case of diffusion of a pulse *without drift, neglecting recombination* (Fig. 4–12). The equation which the hole distribution must satisfy is the time-dependent diffusion equation, Eq. (4–33b). For the case of negligible recombination (τ_p long compared with the times involved in the diffusion), we can write the diffusion equation as

$$\frac{\partial \delta p(x, t)}{\partial t} = D_p \frac{\partial^2 \delta p(x, t)}{\partial x^2} \tag{4-43}$$

The function which satisfies this equation is called a *gaussian distribution,*

$$\delta p(x, t) = \left[\frac{\Delta P}{2\sqrt{\pi D_p t}}\right] e^{-x^2/4D_p t} \tag{4-44}$$

where ΔP is the number of holes per unit area created over a negligibly small distance at $t = 0$. The factor in brackets indicates that the peak value of the pulse (at $x = 0$) decreases with time, and the exponential factor predicts the spread of the pulse in the positive and negative x-directions (Fig. 4–19). If we designate the peak value of the pulse as $\delta\hat{p}$ at any time (say t_d), we can use Eq. (4–44) to calculate D_p from the value of δp at some point x. The most convenient choice is the point $\Delta x/2$, at which δp is down by $1/e$ of its peak value $\delta\hat{p}$. At this point we can write

$$e^{-1}\delta\hat{p} = \delta\hat{p} e^{-(\Delta x/2)^2/4D_p t_d} \tag{4-45}$$

$$D_p = \frac{(\Delta x)^2}{16 t_d} \tag{4-46}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 145

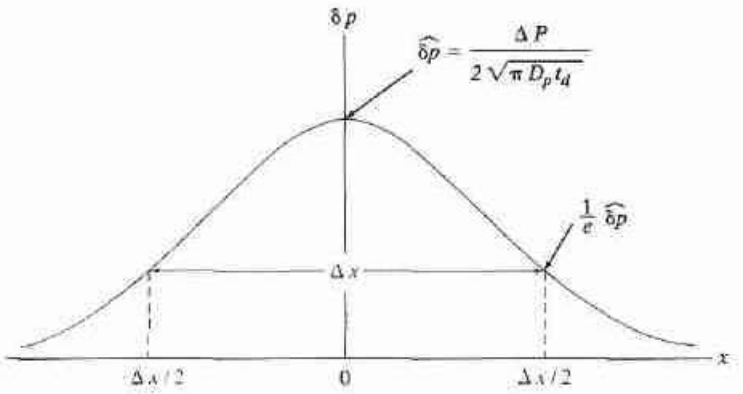


Figure 4–19 Calculation of D_p from the shape of the δp distribution after time t_d. No drift or recombination is included.

Since Δx cannot be measured directly, we use an experimental setup such as Fig. 4–20, which allows us to display the pulse on an oscilloscope as the carriers pass under a detector. As we shall see in Chapter 5, a forward-biased p-n junction serves as an excellent injector of minority carriers, and a reverse-biased junction serves as a detector. The measured quantity in Fig. 4–20 is the pulse width Δt displayed on the oscilloscope in time. It is related to Δx by the drift velocity, as the pulse drifts past the detector point (2)

$$\Delta x = \Delta t v_d = \Delta t \frac{L}{t_d} \tag{4-47}$$

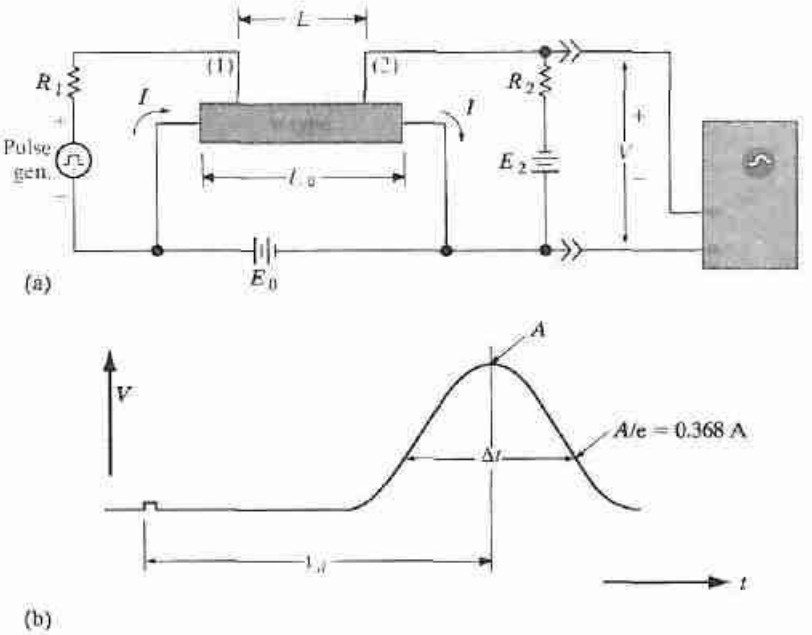


Figure 4–20 The Haynes–Shockley experiment: (a) circuit schematic; (b) typical trace on the oscilloscope screen.

EXAMPLE 4–6

An n-type Ge sample is used in the Haynes-Shockley experiment shown in Fig. 4–20. The length of the sample is 1 cm, and the probes (1) and (2) are separated by 0.95 cm. The battery voltage E_0 is 2 V. A pulse arrives at point (2) 0.25 ms after injection at (1); the width of the pulse Δt is 117 μs. Calculate the hole mobility and diffusion coefficient, and check the results against the Einstein relation.

SOLUTION

$$\mu_p = \frac{v_d}{\mathscr{E}} = \frac{0.95/(0.25 \times 10^{-3})}{2/1} = 1900 \text{ cm}^2/(\text{V-s})$$

$$D_p = \frac{(\Delta x)^2}{16 t_d} = \frac{(\Delta t L)^2}{16 t_d^3}$$

$$= \frac{(117 \times 0.95)^2 \times 10^{-12}}{16(0.25)^3 \times 10^{-9}} = 49.4 \text{ cm}^2/\text{s}$$

$$\frac{D_p}{\mu_p} = \frac{49.4}{1900} = 0.026 = \frac{kT}{q}$$

4.4.6 Gradients in the Quasi-Fermi Levels

In Section 3.5 we saw that equilibrium implies no gradient in the Fermi level E_F. In contrast, any combination of drift and diffusion implies a gradient in the steady state quasi-Fermi level.

We can use the results of Eqs. (4–23), (4–26), and (4–29) to demonstrate the power of the concept of quasi-Fermi levels in semiconductors [see Eq. (4–15)]. If we take the general case of nonequilibrium electron concentration with drift and diffusion, we must write the total electron current as

$$J_n(x) = q\mu_n n(x)\mathscr{E}(x) + qD_n \frac{dn(x)}{dx} \tag{4–48}$$

where the gradient in electron concentration is

$$\frac{dn(x)}{dx} = \frac{d}{dx}\left[n_i e^{(F_n - E_i)/kT}\right] = \frac{n(x)}{kT}\left(\frac{dF_n}{dx} - \frac{dE_i}{dx}\right) \tag{4–49}$$

Using the Einstein relation, the total electron current becomes

$$J_n(x) = q\mu_n n(x)\mathscr{E}(x) + \mu_n n(x)\left[\frac{dF_n}{dx} - \frac{dE_i}{dx}\right] \tag{4–50}$$

But Eq. (4–26) indicates that the subtractive term in the brackets is just $q\mathscr{E}(x)$, giving a direct cancellation of $q\mu_n n(x)\mathscr{E}(x)$ and leaving

$$J_n(x) = \mu_n n(x)\frac{dF_n}{dx} \tag{4–51}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 147

Thus, the processes of electron drift and diffusion are summed up by the spatial variation of the quasi-Fermi level. The same derivation can be made for holes, and we can write the current due to drift and diffusion in the form of a *modified Ohm's law*

$$J_n(x) = q\mu_n n(x)\frac{d(F_n/q)}{dx} = \sigma_n(x)\frac{d(F_n/q)}{dx} \tag{4-52a}$$

$$J_p(x) = q\mu_p p(x)\frac{d(F_p/q)}{dx} = \sigma_p(x)\frac{d(F_p/q)}{dx} \tag{4-52b}$$

Therefore, any drift, diffusion, or combination of the two in a semiconductor results in currents proportional to the gradients of the two quasi-Fermi levels. Conversely, a lack of current implies constant quasi-Fermi levels. One can use a hydrostatic analogy for quasi-Fermi levels and identify it as water pressure in a system. Just as water flows from a high-pressure region to a low-pressure region, until in equilibrium the water pressure is the same everywhere, similarly electrons flow from a high- to low-electron quasi-Fermi level region, until we get a flat Fermi level in equilibrium. Quasi-Fermi levels are sometimes also known as electrochemical potentials because, as we just saw, the driving force for carriers is governed partly by gradients of electrical potential (or electric field), which determines drift, and partly by gradients of carrier concentration (which is related to a thermodynamic concept called *chemical potential*), giving rise to diffusion.

SUMMARY

4.1 *Excess* carriers, above the equilibrium values contributed by doping, may be created *optically* (or by electrical *biasing* in devices). *Generation–recombination* (G–R) of electron–hole pairs (*EHPs*) can occur by absorption of the photons with energy greater than the band gap, balanced by direct or indirect recombination.

4.2 G–R processes can be mediated by *traps*, especially deep traps near midgap. Band-to-band or trap-assisted G–R processes lead to an average *lifetime* for the excess carriers. Carrier lifetime multiplied by the optical generation rate establishes a steady state excess population of carriers. The square root of carrier lifetime multiplied by the diffusion coefficient determines the diffusion length.

4.3 In *equilibrium*, we have a *constant Fermi level*. In nonequilibrium with excess carriers, Fermi levels are generalized to separate *quasi-Fermi levels* for electrons and holes. The quasi-Fermi level splitting is a measure of the departure from equilibrium. Minority carrier quasi-Fermi levels change more than majority carrier quasi-Fermi levels because the relative change of minority carriers is larger. Gradients in the quasi-Fermi level determine the net drift–diffusion current.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 148

4.4 *Diffusion flux* measures the flow of carriers from *high- to low-concentration* regions and is given by the *diffusivity* times the concentration *gradient*. The direction of diffusion current is opposite to the flux for the negative electrons, but in the same direction for the positive holes. Carrier diffusivity is related to mobility by the thermal voltage kT/q (*Einstein relation*).

4.5 When carriers move in a semiconductor due to drift or diffusion, the time-dependent carrier concentrations at different points is given by the *carrier continuity* equation, which says that if more carriers flow into a point than flow out, the concentration will increase as a function of time and vice versa. G–R processes also affect carrier concentrations.

PROBLEMS

4.1 With E_F located 0.4 eV above the valence band in a Si sample, what charge state would you expect for most Ga atoms in the sample? What would be the predominant charge state of Zn? Au? *Note:* By charge state we mean neutral, singly positive, doubly negative, etc.

4.2 A Si sample with $10^{16}/cm^3$ donors is optically excited such that $10^{19}/cm^3$ electron–hole pairs are generated per second uniformly in the sample. The laser causes the sample to heat up to 450 K. Find the quasi-Fermi levels and the change in conductivity of the sample upon shining the light. Electron and hole lifetimes are both 10 μs. $D_p = 12\ cm^2/s$; $D_n = 36\ cm^2/s$; $n_i = 10^{14}\ cm^{-3}$ at 450 K. What is the change in conductivity upon shining light?

4.3 Construct a semilogarithmic plot such as Fig. 4–7 for Si doped with 2×10^{15} donors/cm^3 and having 4×10^{14} EHP/cm^3 created uniformly at $t = 0$. Assume that $\tau_n = \tau_p = 5$ μs.

4.4 Calculate the recombination coefficient α_r for the low-level excitation described in Prob. 4.3. Assume that this value of α_r applies when the GaAs sample is uniformly exposed to a steady state optical generation rate $g_{op} = 10^{19}$ EHP/cm^3-s. Find the steady state excess carrier concentration $\Delta n = \Delta p$.

4.5 An intrinsic Si sample is doped with donors from one side such that $N_d = N_0 \exp(-ax)$. (a) Find an expression for the built-in electric field at equilibrium over the range for which $N_d \gg n_i$. (b) Evaluate the field when $a = 1\ (\mu m)^{-1}$. (c) Sketch a band diagram such as in Fig. 4–15 and indicate the direction of the field.

4.6 A Si sample with $10^{15}/cm^3$ donors is uniformly optically excited at room temperature such that $10^{19}/cm^3$ electron–hole pairs are generated per second. Find the separation of the quasi-Fermi levels and the change of conductivity upon shining the light. Electron and hole lifetimes are both 10 μs. $D_p = 12\ cm^2/s$.

4.7 An n-type Si sample with $N_d = 10^{15}\ cm^{-3}$ is steadily illuminated such that $g_{op} = 10^{21}$ EHP/cm^3-s. If $\tau_n = \tau_p = 1$ μs for this excitation, calculate the separation in the quasi-Fermi levels, $(F_n - F_p)$. Draw a band diagram such as Fig. 4–11.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 149

4.8 For a 2-cm-long doped Si bar ($N_d = 10^{16}$ cm^{-3}) with a cross-sectional area = 0.05 cm^2, what is the current if we apply 10V across it? If we generate 10^{20} electron-hole pairs per second per cm^3 uniformly in the bar and the lifetime $\tau_n = \tau_p = 10^{-4}$ s, what is the new current? Assume the low-level α, doesn't change for high-level injection. If the voltage is then increased to 100,000 V, what is the new current? Assume $\mu_p = 500$ cm^2/V-s, but you must choose the appropriate value for electrons.

4.9 Design and sketch a photoconductor using a 5-μm-thick film of CdS, assuming that $\tau_n = \tau_p = 10^{-6}$ s and $N_d = 10^{14}$ cm^{-3}. The dark resistance (with $g_{op} = 0$) should be 10 MΩ, and the device must fit in a square 0.5 cm on a side; therefore, some sort of folded or zigzag pattern is in order. With an excitation of $g_{op} = 10^{21}$ EHP/cm^3-s, what is the resistance change?

4.10 A 100-mW laser beam with wavelength $\lambda = 6328$ Å is focused onto a GaAs sample 100 μm thick. The absorption coefficient at this wavelength is 3×10^4 cm^{-1}. Find the number of photons emitted per second by radiative recombination in the GaAs, assuming perfect quantum efficiency. What power is delivered to the sample as heat?

4.11 Assume that a photoconductor in the shape of a bar of length L and area A has a constant voltage V applied, and it is illuminated such that g_{op} EHP/cm^3-s are generated uniformly throughout. If $\mu_n \gg \mu_p$, we can assume the optically induced change in current ΔI is dominated by the mobility μ_n and lifetime τ_n for electrons. Show that $\Delta I = qAL g_{op}\tau_n/\tau_t$ for this photoconductor, where τ_t is the transit time of electrons drifting down the length of the bar.

4.12 For the steady state minority hole distribution shown in Fig. 4–17, find the expression for the hole quasi-Fermi level position $E_i - F_p(x)$ while $p(x) \gg p_0$ (i.e., while F_p is below E_F). On a band diagram, draw the variation of $F_p(x)$. Be careful—when the minority carriers are few (e.g., when δp is n_i), F_p still has a long way to go to reach E_F.

4.13 In an n-type semiconductor bar, there is an increase in electron concentration from left to right and an electric field pointing to the left. With a suitable sketch, indicate the directions of the electron drift and diffusion current flow and explain why. If we double the electron concentration everywhere, what happens to the diffusion current and the drift current? If we add a constant concentration of electrons everywhere, what happens to the drift and diffusion currents? Explain your answers with appropriate equations.

4.14 The current required to feed the hole injection at $x = 0$ in Fig. 4–17 is obtained by evaluating Eq. (4–40) at $x = 0$. The result is $I_p(x = 0) = qAD_p\Delta p/L_p$. Show that this current can be calculated by integrating the charge stored in the steady state hole distribution $\delta p(x)$ and then dividing by the average hole lifetime τ_p. Explain why this approach gives $I_p(x = 0)$.

4.15 The direction of the built-in electric field can be deduced without math by sketching the result of a doping gradient on the band diagram. Starting with a flat Fermi level at equilibrium, place E_i near or far from E_F as the doping is

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 150

varied for the two cases of a gradient in donor or acceptor doping as in Prob. 4.5. Show the electric field direction in each case, based on Eq. (4–26). If a *minority* carrier is injected into the impurity gradient region, in what direction is it accelerated in the two cases? This is an interesting effect that we will use later in discussing bipolar transistors.

4.16 In Prob. 4.5, the direction of the built-in electric field due to a gradient in doping was determined from Eqs. (4–23) and (4–26). In this problem, you are asked to explain qualitatively why the field must arise and find its direction. (a) Sketch a donor doping distribution as in Prob. 4.5, and explain the field required to keep the mobile electrons from diffusing down the gradient. Repeat for acceptors and holes. (b) Sketch a microscopic region of the doping distribution, showing ionized donors and the resulting mobile electrons. Explain the origin and direction of the field as the electrons attempt to diffuse toward lower concentrations. Repeat for acceptors and holes.

4.17 We wish to use the Haynes–Shockley experiment to calculate the hole lifetime τ_p in an n-type sample. Assume that the peak voltage of the pulse displayed on the oscilloscope screen is proportional to the hole concentration under the collector terminal at time t_d and that the displayed pulse can be approximated as a gaussian, as in Eq. (4–44), which decays due to recombination by e^{-t/τ_p}. The electric field is varied and the following data are taken: For $t_d = 200$ μs, the peak is 20 mV; for $t_d = 50$ μs, the peak is 80 mV. What is τ_p?

READING LIST

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

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

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

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

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

Pierret, R. F. *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley, 1996.

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

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

SELF QUIZ

Question 1

Consider a p-type semiconductor that has a band gap of 1.0 eV and a minority electron lifetime of 0.1 μs, and is uniformly illuminated by light having photon energy of 2.0 eV.

(a) What rate of uniform excess carrier generation is required to generate a uniform electron concentration of $10^{10}/cm^3$?

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 151

(b) How much optical power per cm^3 must be absorbed in order to create the excess carrier population of part (a)? (You may leave your answer in units of $eV/s\text{-}cm^3$.)

(c) If the carriers recombine via photon emission, approximately how much optical power per cm^3 will be generated? (You may leave your answer in units of $eV/s\text{-}cm^3$.)

Question 2

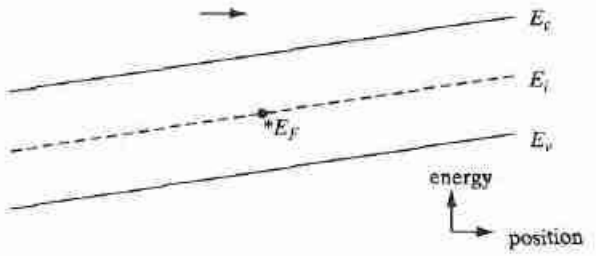
(a) What do we mean by "deep" versus "shallow" traps? Which are more harmful for semiconductor devices and why? What is an example of a deep trap in Si?

(b) Are absorption lengths of slightly above band gap photons longer in Si or GaAs? Why?

(c) Do absorption coefficients of photons increase or decrease with photon energy? Why?

Question 3

Consider the following equilibrium band diagram for a portion of a semiconductor sample with a built-in electric field $\mathscr{E}$:

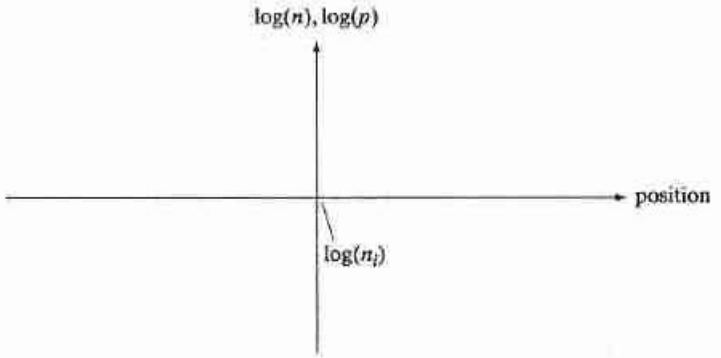


(a) Sketch the Fermi level as a function of position *through* the indicated point, E_F, across the width of the band diagram above.

(b) On the band diagram, sketch the direction of the electric field. Is the field constant or position dependent?

(c) On the following graph, sketch and label both the electron and hole concentrations as a function of position across the full width of the sample. Note that the carrier concentration scale is logarithmic such that exponential variations in the carrier concentration with position appear as straight lines. Note also that the horizontal axis corresponds to the intrinsic carrier concentration of n_i.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 152



Question 4

(a) Indicate the directions of the hole and electron flux densities ϕ due to diffusion and drift under these *equilibrium* conditions corresponding to the previous Question 3.

(b) Indicate the directions of the hole and electron current densities j due to diffusion and drift under these *equilibrium* conditions.

Question 5

(a) What are the relevant equations that must be solved in general for a semiconductor device problem?

(b) In general how many components of conduction current can you have in a semiconductor device? What are they?

Question 6

(a) Consider a region in a semiconductor with an electric field directed toward the right ($\rightarrow$) and carrier concentrations increasing toward the left ($\leftarrow$). Indicate the directions of particle fluxes ϕ (circle one for each) and charge currents j due to drift and diffusion within that region.

(b) Based on your answers to part (a), indicate the directions of the charge currents j due to drift and diffusion within that region.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 153

Chapter 5
Junctions

OBJECTIVES

1. Determine the band diagram of a p–n junction in equilibrium and use the Poisson equation to calculate electric fields and potentials
2. Determine current flow components in an "ideal" diode, and why reverse leakage in an ideal diode is independent of bias; study applications in rectifiers
3. Understand depletion capacitance due to dopant charges, and diffusion capacitance due to mobile carriers
4. Study second-order effects—high-level injection, generation–recombination in depletion region, series resistance, and graded junctions
5. Study metal–semiconductor junctions (Schottky and ohmic) and heterojunctions, in terms of vacuum level, electron affinity, and work function

Most semiconductor devices contain at least one junction between p-type and n-type material. These p-n junctions are fundamental to the performance of functions such as rectification, amplification, switching, and other operations in electronic circuits. In this chapter we shall discuss the equilibrium state of the junction and the flow of electrons and holes across a junction under steady state and transient conditions. This is followed by a discussion of metal–semiconductor junctions and heterojunctions between semiconductors having different band gaps. With the background provided in this chapter on junction properties, we can then discuss specific devices in later chapters.

5.1 FABRICATION OF p-n JUNCTIONS

Although this book deals primarily with how devices work rather than how they are made, it is instructive to have an overview of the fabrication process in order to appreciate device physics. We have already discussed in Chapter 1 how single-crystal substrates and epitaxial layers needed for high quality devices are grown, and how the doping can be varied as a function of depth. However, we have not discussed how doping can be varied laterally across the surface, which is key to making integrated circuits on a wafer. Hence, it is necessary to be able to form patterned masks on the wafer corresponding to the circuitry, and introduce the dopants selectively through windows in the mask. We will first briefly describe the major process steps that form the underpinnings of modern integrated circuit manufacturing. Relatively few

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 154

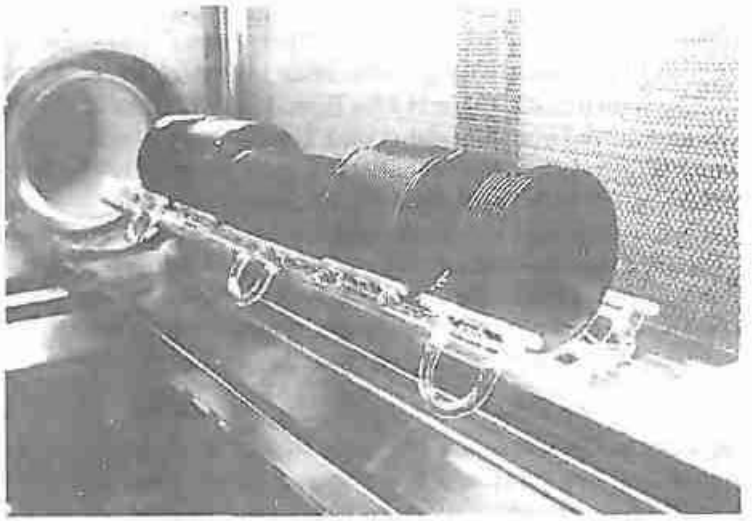


Figure 5–1a
Silicon wafers being loaded into a furnace. For 8-inch and larger wafers, this type of horizontal loading is often replaced by a vertical furnace.

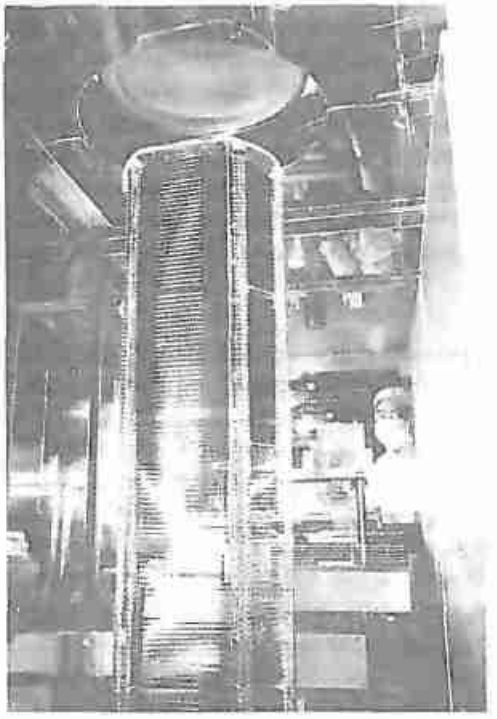


Figure 5–1b
Vertical furnace for large Si wafers. The silica wafer holder is loaded with 8-inch Si wafers and moved into the furnace above for oxidation, diffusion, or deposition operations. (Photograph courtesy of Tokyo Electron Ltd.)

unit process steps can be used in different permutations and combinations to make everything from simple diodes to the most complex microprocessors.

5.1.1 Thermal Oxidation

Many fabrication steps involve heating up the wafer in order to enhance a chemical process. An important example of this is thermal oxidation of Si to

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 155

form SiO_2. This involves placing a batch of wafers in a clean silica (quartz) tube which can be heated to very high temperatures (~800–1000°C) using heating coils in a furnace with ceramic brick insulating liners. An oxygen-containing gas such as dry O_2 or H_2O is flowed into the tube at atmospheric pressure, and flowed out at the other end. Traditionally, horizontal furnaces were used (Fig. 5-1a). More recently, it has become common to employ vertical furnaces (Fig. 5-1b). A batch of Si wafers is placed in the silica wafer holders, each facing down to minimize particulate contamination. The wafers are then moved into the furnace. The gases flow in from the top and flow out at the bottom, providing more uniform flow than in conventional horizontal furnaces. The overall reactions that occur during oxidation are:

$$Si + O_2 \rightarrow SiO_2 \text{ (dry oxidation)}$$

$$Si + 2H_2O \rightarrow SiO_2 + 2H_2 \text{ (wet oxidation)}$$

In both cases, Si is consumed from the surface of the substrate. For every micron of SiO_2 grown, 0.44 μm of Si is consumed, leading to a 2.2× volume expansion of the consumed layer upon oxidation. The oxidation proceeds by having the oxidant (O_2 or H_2O) molecules diffuse through the already grown oxide to the Si-SiO_2 interface, where the above reactions take place. One of the very important reasons why Si integrated circuits exist (and by extension why modern computers exist) is that a stable thermal oxide can be grown on Si with excellent interface electrical properties. Other semiconductor materials do not have such a useful native oxide. We can argue that modern electronics and computer technology owe their existence to this simple oxidation process.

Plots of oxide thickness as a function of time, at different temperatures, are shown for dry and wet oxidation of (100) Si in Appendix VI.

5.1.2 Diffusion

Another thermal process that was used extensively in IC fabrication in the past is thermal in-diffusion of dopants in furnaces such as those shown in Fig. 5-1a. The wafers are first oxidized and windows are opened in the oxide using the photolithography and etching steps described in Sections 5.1.6 and 5.1.7, respectively. Dopants such as B, P, or As are introduced into these patterned wafers in a high temperature (~800–1100°C) diffusion furnace, generally using a gas or vapor source. The dopants are gradually transported from the high concentration region near the surface into the substrate through diffusion, similar to that described for carriers in Section 4.4. The maximum number of impurities that can be dissolved (the solid solubility) in Si is shown for various impurities as a function of temperature in Appendix VII. The diffusivity of dopants in solids, D, has a strong Arrhenius dependence on temperature, T. It is given by $D = D_0 \exp -(E_A/kT)$, where D_0 is a constant depending on the material and the dopant, and E_A is the activation energy. The average distance the dopants diffuse is related to the diffusion length as in Section 4.4.4. In this case, the diffusion length is $\sqrt{Dt}$, where t is the processing time. The product Dt is sometimes called the *thermal budget*. The Arrhenius dependence of diffusivity on temperature explains why high temperatures are required for diffusion; otherwise, the diffusivities are far too

of f download.ir
Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 156

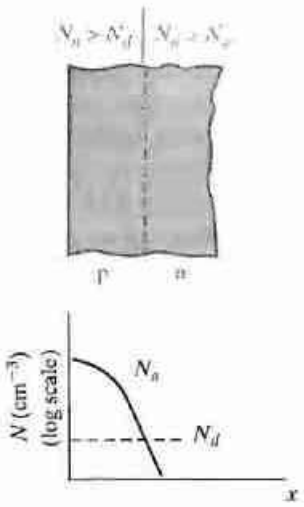


Figure 5–2
Impurity concentration profile for fabricating a p-n junction by diffusion.

low. Since D varies exponentially with T, it is critical to have very precise control over the furnace temperatures, within several degrees, in order to have control over the diffusion profiles (Fig. 5-2). The dopants are effectively blocked or masked by the oxide because their diffusivity in oxide is very low. The diffusivities of various dopants in Si and SiO_2 are shown as a function of temperature in Appendix VIII. Difficulty with profile control and the very high temperature requirement has led to diffusion being supplanted by ion implantation as a doping technique, as discussed in Section 5.1.4.

The trend of using larger Si wafers has changed many processing steps. For example, eight-inch and larger wafers are best handled in a vertical furnace (Fig. 5-1b) rather than the traditional horizontal furnace (Fig. 5-1a). Also, large wafers are often handled individually for a variety of deposition, etching, and implantation processes. Such single-wafer processing has led to development of robotic systems for fast and accurate wafer handling.

The distribution of impurities in the sample at any time during the diffusion can be calculated from a solution of the diffusion equation with appropriate boundary conditions. If the source of dopant atoms at the surface of the sample is limited (e.g., a given number of atoms deposited on the Si surface before diffusion), a gaussian distribution as described by Eq. (4-44) (for $x > 0$) is obtained. On the other hand, if the dopant atoms are supplied continuously, such that the concentration at the surface is maintained at a constant value, the distribution follows what is called a *complementary error function*. In Fig. 5-2, there is some point in the sample at which the introduced acceptor concentration just equals the background donor concentration in the originally n-type sample. This point is the location of the p-n junction. To the left of this point in the sample of Fig. 5-2, acceptor atoms predominate and the material is p-type, whereas to the right of the junction, the background donor atoms predominate and the material is n-type. The depth of the junction beneath the surface of the sample can be controlled by the time and temperature of the diffusion (Prob. 5.2).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 157

In the horizontal diffusion furnace shown in Fig. 5-1a, Si wafers are placed in the tube during diffusion, and the impurity atoms are introduced into the gas which flows through the silica tube. Common impurity source materials for diffusions in Si are B_2O_3, BBr_3, and BCl_3 for boron; phosphorus sources include PH_3, P_2O_5, and $POCl_3$. Solid sources are placed in the silica tube upstream from the sample or in a separate heating zone of the furnace; gaseous sources can be metered directly into the gas flow system; and with liquid sources inert carrier gas is bubbled through the liquid before being introduced into the furnace tube. The Si wafers are held in a silica "boat" (Fig. 5-1a) which can be pushed into position in the furnace and removed by a silica rod.

It is important to remember the degree of cleanliness required in these processing steps. Since typical doping concentrations represent one part per million or less, cleanliness and purity of materials are critically important. Thus the impurity source and carrier gas must be extremely pure; the silica tube, sample holder, and pushrod must be cleaned and etched in hydrofluoric acid (HF) before use (once in use, the tube cleanliness can be maintained if no unwanted impurities are introduced); finally, the Si wafers themselves must undergo an elaborate cleaning procedure before diffusion, including a final etch containing HF to remove any unwanted SiO_2 from the surface.

5.1.3 Rapid Thermal Processing

Increasingly, many thermal steps formerly performed in furnaces are being done using what is called *rapid thermal processing* (RTP). This includes rapid thermal oxidation, annealing of ion implantation, and chemical vapor deposition, which are discussed in the following paragraphs. A simple RTP system is shown in Fig. 5-3. Instead of having a large batch of wafers in a conventional furnace where the temperature cannot be changed rapidly, a single wafer is held (face down to minimize particulates) on low-thermal-mass quartz pins, surrounded by a bank of high-intensity (tens of kW) tungsten-halogen infrared lamps, with gold-plated reflectors around them. By turning on the lamps, the high intensity infrared radiation shines through the quartz chamber and is absorbed by the wafer, causing its temperature

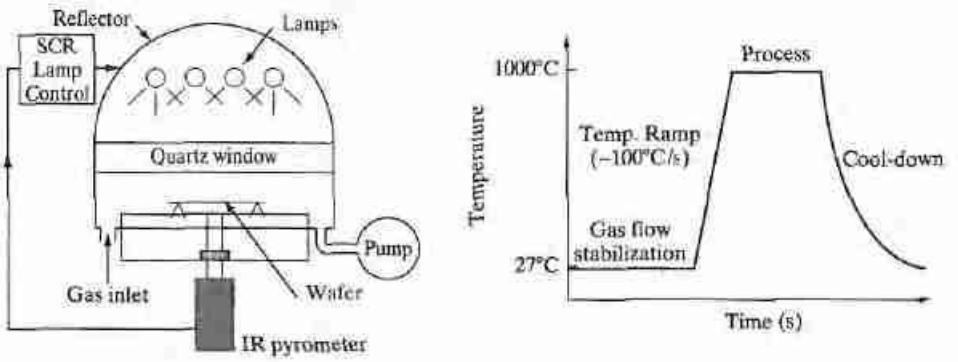


Figure 5-3
Schematic diagram of a rapid thermal processor and typical time–temperature profile.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 158

to rise *very* rapidly (~50−100°C/s). The processing temperature can be reached quickly, after the gas flows have been stabilized in the chamber. At the end of the process, the lamps are turned off, allowing the wafer temperature to drop rapidly, once again because of the much lower thermal mass of an RTP system compared to a furnace. In RTP, therefore, temperature is essentially used as a "switch" to start or quench the reaction. Two critical aspects of RTP are ensuring temperature uniformity across large wafers, and accurate temperature measurement, for example with thermocouples or pyrometers.

A key parameter in all thermal processing steps is the thermal budget, Dt. Generally speaking, we try to minimize this quantity because an excessive Dt product leads to loss of control over compact doping profiles, which is detrimental to ultra-small devices. In furnace processing, thermal budgets are minimized by operating at as low a temperature as feasible so that D is small. On the other hand, RTP operates at higher temperatures (~1000°C) but does so for only a few seconds (compared to minutes or hours in a furnace).

5.1.4 Ion Implantation

A useful alternative to high-temperature diffusion is the direct implantation of energetic ions into the semiconductor. In this process a beam of impurity ions is accelerated to kinetic energies ranging from several keV to several MeV and is directed onto the surface of the semiconductor. As the impurity atoms enter the crystal, they give up their energy to the lattice in collisions and finally come to rest at some average penetration depth, called the *projected range*. Depending on the impurity and its implantation energy, the range in a given semiconductor may vary from a few hundred angstroms to about 1 μm. For most implantations the ions come to rest distributed almost evenly about the projected range R_p, as shown in Fig. 5-4. An implanted dose of ϕ ions/cm^2 is distributed approximately by a gaussian formula

$$N(x) = \frac{\phi}{\sqrt{2\pi}\Delta R_p} \exp\left[-\frac{1}{2}\left(\frac{x - R_p}{\Delta R_p}\right)^2\right] \qquad (5\text{-}12)$$

where ΔR_p, called the *straggle*, measures the half-width of the distribution at $e^{-1/2}$ of the peak (Fig. 5-4). Both R_p and ΔR_p increase with increasing implantation energy. These parameters are shown as a function of energy for various implant species into Si in Appendix IX.

An ion implanter is shown schematically in Fig. 5-5. A gas containing the desired impurity is ionized within the *source* and is then extracted into the *acceleration tube*. After acceleration to the desired kinetic energy, the ions are passed through a *mass separator* to ensure that only the desired ion species enters the *drift tube*.[1] The ion beam is then focused and scanned electrostatically over the surface of the wafer in the *target chamber*. Repetitive scanning in a raster pattern provides exceptionally uniform doping of the wafer surface. The target chamber commonly includes automatic wafer-handling facilities to speed up the process of implanting many wafers per hour.

[1] In many ion implanters the mass separation occurs before the ions are accelerated to high energy.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 159

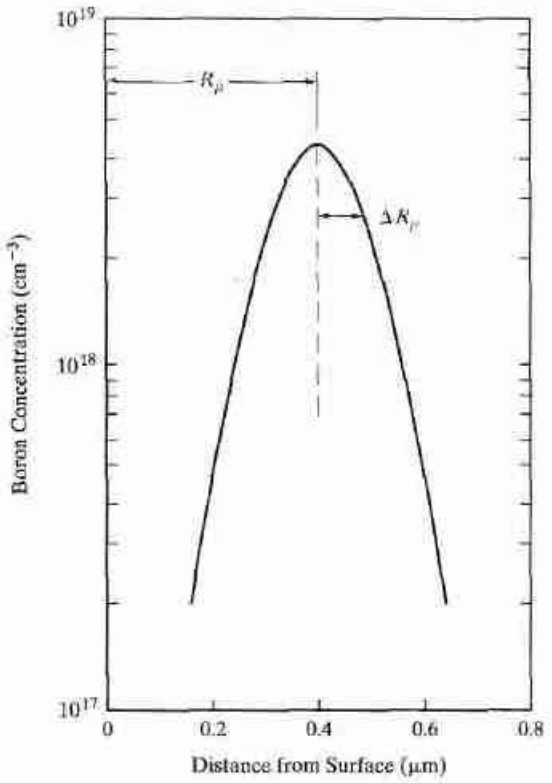


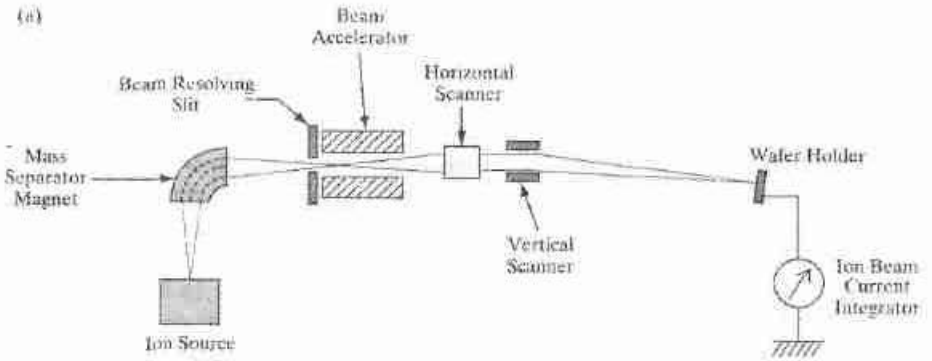
Figure 5–4 Distributions of implanted impurities: gaussian distribution of boron atoms about a projected range R_p (in this example, a dose of 10^{14} B atoms/cm^2 implanted at 140 keV).

An obvious advantage of implantation is that it can be done at relatively low temperatures; this means that doping layers can be implanted without disturbing previously diffused regions. The ions can be blocked by metal or photoresist layers; therefore, the photolithographic techniques described in Section 5.1.6 can be used to define ion implanted doping patterns. Very shallow (tenths of a micron) and well-defined doping layers can be achieved by this method. As we shall see in later chapters, many devices require thin doping regions and may be improved by ion implantation techniques. Furthermore, it is possible to implant impurities which do not diffuse conveniently into semiconductors.

One of the major advantages of implantation is the precise control of doping concentration it provides. Since the ion beam current can be measured accurately during implantation, a precise quantity of impurity can be introduced. This control over doping level, along with the uniformity of the implant over the wafer surface, make ion implantation particularly attractive for the fabrication of Si integrated circuits (Chapter 9).

One problem with this doping method is the lattice damage which results from collisions between the ions and the lattice atoms. However, most of this damage can be removed in Si by heating the crystal after the implantation. This process is called *annealing*. Although Si can be heated to

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 160



(b)

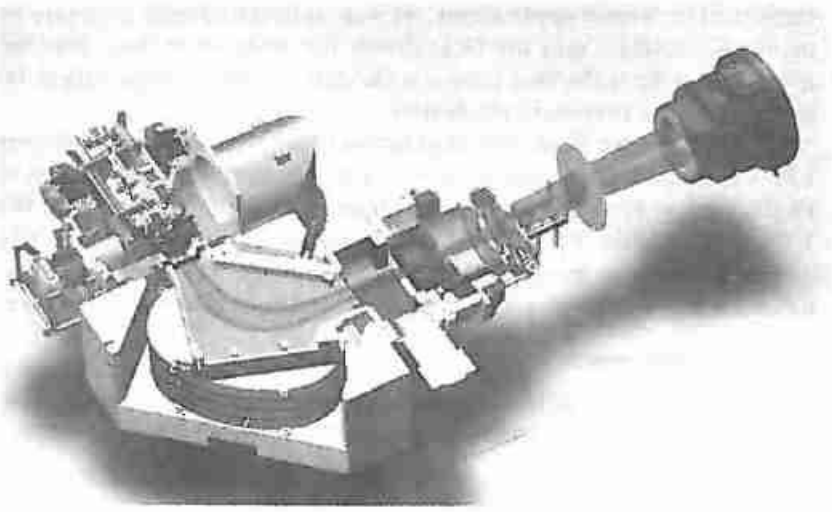


Figure 5–5
(a) Schematic diagram of an ion implantation system; (b) schematic of ion beam path through mass analyzer magnet. (Courtesy of Applied Materials.)

temperatures in excess of 1000°C without difficulty, GaAs and some other compounds tend to dissociate at high temperatures. For example, As evaporation from the surface of GaAs during annealing damages the sample. Therefore, it is common to encapsulate the GaAs with a thin layer of silicon nitride during the anneal. Another approach to annealing either Si or compounds is to heat the sample only briefly (e.g., 10 s) using RTP, rather than a conventional furnace. Annealing leads to some unintended diffusion of the implanted species. It is desirable to minimize this diffusion by optimizing the annealing time and temperature. The profile after annealing is given by

$$N(x) = \frac{\phi}{\sqrt{2\pi}(\Delta R_p^2 + 2Dt)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{(x - R_p)^2}{\Delta R_p^2 + 2Dt}\right)\right] \quad (5\text{-}1b)$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 161

5.1.5 Chemical Vapor Deposition (CVD)

At various stages of device fabrication, thin films of dielectrics, semiconductors and metals have to be formed on the wafer and then patterned and etched. We have already discussed one important example of this involving thermal oxidation of Si. SiO_2 films can also be formed by *low pressure* (~100 mTorr)[2] chemical vapor deposition (LPCVD) (Fig. 5-6) or plasma-enhanced CVD (PECVD). The key differences are that thermal oxidation consumes Si from the substrate, and very high temperatures are required, whereas CVD of SiO_2 does not consume Si from the substrate and can be done at much lower temperatures. The CVD process reacts a Si-containing gas such as SiH_4 with an oxygen-containing precursor, causing a chemical reaction, leading to the deposition of SiO_2 on the substrate. Being able to deposit SiO_2 is very important in certain applications. As a complicated device structure is built up, the Si substrate may not be available for reaction, or there may be metallization on the wafer that cannot withstand very high temperatures. In such cases, CVD is a necessary alternative.

Although we have used deposition of SiO_2 as an important example, LPCVD is also widely used to deposit other dielectrics such as silicon nitride (Si_3N_4), and polycrystalline or amorphous Si. It should also be clear that the VPE of Si or MOCVD of compound semiconductors discussed in Chapter 1 is really a special, more challenging example of CVD where not only must a film be deposited, but single-crystal growth must also be maintained.

5.1.6 Photolithography

Patterns corresponding to complex circuitry are formed on a wafer using *photolithography*. This involves first generating a *reticle* which is a transparent

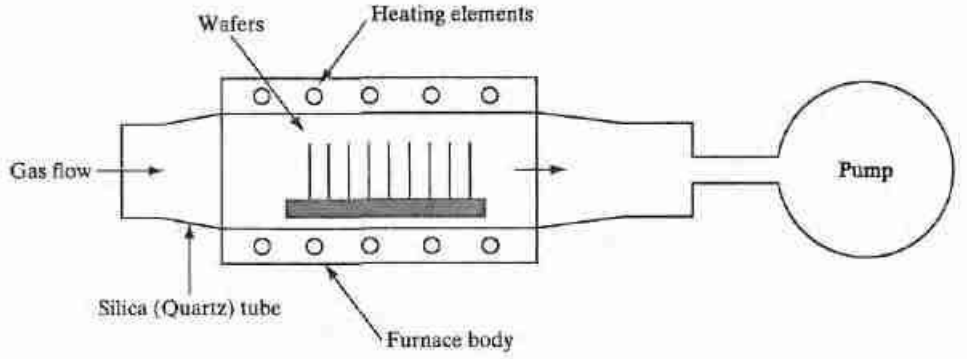


Figure 5-6
Low-pressure chemical vapor deposition (LPCVD) reactor.

[2]Torr or Torricelli = 1 mm Hg or 133 Pa.

silica (quartz) plate containing the pattern (Fig. 5-7a). Opaque regions on the mask are made up of an ultraviolet light-absorbing layer, such as iron oxide. The reticle typically contains the patterns corresponding to a single *chip* or *die*, rather than the entire wafer (in which case it would be called a *mask*). It is usually created by a computer controlled electron beam driven by the circuit layout data, using pattern generation software. A thin layer of electron beam sensitive material called electron beam resist is placed on the iron-oxide-covered quartz plate, and the resist is exposed by the electron beam. A resist is a thin organic polymer layer that undergoes chemical changes if it is exposed to energetic particles such as electrons or photons. The resist is exposed selectively, corresponding to the patterns that are required. After exposure, the resist is *developed* in a chemical solution. There are two types of resist. The developer is either used to remove the exposed (*positive* resist) or unexposed (*negative* resist) material. The iron oxide layer is then selectively etched off in a plasma to generate the appropriate patterns. The reticle can be used repeatedly to pattern Si wafers. To make a typical integrated circuit, a dozen or more reticles are required, corresponding to different process steps.

The Si wafers are first covered with an ultraviolet light-sensitive organic material or photoemulsion called *photoresist* by dispensing the liquid resist onto the wafer and spinning it rapidly (~3000 rpm) to form a uniform coating (~0.5 μm). As mentioned above, there are two types of resist—negative, which forms the opposite polarity image on the wafer compared to that on the reticle, and positive (same polarity). Currently, positive resist has supplanted negative because it can achieve far better resolution, down



Figure 5-7a
A photolithographic reticle used for one step in the processing of a 16-Mb dynamic random-access memory (DRAM). In a "stepper" projection exposure system, ultraviolet light shines through the glass plate and the image is projected onto the wafer to expose photoresist for one die in the array of circuits and then steps to the next. (Photograph courtesy of IBM Corp.)

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 163

to ~0.25 μm using ultraviolet light. The light shines on the resist-covered wafer through the reticle, causing the exposed regions to become acidified. Subsequently, the exposed wafers are developed in a basic solution of NaOH, which causes the exposed resist to etch away. Thereby, the pattern on the reticle is transferred to the die on the wafer. After the remaining resist is cured by baking at ~125°C in order to harden it, the appropriate process step can be performed, such as implanting dopants through windows in the resist pattern or plasma etching of the underlying layers.

The exposure of the wafers is achieved die-by-die in a step-and-repeat system called a *stepper* (Fig. 5-7b). As the name implies, the ultraviolet light shines selectively through the reticle onto a single die location. After the photoexposure is done, the wafer mechanically translates on a precisely controlled x-y translation stage to the next die location and is exposed again. It

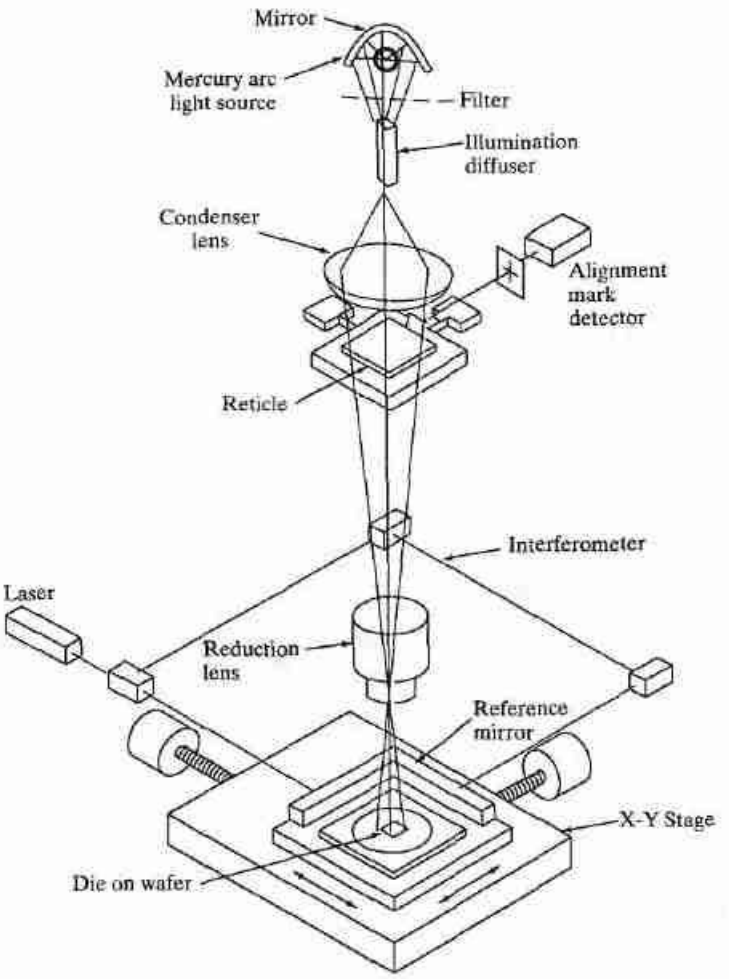


Figure 5-7b Schematic diagram of an optical stepper.

is very important to be able to precisely align the patterns on the reticle with respect to pre-existing patterns on the wafer, which is why these tools are also sometimes known as *mask aligners*. An advantage of such a "stepper" projection system is that refocusing and realignment can be done at each die to accommodate slight variations in surface flatness across the wafer. This is especially important in printing ultra-small linewidths over a very large wafer. The success of modern IC manufacture has depended on numerous advances in deep ultraviolet light sources, precision optical projection systems, techniques for registration between masking layers, and stepper design.

What makes photolithography (along with etching) so critical is that it obviously determines how small and closely packed the individual devices (e.g. transistors) can be made. We shall see that smaller devices operate better in terms of higher speed and lower power dissipation. What makes modern lithography so challenging is the fact that pattern dimensions are comparable to the wavelength of light that is used. Under these circumstances we cannot treat light propagation using simple geometrical ray optics; rather, the wave nature of light is manifested in terms of diffraction, which makes it harder to control the patterns. The *diffraction-limited minimum geometry* is given by

$$l_{min} = 0.8\,\lambda/NA \tag{5-2a}$$

where λ is the wavelength of the light and NA (~0.5) is the numerical aperture or "size" of the lens used in the aligner. This expression implies that for finer patterns, we should work with larger (and, therefore, more expensive) lenses and shorter wavelengths. As a result, smaller geometries require shorter wavelengths. This has led the push to replace UV mercury lamp sources (0.365 μm), shown in Fig. 5-7b, with argon fluoride (ArF) excimer lasers (λ = 0.193 μm). Novel exposure techniques employing phase-shift masks, optical proximity correction, and off-axis illumination all exploit Fourier optics to allow resolution near or below the dimension of the wavelength being used. There is also interest in extreme ultraviolet (EUV) sources (13 nm), which use a plasma to generate such short wavelengths, for next-generation lithography. X-ray lithography involving even shorter wavelengths has been the source of much research for many years, but does not seem practical in a manufacturing environment. A promising advancement involves imprinting patterns in a photoresist using a physical pattern, or mold, thereby overcoming the diffraction limits of optical lithography.

The other key parameter in lithography is the so-called *depth-of-focus* (DOF), which is given by

$$\mathrm{DOF} = \frac{\lambda}{2\,(NA)^2} \tag{5-2b}$$

The DOF tells us the range of distances around the focal plane where the image quality is sharp. Unfortunately, this expression implies that exposure

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 165

with very short wavelengths leads to poor DOF. This is a big challenge because the topography or the "hills and valleys" on a chip during processing can be larger than the DOF allowed by the optics.

We must therefore add steps in the fabrication process to planarize the surface using *chemical mechanical polishing* (CMP). As the name implies, the planarizing process is partly chemical in nature (using a basic solution), and partly mechanical grinding of the layers using an abrasive slurry. As described in Section 1.3.3, CMP can be achieved using a slurry of fine SiO_2 particles in an NaOH solution.

The expression for diffraction-limited geometry [Eq. (5-2a)] explains why there is interest in electron beam lithography. The de Broglie relation states that the wavelength of a particle varies inversely with its momentum:

$$\lambda = \frac{h}{p} \tag{5-2c}$$

Thus, more massive or energetic particles have shorter wavelengths. Electron beams are easily generated, focused, and deflected. Since a 10-keV electron has a wavelength of about 0.1 Å, the linewidth limits become the size of the focused beam and its interaction with the photoresist layer. It is possible to achieve linewidths of 0.1 μm by direct electron-beam writing on the wafer photoresist. Furthermore, the computer-controlled electron-beam exposure requires no masks. This capability allows extremely dense packing of circuit elements on the chip, but direct writing of complex patterns is slow. Because of the time required for electron-beam wafer exposure, it is usually advantageous to use electron-beam writing to make the reticle (Fig. 5-7a) and then to expose the wafer photoresist by using photons. Another approach being considered is electron projection lithography (EPL), using a mask, instead of steering a focused electron beam, in order to solve the throughput problem.

5.1.7 Etching

After the photoresist pattern is formed, it can be used as a mask to etch the material underneath. In the early days of Si technology, etching was done using wet chemicals. For example, dilute HF can be used to etch SiO_2 layers grown on a Si substrate with excellent *selectivity*. The term selectivity here refers to the fact that HF attacks SiO_2, but does not affect the Si substrate underneath or the photoresist mask. Although many wet etches are selective, they are unfortunately *isotropic*, which means that they etch as fast laterally as they etch vertically. This is unacceptable for ultra-small features. Hence, wet etching has been largely supplanted by dry, plasma-based etching which can be made both selective and *anisotropic* (etches vertically but not laterally along the surface). In modern IC processing the main use of wet chemical processing is in cleaning the wafers.

Plasmas are ubiquitous in IC processing. The most popular type of plasma-based etching is known as *reactive ion etching* (RIE) (Fig. 5-8). In a typical process, appropriate etch gases such as chlorofluorocarbons (CFCs) flow into the chamber at reduced pressure (~1 – 100 mTorr), and a plasma is struck

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 166

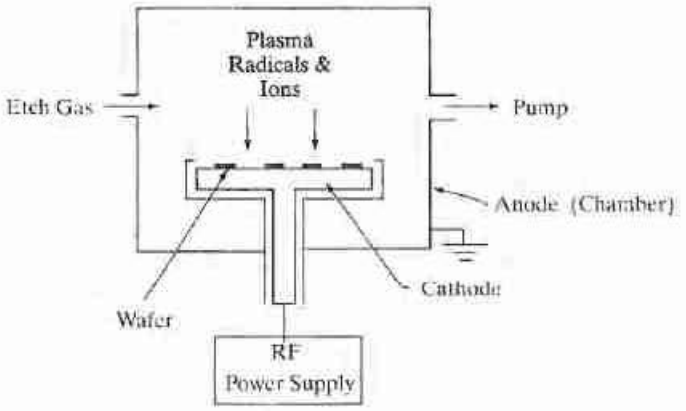


Figure 5–8
Reactive ion etcher. Single or multiple wafers are placed on the rf powered cathode to maximize the ion bombardment. Shown in the figure is a simple *diode* etcher in which we have just two electrodes. We can also use a third electrode to supply rf power separately to the etch gases in a *triode* etcher. The most commonly used rf frequency is 13.56 MHz, which is a frequency dedicated to industrial use so that there is minimal interference with radio communications.

by applying an rf voltage across a cathode and an anode. The rf voltage accelerates the light electrons in the system to much higher kinetic energies (~10 eV) than the heavier ions. The high energy electrons collide with neutral atoms and molecules to create ions and molecular fragments called radicals. The wafers are held on the rf powered cathode, while the grounded chamber walls act as the anode. From a study of plasma physics, we can show that although the bulk of the plasma is a highly conducting, equi-potential region, less conducting *sheath* regions form next to the two electrodes. It can also be shown that the sheath voltage next to the cathode can be increased by making the (powered) cathode smaller in area than the (grounded) anode. A high d-c voltage (~100–1000 V) develops across the sheath next to the rf powered cathode, such that positive ions gain kinetic energy by being accelerated in this region, and bombard the wafer normal to the surface. This bombardment at normal incidence contributes a physical component to the etch that makes it anisotropic. Physical etching, however, is rather unselective. Simultaneously, the highly reactive radicals in the system give rise to a chemical etch component that is very selective, but not anisotropic. The result is that RIE achieves a good compromise between anisotropy and selectivity, and has become the mainstay of modern IC etch technology.

5.1.8 Metallization

After the semiconductor devices are made by the processing methods described previously, they have to be connected to each other, and ultimately to the IC package, by metallization. Metal films are generally deposited by a physical vapor deposition technique such as evaporation (e.g., Au on GaAs) or sputtering (e.g., Al on Si). Sputtering of Al is achieved by immersing an Al target (typically alloyed with ~1% Si and ~4% Cu to improve the electrical and metallurgical properties of the Al, as described in Section 9.3.1) in an Ar plasma. Argon ions bombard the Al and physically dislodge Al atoms by

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 167

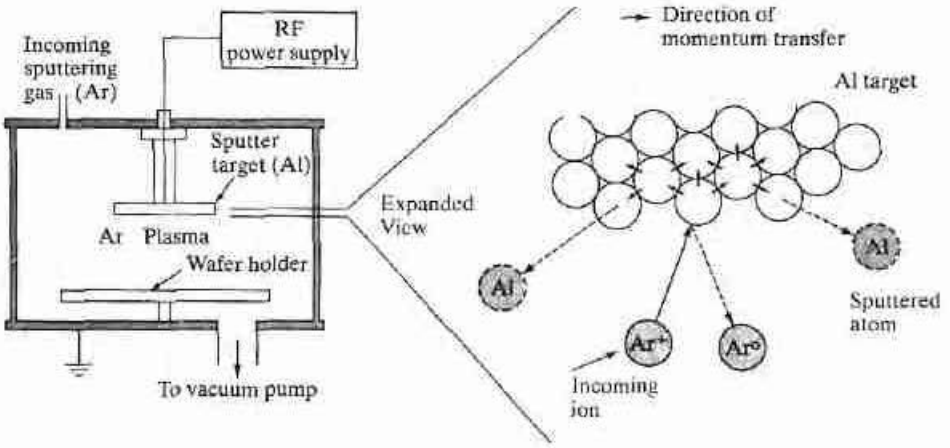


Figure 5–9
Aluminum sputtering by Ar^+ ions. The Ar^+ ions with energies of ~1–3 keV physically dislodge Al atoms which end up depositing on the Si wafers held in close proximity. The chamber pressures are kept low such that the mean free path of the ejected Al atoms is long compared to the target-to-wafer separation.

momentum transfer (Fig. 5–9). Many of the Al atoms ejected from the target deposit on the Si wafers held in close proximity to the target. The Al is then patterned using the metallization reticle and subsequently etched by RIE. Finally, it is sintered at ~450°C for ~30 minutes to form a good electrical, ohmic contact to the Si.

After the interconnection metallization is complete, a protective overcoat of silicon nitride is deposited using plasma-enhanced CVD. Then the individual integrated circuits can be separated by sawing or by scribing and breaking the wafer. The final steps of the process are mounting individual devices in appropriate packages and connecting leads to the Al contact regions. Very precise lead bonders are available for bonding Au or Al wire (about one thousandth of an inch in diameter) to the device and then to the package leads. This phase of device fabrication is called back-end processing, and is discussed in more detail in Chapter 9.

The main steps in making p-n junctions using some of these unit processes are illustrated in Fig. 5–10. Similarly, we will discuss how the key semiconductor devices are made using these same unit processes in subsequent chapters.

5.2 EQUILIBRIUM CONDITIONS

In this chapter we wish to develop both a useful mathematical description of the p-n junction and a strong qualitative understanding of its properties. There must be some compromise in these two goals, since a complete mathematical treatment would obscure the essentially simple physical features of

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 168

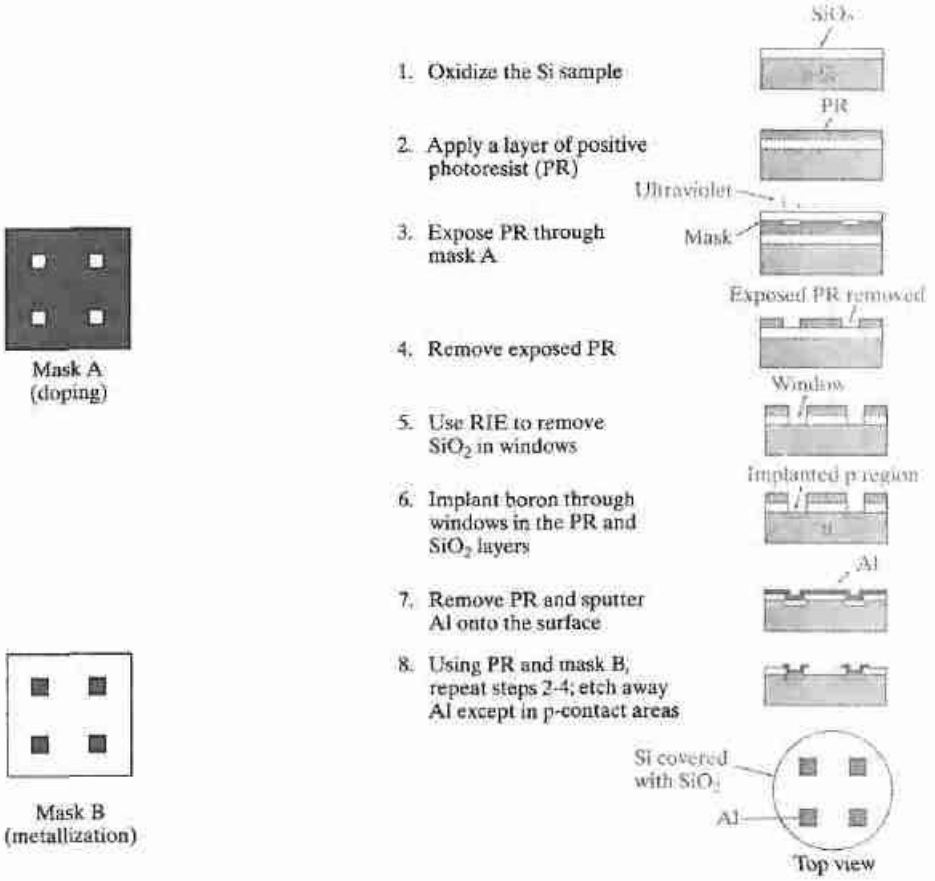


Figure 5–10
Simplified description of steps in the fabrication of p-n junctions. For simplicity, only four diodes per wafer are shown, and the relative thicknesses of the oxide, PR, and the Al layers are exaggerated.

junction operation, while a completely qualitative description would not be useful in making calculations. The approach, therefore, will be to describe the junction mathematically while neglecting small effects which add little to the basic solution. In Section 5.6 we shall include several deviations from the simple theory.

The mathematics of p-n junctions is greatly simplified for the case of the *step junction*, which has uniform p doping on one side of a sharp junction and uniform n doping on the other side. This model represents epitaxial junctions quite well; diffused or implanted junctions, however, are actually *graded* ($N_d - N_a$ varies over a significant distance on either side of the junction). After the basic ideas of junction theory are explored for the step junction, we can make the appropriate corrections to extend the theory to the graded junction. In these discussions we shall assume one-dimensional current flow in samples of uniform cross-sectional area.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 169

In this section we investigate the properties of the step junction at equilibrium (i.e., with no external excitation and no net currents flowing in the device). We shall find that the difference in doping on each side of the junction causes a potential difference between the two types of material. This is a reasonable result, since we would expect some charge transfer because of diffusion between the p material (many holes) and the n material (many electrons). In addition, we shall find that there are four components of current which flow across the junction due to the drift and diffusion of electrons and holes. These four components combine to give zero net current for the equilibrium case. However, the application of bias to the junction increases some of these current components with respect to others, giving net current flow. If we understand the nature of these four current components, a sound view of p-n junction operation, with or without bias, will follow.

5.2.1 The Contact Potential

Let us consider separate regions of p- and n-type semiconductor material, brought together to form a junction (Fig. 5-11). This is not a practical way of forming a device, but this "thought experiment" does allow us to discover the requirements of equilibrium at a junction. Before they are joined, the n material has a large concentration of electrons and few holes, whereas the converse is true for the p material. Upon joining the two regions (Fig. 5-11), we expect diffusion of carriers to take place because of the large carrier concentration gradients at the junction. Thus holes diffuse from the p side into the n side, and electrons diffuse from n to p. The resulting diffusion current cannot build up indefinitely, however, because an opposing electric field is created at the junction (Fig. 5-11b). If the two regions were boxes of red air molecules and green molecules (perhaps due to appropriate types of pollution), eventually there would be a homogeneous mixture of the two after the boxes were joined. This cannot occur in the case of the charged particles in a p-n junction because of the development of space charge and the electric field $\mathscr{E}$. If we consider that electrons diffusing from n to p leave behind uncompensated[3] donor ions (N_d^+) in the n material, and holes leaving the p region leave behind uncompensated acceptors (N_a^-), it is easy to visualize the development of a region of positive space charge near the n side of the junction and negative charge near the p side. The resulting electric field is directed from the positive charge toward the negative charge. Thus $\mathscr{E}$ is in the direction opposite to that of diffusion current for each type of carrier (recall electron current is opposite to the direction of electron flow). Therefore, the

[3]We recall that neutrality is maintained in the bulk materials of Fig. 5-11a by the presence of one electron for each ionized donor ($n = N_d^+$) in the n material and one hole for each ionized acceptor ($p = N_a^-$) in the p material (neglecting minority carriers). Thus, if electrons leave n, some of the positive donor ions near the junction are left uncompensated, as in Fig. 5-11b. The donors and acceptors are fixed in the lattice, in contrast to the mobile electrons and holes.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 170

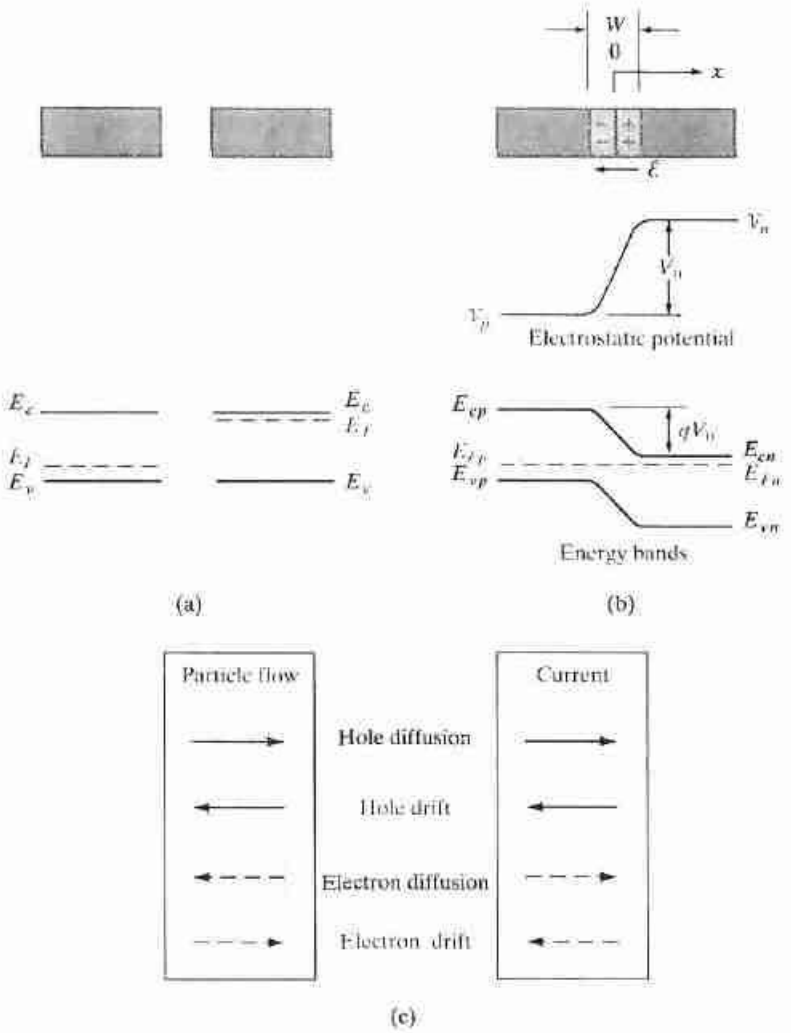


Figure 5–11
Properties of an equilibrium p-n junction: (a) isolated, neutral regions of p-type and n-type material and energy bands for the isolated regions; (b) junction, showing space charge in the transition region W, the resulting electric field $\mathscr{E}$ and contact potential V_0, and the separation of the energy bands; (c) directions of the four components of particle flow within the transition region, and the resulting current directions.

field creates a drift component of current from n to p, opposing the diffusion current (Fig. 5–11c).

Since we know that no *net* current can flow across the junction at equilibrium, the current due to the drift of carriers in the $\mathscr{E}$ field must exactly cancel the diffusion current. Furthermore, since there can be no net buildup of electrons or holes on either side as a function of time, the drift and diffusion currents must cancel for *each* type of carrier:

$$J_p(\text{drift}) + J_p(\text{diff.}) = 0 \tag{5-3a}$$

$$J_n(\text{drift}) + J_n(\text{diff.}) = 0 \tag{5-3b}$$

Therefore, the electric field $\mathscr{E}$ builds up to the point where the net current is zero at equilibrium. The electric field appears in some region W about the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 171

junction, and there is an equilibrium potential difference V_0 across W. In the electrostatic potential diagram of Fig. 5-11b, there is a gradient in potential in the direction opposite to $\mathscr{E}$, in accordance with the fundamental relation[4] $\mathscr{E}(x) = -d\mathscr{V}(x)/dx$. We assume the electric field is zero in the neutral regions outside W. Thus there is a constant potential $\mathscr{V}_n$ in the neutral n material, a constant $\mathscr{V}_p$ in the neutral p material, and a potential difference $V_0 = \mathscr{V}_n - \mathscr{V}_p$ between the two. The region W is called the *transition region*,[5] and the potential difference V_0 is called the *contact potential*. The contact potential appearing across W is a *built-in* potential barrier, in that it is necessary to the maintenance of equilibrium at the junction; it does not imply any external potential. Indeed, the contact potential cannot be measured by placing a voltmeter across the devices, because new contact potentials are formed at each probe, just canceling V_0. By definition V_0 is an equilibrium quantity, and no net current can result from it.

The contact potential separates the bands as in Fig. 5-11b; the valence and conduction energy bands are higher on the p side of the junction than on the n side[6] by the amount qV_0. The separation of the bands at equilibrium is just that required to make the Fermi level constant throughout the device. We discussed the lack of spatial variation of the Fermi level at equilibrium in Section 3.5. Thus if we know the band diagram, including E_F, for each separate material (Fig. 5-11a), we can find the band separation for the junction at equilibrium simply by drawing a diagram such as Fig. 5-11b with the Fermi levels aligned.

To obtain a quantitative relationship between V_0 and the doping concentrations on each side of the junction, we must use the requirements for equilibrium in the drift and diffusion current equations. For example, the drift and diffusion components of the hole current just cancel at equilibrium:

$$J_p(x) = q\left[\mu_p p(x)\mathscr{E}(x) - D_p\frac{dp(x)}{dx}\right] = 0 \tag{5-4a}$$

This equation can be rearranged to obtain

$$\frac{\mu_p}{D_p}\mathscr{E}(x) = \frac{1}{p(x)}\frac{dp(x)}{dx} \tag{5-4b}$$

[4]When we write $\mathscr{E}(x)$, we refer to the value of $\mathscr{E}$ as computed in the x-direction. This value will of course be negative, since it is directed opposite to the true direction of $\mathscr{E}$ as shown in Fig. 5-11b.

[5]Other names for this region are the *space charge region*, since space charge exists within W while neutrality is maintained outside this region, and the *depletion region*, since W is almost depleted of carriers compared with the rest of the crystal. The contact potential V_0 is also called the *diffusion potential*, since it represents a potential barrier which diffusing carriers must surmount in going from one side of the junction to the other.

[6]The electron energy diagram of Fig. 5-11b is related to the electrostatic potential diagram by $-q$, the negative charge on the electron. Since $\mathscr{V}_n$ is a higher potential than $\mathscr{V}_p$ by the amount V_0, the electron energies on the n side are *lower* than those on the p side by qV_0.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 172

where the x-direction is arbitrarily taken from p to n. The electric field can be written in terms of the gradient in the potential, $\mathscr{E}(x) = -d\mathscr{V}(x)/dx$, so that Eq. (5-4b) becomes

$$-\frac{q}{kT}\frac{d\mathscr{V}(x)}{dx} = \frac{1}{p(x)}\frac{dp(x)}{dx} \tag{5-5}$$

with the use of the Einstein relation for μ_p/D_p. This equation can be solved by integration over the appropriate limits. In this case we are interested in the potential on either side of the junction, $\mathscr{V}_p$ and $\mathscr{V}_n$, and the hole concentration just at the edge of the transition region on either side, p_p and p_n. For a step junction it is reasonable to take the electron and hole concentration in the neutral regions outside the transition region as their equilibrium values. Since we have assumed a one-dimensional geometry, p and $\mathscr{V}$ can be taken reasonably as functions of x only. Integration of Eq. (5-5) gives

$$-\frac{q}{kT}\int_{\mathscr{V}_p}^{\mathscr{V}_n} d\mathscr{V} = \int_{p_p}^{p_n}\frac{1}{p}dp$$

$$-\frac{q}{kT}(\mathscr{V}_n - \mathscr{V}_p) = \ln p_n - \ln p_p = \ln\frac{p_n}{p_p} \tag{5-6}$$

The potential difference $\mathscr{V}_n - \mathscr{V}_p$ is the contact potential V_0 (Fig. 5-11b). Thus we can write V_0 in terms of the equilibrium hole concentrations on either side of the junction:

$$V_0 = \frac{kT}{q}\ln\frac{p_p}{p_n} \tag{5-7}$$

If we consider the step junction to be made up of material with N_a acceptors/cm^3 on the p side and a concentration of N_d donors on the n side, we can write Eq. (5-7) as

$$V_0 = \frac{kT}{q}\ln\frac{N_a}{n_i^2/N_d} = \frac{kT}{q}\ln\frac{N_aN_d}{n_i^2} \tag{5-8}$$

by considering the majority carrier concentration to be the doping concentration on each side.

Another useful form of Eq. (5-7) is

$$\frac{p_p}{p_n} = e^{qV_0/kT} \tag{5-9}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 173

By using the equilibrium condition $p_p n_p = n_i^2 = p_n n_n$, we can extend Eq. (5-9) to include the electron concentrations on either side of the junction:

$$\frac{p_p}{p_n} = \frac{n_n}{n_p} = e^{qV_0/kT} \tag{5-10}$$

This relation will be very valuable in calculation of the I–V characteristics of the junction.

EXAMPLE 5-1 An abrupt Si p-n junction has $N_a = 10^{18}$ cm^{-3} on one side and $N_d = 5 \times 10^{15}$ cm^{-3} on the other.

(a) Calculate the Fermi level positions at 300 K in the p and n regions.
(b) Draw an equilibrium band diagram for the junction and determine the contact potential V_0 from the diagram.
(c) Compare the results of part (b) with V_0 as calculated from Eq. (5-8).

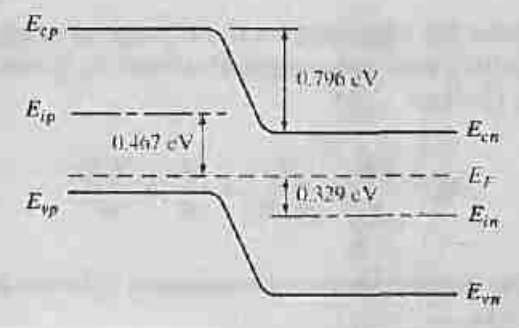
SOLUTION

(a) $$E_{ip} - E_F = kT \ln \frac{p_p}{n_i} = 0.0259 \ln \frac{10^{18}}{(1.5 \times 10^{10})} = \mathbf{0.467\ eV}$$

$$E_F - E_{in} = kT \ln \frac{n_n}{n_i} = 0.0259 \ln \frac{5 \times 10^{15}}{(1.5 \times 10^{10})} = \mathbf{0.329\ eV}$$

(b) $qV_0 = 0.467 + 0.329 = \mathbf{0.796\ eV}$

(c) $$qV_0 = kT \ln \frac{N_a N_d}{n_i^2} = 0.0259 \ln \frac{5 \times 10^{33}}{2.25 \times 10^{20}} = \mathbf{0.796\ eV}$$



5.2.2 Equilibrium Fermi Levels

We have observed that the Fermi level must be constant throughout the device at equilibrium. This observation can be easily related to the results of the previous section. Since we have assumed that p_n and p_p are given by their

equilibrium values outside the transition region, we can write Eq. (5-9) in terms of the basic definitions of these quantities using Eq. (3-19):

$$\frac{p_p}{p_n} = e^{qV_0/kT} = \frac{N_v e^{-(E_{Fp}-E_{vp})/kT}}{N_v e^{-(E_{Fn}-E_{vn})/kT}} \tag{5-11a}$$

$$e^{qV_0/kT} = e^{(E_{Fn}-E_{Fp})/kT} e^{(E_{vp}-E_{vn})/kT} \tag{5-11b}$$

$$qV_0 = E_{vp} - E_{vn} \tag{5-12}$$

The Fermi level and valence band energies are written with subscripts to indicate the p side and the n side of the junction.

From Fig. 5-11b the energy bands on either side of the junction are separated by the contact potential V_0 times the electronic charge q; thus the energy difference $E_{vp} - E_{vn}$ is just qV_0. Equation (5-12) results from the fact that the Fermi levels on either side of the junction are equal at equilibrium ($E_{Fn} - E_{Fp} = 0$). When bias is applied to the junction, the potential barrier is raised or lowered from the value of the contact potential, and the Fermi levels on either side of the junction are shifted with respect to each other by an energy in electron volts numerically equal to the applied voltage in volts.

5.2.3 Space Charge at a Junction

Within the transition region, electrons and holes are in transit from one side of the junction to the other. Some electrons diffuse from n to p, and some are swept by the electric field from p to n (and conversely for holes); there are, however, very few carriers within the transition region at any given time, since the electric field serves to sweep out carriers which have wandered into W. To a good approximation, we can consider the space charge within the transition region as due only to the uncompensated donor and acceptor ions. The charge density within W is plotted in Fig. 5-12b. Neglecting carriers within the space charge region, the charge density on the n side is just q times the concentration of donor ions N_d, and the negative charge density on the p side is $-q$ times the concentration of acceptors N_a. The assumption of carrier depletion within W and neutrality outside W is known as the *depletion approximation*.

Since the dipole about the junction must have an equal number of charges on either side,[7] ($Q_+ = |Q_-|$), the transition region may extend into the p and n regions unequally, depending on the relative doping of the two sides. For example, if the p side is more lightly doped than the n side ($N_a < N_d$), the space charge region must extend farther into the p material than into the n,

[7] A simple way of remembering this equal charge requirement is to note that electric flux lines must begin and end on charges of opposite sign. Therefore, if Q_+ and Q_- were not of equal magnitude, the electric field would not be contained within W but would extend farther into the p or n regions until the enclosed charges became equal.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 175

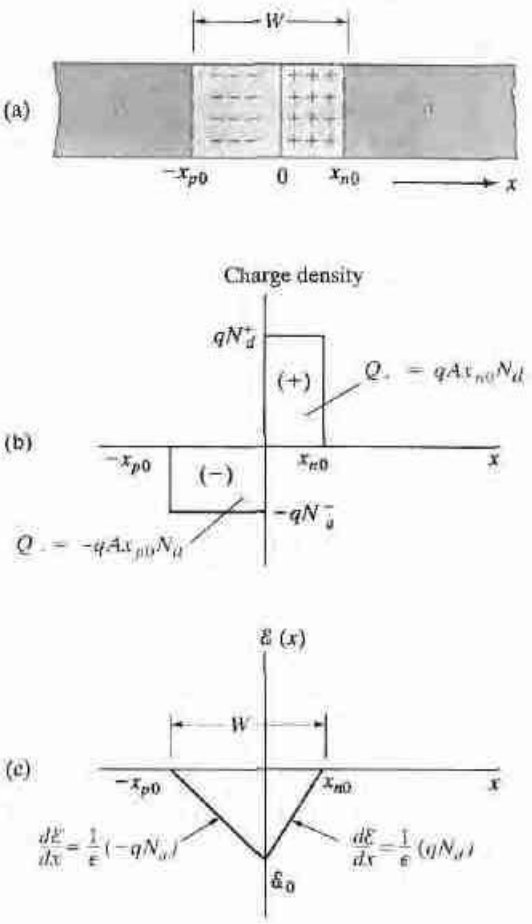


Figure 5–12
Space charge and electric field distribution within the transition region of a p-n junction with $N_d > N_a$: (a) the transition region, with $x = 0$ defined at the metallurgical junction; (b) charge density within the transition region, neglecting the free carriers; (c) the electric field distribution, where the reference direction for $\mathcal{E}$ is arbitrarily taken as the +x-direction.

to "uncover" an equivalent amount of charge. For a sample of cross-sectional area A, the total uncompensated charge on either side of the junction is

$$qAx_{p0}N_a = qAx_{n0}N_d \tag{5-13}$$

where x_{p0} is the penetration of the space charge region into the p material, and x_{n0} is the penetration into n. The total width of the transition region (W) is the sum of x_{p0} and x_{n0}.

To calculate the electric field distribution within the transition region, we begin with *Poisson's equation*, which relates the gradient of the electric field to the local space charge at any point x:

$$\frac{d\mathcal{E}(x)}{dx} = \frac{q}{\epsilon}(p - n + N_d^+ - N_a^-) \tag{5-14}$$

This equation is greatly simplified within the transition region if we neglect the contribution of the carriers $(p - n)$ to the space charge. With

this approximation we have two regions of constant space charge:

$$\frac{d\mathscr{E}}{dx} = \frac{q}{\epsilon}N_d, \qquad 0 < x < x_{n0} \tag{5-15a}$$

$$\frac{d\mathscr{E}}{dx} = -\frac{q}{\epsilon}N_a, \quad -x_{p0} < x < 0 \tag{5-15b}$$

assuming complete ionization of the impurities ($N_d^+ = N_d$) and ($N_a^- = N_a$). We can see from these two equations that a plot of $\mathscr{E}(x)$ vs. x within the transition region has two slopes, positive ($\mathscr{E}$ increasing with x) on the n side and negative ($\mathscr{E}$ becoming more negative as x increases) on the p side. There is some maximum value of the field $\mathscr{E}_0$ at $x = 0$ (the metallurgical junction between the p and n materials), and $\mathscr{E}(x)$ is everywhere negative within the transition region (Fig. 5-12c). These conclusions come from Gauss's law, but we could predict the qualitative features of Fig. 5-12 without equations. We expect the electric field $\mathscr{E}(x)$ to be negative throughout W, since we know that the $\mathscr{E}$ field actually points in the $-x$-direction, from n to p (i.e., from the positive charges of the transition region dipole toward the negative charges). The electric field is assumed to go to zero at the edges of the transition region, since we are neglecting any small $\mathscr{E}$ field in the neutral n or p regions. Finally, there must be a maximum $\mathscr{E}_0$ at the junction, since this point is between the charges Q_+ and Q_- on either side of the transition region. All the electric flux lines pass through the $x = 0$ plane, so this is the obvious point of maximum electric field.

The value of $\mathscr{E}_0$ can be found by integrating either part of Eq. (5-15) with appropriate limits (see Fig. 5-12c in choosing the limits of integration).

$$\int_{\mathscr{E}_0}^{0} d\mathscr{E} = \frac{q}{\epsilon}N_d\int_{0}^{x_{n0}} dx, \qquad 0 < x < x_{n0} \tag{5-16a}$$

$$\int_{0}^{\mathscr{E}_0} d\mathscr{E} = -\frac{q}{\epsilon}N_a\int_{-x_{p0}}^{0} dx, \quad -x_{p0} < x < 0 \tag{5-16b}$$

Therefore, the maximum value of the electric field is

$$\mathscr{E}_0 = -\frac{q}{\epsilon}N_d x_{n0} = -\frac{q}{\epsilon}N_a x_{p0} \tag{5-17}$$

It is simple to relate the electric field to the contact potential V_0, since the $\mathscr{E}$ field at any x is the negative of the potential gradient at that point:

$$\mathscr{E}(x) = -\frac{dV(x)}{dx} \quad \text{or} \quad -V_0 = \int_{-x_{p0}}^{x_{n0}} \mathscr{E}(x)dx \tag{5-18}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 177

Thus the negative of the contact potential is simply the area under the $\mathscr{E}(x)$ vs. x triangle. This relates the contact potential to the width of the depletion region:

$$V_0 = -\frac{1}{2}\mathscr{E}_0 W = \frac{1}{2}\frac{q}{\epsilon}N_d x_{n0} W \tag{5-19}$$

Since the balance of charge requirement is $x_{n0}N_d = x_{p0}N_a$, and W is simply $x_{p0} + x_{n0}$, we can write $x_{n0} = WN_a/(N_a + N_d)$ in Eq. (5-19):

$$V_0 = \frac{1}{2}\frac{q}{\epsilon}\frac{N_a N_d}{N_a + N_d}W^2 \tag{5-20}$$

By solving for W, we have an expression for the width of the transition region in terms of the contact potential, the doping concentrations, and known constants q and ϵ

$$\boxed{W = \left[\frac{2\epsilon V_0}{q}\left(\frac{N_a + N_d}{N_a N_d}\right)\right]^{1/2} = \left[\frac{2\epsilon V_0}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]^{1/2}} \tag{5-21}$$

There are several useful variations of Eq. (5-21); for example, V_0 can be written in terms of the doping concentrations with the aid of Eq. (5-8):

$$W = \left[\frac{2\epsilon kT}{q^2}\left(\ln\frac{N_a N_d}{n_i^2}\right)\left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]^{1/2} \tag{5-22}$$

We can also calculate the penetration of the transition region into the n and p materials:

$$x_{p0} = \frac{WN_d}{N_a + N_d} = \frac{W}{1 + N_a/N_d} = \left\{\frac{2\epsilon V_0}{q}\left[\frac{N_d}{N_a(N_a + N_d)}\right]\right\}^{1/2} \tag{5-23a}$$

$$x_{n0} = \frac{WN_a}{N_a + N_d} = \frac{W}{1 + N_d/N_a} = \left\{\frac{2\epsilon V_0}{q}\left[\frac{N_a}{N_d(N_a + N_d)}\right]\right\}^{1/2} \tag{5-23b}$$

As expected, Eqs. (5-23) predict that the transition region extends farther into the side with the lighter doping. For example, if $N_a \ll N_d$, x_{p0} is large compared with x_{n0}. This agrees with our qualitative argument that a deep penetration is necessary in lightly doped material to "uncover" the same amount of space charge as for a short penetration into heavily doped material.

Another important result of Eq. (5-21) is that the transition width W varies as the square root of the potential across the region. In the derivation to this point, we have considered only the equilibrium contact potential V_0. In

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 178

Section 5.3 we shall see that an applied voltage can increase or decrease the potential across the transition region by aiding or opposing the equilibrium electric field. Therefore, Eq. (5-21) predicts that an applied voltage will increase or decrease the width of the transition region as well.

EXAMPLE 5-2

The junction described in Example 5-1 has a circular cross section with a diameter of 10 μm. Calculate x_{n0}, x_{p0}, Q_+, and $\mathscr{E}_0$ for this junction at equilibrium (300 K). Sketch $\mathscr{E}(x)$ and charge density to scale, as in Fig. 5-12.

SOLUTION

$$A = \pi(5 \times 10^{-4})^2 = 7.85 \times 10^{-7}\text{cm}^2$$

$$W = \left[\frac{2\epsilon V_0}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]^{1/2}$$

$$= \left[\frac{2(11.8)(8.85 \times 10^{-14})(0.796)}{1.6 \times 10^{-19}}(10^{-18} + 2 \times 10^{-16})\right]^{1/2} = \mathbf{0.457\ \mu m}$$

$$x_{n_0} = \frac{W}{1 + N_d/N_a} = \frac{0.457}{1 + 5 \times 10^{-3}} = \mathbf{0.455\ \mu m}$$

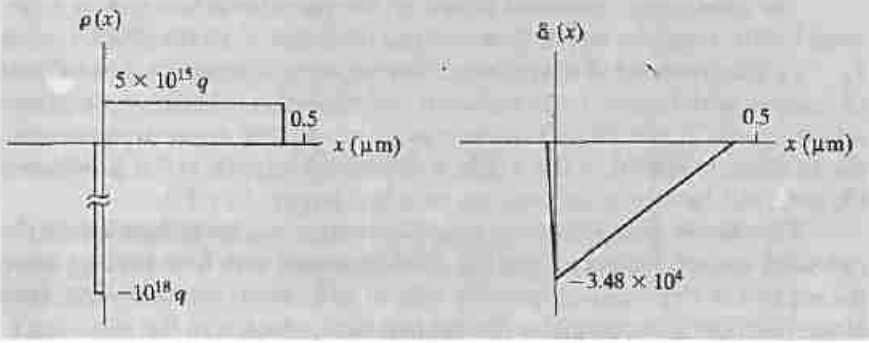
$$x_{p_0} = \frac{0.457}{1 + 200} = \mathbf{2.27 \times 10^{-3}\ \mu m}$$

$$Q_+ = qAx_{n_0}N_d = qAx_{p_0}N_a = (1.6 \times 10^{-19})(7.85 \times 10^{-7})(2.27 \times 10^{11})$$

$$= \mathbf{2.85 \times 10^{-14}C}$$

$$\mathscr{E}_0 = -\frac{q}{\epsilon}x_{n_0}N_d = -\frac{q}{\epsilon}x_{p_0}N_a = \frac{1.6 \times 10^{-19}}{(11.8)(8.85 \times 10^{-14})}(2.27 \times 10^{11})$$

$$= \mathbf{-3.48 \times 10^4\ V/cm}$$



5.3 FORWARD- AND REVERSE-BIASED JUNCTIONS; STEADY STATE CONDITIONS

One useful feature of a p-n junction is that current flows quite freely in the p to n direction when the p region has a positive external voltage bias relative to n (forward bias and forward current), whereas virtually no current flows when p is made negative relative to n (reverse bias and reverse current). This asymmetry of the current flow makes the p-n junction diode very useful as a *rectifier*. While rectification is an important application, it is only the beginning of a host of uses for the biased junction. Biased p-n junctions can be used as voltage-variable capacitors, photocells, light emitters, and many more devices which are basic to modern electronics. Two or more junctions can be used to form transistors and controlled switches.

In this section we begin with a qualitative description of current flow in a biased junction. With the background of the previous section, the basic features of current flow are relatively simple to understand, and these qualitative concepts form the basis for the analytical description of forward and reverse currents in a junction.

5.3.1 Qualitative Description of Current Flow at a Junction

We assume that an applied voltage bias V appears across the transition region of the junction rather than in the neutral n and p regions. Of course, there will be some voltage drop in the neutral material, if a current flows through it. But in most p-n junction devices, the length of each region is small compared with its area, and the doping is usually moderate to heavy; thus the resistance is small in each neutral region, and only a small voltage drop can be maintained outside the space charge (transition) region. For almost all calculations it is valid to assume that an applied voltage appears entirely across the transition region. We shall take V to be positive when the external bias is positive on the p side relative to the n side.

Since an applied voltage changes the electrostatic potential barrier and thus the electric field within the transition region, we would expect changes in the various components of current at the junction (Fig. 5–13). In addition, the separation of the energy bands is affected by the applied bias, along with the width of the depletion region. Let us begin by examining qualitatively the effects of bias on the important features of the junction.

The *electrostatic potential barrier* at the junction is lowered by a forward bias V_f from the equilibrium contact potential V_0 to the smaller value $V_0 - V_f$. This lowering of the potential barrier occurs because a forward bias (p positive with respect to n) raises the electrostatic potential on the p side relative to the n side. For a reverse bias ($V = -V_r$) the opposite occurs; the electrostatic potential of the p side is depressed relative to the n side, and the potential barrier at the junction becomes larger ($V_0 + V_r$).

The *electric field* within the transition region can be deduced from the potential barrier. We notice that the field decreases with forward bias, since the applied electric field opposes the built-in field. With reverse bias the field at the junction is increased by the applied field, which is in the same direction as the equilibrium field.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 180

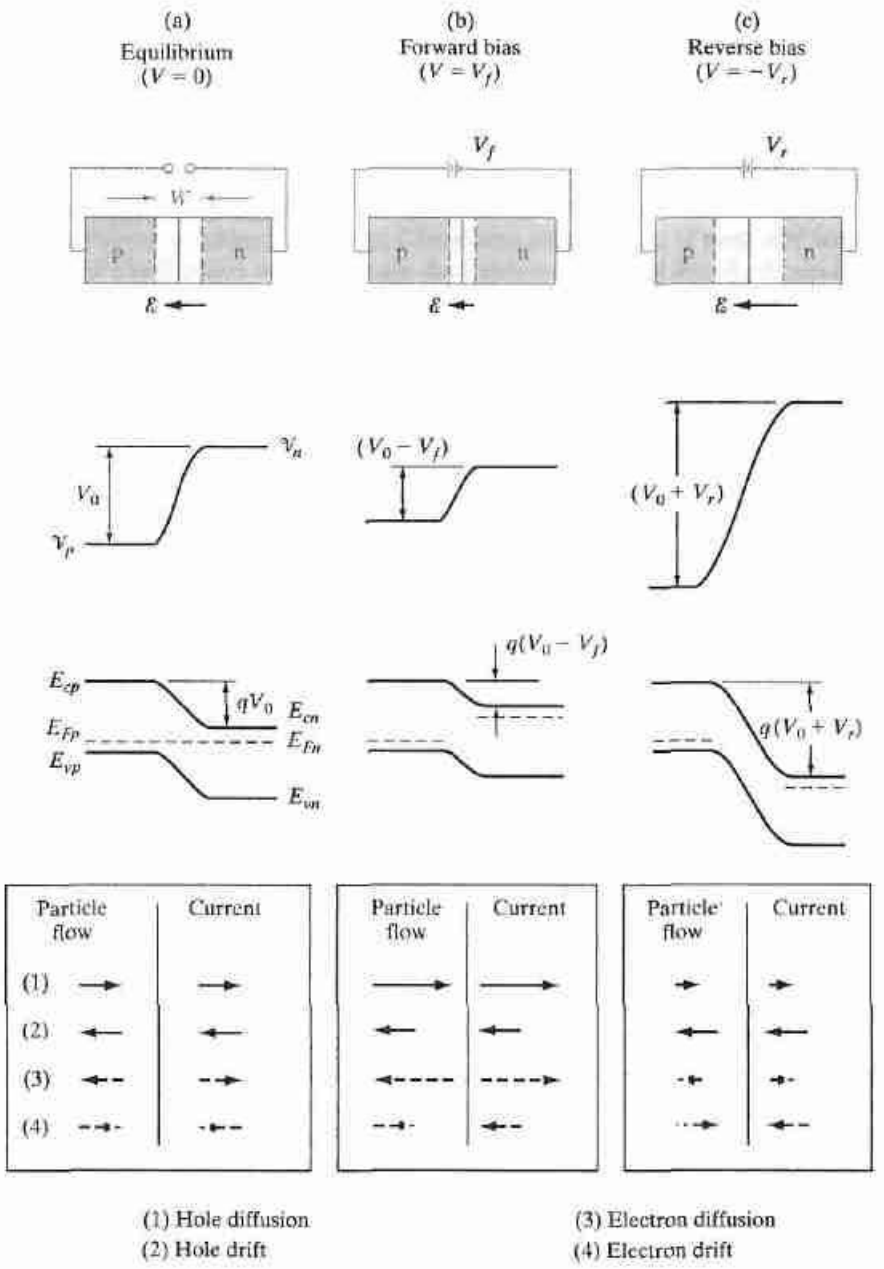


Figure 5–13
Effects of a bias at a p-n junction; transition region width and electric field, electrostatic potential, energy band diagram, and particle flow and current directions within W for (a) equilibrium, (b) forward bias, and (c) reverse bias.

The change in electric field at the junction calls for a change in the *transition region width* W, since it is still necessary that a proper number of positive and negative charges (in the form of uncompensated donor and acceptor ions) be exposed for a given value of the $\mathscr{E}$ field. Thus we would expect the width W to decrease under forward bias (smaller $\mathscr{E}$, fewer uncompensated

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 181

charges) and to increase under reverse bias. Equations (5-21) and (5-23) can be used to calculate W, x_{p0}, and x_{n0} if V_0 is replaced by the new barrier height[8] $V_0 - V$.

The *separation of the energy bands* is a direct function of the electrostatic potential barrier at the junction. The height of the electron energy barrier is simply the electronic charge q times the height of the electrostatic potential barrier. Thus the bands are separated less $[q(V_0 - V_f)]$ under forward bias than at equilibrium, and more $[q(V_0 + V_r)]$ under reverse bias. We assume the Fermi level deep inside each neutral region is essentially the equilibrium value (we shall return to this assumption later); therefore, the shifting of the energy bands under bias implies a separation of the Fermi levels on either side of the junction, as mentioned in Section 5.2.2. Under forward bias, the Fermi level on the n side E_{Fn} is above E_{Fp} by the energy qV_f; for reverse bias, E_{Fp} is qV_r joules higher than E_{Fn}. *In energy units of electron volts, the Fermi levels in the two neutral regions are separated by an energy (eV) numerically equal to the applied voltage (V).*

The *diffusion current* is composed of majority carrier electrons on the n side surmounting the potential energy barrier to diffuse to the p side, and holes surmounting their barrier from p to n.[9] There is a distribution of energies for electrons in the n-side conduction band (Fig. 3-16), and some electrons in the high-energy "tail" of the distribution have enough energy to diffuse from n to p at equilibrium in spite of the barrier. With forward bias, however, the barrier is lowered (to $V_0 - V_f$), and many more electrons in the n-side conduction band have sufficient energy to diffuse from n to p over the smaller barrier. Therefore, the electron diffusion current can be quite large with forward bias. Similarly, more holes can diffuse from p to n under forward bias because of the lowered barrier. For reverse bias the barrier becomes so large $(V_0 + V_r)$ that virtually no electrons in the n-side conduction band or holes in the p-side valence band have enough energy to surmount it. Therefore, the diffusion current is usually negligible for reverse bias.

The *drift current* is relatively insensitive to the height of the potential barrier. This sounds strange at first, since we normally think in terms of material with ample carriers, and therefore we expect drift current to be simply proportional to the applied field. The reason for this apparent anomaly is the fact that the drift current is limited *not* by *how fast* carriers are swept down the barrier, *but* rather *how often*. For example, minority carrier electrons on the p side which wander into the transition region will be swept down the barrier by

[8]With bias applied to the junction, the 0 in the subscripts of x_{n0} and x_{p0} does not imply equilibrium. Instead, it signifies the origin of a new set of coordinates, $x_n = 0$ and $x_p = 0$, as defined later in Fig. 5-15.

[9]Remember that the potential energy barriers for electrons and holes are directed oppositely. The barrier for electrons is apparent from the energy band diagram, which is always drawn for electron energies. For holes, the potential energy barrier at the junction has the same shape as the electrostatic potential barrier (the conversion factor between electrostatic potential and hole energy is $+q$). A simple check of these two barrier directions can be made by asking the directions in which carriers are swept by the $\mathscr{E}$ field within the transition region—a hole is swept in the direction of $\mathscr{E}$, from n to p (swept down the potential "hill" for holes); an electron is swept opposite to $\mathscr{E}$, from p to n (swept down the potential energy "hill" for electrons).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 182

the $\mathscr{E}$ field, giving rise to the electron component of drift current. However, this current is small not because of the size of the barrier, but because there are very few minority electrons in the p side to participate. Every electron on the p side which diffuses to the transition region will be swept down the potential energy hill, whether the hill is large or small. The electron drift current does not depend on how fast an individual electron is swept from p to n, but rather on how many electrons are swept down the barrier per second. Similar comments apply regarding the drift of minority holes from the n side to the p side of the junction. To a good approximation, therefore, the electron and hole drift currents at the junction are independent of the applied voltage.

The supply of minority carriers on each side of the junction required to participate in the drift component of current is generated by thermal excitation of electron-hole pairs. For example, an EHP created near the junction on the p side provides a minority electron in the p material. If the EHP is generated within a diffusion length L_n of the transition region, this electron can diffuse to the junction and be swept down the barrier to the n side. The resulting current due to drift of generated carriers across the junction is commonly called the *generation current* since its magnitude depends entirely on the rate of generation of EHPs. As we shall discuss later, this generation current can be increased greatly by optical excitation of EHPs near the junction (the p-n junction *photodiode*).

The *total current* crossing the junction is composed of the sum of the diffusion and drift components. As Fig. 5-13 indicates, the electron and hole diffusion currents are both directed from p to n (although the particle flow directions are opposite to each other), and the drift currents are from n to p. The *net* current crossing the junction is zero at equilibrium, since the drift and diffusion components cancel for each type of carrier (the equilibrium electron and hole components need not be equal, as in Fig. 5-13, as long as the net hole current and the net electron current are each zero). Under reverse bias, both diffusion components are negligible because of the large barrier at the junction, and the only current is the relatively small (and essentially voltage-independent) generation current from n to p. This generation current is shown in Fig. 5-14, in a sketch of a typical I–V plot for a p-n junction. In this figure the positive direction for the current I is taken from p to n, and the applied voltage V is positive when the positive battery terminal is connected to p and the negative terminal to n. The only current flowing in this p-n junction diode for negative V is the small current I(gen.) due to carriers generated in the transition region or minority carriers which diffuse to the junction and are collected. The current at $V = 0$ (equilibrium) is zero since the generation and diffusion currents cancel:[10]

$$I = I(\text{diff.}) - |I(\text{gen.})| = 0 \quad \text{for } V = 0 \tag{5-24}$$

[10]The total current I is the sum of the generation and diffusion components. However, these components are oppositely directed, with I(diff.) being positive and I(gen.) being negative for the chosen reference direction. To avoid confusion of signs, we use here the magnitude of the drift current $|I(\text{gen.})|$ and include its negative sign in Eq. (5-24). Thus, when we write the term $-|I(\text{gen.})|$, there is no doubt that the generation current is in the negative current direction. This approach emphasizes the fact that the two components of current add with opposite signs to give the total current.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 183

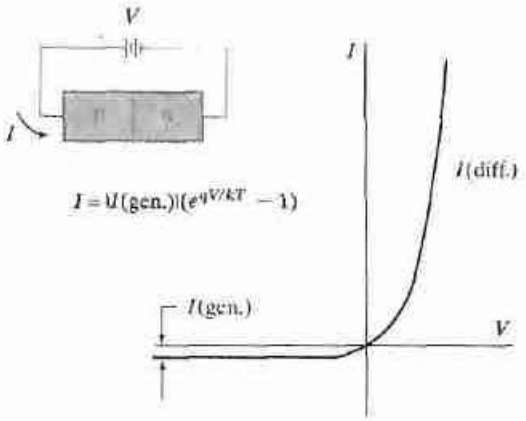


Figure 5–14
I–V characteristic of a p-n junction.

As we shall see in the next section, an applied forward bias $V = V_f$ increases the probability that a carrier can diffuse across the junction, by the factor $\exp(qV_f/kT)$. Thus the diffusion current under forward bias is given by its equilibrium value multiplied by $\exp(qV/kT)$; similarly, for reverse bias the diffusion current is the equilibrium value reduced by the same factor, with $V = -V_r$. Since the equilibrium diffusion current is equal in magnitude to $|I(\text{gen.})|$, the diffusion current with applied bias is simply $|I(\text{gen.})| \exp(qV/kT)$. The total current I is then the diffusion current minus the absolute value of the generation current, which we will now refer to as I_0:

$$I = I_0(e^{qV/kT} - 1) \tag{5-25}$$

In Eq. (5-25) the applied voltage V can be positive or negative, $V = V_f$ or $V = -V_r$. When V is positive and greater than a few kT/q ($kT/q = 0.0259$ V at room temperature), the exponential term is much greater than unity. The current thus increases exponentially with forward bias. When V is negative (reverse bias), the exponential term approaches zero and the current is $-I_0$, which is in the n to p (negative) direction. This negative generation current is also called the *reverse saturation current*. The striking feature of Fig. 5-14 is the nonlinearity of the *I–V* characteristic. Current flows relatively freely in the forward direction of the diode, but almost no current flows in the reverse direction.

5.3.2 Carrier Injection

From the discussion in the previous section, we expect the minority carrier concentration on each side of a p-n junction to vary with the applied bias because of variations in the diffusion of carriers across the junction. The equilibrium ratio of hole concentrations on each side

$$\frac{p_p}{p_n} = e^{qV_0/kT} \tag{5-26}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 184

becomes with bias (Fig. 5-13)

$$\frac{p(-x_{p0})}{p(x_{n0})} = e^{q(V_0 - V)/kT} \tag{5-27}$$

This equation uses the altered barrier $V_0 - V$ to relate the steady state hole concentrations on the two sides of the transition region with either forward or reverse bias (V positive or negative). For low-level injection we can neglect changes in the majority carrier concentrations. Although the absolute increase of the majority carrier concentration is equal to the increase of the minority carrier concentration in order to maintain space charge neutrality, the relative change in majority carrier concentration can be assumed to vary only slightly with bias compared with equilibrium values. With this simplification we can write the ratio of Eq. (5-26) to (5-27) as

$$\frac{p(x_{n0})}{p_n} = e^{qV/kT} \quad \text{taking } p(-x_{p0}) = p_p \tag{5-28}$$

With forward bias, Eq. (5-28) suggests a greatly increased minority carrier hole concentration at the edge of the transition region on the n side $p(x_{n0})$ than was the case at equilibrium. Conversely, the hole concentration $p(x_{n0})$ under reverse bias (V negative) is reduced below the equilibrium value p_n. The exponential increase of the hole concentration at x_{n0} with forward bias is an example of *minority carrier injection*. As Fig. 5-15 suggests, a forward bias V results in a steady state injection of excess holes into the n region and electrons into the p region. We can easily calculate the excess hole concentration Δp_n at the edge of the transition region x_{n0} by subtracting the equilibrium hole concentration from Eq. (5-28),

$$\boxed{\Delta p_n = p(x_{n0}) - p_n = p_n(e^{qV/kT} - 1)} \tag{5-29}$$

and similarly for excess electrons on the p side,

$$\boxed{\Delta n_p = n(-x_{p0}) - n_p = n_p(e^{qV/kT} - 1)} \tag{5-30}$$

From our study of diffusion of excess carriers in Section 4.4.4, we expect that injection leading to a steady concentration of Δp_n excess holes at x_{n0} will produce a *distribution* of excess holes in the n material. As the holes diffuse deeper into the n region, they recombine with electrons in the n material, and the resulting excess hole distribution is obtained as a solution of the diffusion equation, Eq. (4-34b). If the n region is long compared with the hole diffusion length L_p, the solution is exponential, as in Eq. (4-36). Similarly, the injected electrons in the p material diffuse and recombine, giving an exponential distribution of excess electrons. For convenience, let us define two new coordinates (Fig. 5-15): Distances measured in the x-direction in

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 185

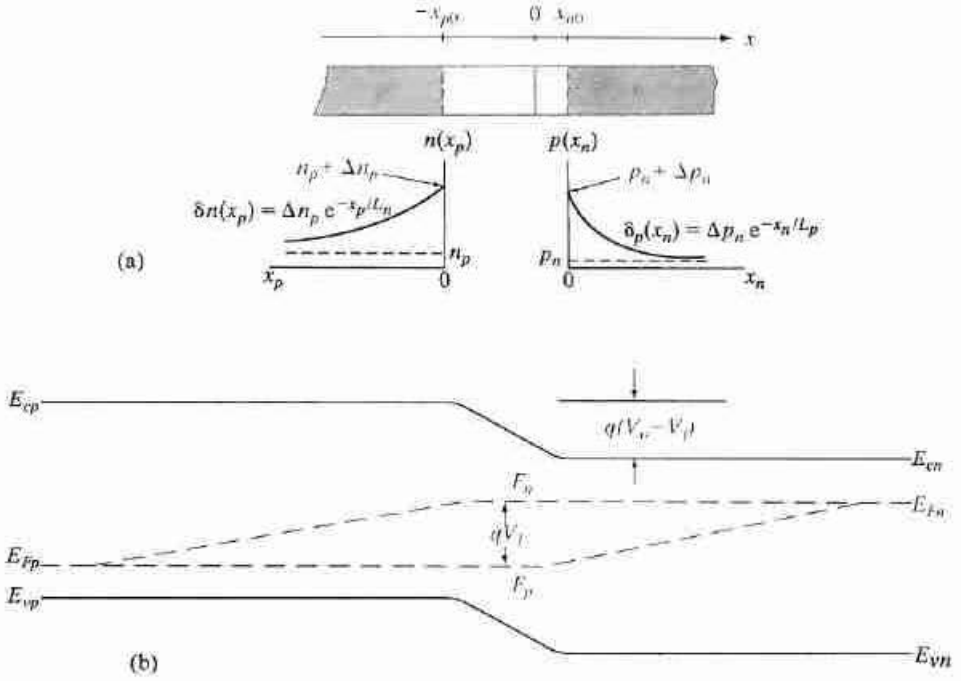


Figure 5–15
Forward-biased junction: (a) minority carrier distributions on the two sides of the transition region and definitions of distances x_n and x_p measured from the transition region edges; (b) variation of the quasi-Fermi levels with position.

the n material from x_{n0} will be designated x_n; distances in the p material measured in the $-x$-direction with $-x_{p0}$ as the origin will be called x_p. This convention will simplify the mathematics considerably. We can write the diffusion equation as in Eq. (4–34) for each side of the junction and solve for the distributions of excess carriers (δn and δp) assuming long p and n regions:

$$\boxed{\delta n(x_p) = \Delta n_p e^{-x_p/L_n} = n_p(e^{qV/kT} - 1)e^{-x_p/L_n}} \quad (5\text{–}31a)$$

$$\boxed{\delta p(x_n) = \Delta p_n e^{-x_n/L_p} = p_n(e^{qV/kT} - 1)e^{-x_n/L_p}} \quad (5\text{–}31b)$$

The hole diffusion current at any point x_n in the n material can be calculated from Eq. (4–40):

$$I_p(x_n) = -qAD_p\frac{d\delta p(x_n)}{dx_n} = qA\frac{D_p}{L_p}\Delta p_n e^{-x_n/L_p} = qA\frac{D_p}{L_p}\delta p(x_n) \quad (5\text{–}32)$$

where A is the cross-sectional area of the junction. Thus the hole diffusion current at each position x_n is proportional to the excess hole concentration

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 186

at that point.[11] The total hole current injected into the n material at the junction can be obtained simply by evaluating Eq. (5-32) at x_{n0}:

$$I_p(x_n = 0) = \frac{qAD_p}{L_p}\Delta p_n = \frac{qAD_p}{L_p}p_n(e^{qV/kT} - 1) \qquad (5\text{-}33)$$

By a similar analysis, the injection of electrons into the p material leads to an electron current at the junction of

$$I_n(x_p = 0) = -\frac{qAD_n}{L_n}\Delta n_p = -\frac{qAD_n}{L_n}n_p(e^{qV/kT} - 1) \qquad (5\text{-}34)$$

The minus sign in Eq. (5-34) means that the electron current is opposite to the x_p-direction; that is, the true direction of I_n is in the $+x$-direction, adding to I_p in the total current (Fig. 5-16). If we neglect recombination in the transition region, which is known as the Shockley ideal diode approximation, we can consider that each injected electron reaching $-x_{p0}$ must pass through x_{n0}. Thus the total diode current I at x_{n0} can be calculated as the sum of $I_p(x_n = 0)$ and $-I_n(x_p = 0)$. If we take the $+x$-direction as the reference direction for the total current I, we must use a minus sign with $I_n(x_p)$ to account for the fact that x_p is defined in the $-x$-direction:

$$I = I_p(x_n = 0) - I_n(x_p = 0) = \frac{qAD_p}{L_p}\Delta p_n + \frac{qAD_n}{L_n}\Delta n_p \qquad (5\text{-}35)$$

$$\boxed{I = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)(e^{qV/kT} - 1) = I_0(e^{qV/kT} - 1)} \qquad (5\text{-}36)$$

Equation (5-36) is the *diode equation*, having the same form as the qualitative relation Eq. (5-25). Nothing in the derivation excludes the possibility that the bias voltage V can be negative; thus the diode equation describes the total current through the diode for either forward or reverse bias. We can calculate the current for reverse bias by letting $V = -V_r$:

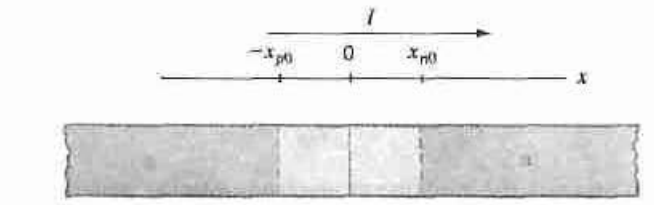
$$I = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)(e^{-qV_r/kT} - 1) \qquad (5\text{-}37a)$$

If V_r is larger than a few kT/q, the total current is just the reverse saturation current

$$I = -qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right) = -I_0 \qquad (5\text{-}37b)$$

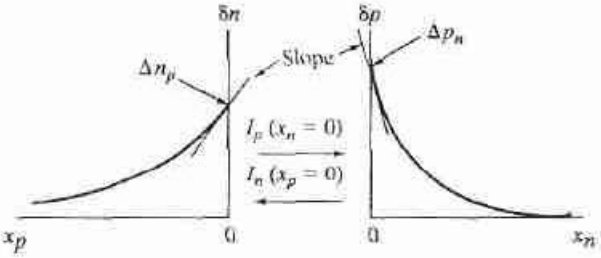
[11] With carrier injection due to bias, it is clear that the equilibrium Fermi levels cannot be used to describe carrier concentrations in the device. It is necessary to use the concept of quasi-Fermi levels, taking into account the spatial variations of the carrier concentrations.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 187



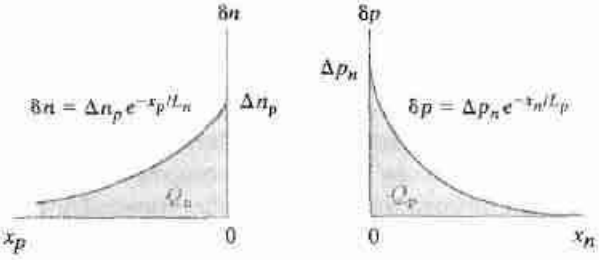
I

(a)



$$I_n(x_p=0) = qAD_n \left.\frac{d\delta n}{dx_p}\right|_{x_p=0} = -qA\frac{D_n}{L_n}\Delta n_p \qquad I_p(x_n=0) = -qAD_p \left.\frac{d\delta p}{dx_n}\right|_{x_n=0} = qA\frac{D_p}{L_p}\Delta p_n$$

(b)



$$Q_n = -qA\int_{-\infty}^{0}\delta n(x_p)\,dx_p \qquad Q_p = qA\int_{0}^{\infty}\delta p(x_n)\,dx_n$$

$$I_n(x_p=0) = \frac{Q_n}{\tau_n} = \frac{-qAL_n}{\tau_n}\Delta n_p \qquad I_p(x_n=0) = \frac{Q_p}{\tau_p} = \frac{qAL_p}{\tau_p}\Delta p_n$$

(c)

$$I = I_p(x_n=0) - I_n(x_p=0) = qA\left(\frac{D_p}{L_p}\Delta p_n + \frac{D_n}{L_n}\Delta n_p\right)$$

$$= qA\left(\frac{D_p p_n}{L_p} + \frac{D_n n_p}{L_n}\right)(e^{qV/kT} - 1)$$

Figure 5–16 Two methods for calculating junction current from the excess minority carrier distributions: (a) diffusion currents at the edges of the transition region; (b) charge in the distributions divided by the minority carrier lifetimes; (c) the diode equation.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 188

One implication of Eq. (5-36) is that the total current at the junction is dominated by injection of carriers from the more heavily doped side into the side with lesser doping. For example, if the p material is very heavily doped and the n region is lightly doped, the minority carrier concentration on the p side (n_p) is negligible compared with the minority carrier concentration on the n side (p_n). Thus the diode equation can be approximated by injection of holes only, as in Eq. (5-33). This means that the charge stored in the minority carrier distributions is due mostly to holes on the n side. For example, to double the hole current in this p^+-n junction one should not double the p^+ doping, but rather reduce the n-type doping by a factor of two. This structure is called a p^+-n junction, where the + superscript simply means heavy doping. Another characteristic of the p^+-n or n^+-p structure is that the transition region extends primarily into the lightly doped region, as we found in the discussion of Eq. (5-23). Having one side heavily doped is a useful arrangement for many practical devices, as we shall see in our discussions of switching diodes and transistors. This type of junction is common in devices which are fabricated by counterdoping. For example, an n-type Si sample with $N_d = 10^{14}$ cm^{-3} can be used as the substrate for an implanted or diffused junction. If the doping of the p region is greater than 10^{19} cm^{-3} (typical of diffused junctions), the structure is definitely p^+-n, with n_p more than five orders of magnitude smaller than p_n. Since this configuration is common in device technology, we shall return to it in much of the following discussion.

Figure 5-15b shows the quasi-Fermi levels as a function of position for a p-n junction in forward bias. The equilibrium E_F is split into the quasi-Fermi levels F_n and F_p which are separated within W by an energy qV caused by the applied bias, V. This energy represents the deviation from equilibrium (see Section 4.3.3). In forward bias in the depletion region we thus get

$$pn = n_i^2 e^{(F_n - F_p)/kT} = n_i^2 e^{(qV/kT)} \tag{5-38}$$

On either side of the junction, it is the minority carrier quasi-Fermi level that varies the most. The majority carrier concentration is not affected much, so the majority carrier quasi-Fermi level is close to the original E_F. Although there is some variation in E_F and F_p within W, it doesn't show up on the scale used in Fig. 5-15. Outside of the depletion regions, the quasi-Fermi levels for the minority carriers vary linearly and eventually merge with the Fermi levels. In contrast, the minority carrier concentrations decay exponentially with distance. In fact it takes many diffusion lengths for the quasi-Fermi level to cross E_i, where the minority carrier concentration is equal to the intrinsic carrier concentration, let alone approach E_F, where for example $\delta p(x_n) \simeq p_n$.

Another simple and instructive way of calculating the total current is to consider the injected current as supplying the carriers for the excess distributions (Fig. 5-16b). For example, $J_p(x_n = 0)$ must supply enough holes

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 189

per second to maintain the steady state exponential distribution $\delta p(x_n)$ as the holes recombine. The total positive charge stored in the excess carrier distribution at any instant of time is

$$Q_p = qA\int_0^\infty \delta p(x_n)dx_n = qA\Delta p_n\int_0^\infty e^{-x_n/L_p}dx_n = qAL_p\Delta p_n \quad (5\text{-}39)$$

The average lifetime of a hole in the n-type material is τ_p. Thus, on the average, this entire charge distribution recombines and must be replenished every τ_p seconds. The injected hole current at $x_n = 0$ needed to maintain the distribution is simply the total charge divided by the average time of replacement:

$$I_p(x_n = 0) = \frac{Q_p}{\tau_p} = qA\frac{L_p}{\tau_p}\Delta p_n = qA\frac{D_p}{L_p}\Delta p_n \quad (5\text{-}40)$$

using $D_p/L_p = L_p/\tau_p$.

This is the same result as Eq. (5-33), which was calculated from the diffusion currents. Similarly, we can calculate the negative charge stored in the distribution $\delta n(x_p)$ and divide by τ_n to obtain the injected electron current in the p material. This method, called the *charge control approximation*, illustrates the important fact that the minority carriers injected into either side of a p-n junction diffuse into the neutral material and recombine with the majority carriers. The minority carrier current [for example, $I_p(x_n)$] decreases exponentially with distance into the neutral region. Thus several diffusion lengths away from the junction, most of the total current is carried by the majority carriers. We shall discuss this point in more detail later in this section.

In summary, we can calculate the current at a p-n junction in two ways (Fig. 5-16): (a) from the slopes of the excess minority carrier distributions at the two edges of the transition regions and (b) from the steady state charge stored in each distribution. We add the hole current injected into the n material $I_p(x_n = 0)$ to the electron current injected into the p material $I_n(x_p = 0)$, after including a minus sign with $I_n(x_p)$ to conform with the conventional definition of positive current in the $+x$-direction. We are able to add these two currents because of the assumption that no recombination takes place within the transition region. Thus we effectively have the total electron and hole current at one point in the device (x_{n0}). Since the total current must be constant throughout the device (despite variations in the current components), I as described by Eq. (5-36) is the total current at every position x in the diode.

The drift of minority carriers can be neglected in the neutral regions *outside W*, because the minority carrier concentration is small compared with that of the majority carriers. If the minority carriers contribute to the total current at all, their contribution must be through diffusion (dependent on the *gradient*

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 190

of the carrier concentration). Even a very small concentration of minority carriers can have an appreciable effect on the current if the spatial variation is large.

Calculation of the majority carrier currents in the two neutral regions is simple, once we have found the minority carrier current. Since the total current I must be constant throughout the device, the majority carrier component of current at any point is just the difference between I and the minority component (Fig. 5-17). For example, since $I_p(x_n)$ is proportional to the excess hole concentration at each position in the n material [Eq. (5-32)], it decreases exponentially in x_n with the decreasing $\delta p(x_n)$. Thus the electron component of current must increase appropriately with x_n to maintain the total current I. Far from the junction, the current in the n material is carried almost entirely by electrons. The physical explanation of this is that electrons must flow in from the n material (and ultimately from the negative terminal of the battery), to resupply electrons lost by recombination in the excess hole distribution near the junction. The electron current $I_n(x_n)$ includes sufficient

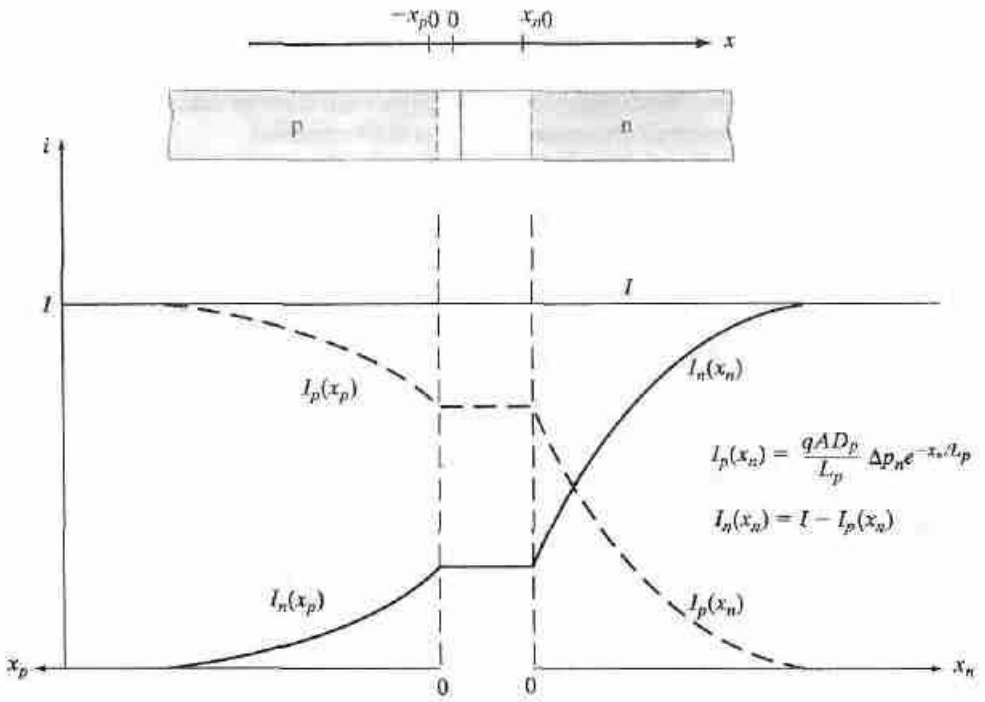


Figure 5–17
Electron and hole components of current in a forward-biased p-n junction. In this example, we have a higher injected minority hole current on the n-side than electron current on the p side because we have a lower n doping than p doping.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 191

electron flow to supply not only recombination near x_{n0}, but also injection of electrons into the p region. Of course, the flow of electrons in the n material toward the junction constitutes a current in the $+x$-direction, contributing to the total current I.

One question that still remains to be answered is whether the majority carrier current is due to drift or diffusion or both, at different points in the diode. Near the junction (just outside of the depletion regions) the majority carrier concentration changes by exactly the same amount as minority carriers in order to maintain space charge neutrality. The majority carrier concentration can change rather fast, in a very short time scale known as the dielectric relaxation time, τ_D $(=\rho\epsilon)$, where ρ is the resistivity and ϵ is the dielectric constant. The relaxation time τ_D is the analog of the RC time constant in a circuit. Very far away from the junction (more than 3 to 5 diffusion lengths), the minority carrier concentration decays to a low, constant background value. Hence, the majority carrier concentration also becomes independent of position. Here, clearly the only possible current component is majority carrier drift current. When approaching the junction there is a spatially varying majority (and minority) carrier concentration and the majority carrier current changes from pure drift to drift and diffusion, although drift always dominates for majority carriers except in cases of very high levels of injection. Throughout the diode, the total current due to majority and minority carriers at any cross section is kept constant.

We thus note that the electric field in the neutral regions cannot be zero as we previously assumed; otherwise, there would be no drift currents. Thus our assumption that all of the applied voltage appears across the transition region is not completely accurate. On the other hand, the majority carrier concentrations are usually large in the neutral regions, so that only a small field is needed to drive the drift currents. Thus the assumption that junction voltage equals applied voltage is acceptable for most calculations.

EXAMPLE 5-3 Find an expression for the electron current in the n-type material of a forward-biased p-n junction.

SOLUTION The total current is

$$I = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)(e^{qV/kT} - 1)$$

The hole current on the n side is

$$I_p(x_n) = qA\frac{D_p}{L_p}p_n e^{-x_n/L_p}(e^{qV/kT} - 1)$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 192

Thus the electron current in the n material is

$$I_n(x_n) = I - I_p(x_n) = qA\left[\frac{D_p}{L_p}(1 - e^{-x_n/L_p})p_n + \frac{D_n}{L_n}n_p\right](e^{qV/kT} - 1)$$

This expression includes the supplying of electrons for recombination with the injected holes, and the injection of electrons across the junction into the p side.

5.3.3 Reverse Bias

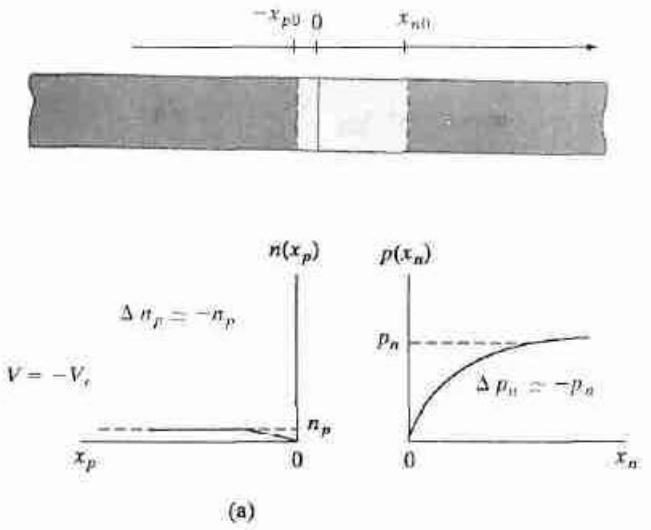
In our discussion of carrier injection and minority carrier distributions, we have primarily assumed forward bias. The distributions for reverse bias can be obtained from the same equations (Fig. 5-18), if a negative value of V is introduced. For example, if $V = -V_r$ (p negatively biased with respect to n), we can approximate Eq. (5-29) as

$$\Delta p_n = p_n(e^{q(-V_r)/kT} - 1) \simeq -p_n \quad \text{for } V_r \gg kT/q \qquad (5\text{-}41)$$

and similarly $\Delta n_p \simeq -n_p$.

Thus for a reverse bias of more than a few tenths of a volt, the minority carrier concentration at each edge of the transition region becomes essentially zero as the excess concentration approaches the negative of the equilibrium concentration. The excess minority carrier concentrations in the neutral regions are still given by Eq. (5-31), so that depletion of carriers below the equilibrium values extends approximately a diffusion length beyond each side of the transition region. This reverse-bias depletion of minority carriers can be thought of as *minority carrier extraction*, analogous to the injection of forward bias. Physically, extraction occurs because minority carriers at the edges of the depletion region are swept down the barrier at the junction to the other side and are not replaced by an opposing diffusion of carriers. For example, when holes at x_{n0} are swept across the junction to the p side by the $\mathscr{E}$ field, a gradient in the hole distribution in the n material exists, and holes in the n region diffuse toward the junction. The steady state hole distribution in the n region has the inverted exponential shape of Fig. 5-18a. It is important to remember that although the reverse saturation current occurs at the junction by drift of carriers down the barrier, this current is fed from each side by diffusion toward the junction of minority carriers in the neutral regions. The rate of carrier drift across the junction (reverse saturation current) depends on the rate at which holes arrive at x_{n0} (and electrons at x_{p0}) by diffusion from the neutral material. These minority carriers are supplied by thermal generation, and we can show that the expression for the reverse saturation current, Eq. (5-38), represents the rate at which carriers are generated thermally within a diffusion length of each side of the transition region.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 193



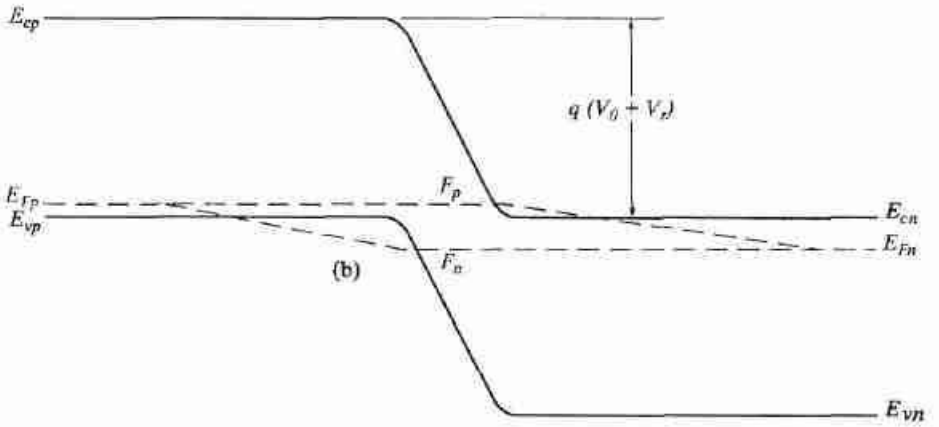


Figure 5–18
Reverse-biased p-n junction: (a) minority carrier distributions near the reverse-biased junction; (b) variation of the quasi-Fermi levels.

In reverse bias, the quasi-Fermi levels split in the opposite sense than in forward bias (Fig. 5-18b). The F_n moves farther away from E_c (close to E_v) and F_p moves farther away from E_v, reflecting the fact that in reverse bias we have fewer carriers than in equilibrium, unlike the forward bias case where we have an excess of carriers. In reverse bias, in the depletion region, we have

$$pn = n_i^2 e^{(F_n - F_p)/kT} \approx 0 \tag{5-42}$$

It is interesting to note that the quasi-Fermi levels in reverse bias can go inside the bands. For example, F_p goes inside the conduction band on the n-side of the depletion region. However, we must remember that F_p is a measure of the hole concentration, and should be correlated with the valence band edge, E_v, and not with E_c. Hence, the band diagram simply reflects the fact that we have very few holes in this region, even fewer than the already small equilibrium minority carrier hole concentration (Fig. 5-18a). Similar observations can be made about the electrons.

EXAMPLE 5-4

An abrupt Si p-n junction ($A = 10^{-4}$ cm^2) has the following properties at 300 K:

p side	*n side*
$N_a = 10^{17}$ cm^{-3}	$N_d = 10^{15}$
$\tau_n = 0.1$ μs	$\tau_p = 10$ μs
$\mu_p = 200$ cm^2/V–s	$\mu_n = 1300$
$\mu_n = 700$	$\mu_p = 450$

The junction is forward biased by 0.5 V. What is the forward current? What is the current at a reverse bias of −0.5 V?

SOLUTION

$$I = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)(e^{qV/kT} - 1) = I_0(e^{qV/kT} - 1)$$

$$p_n = \frac{n_i^2}{n_n} = \frac{(1.5 \times 10^{10})^2}{10^{15}} = 2.25 \times 10^5 \text{ cm}^{-3}$$

$$n_p = \frac{n_i^2}{p_p} = \frac{(1.5 \times 10^{10})^2}{10^{17}} = 2.25 \times 10^3 \text{ cm}^{-3}$$

For minority carriers,

$$D_p = \frac{kT}{q}\mu_p = 0.0259 \times 450 = 11.66 \text{ cm}^2/\text{s on the n side}$$

$$D_n = \frac{kT}{q}\mu_n = 0.0259 \times 700 = 18.13 \text{ cm}^2/\text{s on the p side}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 195

$$L_p = \sqrt{D_p \tau_p} = \sqrt{11.66 \times 10 \times 10^{-6}} = 1.08 \times 10^{-2}\ \text{cm}$$

$$L_n = \sqrt{D_n \tau_n} = \sqrt{18.13 \times 0.1 \times 10^{-6}} = 1.35 \times 10^{-3}\ \text{cm}$$

$$I_0 = qA\left(\frac{D_p}{L_p} p_n + \frac{D_n}{L_n} n_p\right)$$

$$= 1.6 \times 10^{-19} \times 0.0001\left(\frac{11.66}{0.0108} 2.25 \times 10^5 + \frac{18.13}{0.00135} 2.25 \times 10^3\right)$$

$$= 4.370 \times 10^{-15}\ \text{A}$$

$$I = I_0(e^{0.5/0.0259} - 1) \approx \mathbf{1.058 \times 10^{-6}\ A}\ \text{in forward bias.}$$

$$I = -I_0 = \mathbf{-4.37 \times 10^{-15}\ A}\ \text{in reverse bias.}$$

5.4 REVERSE-BIAS BREAKDOWN

We have found that a p-n junction biased in the reverse direction exhibits a small, essentially voltage-independent saturation current. This is true until a critical reverse bias is reached, for which *reverse breakdown* occurs (Fig. 5-19). At this critical voltage (V_{br}) the reverse current through the diode increases sharply, and relatively large currents can flow with little further increase in voltage. The existence of a critical breakdown voltage introduces almost a right-angle appearance to the reverse characteristic of most diodes.

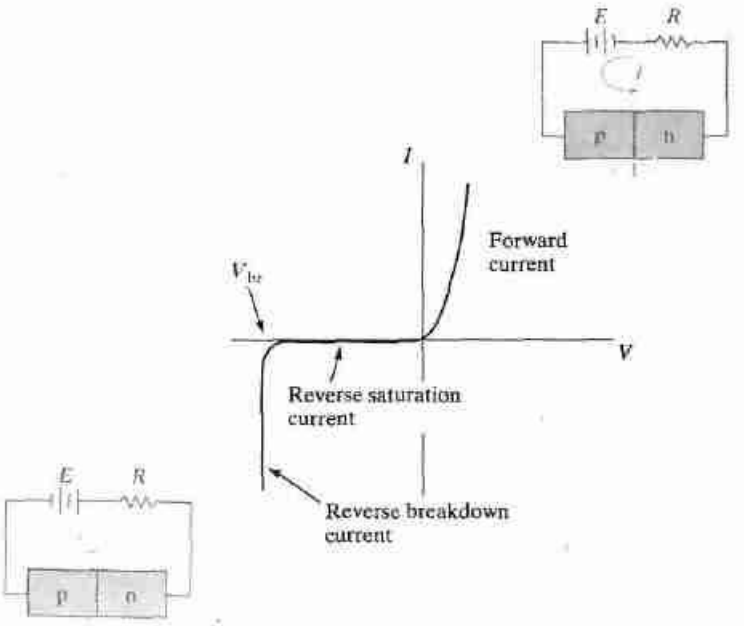


Figure 5–19 Reverse breakdown in a p-n junction.

There is nothing inherently destructive about reverse breakdown. If the current is limited to a reasonable value by the external circuit, the p-n junction can be operated in reverse breakdown as safely as in the forward-bias condition. For example, the maximum reverse current which can flow in the device of Fig. 5-19 is $(E - V_{br})/R$; the series resistance R can be chosen to limit the current to a safe level for the particular diode used. If the current is not limited externally, the junction can be damaged by excessive reverse current, which overheats the device as the maximum power rating is exceeded. It is important to remember, however, that such destruction of the device is not necessarily due to mechanisms unique to reverse breakdown; similar results occur if the device passes excessive current in the forward direction.[12] As we shall see in Section 5.4.4, useful devices called *breakdown diodes* are designed to operate in the reverse breakdown regime of their characteristics.

Reverse breakdown can occur by two mechanisms, each of which requires a critical electric field in the junction transition region. The first mechanism, called the *Zener effect*, is operative at low voltages (up to a few volts reverse bias). If the breakdown occurs at higher voltages (from a few volts to thousands of volts), the mechanism is *avalanche breakdown*. We shall discuss these two mechanisms in this section.

5.4.1 Zener Breakdown

When a heavily doped junction is reverse biased, the energy bands become crossed at relatively low voltages (i.e., the n-side conduction band appears opposite the p-side valence band). As Fig. 5-20 indicates, the crossing of the bands aligns the large number of empty states in the n-side conduction band opposite the many filled states of the p-side valence band. If the barrier separating these two bands is narrow, tunneling of electrons can occur, as discussed in Section 2.4.4. Tunneling of electrons from the p-side valence band to the n-side conduction band constitutes a reverse current from n to p; this is the *Zener effect*.

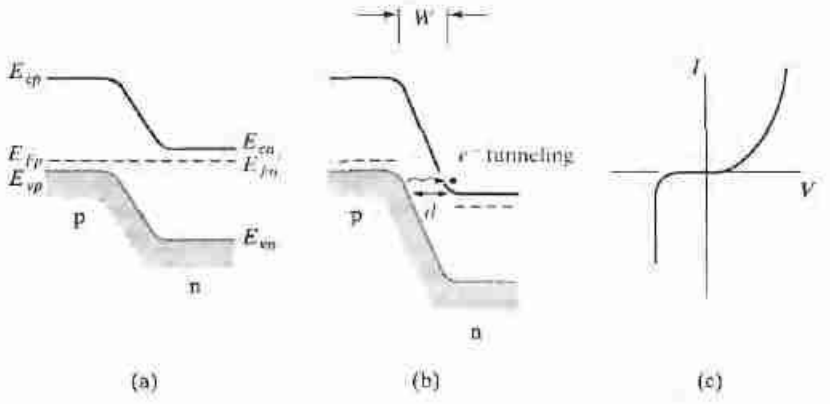


Figure 5-20
The Zener effect: (a) heavily doped junction at equilibrium; (b) reverse bias with electron tunneling from p to n; (c) I–V characteristic.

[12]The dissipated power (IV) in the junction is, of course, greater for a given current in the breakdown regime than it would be for the forward bias, simply because V is greater.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 197

The basic requirements for tunneling current are a large number of electrons separated from a large number of empty states by a narrow barrier of finite height. Since the tunneling probability depends upon the width of the barrier (d in Fig. 5-20), it is important that the metallurgical junction be sharp and the doping high, so that the transition region W extends only a very short distance from each side of the junction. If the junction is not abrupt, or if either side of the junction is lightly doped, the transition region W will be too wide for tunneling.

As the bands are crossed (at a few tenths of a volt for a heavily doped junction), the tunneling distance d may be too large for appreciable tunneling. However, d becomes smaller as the reverse bias is increased, because the higher electric fields result in steeper slopes for the band edges. This assumes that the transition region width W does not increase appreciably with reverse bias. For low voltages and heavy doping on each side of the junction, this is a good assumption. However, if Zener breakdown does not occur with reverse bias of a few volts, avalanche breakdown will become dominant.

In the simple covalent bonding model (Fig. 3-1), the Zener effect can be thought of as *field ionization* of the host atoms at the junction. That is, the reverse bias of a heavily doped junction causes a large electric field within W; at a critical field strength, electrons participating in covalent bonds may be torn from the bonds by the field and accelerated to the n side of the junction. The electric field required for this type of ionization is on the order of 10^6 V/cm.

5.4.2 Avalanche Breakdown

For lightly doped junctions electron tunneling is negligible, and instead, the breakdown mechanism involves the *impact ionization* of host atoms by energetic carriers. Normal lattice-scattering events can result in the creation of EHPs if the carrier being scattered has sufficient energy. For example, if the electric field $\mathscr{E}$ in the transition region is large, an electron entering from the p side may be accelerated to high enough kinetic energy to cause an ionizing collision with the lattice (Fig. 5-21a). A single such interaction results in *carrier multiplication*; the original electron and the generated electron are both swept to the n side of the junction, and the generated hole is swept to the p side (Fig. 5-21b). The degree of multiplication can become very high if carriers generated within the transition region also have ionizing collisions with the lattice. For example, an incoming electron may have a collision with the lattice and create an EHP; each of these carriers has a chance of creating a new EHP, and each of those can also create an EHP, and so forth (Fig. 5-21c). This is an *avalanche* process, since each incoming carrier can initiate the creation of a large number of new carriers.

We can make an approximate analysis of avalanche multiplication by assuming that a carrier of either type has a probability P of having an ionizing collision with the lattice while being accelerated a distance W through the transition region. Thus for n_{in} electrons entering from the p side, there will be Pn_{in} ionizing collisions and an EHP (secondary carriers) for each collision.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 198

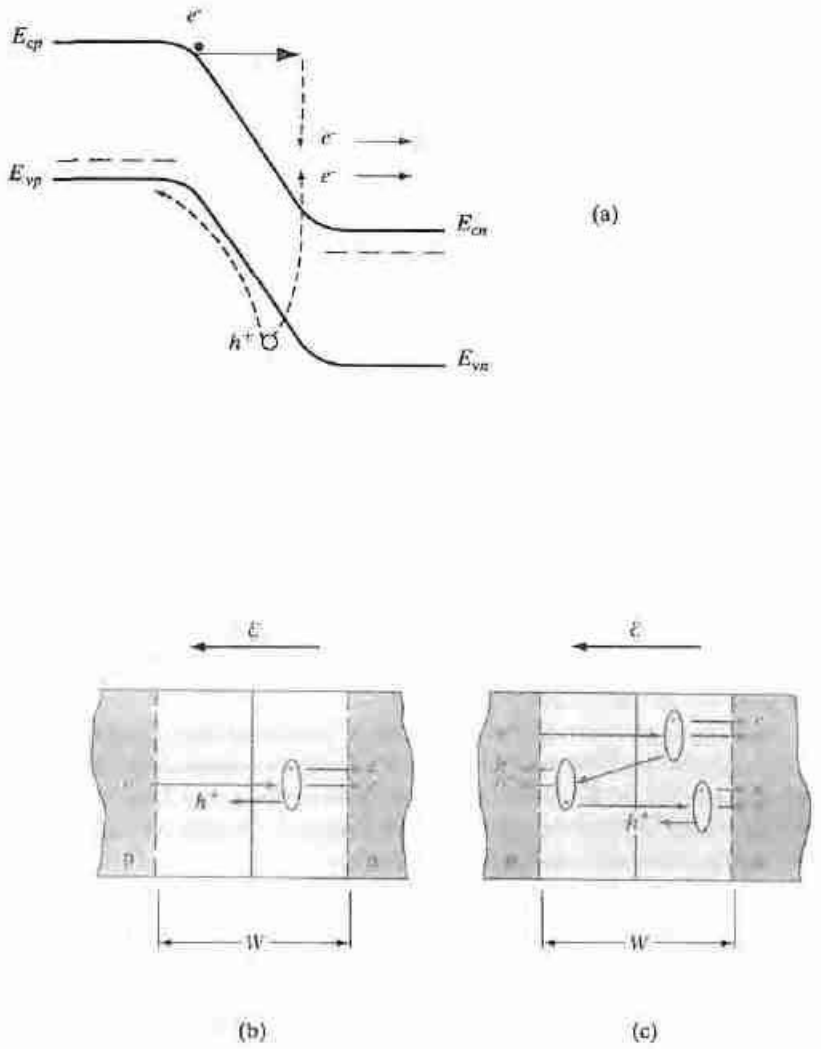


Figure 5–21
Electron-hole pairs created by impact ionization: (a) band diagram of a p-n junction in reverse bias showing (primary) electron gaining kinetic energy in the field of the depletion region, and creating a (secondary) electron–hole pair by impact ionization, the primary electron losing most of its kinetic energy in the process; (b) a single ionizing collision by an incoming electron in the depletion region of the junction; (c) primary, secondary, and tertiary collisions.

After the Pn_{in} collisions by the primary electrons, we have the primary plus the secondary electrons, $n_{in}(1 + P)$. After a collision, each EHP moves effectively a distance of W within the transition region. For example, if an EHP is created at the center of the region, the electron drifts a distance $W/2$ to n and the hole $W/2$ to p. Thus the probability that an ionizing collision will occur due to the motion of the secondary carriers is still P in this simplified model. For $n_{in}P$ secondary pairs there will be $(n_{in}P)P$ ionizing collisions and $n_{in}P^2$ tertiary pairs. Summing up the total number of electrons out of the region at n after many collisions, we have

$$n_{out} = n_{in}(1 + P + P^2 + P^3 + \cdots) \tag{5–43}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 199

assuming no recombination. In a more comprehensive theory we would include recombination as well as different probabilities for ionizing collisions by electrons and holes. In our simple theory, the electron multiplication M_n is

$$M_n = \frac{n_{out}}{n_{in}} = 1 + P + P^2 + P^3 + \cdots = \frac{1}{1 - P} \qquad (5\text{-}44a)$$

as can be verified by direct division. As the probability of ionization P approaches unity, the carrier multiplication (and therefore the reverse current through the junction) increases without limit. Actually, the limit on the current will be dictated by the external circuit.

The relation between multiplication and P was easy to write in Eq. (5-44a); however, the relation of P to parameters of the junction is much more complicated. Physically, we expect the ionization probability to increase with increasing electric field, and therefore to depend on the reverse bias. Measurements of carrier multiplication M in junctions near breakdown lead to an empirical relation

$$M = \frac{1}{1 - (V/V_{br})^n} \qquad (5\text{-}44b)$$

where the exponent **n** varies from about 3 to 6, depending on the type of material used for the junction.

In general, the critical reverse voltage for breakdown increases with the band gap of the material, since more energy is required for an ionizing collision. Also, the peak electric field within W increases with increased doping on the more lightly doped side of the junction. Therefore, V_{br} decreases as the doping increases, as Fig. 5-22 indicates.

5.4.3 Rectifiers

The most obvious property of a p-n junction is its *unilateral* nature; that is, to a good approximation it conducts current in only one direction. We can think of an *ideal diode* as a short circuit when forward biased and as an open circuit when reverse biased (Fig. 5-23a). The p-n junction diode does not quite fit this description, but the I–V characteristics of many junctions can be approximated by the ideal diode in series with other circuit elements to form an equivalent circuit. For example, most forward-biased diodes exhibit an *offset voltage* E_0 (see Fig. 5-33), which can be approximated in a circuit model by a battery in series with the ideal diode (Fig. 5-23b). The series battery in the model keeps the ideal diode turned off for applied voltages less than E_0. From Section 5.6.1 we expect E_0 to be approximately the contact potential of the junction. In some cases the approximation to the actual diode characteristic is improved by adding a series resistor R to the circuit equivalent (Fig. 5-23c). The equivalent circuit approximations illustrated in Fig. 5-23

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 200

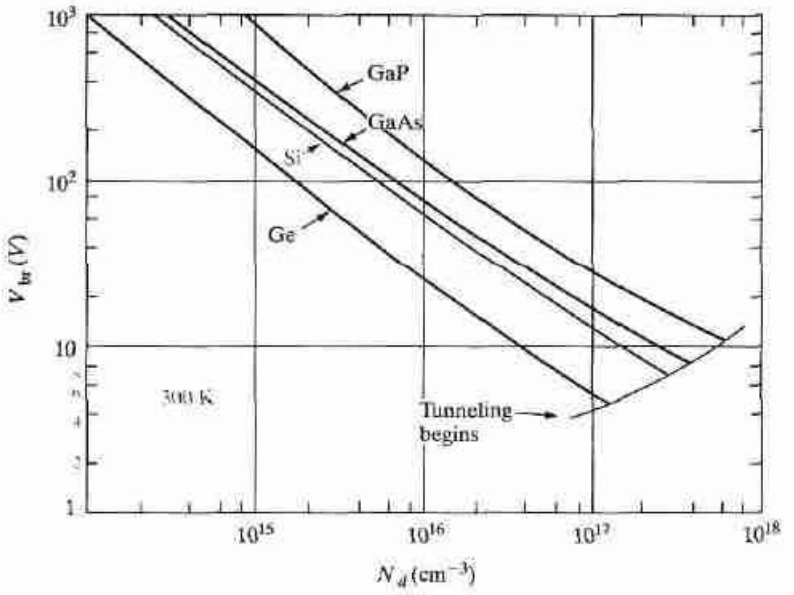


Figure 5–22 Variation of avalanche breakdown voltage in abrupt p⁺-n junctions, as a function of donor concentration on the n side, for several semiconductors. [After S. M. Sze and G. Gibbons, *Applied Physics Letters*, vol. 8, p. 111 (1966).]

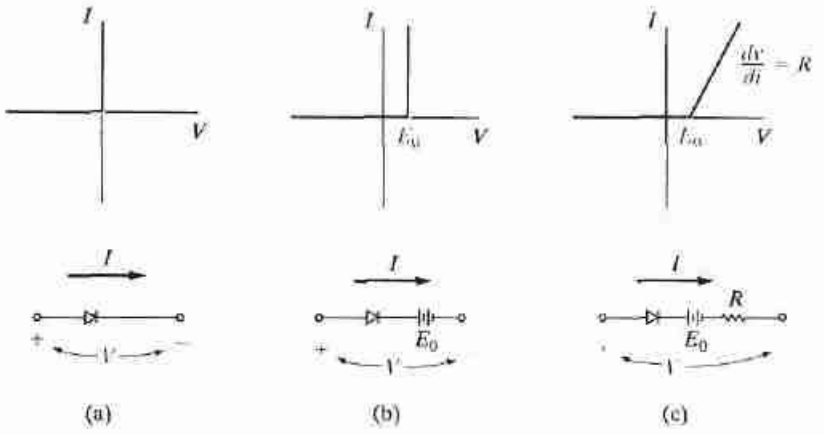


Figure 5–23 Piecewise-linear approximations of junction diode characteristics: (a) the ideal diode; (b) ideal diode with an offset voltage; (c) ideal diode with an offset voltage and a resistance to account for slope in the forward characteristic.

are called *piecewise-linear equivalents*, since the approximate characteristics are linear over specific ranges of voltage and current.

An ideal diode can be placed in series with an a-c voltage source to provide *rectification* of the signal. Since current can flow only in the forward direction through the diode, only the positive half-cycles of the input sine wave are passed. The output voltage is a *half-rectified sine wave*. Whereas the input sinusoid has zero average value, the rectified signal has a positive average value and therefore contains a d-c component. By appropriate filtering, this d-c level can be extracted from the rectified signal.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 201

The unilateral nature of diodes is useful for many other circuit applications that require *waveshaping*. This involves alteration of a-c signals by passing only certain portions of the signal while blocking other portions.

Junction diodes designed for use as rectifiers should have I–V characteristics as close as possible to that of the ideal diode. The reverse current should be negligible, and the forward current should exhibit little voltage dependence (negligible *forward resistance* R). The reverse breakdown voltage should be large, and the offset voltage E_0 in the forward direction should be small. Unfortunately, not all of these requirements can be met by a single device; compromises must be made in the design of the junction to provide the best diode for the intended application.

From the theory derived in Section 5.3 we can easily list the various requirements for good rectifier junctions. *Band gap* is obviously an important consideration in choosing a material for rectifier diodes. Since n_i is small for large band gap materials, the reverse saturation current (which depends on thermally generated carriers) decreases with increasing E_g. A rectifier made with a wide band gap material can be operated at higher temperatures, because thermal excitation of EHPs is reduced by the increased band gap. Such temperature effects are critically important in rectifiers, which must carry large currents in the forward direction and are thereby subjected to appreciable heating. On the other hand, the contact potential and offset voltage E_0 generally increase with E_g. This drawback is usually outweighed by the advantages of low n_i; for example, Si is generally preferred over Ge for power rectifiers because of its wider band gap, lower leakage current, and higher breakdown voltage, as well as its more convenient fabrication properties.

The *doping concentration* on each side of the junction influences the avalanche breakdown voltage, the contact potential, and the series resistance of the diode. If the junction has one highly doped side and one lightly doped side (such as a p^+-n junction), the lightly doped region determines many of the properties of the junction. From Fig. 5-22 we see that a high-resistivity region should be used for at least one side of the junction to increase the breakdown voltage V_{br}. However, this approach tends to increase the forward resistance R of Fig. 5-23c, and therefore contributes to the problems of thermal effects due to I^2R heating. To reduce the resistance of the lightly doped region, it is necessary to make its area large and reduce its length. Therefore, the physical *geometry* of the diode is another important design variable. Limitations on the practical area for a diode include problems of obtaining uniform starting material and junction processing over large areas. Localized flaws in junction uniformity can cause premature reverse breakdown in a small region of the device. Similarly, the lightly doped region of the junction cannot be made arbitrarily short. One of the primary problems with a short, lightly doped region is an effect called *punch-through*. Since the transition region width W increases with reverse bias and extends primarily into the lightly doped region, it is possible for W to increase until it fills the entire length of this region (Prob. 5.25). The result of punch-through is a breakdown below the value of V_{br} expected from Fig. 5-22.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 202

In devices designed for use at high reverse bias, care must be taken to avoid premature breakdown across the edge of the sample. This effect can be reduced by *beveling* the edge or by diffusing a *guard ring* to isolate the junction from the edge of the sample (Fig. 5-24). The electric field is lower at the beveled edge of the sample in Fig. 5-24b than it is in the main body of the device. Similarly, the junction at the lightly doped p guard ring of Fig. 5-24c breaks down at higher voltage than the p^+-n junction. Since the depletion region is wider in the p ring than in the p^+ region, the average electric field is smaller at the ring for a given diode reverse voltage.

In fabricating a p^+-n or a p-n^+ junction, it is common to terminate the lightly doped region with a heavily doped layer of the same type (Fig. 5-25a), to ease the problem of making ohmic contact to the device. The result is a p^+-n-n^+ structure with the p^+-n layer serving as the active junction, or a p^+-p-n^+ device with an active p-n^+ junction. The lightly doped center region determines the avalanche breakdown voltage. If this region is short compared with the minority carrier diffusion length, the excess carrier injection for large forward currents can increase the conductivity of the region significantly. This type of *conductivity modulation*, which reduces the forward resistance R,

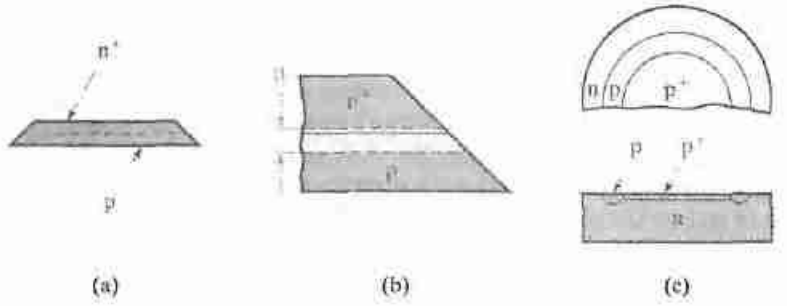


Figure 5–24
Beveled edge and guard ring to prevent edge breakdown under reverse bias: (a) diode with beveled edge; (b) closeup view of edge, showing reduction of depletion region near the level; (c) guard ring.

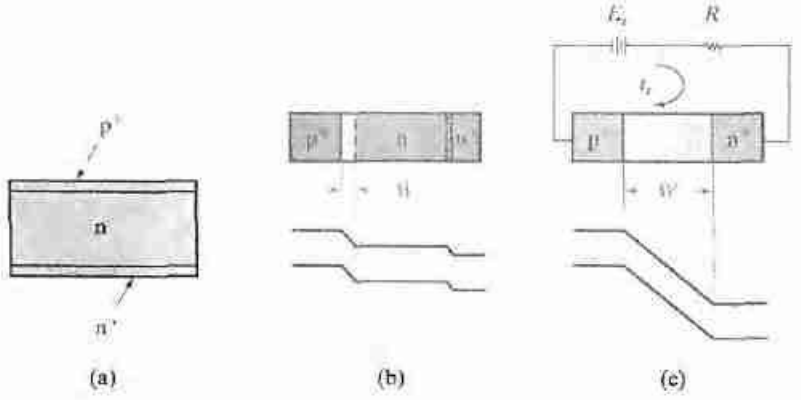


Figure 5–25
A p^+-n-n^+ junction diode: (a) device configuration; (b) zero-bias condition; (c) reverse-biased to punch-through.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 203

can be very useful for high-current devices. On the other hand, a short, lightly doped center region can also lead to punch-through under reverse bias, as in Fig. 5-25c.

The mounting of a rectifier junction is critical to its ability to handle power. For diodes used in low-power circuits, glass or plastic encapsulation or a simple header mounting is adequate. However, high-current devices that must dissipate large amounts of heat require special mountings to transfer thermal energy away from the junction. A typical Si power rectifier is mounted on a molybdenum or tungsten disk to match the thermal expansion properties of the Si. This disk is fastened to a large stud of copper or other thermally conductive material that can be bolted to a heat sink with appropriate cooling.

5.4.4 The Breakdown Diode

As we discussed earlier in this section, the reverse-bias breakdown voltage of a junction can be varied by choice of junction doping concentrations. The breakdown mechanism is the Zener effect (tunneling) for abrupt junctions with extremely heavy doping; however, the more common breakdown is avalanche (impact ionization), typical of more lightly doped or graded junctions. By varying the doping we can fabricate diodes with specific breakdown voltages ranging from less than one volt to several hundred volts. If the junction is well designed, the breakdown will be sharp and the current after breakdown will be essentially independent of voltage (Fig. 5-26a). When a diode is designed for a specific breakdown voltage, it is called a *breakdown diode*. Such diodes are also called *Zener diodes*, despite the fact that the actual breakdown mechanism is usually the avalanche effect. This error in terminology is due to an early mistake in identifying the first observations of breakdown in p-n junctions.

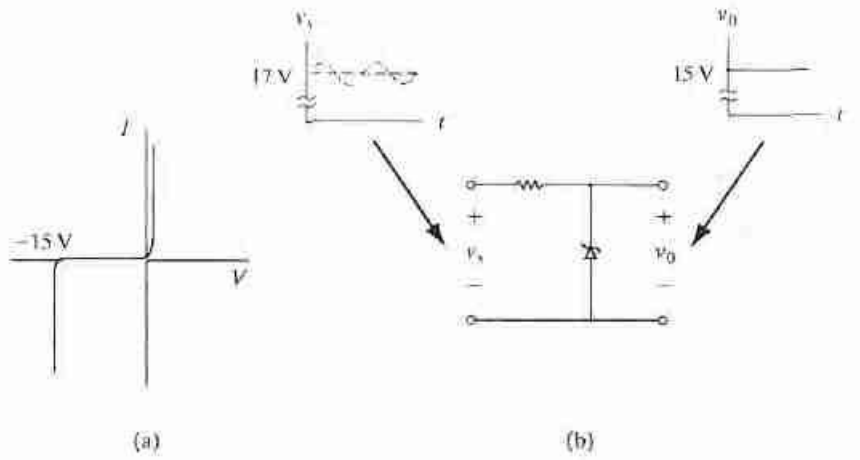


Figure 5-26
A breakdown diode: (a) *I–V* characteristic; (b) application as a voltage regulator.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 204

Breakdown diodes can be used as *voltage regulators* in circuits with varying inputs. The 15-V breakdown diode of Fig. 5-26 holds the circuit output voltage v_0 constant at 15 V, while the input varies at voltages greater than 15 V. For example, if v_s is a rectified and filtered signal composed of a 17-V d-c component and a 1-V ripple variation above and below 17 V, the output v_0 will remain constant at 15 V. More complicated voltage regulator circuits can be designed using breakdown diodes, depending on the type of signal being regulated and the nature of the output load. In a similar application, such a device can be used as a *reference diode*; since the breakdown voltage of a particular diode is known, the voltage across it during breakdown can be used as a reference in circuits that require a known value of voltage.

5.5 TRANSIENT AND A-C CONDITIONS

We have considered the properties of p-n junctions under equilibrium conditions and with steady state current flow. Most of the basic concepts of junction devices can be obtained from these properties, except for the important behavior of junctions under transient or a-c conditions. Since most solid state devices are used for switching or for processing a-c signals, we cannot claim to understand p-n junctions without knowing at least the basics of time-dependent processes. Unfortunately, a complete analysis of these effects involves more mathematical manipulation than is appropriate for an introductory discussion. Basically, the problem involves solving the various current flow equations in two simultaneous variables, space and time. We can, however, obtain the basic results for several special cases which represent typical time-dependent applications of junction devices.

In this section we investigate the important influence of excess carriers in transient and a-c problems. The switching of a diode from its forward state to its reverse state is analyzed to illustrate a typical transient problem. Finally, these concepts are applied to the case of small a-c signals to determine the equivalent capacitance of a p-n junction.

5.5.1 Time Variation of Stored Charge

Another look at the excess carrier distributions of a p-n junction under bias (e.g., Fig. 5-15) tells us that any change in current must lead to a change of charge stored in the carrier distributions. Since time is required in building up or depleting a charge distribution, however, the stored charge must inevitably lag behind the current in a time-dependent problem. This is inherently a capacitive effect, as we shall see in Section 5.5.4.

For a proper solution of a transient problem, we must use the time-dependent continuity equations, Eqs. (4-31). We can obtain each component of the current at position x and time t from these equations; for example, from Eq. (4-31a) we can write

$$-\frac{\partial J_p(x,t)}{\partial x} = q\frac{\delta p(x,t)}{\tau_p} + q\frac{\partial p(x,t)}{\partial t} \qquad (5\text{-}45)$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 205

To obtain the instantaneous current density, we can integrate both sides at time t to obtain

$$J_p(0) - J_p(x) = q\int_0^x \left[\frac{\delta p(x,t)}{\tau_p} + \frac{\partial p(x,t)}{\partial t}\right] dx \tag{5-46}$$

For injection into a long n region from a p^+ region, we can take the current at $x_n = 0$ to be all hole current, and J_p at $x_n = \infty$ to be zero. Then the total injected current, including time variations, is

$$i(t) = i_p(x_n = 0, t) = \frac{qA}{\tau_p}\int_0^\infty \delta p(x_n, t)dx_n + qA\frac{\partial}{\partial t}\int_0^\infty \delta p(x_n, t)dx_n$$

$$\boxed{i(t) = \frac{Q_p(t)}{\tau_p} + \frac{dQ_p(t)}{dt}} \tag{5-47}$$

This result indicates that the hole current injected across the p^+-n junction (and therefore approximately the total diode current) is determined by two charge storage effects: (1) the usual recombination term Q_p/τ_p in which the excess carrier distribution is replaced every τ_p seconds, and (2) a charge buildup (or depletion) term dQ_p/dt, which allows for the fact that the distribution of excess carriers can be increasing or decreasing in a time-dependent problem. For steady state the dQ_p/dt term is zero, and Eq. (5-47) reduces to Eq. (5-40), as expected. In fact, we could have written Eq. (5-47) intuitively rather than having obtained it from the continuity equation, since it is reasonable that the hole current injected at any given time must supply minority carriers for recombination and for whatever variations occur in the total stored charge.

We can solve for the stored charge as a function of time for a given current transient. For example, the step turn-off transient (Fig. 5-27a), in which a current I is suddenly removed at $t = 0$, leaves the diode with stored charge. Since the

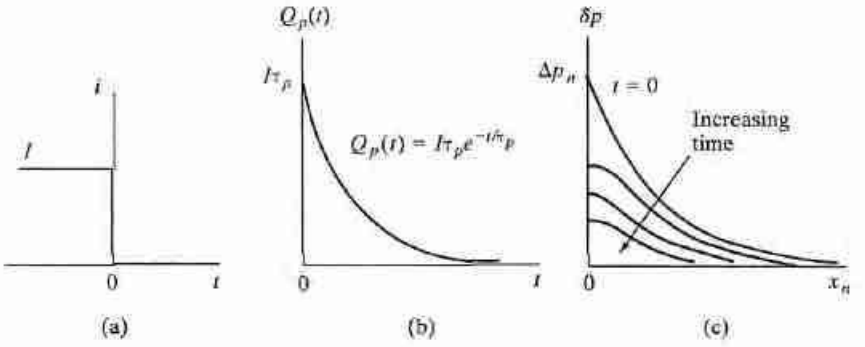


Figure 5-27 Effects of a step turn-off transient in a p^+-n diode: (a) current through the diode; (b) decay of stored charge in the n-region; (c) excess hole distribution in the n-region as a function of time during the transient.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 206

excess holes in the n region must die out by recombination with the matching excess electron population, some time is required for $Q_p(t)$ to reach zero. Solving Eq. (5-47) with Laplace transforms, with $i(t > 0) = 0$ and $Q_p(0) = I\tau_p$, we obtain

$$0 = \frac{1}{\tau_p} Q_p(s) + sQ_p(s) - I\tau_p$$

$$Q_p(s) = \frac{I\tau_p}{s + 1/\tau_p}$$

$$Q_p(t) = I\tau_p e^{-t/\tau_p} \tag{5-48}$$

As expected, the stored charge dies out exponentially from its initial value $I\tau_p$ with a time constant equal to the hole lifetime in the n material.

An important implication of Fig. 5-27 is that even though the current is suddenly terminated, the voltage across the junction persists until Q_p disappears. Since the excess hole concentration can be related to junction voltage by formulas derived in Section 5.3.2, we can presumably solve for $v(t)$. We already know that at any time during the transient, the excess hole concentration at $x_n = 0$ is

$$\Delta p_n(t) = p_n(e^{qv(t)/kT} - 1) \tag{5-49}$$

so that finding $\Delta p_n(t)$ will easily give us the transient voltage. Unfortunately, it is not simple to obtain $\Delta p_n(t)$ exactly from our expression for $Q_p(t)$. The problem is that the hole distribution does not remain in the convenient exponential form it has in steady state. As Fig. 5-27c suggests, the quantity $\delta p(x_n, t)$ becomes markedly nonexponential as the transient proceeds. For example, since the injected hole current is proportional to the gradient of the hole distribution at $x_n = 0$ (Fig. 5-16a), zero current implies zero gradient. Thus the slope of the distribution must be exactly zero at $x_n = 0$ throughout the transient.[13] This zero slope at the point of injection distorts the exponential distribution, particularly in the region near the junction. As time progresses in Fig. 5-27c, δp (and therefore δn) decreases as the excess electrons and holes recombine. To find the exact expression for $\delta p(x_n, t)$ during the transient would require a rather difficult solution of the time-dependent continuity equation.

An approximate solution for $v(t)$ can be obtained by assuming an exponential distribution for δp at every instant during the decay. This type of *quasi-steady state* approximation neglects distortion due to the slope requirement at $x_n = 0$ and the effects of diffusion during the transient. Thus we would expect the calculation to give rather crude results. On the other hand, such a solution can give us a feeling for the variation of junction voltage during the transient. If we take

$$\delta p(x_n, t) = \Delta p_n(t) e^{-x_n/L_p} \tag{5-50}$$

[13]We notice that, while the magnitude of δp cannot change instantaneously, the slope must go to zero immediately. This can occur in a small region near the junction with negligible redistribution of charge at $t = 0$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 207

we have for the stored charge at any instant

$$Q_p(t) = qA\int_0^\infty \Delta p_n(t)e^{-x_n/L_p}dx_n = qAL_p\Delta p_n(t) \tag{5-51}$$

Relating $\Delta p_n(t)$ to $v(t)$ by Eq. (5-49) we have

$$\Delta p_n(t) = p_n(e^{qv(t)/kT} - 1) = \frac{Q_p(t)}{qAL_p} \tag{5-52}$$

Thus in the quasi-steady state approximation, the junction voltage varies according to

$$v(t) = \frac{kT}{q}\ln\left(\frac{I\tau_p}{qAL_pp_n}e^{-t/\tau_p} + 1\right) \tag{5-53}$$

during the turn-off transient of Fig. 5-27. This analysis, while not accurate in its details, does indicate clearly that the voltage across a p-n junction cannot be changed instantaneously, and that stored charge can present a problem in a diode intended for switching applications.

Many of the problems of stored charge can be reduced by designing a p^+-n diode (for example) with a very narrow n region. If the n region is shorter than a hole diffusion length, very little charge is stored. Thus, little time is required to switch the diode on and off. This type of structure, called the *narrow base diode*, is considered in Prob. 5.35. The switching process can be made still faster by purposely adding recombination centers, such as Au atoms in Si, to increase the recombination rate.

5.5.2 Reverse Recovery Transient

In most switching applications a diode is switched from forward conduction to a reverse-biased state, and vice versa. The resulting stored charge transient is somewhat more complicated than for a simple turn-off transient, and therefore it requires slightly more analysis. An important result of this example is that a reverse current much larger than the normal reverse saturation current can flow in a junction during the time required for readjustment of the stored charge.

Let us assume a p^+-n junction is driven by a square wave generator that periodically switches from $+E$ to $-E$ volts (Fig. 5-28a). While E is positive the diode is forward biased, and in steady state the current I_f flows through the junction. If E is much larger than the small forward voltage of the junction, the source voltage appears almost entirely across the resistor, and the current is approximately $i = I_f \simeq E/R$. After the generator voltage is reversed ($t > 0$), the current must initially reverse to $i = I_r \simeq -E/R$. The reason for this unusually large reverse current through the diode is that the stored charge (and hence the junction voltage) cannot be changed instantaneously. Therefore, just as the current is reversed, the junction voltage remains at the small

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 208

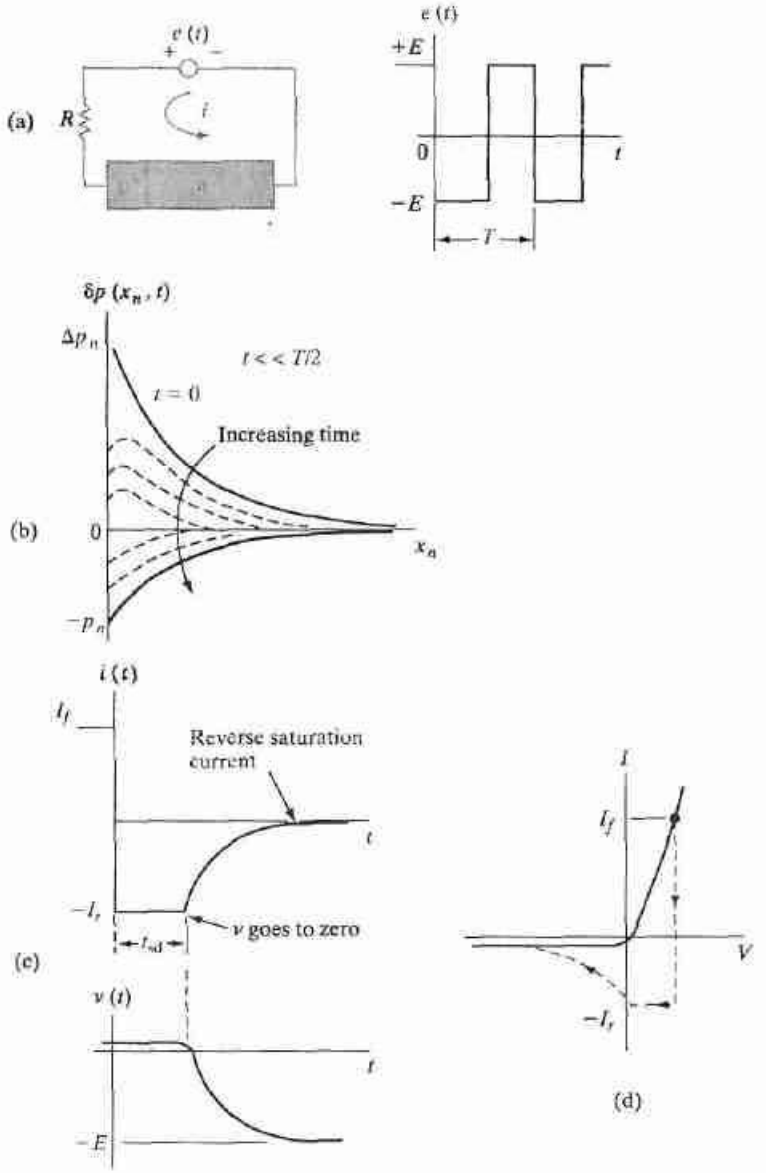


Figure 5–28
Storage delay time in a p^+-n diode: (a) circuit and input square wave; (b) hole distribution in the n-region as a function of time during the transient; (c) variation of current and voltage with time; (d) sketch of transient current and voltage on the device I–V characteristic.

forward-bias value it had before $t = 0$. A voltage loop equation then tells us that the large reverse current $-E/R$ must flow temporarily. While the current is negative through the junction, the slope of the $\delta p(x_n)$ distribution must be positive at $x_n = 0$.

As the stored charge is depleted from the neighborhood of the junction (Fig. 5-28b), we can find the junction voltage again from Eq. (5-49).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 209

As long as Δp_n is positive, the junction voltage $v(t)$ is positive and small; thus $i \simeq -E/R$ until Δp_n goes to zero. When the stored charge is depleted and Δp_n becomes negative, the junction exhibits a negative voltage. Since the reverse-bias voltage of a junction can be large, the source voltage begins to divide between R and the junction. As time proceeds, the magnitude of the reverse current becomes smaller as more of $-E$ appears across the reverse-biased junction, until finally the only current is the small reverse saturation current which is characteristic of the diode. The time t_{sd} required for the stored charge (and therefore the junction voltage) to become zero is called the *storage delay time*. This delay time is an important figure of merit in evaluating diodes for switching applications. It is usually desirable that t_{sd} be small compared with the switching times required (Fig. 5-29). The critical parameter determining t_{sd} is the carrier lifetime (τ_p for the example of the p^+-n junction). Since the recombination rate determines the speed with which excess holes can disappear from the n region, we would expect t_{sd} to be proportional to τ_p. In fact, an exact analysis of the problem of Fig. 5-28 leads to the result

$$t_{sd} = \tau_p \left[\text{erf}^{-1}\left(\frac{I_f}{I_f + I_r}\right)\right]^2 \tag{5-54}$$

where the error function (erf) is a tabulated function. Although the exact solution leading to Eq. (5-54) is too lengthy for us to consider here, an approximate result can be obtained from the quasi-steady state assumption.

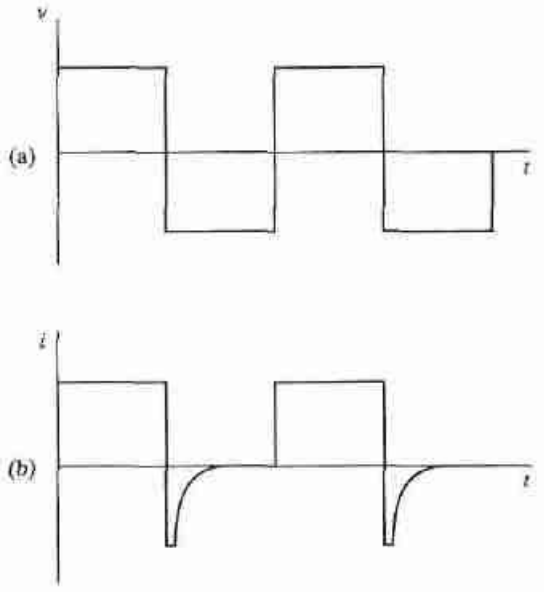


Figure 5-29
Effects of storage delay time on switching signal: (a) switching voltage; (b) diode current.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 210

EXAMPLE 5–5

Assume a p^+-n diode is biased in the forward direction, with a current I_f. At time $t = 0$ the current is switched to $-I_r$. Use the appropriate boundary conditions to solve Eq. (5-47) for $Q_p(t)$. Apply the quasi-steady state approximation to find the storage delay time t_{sd}.

SOLUTION

From Eq. (5-47),

$$i(t) = \frac{Q_p(t)}{\tau_p} + \frac{dQ_p(t)}{dt} \quad for\ t < 0, Q_p = I_f\tau_p$$

Using Laplace transforms,

$$-\frac{I_r}{s} = \frac{Q_p(s)}{\tau_p} + sQ_p(s) - I_f\tau_p$$

$$Q_p(s) = \frac{I_f\tau_p}{s + 1/\tau_p} - \frac{I_r}{s(s + 1/\tau_p)}$$

$$Q_p(t) = I_f\tau_p e^{-t/\tau_p} + I_r\tau_p(e^{-t/\tau_p} - 1) = \tau_p[-I_r + (I_f + I_r)e^{-t/\tau_p}]$$

Assuming that $Q_p(t) = qAL_p\Delta p_n(t)$ as in Eq. (5-52),

$$\Delta p_n(t) = \frac{\tau_p}{qAL_p}[-I_r + (I_f + I_r)e^{-t/\tau_p}]$$

This is set to equal zero when $t = t_{sd}$, and we obtain:

$$t_{sd} = -\tau_p \ln\left[\frac{I_r}{I_f + I_r}\right] = \tau_p \ln\left(1 + \frac{I_f}{I_r}\right)$$

An important result of Eq. (5-54) is that τ_p can be calculated in a straightforward way from a measurement of storage delay time. In fact, measurement of t_{sd} from an experimental arrangement such as Fig. 5-28a is a common method of measuring lifetimes. In some cases this is a more convenient technique than the photoconductive decay measurement discussed in Section 4.3.2.

As in the case of the turn-off transient of the previous section, the storage delay time can be reduced by introducing recombination centers into the diode material, thus reducing the carrier lifetimes, or by utilizing the narrow base diode configuration.

5.5.3 Switching Diodes

In discussing rectifiers we emphasized the importance of minimizing the reverse-bias current and the power losses under forward bias. In many applications, time response can be important as well. If a junction diode is to be

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 211

used to switch rapidly from the conducting to the nonconducting state and back again, special consideration must be given to its charge control properties. We have discussed the equations governing the turn-on time and the reverse recovery time of a junction. From Eqs. (5-47) and (5-54) it is clear that a diode with fast switching properties must either store very little charge in the neutral regions for steady forward currents, or have a very short carrier lifetime, or both.

As mentioned above, we can improve the switching speed of a diode by adding efficient recombination centers to the bulk material. For Si diodes, Au doping is useful for this purpose. To a good approximation the carrier lifetime varies with the reciprocal of the recombination center concentration. Thus, for example, a p^+-n Si diode may have $\tau_p = 1$ μs and a reverse recovery time of 0.1 μs before Au doping. If the addition of 10^{14} Au atoms/cm^3 reduces the lifetime to 0.1 μs and t_{sd} to 0.01 μs, 10^{15} cm^{-3} Au atoms could reduce τ_p to 0.01 μs and t_{sd} to 1 ns (10^{-9} s). This process cannot be continued indefinitely, however. The reverse current due to generation of carriers from the Au centers in the depletion region becomes appreciable with large Au concentration (Section 5.6.2). In addition, as the Au concentration approaches the lightest doping of the junction, the equilibrium carrier concentration of that region can be affected.

A second approach to improving the diode switching time is to make the lightly doped neutral region shorter than a minority carrier diffusion length. This is the *narrow base diode* (Prob. 5.35). In this case the stored charge for forward conduction is very small, since most of the injected carriers diffuse through the lightly doped region to the end contact. When such a diode is switched to reverse conduction, very little time is required to eliminate the stored charge in the narrow neutral region. The mathematics involved in Prob. 5.35 is particularly interesting, because it closely resembles the calculations we shall make in analyzing the bipolar junction transistor in Chapter 7.

5.5.4 Capacitance of p-n Junctions

There are basically two types of capacitance associated with a junction: (1) the *junction capacitance* due to the dipole in the transition region and (2) the *charge storage capacitance* arising from the lagging behind of voltage as current changes, due to charge storage effects.[14] Both of these capacitances are important, and they must be considered in designing p-n junction devices for use with time-varying signals. The junction capacitance (1) is dominant under reverse-bias conditions, and the charge storage capacitance (2) is dominant when the junction is forward biased. In many applications of p-n junctions, the capacitance is a limiting factor in the usefulness of the device; on the other hand, there are important applications in which the capacitance discussed here can be useful in circuit applications and in providing important information about the structure of the p-n junction.

[14]The capacitance (1) above is also referred to as the transition region capacitance or depletion layer capacitance; (2) is often called the diffusion capacitance.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 212

The junction capacitance of a diode is easy to visualize from the charge distribution in the transition region (Fig. 5-12). The uncompensated acceptor ions on the p side provide a negative charge, and an equal positive charge results from the ionized donors on the n side of the transition region. The capacitance of the resulting dipole is slightly more difficult to calculate than is the usual parallel plate capacitance, but we can obtain it in a few steps.

Instead of the common expression $C = |Q/V|$, which applies to capacitors in which charge is a linear function of voltage, we must use the more general definition

$$C = \left|\frac{dQ}{dV}\right| \tag{5-55}$$

since the charge Q on each side of the transition region varies nonlinearly with the applied voltage (Fig. 5-30a). We can demonstrate this nonlinear dependence by reviewing the equations for the width of the transition region (W) and the resulting charge. The equilibrium value of W was found in Eq. (5-21) to be

$$W = \left[\frac{2\epsilon V_0}{q}\left(\frac{N_a + N_d}{N_a N_d}\right)\right]^{1/2} \quad (equilibrium) \tag{5-56}$$

Since we are dealing with the nonequilibrium case with voltage V applied, we must use the altered value of the electrostatic potential barrier $(V_0 - V)$, as discussed in relation to Fig. 5-13. The proper expression for the width of the transition region is then

$$W = \left[\frac{2\epsilon(V_0 - V)}{q}\left(\frac{N_a + N_d}{N_a N_d}\right)\right]^{1/2} \quad (with\ bias) \tag{5-57}$$

In this expression the applied voltage V can be either positive or negative to account for forward or reverse bias. As expected, the width of the transition region is increased for reverse bias and is decreased under forward bias. Since the uncompensated charge Q on each side of the junction varies with the transition region width, variations in the applied voltage result in corresponding variations in the charge, as required for a capacitor. The value of Q can be written in terms of the doping concentration and transition region width on each side of the junction (Fig. 5-12):

$$|Q| = qAx_{n0}N_d = qAx_{p0}N_a \tag{5-58}$$

Relating the total width of the transition region W to the individual widths x_{n0} and x_{p0} from Eqs. (5-23) we have

$$x_{n0} = \frac{N_a}{N_a + N_d}W, \quad x_{p0} = \frac{N_d}{N_a + N_d}W \tag{5-59}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 213

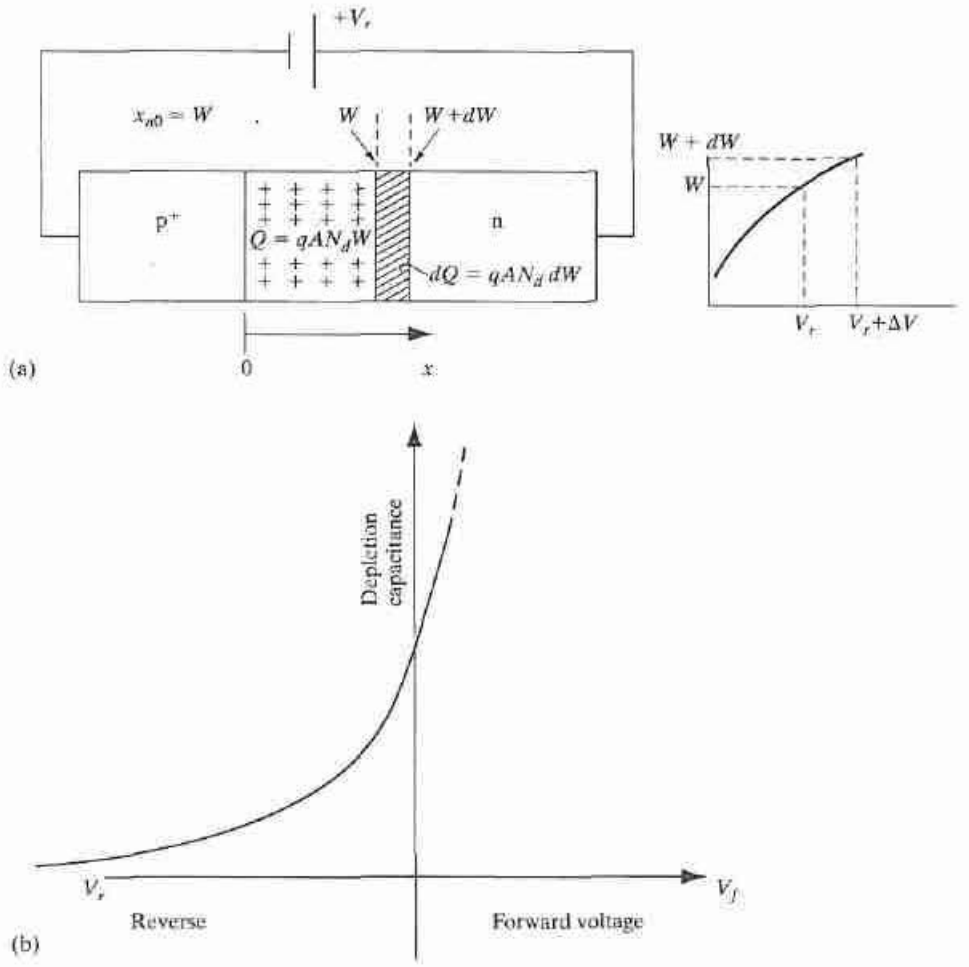


Figure 5–30
Depletion capacitance of a junction: (a) p^+-n junction showing variation of depletion edge on n side with reverse bias. Electrically, the structure looks like a parallel plate capacitor whose dielectric is the depletion region, and the plates are the space charge neutral regions; (b) variation of depletion capacitance with reverse bias [Eq. (5–63)]. We neglect x_{p0} in the heavily-doped p^+ material.

and therefore the charge on each side of the dipole is

$$|Q| = qA\frac{N_d N_a}{N_d + N_a}W = A\left[2q\epsilon(V_0 - V)\frac{N_d N_a}{N_d + N_a}\right]^{1/2} \tag{5-60}$$

Thus the charge is indeed a nonlinear function of applied voltage. From this expression and the definition of capacitance in Eq. (5-55), we

can calculate the junction capacitance C_j. Since the voltage that varies the charge in the transition region is the barrier height $(V_0 - V)$, we must take the derivative with respect to this potential difference:

$$C_j = \left|\frac{dQ}{d(V_0 - V)}\right| = \frac{A}{2}\left[\frac{2q\epsilon}{(V_0 - V)}\frac{N_d N_a}{N_d + N_a}\right]^{1/2} \tag{5-61}$$

The quantity C_j is a *voltage-variable capacitance*, since C_j is proportional to $(V_0 - V)^{-1/2}$. There are several important applications for variable capacitors, including use in tuned circuits. The p-n junction device which makes use of the voltage-variable properties of C_j is called a *varactor*. We shall discuss this device further in Section 5.5.5.

Although the dipole charge is distributed in the transition region of the junction, the form of the parallel plate capacitor formula is obtained from the expressions for C_j and W (Fig. 5-30a):

$$C_j = \epsilon A\left[\frac{q}{2\epsilon(V_0 - V)}\frac{N_d N_a}{N_d + N_a}\right]^{1/2} = \frac{\epsilon A}{W} \tag{5-62}$$

In analogy with the parallel plate capacitor, the transition region width W corresponds with the plate separation of the conventional capacitor.

In the case of an asymmetrically doped junction, the transition region extends primarily into the less heavily doped side, and the capacitance is determined by only one of the doping concentrations (Fig. 5-30a). For a p^+-n junction, $N_a \gg N_d$ and $x_{n0} \simeq W$, while x_{p0} is negligible. The capacitance is then (Fig. 5-30b)

$$\boxed{C_j = \frac{A}{2}\left[\frac{2q\epsilon}{V_0 - V}N_d\right]^{1/2} \qquad \text{for } p^+\text{-}n} \tag{5-63}$$

It is therefore possible to obtain the doping concentration of the lightly doped n region from a measurement of capacitance. For example, in a reverse-biased junction the applied voltage $V = -V_r$ can be made much larger than the contact potential V_0, so that the latter becomes negligible. If the area of the junction can be measured, a reliable value of N_d results from a measurement of C_j. However, these equations were obtained by assuming a sharp step junction. Certain modifications must be made in the case of a graded junction (Section 5.6.4 and Prob. 5.37).

The junction capacitance dominates the reactance of a p-n junction under reverse bias; for forward bias, however, the charge storage, or diffusion

[15] **Laux, S., and K. Hess,** "Revisiting the Analytic Theory of P-N Junction Impedance: Improvements Guided by Computer Simulation Leading to a New Equivalent Circuit," *IEEE Trans. Elec. Dev.* 46(2) (1999), 396.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 215

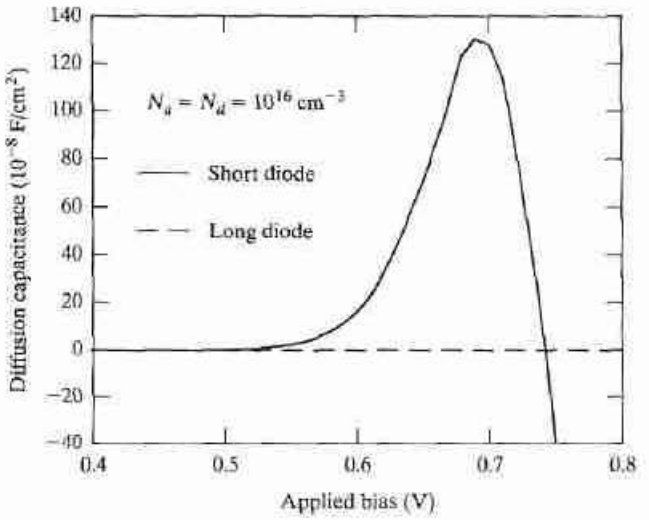


Figure 5–31
Diffusion capacitance as a function of forward bias in long and short diodes.

capacitance, C_s becomes dominant. It has been recently shown[15] that the various time-dependent current components, as well as the boundary conditions, affect the diffusion capacitance in forward bias. We need to specify at which terminals the stored charges are extracted or "reclaimed" and where the relevant voltage drops occur. For *long* diodes whose dimensions are larger than the diffusion lengths, there is no diffusion capacitance. For a *short* diode, the reclaimable charge is two-thirds of the total stored charge. The diffusion capacitance due to stored holes on the n side of length c is:

$$C_s = \frac{dQ_p}{dV} = \frac{1}{3}\frac{q^2}{kT} Acp_n e^{qV/kT} \tag{5–64}$$

There is a similar contribution from the stored electrons on the p side (Fig. 5–31). In practice, most Si p-n junctions behave like short diodes, while laser diodes made in direct band gap (short lifetime) semiconductors often correspond to the long-diode case.

Similarly, we can determine the *a-c conductance* by allowing small changes in the current. For example, for a long diode, we get

$$G_s = \frac{dI}{dV} = \frac{qAL_p p_n}{\tau_p}\frac{d}{dV}(e^{qV/kT}) = \frac{q}{kT} I \tag{5–65}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 216

EXAMPLE 5–6

For the diode of Example 5–4, what is the total depletion capacitance at –4 V?

SOLUTION

$$C_j = \sqrt{\epsilon A}\left[\frac{q}{2(V_0 - V)}\frac{N_d N_a}{N_d + N_a}\right]^{1/2}$$

$$= \sqrt{(8.85 \times 10^{-14} \times 11.8)(10^{-4})}\left[\frac{1.6 \times 10^{-19}}{2(0.695 + 4)}\left(\frac{10^{15} \times 10^{17}}{10^{15} + 10^{17}}\right)\right]^{1/2}$$

$$= \mathbf{4.198 \times 10^{-13}\ F}$$

5.5.5 The Varactor Diode

The term *varactor* is a shortened form of *var*iable re*actor*, referring to the voltage-variable capacitance of a reverse-biased p-n junction. The equations derived in Section 5.5.4 indicate that junction capacitance depends on the applied voltage and the design of the junction. In some cases a junction with fixed reverse bias may be used as a capacitance of a set value. More commonly the varactor diode is designed to exploit the voltage-variable properties of the junction capacitance. For example, a varactor (or a set of varactors) may be used in the tuning stage of a radio receiver to replace the bulky variable plate capacitor. The size of the resulting circuit can be greatly reduced, and its dependability is improved. Other applications of varactors include use in harmonic generation, microwave frequency multiplication, and active filters.

If the p-n junction is abrupt, the capacitance varies as the square root of the reverse bias V_r [Eq. (5-61)]. In a graded junction, however, the capacitance can usually be written in the form

$$C_j \propto V_r^{-n} \quad \text{for } V_r \gg V_0 \tag{5-66}$$

For example, in a linearly graded junction the exponent **n** is one-third (Prob. 5.37). Thus the voltage sensitivity of C_j is greater for an abrupt junction than for a linearly graded junction. For this reason, varactor diodes are often made by epitaxial growth techniques, or by ion implantation. The epitaxial layer and the substrate doping profile can be designed to obtain junctions for which the exponent **n** in Eq. (5-66) is greater than one-half. Such junctions are called *hyperabrupt junctions.*

In the set of doping profiles shown in Fig. 5–32, the junction is assumed p^+-n so that the depletion layer width W extends primarily into the n side. Three types of doping profiles on the n side are illustrated, with the donor distribution $N_d(x)$ given by $Gx^{\mathbf{m}}$, where G is a constant and the exponent **m** is 0, 1, or $-\frac{3}{2}$. We can show (Prob. 5.37) that the exponent **n** in Eq. (5-66) is $1/(\mathbf{m} + 2)$ for the p^+-n junction. Thus for the profiles of Fig. 5–32, **n** is $\frac{1}{2}$ for the abrupt junction and $\frac{1}{3}$ for the linearly graded junction. The hyperabrupt

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 217

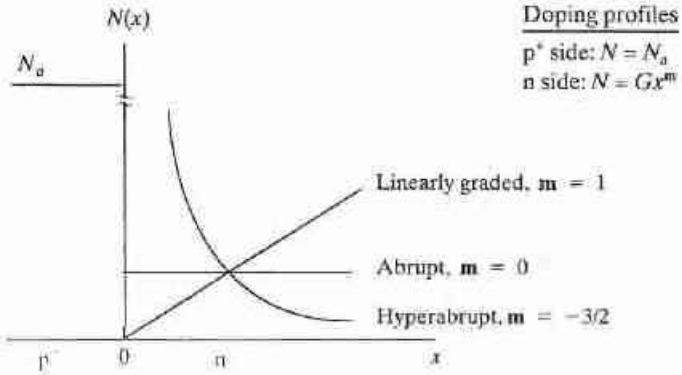


Figure 5–32
Graded junction profiles: linearly graded, abrupt, hyperabrupt.

junction[16] with $\mathbf{m} = -\frac{3}{2}$ is particularly interesting for certain varactor applications, since for this case $\mathbf{n} = 2$ and the capacitance is proportional to V_r^{-2}. When such a capacitor is used with an inductor L in a resonant circuit, the resonant frequency varies linearly with the voltage applied to the varactor.

$$\omega_r = \frac{1}{\sqrt{LC}} \propto \frac{1}{\sqrt{V_r^{-n}}} \propto V_r, \quad \text{for } \mathbf{n} = 2 \tag{5-67}$$

Because of the wide variety of C_j vs. V_r dependencies available by choosing doping profiles, varactor diodes can be designed for specific applications. For some high-frequency applications, varactors can be designed to exploit the forward-bias charge storage capacitance in short diodes.

5.6 DEVIATIONS FROM THE SIMPLE THEORY

The approach we have taken in studying p-n junctions has focused on the basic principles of operation, neglecting secondary effects. This allows for a relatively uncluttered view of carrier injection and other junction properties, and illuminates the essential features of diode operation. To complete the description, however, we must now fill in a few details which can affect the operation of junction devices under special circumstances.

Most of the deviations from the simple theory can be treated by fairly straightforward modifications of the basic equations. In this section we shall investigate the most important deviations and alter the theory wherever possible. In a few cases, we shall simply indicate the approach to be taken and the result. The most important alterations to the simple diode theory are the effects of contact potential and changes in majority carrier concentration on carrier injection, recombination and generation within the transition region, ohmic effects, and the effects of graded junctions.

[16] It is clear that $N_a(x)$ cannot become arbitrarily large at $x = 0$. However, the $\mathbf{m} = -\frac{3}{2}$ profile can be approximated a short distance away from the junction.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 218

5.6.1 Effects of Contact Potential on Carrier Injection

If the forward-bias I–V characteristics of various semiconductor diodes are compared, it becomes clear that the band gap has an important influence on carrier injection. For example, Fig. 5-33 compares the low-temperature characteristics of heavily doped diodes having various band gaps. One obvious feature of this figure is that the I–V characteristics appear "square"; that is, the current is very small until a critical forward bias is reached, and then the current increases rapidly. This is typical of exponentials plotted on such a scale. However, it is significant that the limiting voltage is slightly less than the value of the band gap in electron volts.

The reason for the small current at low voltages for these devices can be understood from a simple rearrangement of the diode equation. If we rewrite Eq. (5-36) for a forward-biased p$^+$-n diode (with $V \gg kT/q$) and include the exponential form for the minority carrier concentration p_n, we obtain

$$I = \frac{qAD_p}{L_p} p_n e^{qV/kT} = \frac{qAD_p}{L_p} N_v e^{[qV-(E_{Fn}-E_{vn})]/kT} \tag{5-68}$$

Hole injection into the n material is small if the forward bias V is much less than $(E_{Fn} - E_{vn})/q$. For a p$^+$-n diode, this quantity is essentially the contact potential, since the Fermi level is near the valence band on the p side. If the n region is also heavily doped, the contact potential is almost equal to the band gap (Fig. 5-34). This accounts for the dramatic increase in diode current near the band gap voltage in Fig. 5-33. Contributing to the small current at lower voltages is the fact that the minority carrier concentration $p_n = n_i^2/N_d$ is very small at low temperature (n_i small) and with heavy doping (N_d large).

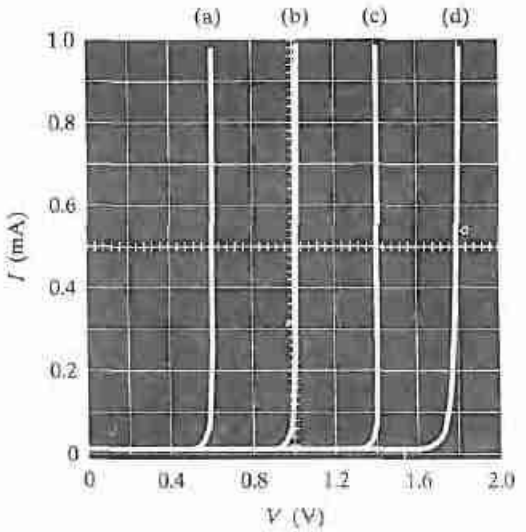


Figure 5-33
I–V characteristics of heavily doped p-n junction diodes at 77 K, illustrating the effects of contact potential on the forward current: (a) Ge, $E_g = 0.7$ eV; (b) Si, $E_g = 1.1$ eV; (c) GaAs, $E_g = 1.4$ eV; (d) GaAsP, $E_g = 1.9$ eV.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 219

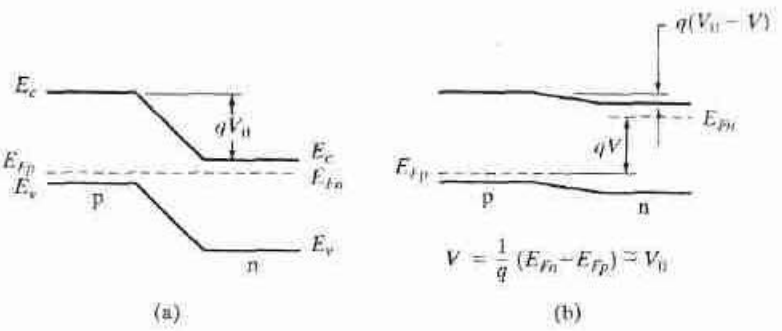


Figure 5–34 Examples of contact potential for a heavily doped p-n junction: (a) at equilibrium; (b) approaching the maximum forward bias $V = V_0$.

The limiting forward bias across a p-n junction is equal to the contact potential, as in Fig. 5–34b. This effect is not predicted by the simple diode equation, for which the current increases exponentially with applied voltage. The reason this important result is excluded in the simple theory is that in Eq. (5-28) we neglect changes in the majority carrier concentrations on either side of the junction. This assumption is valid only for low injection levels; for large injected carrier concentrations, the excess majority carriers become important compared with the majority doping. For example, at low injection $\Delta n_p = \Delta p_p$ is important compared with the equilibrium minority electron concentration n_p, but is negligible compared with the majority hole concentration p_p; this was the basis for neglecting Δp_p in Eq. (5-28). For high injection levels, however, Δp_p can be comparable to p_p and we must write Eq. (5-27) in the form

$$\frac{p(-x_{p0})}{p(x_{n0})} = \frac{p_p + \Delta p_p}{p_n + \Delta p_n} = e^{q(V_0 - V)/kT} = \frac{n_n + \Delta n_n}{n_p + \Delta n_p} \tag{5-69}$$

From Eq. (5-38), we get at either edge of the depletion region,

$$pn = p(-x_{p0})n(-x_{p0}) = p(x_{n0})n(x_{n0}) = n_i^2 e^{\frac{F_n - F_p}{kT}} = n_i^2 e^{qV/kT} \tag{5-70}$$

For example at $-x_{po}$ we then get

$$(p_p + \Delta p_p)(n_p + \Delta n_p) = n_i^2 e^{qV/kT} \tag{5-71}$$

Keeping in mind that $\Delta p_p = \Delta n_p, n_p \ll \Delta n_p$, and in high level injection $p_p < \Delta p_p$, we approximately get

$$\Delta n_p = n_i e^{qV/2kT} \tag{5-72}$$

The rest of the derivation is very similar to that in Section 5.3.2. Hence, the diode current in high-level injection scales as

$$I \propto e^{qV/2kT} \tag{5-73}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 220

5.6.2 Recombination and Generation in the Transition Region

In analyzing the p-n junction, we have assumed that recombination and thermal generation of carriers occur primarily in the neutral p and n regions, outside the transition region. In this model, forward current in the diode is carried by recombination of excess minority carriers injected into each neutral region by the junction. Similarly, the reverse saturation current is due to the thermal generation of EHPs in the neutral regions and the subsequent diffusion of the generated minority carriers to the transition region, where they are swept to the other side by the field. In many devices this model is adequate; however, a more complete description of junction operation should include recombination and generation within the transition region itself.

When a junction is forward biased, the transition region contains excess carriers of both types, which are in transit from one side of the junction to the other. Unless the width of the transition region W is very small compared with the carrier diffusion lengths L_n and L_p, significant recombination can take place within W. An accurate calculation of this recombination current is complicated by the fact that the recombination rate, which depends on the carrier concentrations [Eq. (4–5)], varies with position within the transition region. Analysis of the recombination kinetics shows that the current due to recombination within W is proportional to n_i and increases with forward bias according to approximately $\exp(qV/2kT)$. On the other hand, current due to recombination in the neutral regions is proportional to p_n and n_p [Eq. (5-36)] and therefore to n_i^2/N_d and n_i^2/N_a, and increases according to $\exp(qV/kT)$. The diode equation can be modified to include this effect by including the parameter **n**:

$$\boxed{I = I_0'(e^{qV/\mathbf{n}kT} - 1)} \tag{5-74}$$

where **n** varies between 1 and 2, depending on the material and temperature. Since **n** determines the departure from the ideal diode characteristic, it is often called the *ideality factor*.

The ratio of the two currents

$$\frac{I(\text{recombination in neutral regions})}{I(\text{recombination in transition region})} \propto \frac{n_i^2 e^{qV/kT}}{n_i e^{qV/2kT}} \propto n_i e^{qV/2kT} \tag{5-75}$$

becomes small for wide band gap materials, low temperatures (small n_i), and for low voltage. Thus the forward current for low injection in a Si diode is likely to be dominated by recombination in the transition region, while a Ge diode may follow the usual diode equation. In either case, injection through W into the neutral regions becomes more important with increased voltage. Therefore, **n** in Eq. (5-74) may vary from ~2 at low voltage to ~1 at higher voltage.

Just as recombination within W can affect the forward characteristics, the reverse current through a junction can be influenced by carrier *generation* in the transition region. We found in Section 5.3.3 that the reverse saturation

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 221

current can be accounted for by the thermal generation of EHPs within a diffusion length of either side of the transition region. The generated minority carriers diffuse to the transition region, where they are swept to the other side of the junction by the electric field (Fig. 5–35). However, carrier generation can take place within the transition region itself. If W is small compared with L_n or L_p, band-to-band generation of EHPs within the transition region is not important compared with generation in the neutral regions. However, the lack of free carriers within the space charge of the transition region can create a current due to the net generation of carriers by *emission from recombination centers*. Of the four generation-recombination processes depicted in Fig. 5–36, the two capture rates R_n and R_p are negligible within W because of the very small carrier concentrations in the reverse-bias space charge region. Therefore, a recombination level E_r near the center of the band gap can provide carriers through the thermal generation rates G_n and G_p. Each recombination center alternately emits an electron and a hole; physically, this means that an electron at E_r is thermally excited to the conduction band (G_n) and a valence band electron is subsequently excited thermally to the empty state on the recombination level, leaving a hole behind in the valence band (G_p). The process can then be repeated over and over, providing electrons for the conduction band and holes for the valence band. Normally, these emission processes are exactly balanced by the corresponding capture processes R_n and R_p. However, in the reverse-bias transition region, generated carriers are swept out before recombination can occur, and net generation results.

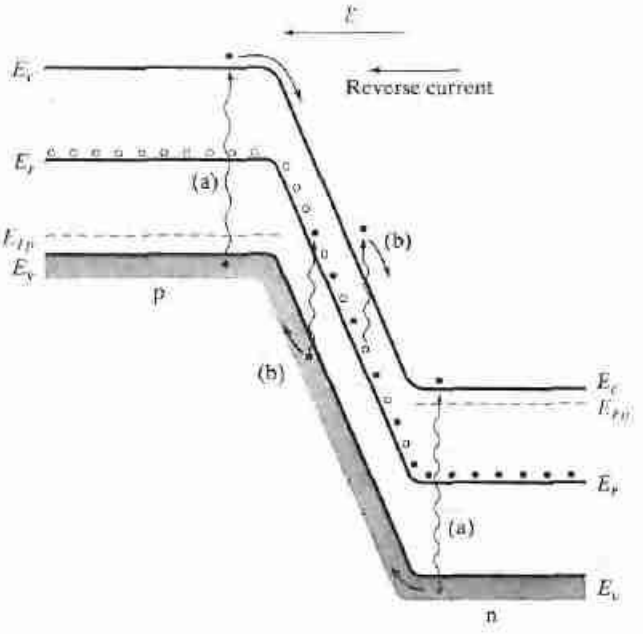


Figure 5–35
Current in a reverse-biased p-n junction due to thermal generation of carriers by (a) band-to-band EHP generation, and (b) generation from a recombination level.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 222

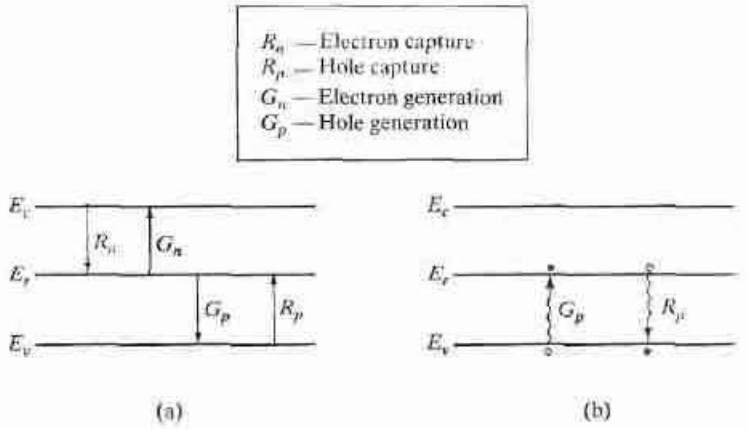


Figure 5–36
Capture and generation of carriers at a recombination center: (a) capture and generation of electrons and holes; (b) hole capture and generation processes redrawn in terms of valence band electron excitation to E_t (hole generation) and electron deexcitation from E_t to E_v (hole capture by E_t).

Of course, the importance of thermal generation within W depends on the temperature and the nature of the recombination centers. A level near the middle of the band gap is most effective, since for such centers neither G_n nor G_p requires thermal excitation of an electron over more than about half the band gap. If no recombination level is available, this type of generation is negligible. However, in most materials recombination centers exist near the middle of the gap due to trace impurities or lattice defects. Generation from centers within W is most important in materials with large band gaps, for which band-to-band generation in the neutral regions is small. Thus for Si, generation within W is generally more important than for a narrower band gap material such as Ge.

The saturation current due to generation in the neutral regions was found to be essentially independent of reverse bias. However, generation within W naturally increases as W increases with reverse bias. As a result, the reverse current can increase almost linearly with W, or with the square root of reverse-bias voltage.

5.6.3 Ohmic Losses

In deriving the diode equation we assumed that the voltage applied to the device appears entirely across the junction. Thus we neglected any voltage drop in the neutral regions or at the external contacts. For most devices this is a valid assumption; the doping is usually fairly high, so that the resistivity of each neutral region is low, and the area of a typical diode is large compared with its length. However, some devices do exhibit ohmic effects, which cause significant deviation from the expected I–V characteristic.

We can seldom represent ohmic losses in a diode accurately by including a simple resistance in series with the junction. The effects of voltage drops outside the transition region are complicated by the fact that the voltage drop depends on the current, which in turn is dictated by the voltage across the junction.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 223

For example, if we represent the series resistance of the p and n regions by R_p and R_n, respectively, we can write the junction voltage V as

$$V = V_a - I[R_p(I) + R_n(I)] \tag{5-76}$$

where V_a is the external voltage applied to the device. As the current increases, there is an increasing voltage drop in R_p and R_n, and the junction voltage V decreases. This reduction in V lowers the level of injection so that the current increases more slowly with increased bias. A further complication in calculating the ohmic loss is that the conductivity of each neutral region increases with increasing carrier injection. Since the effects of Eq. (5-76) are most pronounced at high injection levels, this *conductivity modulation* by the injected excess carriers can reduce R_p and R_n significantly.

Ohmic losses are purposely avoided in properly designed devices by appropriate choices of doping and geometry. Therefore, deviations of the current generally appear only for very high currents, outside the normal operating range of the device.

Figure 5-37 shows the forward and reverse current-voltage characteristics of a p-n junction on a semilog scale, both for an ideal Shockley diode as well as for non-ideal devices. For an ideal forward-biased diode, we get a straight line on a semilog plot reflecting the exponential dependence of current on voltage. On the other hand, taking into account all the second order effects discussed in Section 5.6, we see various regions of operation. At low current levels, we see the enhanced generation–recombination current, leading to a higher diode ideality factor ($\mathbf{n} = 2$). For moderate currents, we get ideal low-level injection and diffusion-limited current ($\mathbf{n} = 1$). At higher currents, we get high-level injection and $\mathbf{n} = 2$, while at even higher currents, the ohmic drops in the space charge neutral regions become important.

Similarly, in reverse bias, in an ideal diode, we have a constant, voltage-independent reverse saturation current. However, in actuality, we get an enhanced, voltage-dependent generation-recombination leakage current. At very high reverse biases, the diode breaks down reversibly due to avalanche or Zener effects.

5.6.4 Graded Junctions

While the abrupt junction approximation accurately describes the properties of many epitaxially grown junctions, it is often inadequate in analyzing diffused or implanted junction devices. For shallow diffusions, in which the diffused impurity profile is very steep (Fig. 5-38a), the abrupt approximation is usually acceptable. If the impurity profile is spread out into the sample, however, a graded junction can result (Fig. 5-38b). Several of the expressions we have derived for the abrupt junction must be modified for this case (see Section 5.5.5).

The graded junction problem can be solved analytically if, for example, we make a linear approximation of the net impurity distribution near the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 224

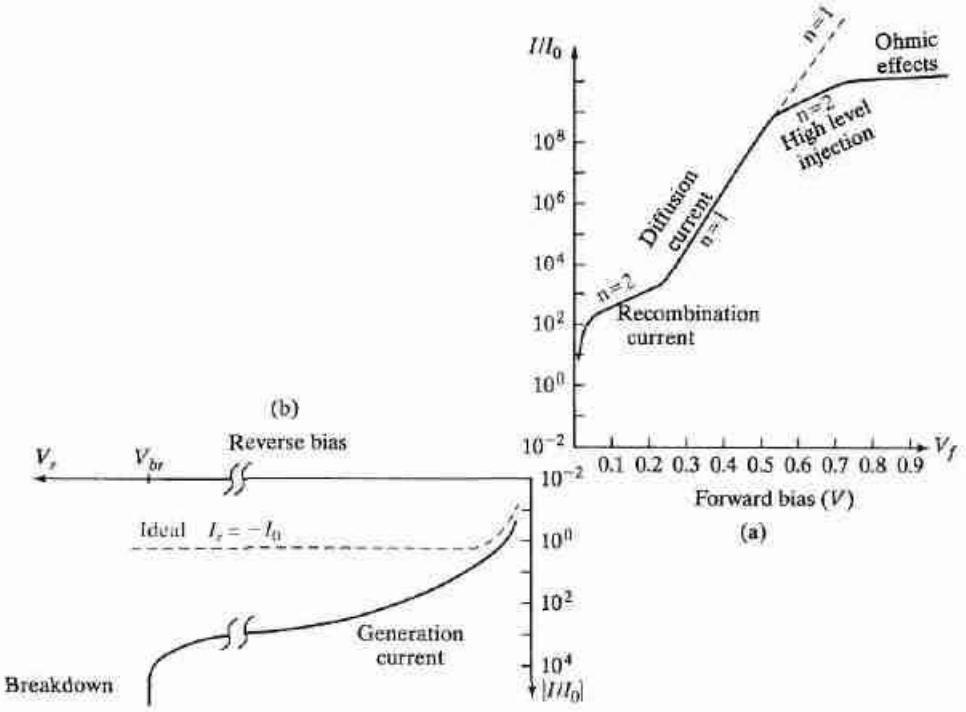


Figure 5–37
Forward and reverse current–voltage characteristics plotted on semilog scales, with current normalized with respect to saturation current, I_0; (a) the ideal forward characteristic is an exponential with an ideality factor **n** = 1 (dashed straight line on log-linear plot). The actual forward characteristics of a typical diode (solid line) have four regimes of operation; (b) ideal reverse characteristic (dashed line) is a voltage-independent current = $-I_0$. Actual leakage characteristics (solid line) are higher due to generation in the depletion region, and also show breakdown at high voltages.

junction (Fig. 5-38c). We assume that the graded region can be described approximately by

$$N_d - N_a = Gx \tag{5-77}$$

where G is a grade constant giving the slope of the net impurity distribution.

In Poisson's equation [Eq. (5-14)], the linear approximation becomes

$$\frac{d\mathscr{E}}{dx} = \frac{q}{\epsilon}(p - n + N_d^+ - N_a^-) \simeq \frac{q}{\epsilon}Gx \tag{5-78}$$

within the transition region. In this approximation we assume complete ionization of the impurities and neglect the carrier concentrations in the transition region, as before. The net space charge varies linearly over W, and the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 225

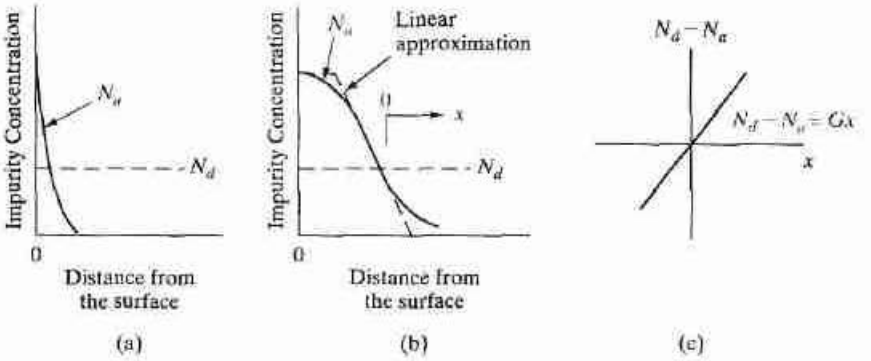


Figure 5–38 Approximations to diffused junctions: (a) shallow diffusion (abrupt); (b) deep drive-in diffusion with source removed (graded); (c) linear approximation to the graded junction.

electric field distribution is therefore parabolic. The expressions for contact potential and junction capacitance are different from the abrupt junction case (Fig. 5–39 and Prob. 5.37), since the electric field is no longer linear on each side of the junction.

In a graded junction the usual depletion approximation is often inaccurate. If the grade constant G is small, the carrier concentrations $(p - n)$ can be important in Eq. (5–78). Similarly, the usual assumption of negligible space charge outside the transition region is questionable for small G. It would be more accurate to refer to the regions just outside the transition region as quasi-neutral rather than neutral. Thus the edges of the transition region are not sharp as Fig. 5–39 implies but are spread out in x. These effects complicate

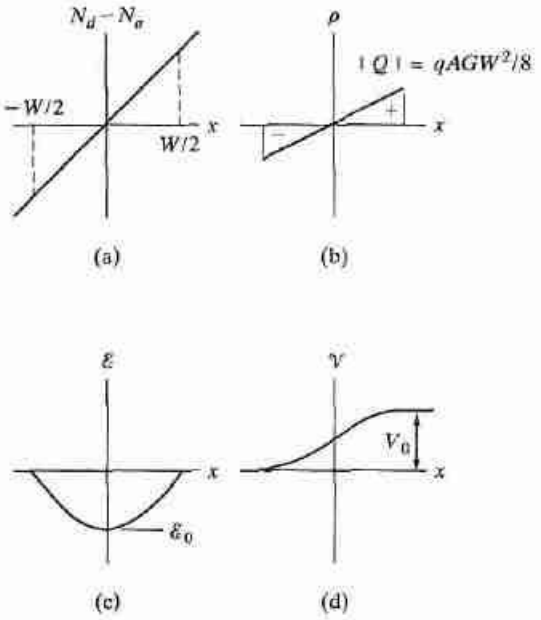


Figure 5–39 Properties of the graded junction transition region: (a) net impurity profile; (b) net charge distribution; (c) electric field; (d) electrostatic potential.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 226

calculations of junction properties, and a computer must be used in solving the problem accurately.

Most of the conclusions we have made regarding carrier injection, recombination and generation currents, and other properties are qualitatively applicable to graded junctions, with some alterations in the functional form of the resulting equations. Therefore, we can apply most of our basic concepts of junction theory to reasonably graded junctions as long as we remember that certain modifications should be made in accurate computations.

5.7 METAL–SEMICONDUCTOR JUNCTIONS

Many of the useful properties of a p-n junction can be achieved by simply forming an appropriate metal-semiconductor contact. This approach is obviously attractive because of its simplicity of fabrication; also, as we shall see in this section, metal-semiconductor junctions are particularly useful when high-speed rectification is required. On the other hand, we also must be able to form nonrectifying (ohmic) contacts to semiconductors. Therefore, this section deals with both rectifying and ohmic contacts.

5.7.1 Schottky Barriers

In Section 2.2.1 we discussed the work function $q\Phi_m$ of a metal in a vacuum. An energy of $q\Phi_m$ is required to remove an electron at the Fermi level to the vacuum outside the metal. Typical values of Φ_m for very clean surfaces are 4.3 V for Al and 4.8 V for Au. When negative charges are brought near the metal surface, positive (image) charges are induced in the metal. When this image force is combined with an applied electric field, the effective work function is somewhat reduced. Such barrier lowering is called the *Schottky effect*, and this terminology is carried over to the discussion of potential barriers arising in metal-semiconductor contacts. Although the Schottky effect is only a part of the explanation of metal-semiconductor contacts, rectifying contacts are generally referred to as *Schottky barrier diodes*. In this section we shall see how such barriers arise in metal-semiconductor contacts. First we consider barriers in ideal metal-semiconductor junctions, and then in Section 5.7.4 we will include effects which alter the barrier height.

When a metal with work function $q\Phi_m$ is brought in contact with a semiconductor having a work function $q\Phi_s$, charge transfer occurs until the Fermi levels align at equilibrium (Fig. 5-40). For example, when $\Phi_m > \Phi_s$, the semiconductor Fermi level is initially higher than that of the metal before contact is made. To align the two Fermi levels, the electrostatic potential of the semiconductor must be raised (i.e., the electron energies must be lowered) relative to that of the metal. In the n-type semiconductor of Fig. 5-40 a depletion region W is formed near the junction. The positive charge due to uncompensated donor ions within W matches the negative charge on the metal. The electric field and the bending of the bands within W are similar to effects already discussed for p-n junctions. For example, the depletion width W in the semiconductor can be

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 227

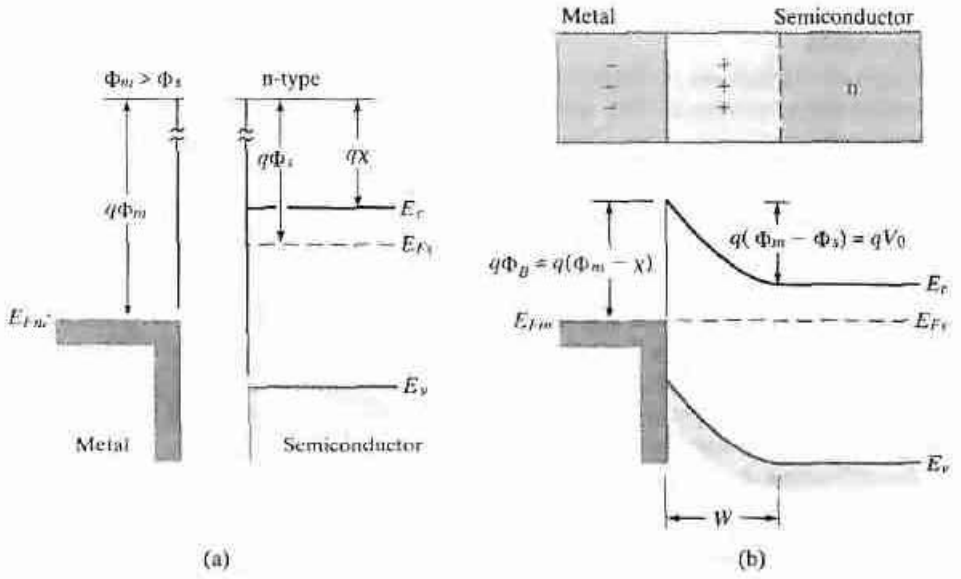


Figure 5–40
A Schottky barrier formed by contacting an n-type semiconductor with a metal having a larger work function: (a) band diagrams for the metal and the semiconductor before joining; (b) equilibrium band diagram for the junction.

calculated from Eq. (5-21) by using the p^+-n approximation (i.e., by assuming the negative charge in the dipole is a thin sheet of charge to the left of the junction). Similarly, the junction capacitance is $A\epsilon_s/W$, as in the p^+-n junction.[17]

The equilibrium contact potential V_0, which prevents further net electron diffusion from the semiconductor conduction band into the metal, is the difference in work function potentials $\Phi_m - \Phi_s$. The potential barrier height Φ_B for electron injection from the metal into the semiconductor conduction band is $\Phi_m - \chi$, where $q\chi$ (called the *electron affinity*) is measured from the vacuum level to the semiconductor conduction band edge. The equilibrium potential difference V_0 can be decreased or increased by the application of either forward- or reverse-bias voltage, as in the p-n junction.

Figure 5-41 illustrates a Schottky barrier on a p-type semiconductor, with $\Phi_m < \Phi_s$. In this case aligning the Fermi levels at equilibrium requires a positive charge on the metal side and a negative charge on the semiconductor side of the junction. The negative charge is accommodated by a depletion region W in which ionized acceptors (N_a^-) are left uncompensated by holes. The potential barrier V_0 retarding hole diffusion from the semiconductor to the metal is $\Phi_s - \Phi_m$, and as before this barrier can be raised or lowered

[17]While the properties of the Schottky barrier depletion region are similar to those of the p^+-n region, it is clear that the analogy does not include forward-bias hole injection, which is dominant for the p^+-n region, but not for the contact of Fig. 5-40.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 228

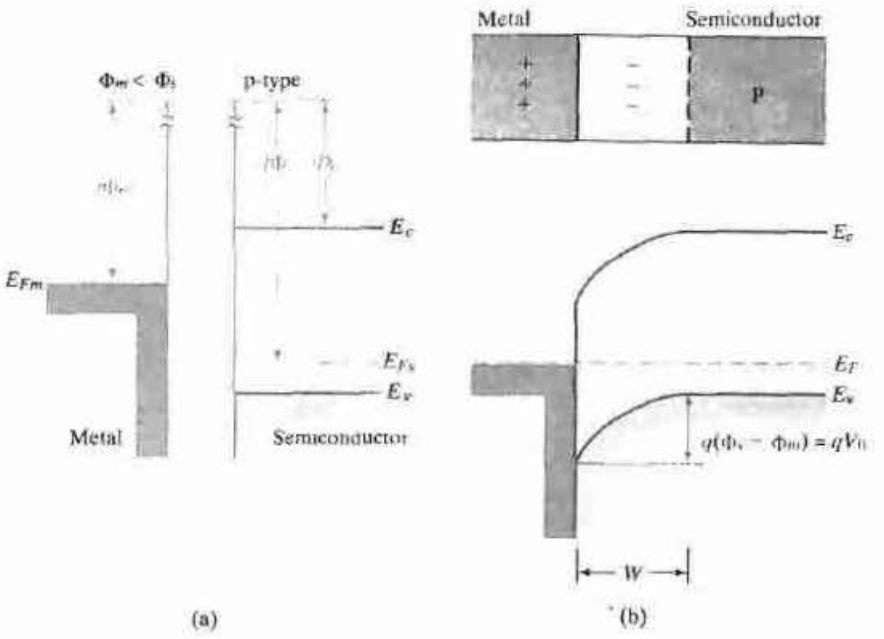


Figure 5–41
Schottky barrier between a p-type semiconductor and a metal having a smaller work function: (a) band diagrams before joining; (b) band diagram for the junction at equilibrium.

by the application of voltage across the junction. In visualizing the barrier for holes, we recall from Fig. 5–11 that the electrostatic potential barrier for positive charge is opposite to the barrier on the electron energy diagram.

The two other cases of ideal metal-semiconductor contacts ($\Phi_m < \Phi_s$ for n-type semiconductors, and $\Phi_m > \Phi_s$ for p-type) result in nonrectifying contacts. We will save treatment of these cases for Section 5.7.3, where ohmic contacts are discussed.

5.7.2 Rectifying Contacts

When a forward-bias voltage V is applied to the Schottky barrier of Fig. 5–40b, the contact potential is reduced from V_0 to $V_0 - V$ (Fig. 5–42a). As a result, electrons in the semiconductor conduction band can diffuse across the depletion region to the metal. This gives rise to a forward current (metal to semiconductor) through the junction. Conversely, a reverse bias increases the barrier to $V_0 + V_r$, and electron flow from semiconductor to metal becomes negligible. In either case flow of electrons from the metal to the semiconductor is retarded by the barrier $\Phi_m - \chi$. The resulting diode equation is similar in form to that of the p-n junction

$$I = I_0(e^{qV/kT} - 1) \tag{5–79}$$

as Fig. 5–42c suggests. In this case the reverse saturation current I_0 is not simply derived as it was for the p-n junction. One important feature we can

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 229

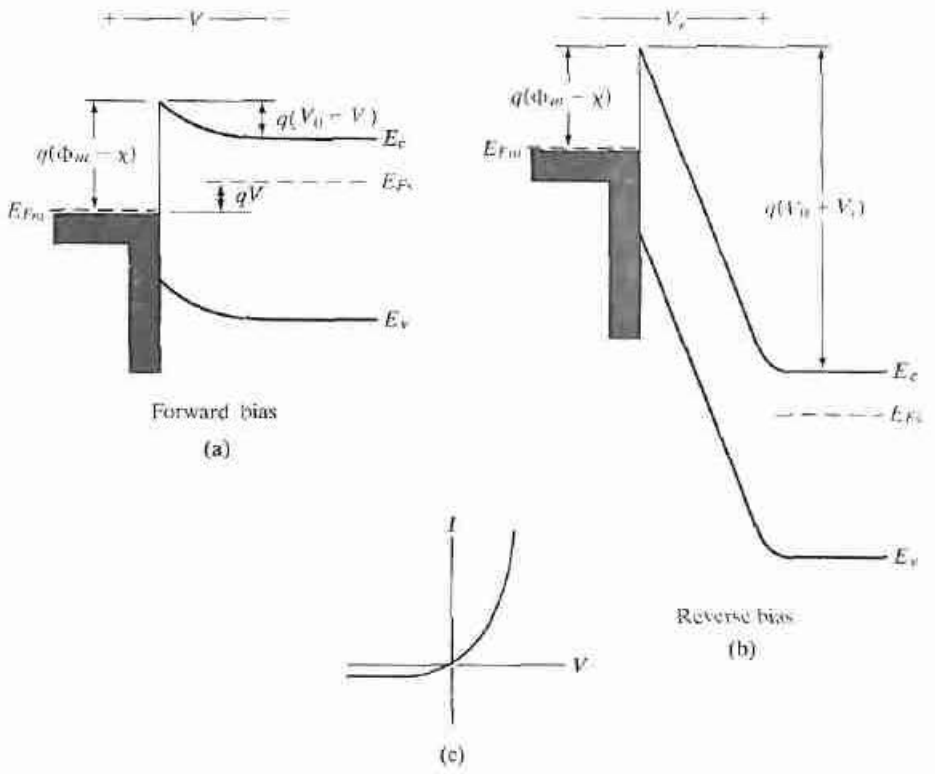


Figure 5–42
Effects of forward and reverse bias on the junction of Fig. 5–40: (a) forward bias; (b) reverse bias; (c) typical current–voltage characteristic.

predict intuitively, however, is that the saturation current should depend upon the size of the barrier Φ_B for electron injection from the metal into the semiconductor. This barrier (which is $\Phi_m - \chi$ for the ideal case shown in Fig. 5–42) is unaffected by the bias voltage. We expect the probability of an electron in the metal surmounting this barrier to be given by a Boltzmann factor. Thus

$$I_0 \propto e^{-q\Phi_B/kT} \tag{5-80}$$

The diode equation (5–79) applies also to the metal-p-type semiconductor junction of Fig. 5–41. In this case forward voltage is defined with the semiconductor biased positively with respect to the metal. Forward current increases as this voltage lowers the potential barrier to $V_0 - V$ and holes flow from the semiconductor to the metal. Of course, a reverse voltage increases the barrier for hole flow and the current becomes negligible.

In both of these cases the Schottky barrier diode is rectifying, with easy current flow in the forward direction and little current in the reverse direction. We also note that the forward current in each case is due to the injection of

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 230

majority carriers from the semiconductor into the metal. The absence of minority carrier injection and the associated storage delay time is an important feature of Schottky barrier diodes. Although some minority carrier injection occurs at high current levels, these are essentially majority carrier devices. Their high-frequency properties and switching speed are therefore generally better than typical p-n junctions.

In the early days of semiconductor technology, rectifying contacts were made simply by pressing a wire against the surface of the semiconductor. In modern devices, however, the metal-semiconductor contact is made by depositing an appropriate metal film on a clean semiconductor surface and defining the contact pattern photolithographically. Schottky barrier devices are particularly well suited for use in densely packed integrated circuits, because fewer photolithographic masking steps are required compared to p-n junction devices.

5.7.3 Ohmic Contacts

In many cases we wish to have an *ohmic* metal-semiconductor contact, having a linear I-V characteristic in both biasing directions. For example, the surface of a typical integrated circuit is a maze of p and n regions, which must be contacted and interconnected. It is important that such contacts be ohmic, with minimal resistance and no tendency to rectify signals.

Ideal metal-semiconductor contacts are ohmic when the charge induced in the semiconductor in aligning the Fermi levels is provided by majority carriers (Fig. 5-43). For example, in the $\Phi_m < \Phi_s$ (n-type) case of Fig. 5-43a, the Fermi levels are aligned at equilibrium by transferring electrons from the metal to the semiconductor. This raises the semiconductor electron energies (lowers the electrostatic potential) relative to the metal at equilibrium (Fig. 5-43b). In this case the barrier to electron flow between the metal and the semiconductor is small and easily overcome by a small voltage. Similarly, the case $\Phi_m > \Phi_s$ (p-type) results in easy hole flow across the junction (Fig. 5-43d). Unlike the rectifying contacts discussed previously, no depletion region occurs in the semiconductor in these cases since the electrostatic potential difference required to align the Fermi levels at equilibrium calls for accumulation of majority carriers in the semiconductor.

A practical method for forming ohmic contacts is by doping the semiconductor heavily in the contact region. Thus if a barrier exists at the interface, the depletion width is small enough to allow carriers to tunnel through the barrier. For example, Au containing a small percentage of Sb can be alloyed to n-type Si, forming an n^+ layer at the semiconductor surface and an excellent ohmic contact. Similarly, p-type material requires a p^+ surface layer in contact with the metal. In the case of Al on p-type Si, the metal contact also provides the acceptor dopant. Thus the required p^+ surface layer is formed during a brief heat treatment of the contact after the Al is deposited.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 231

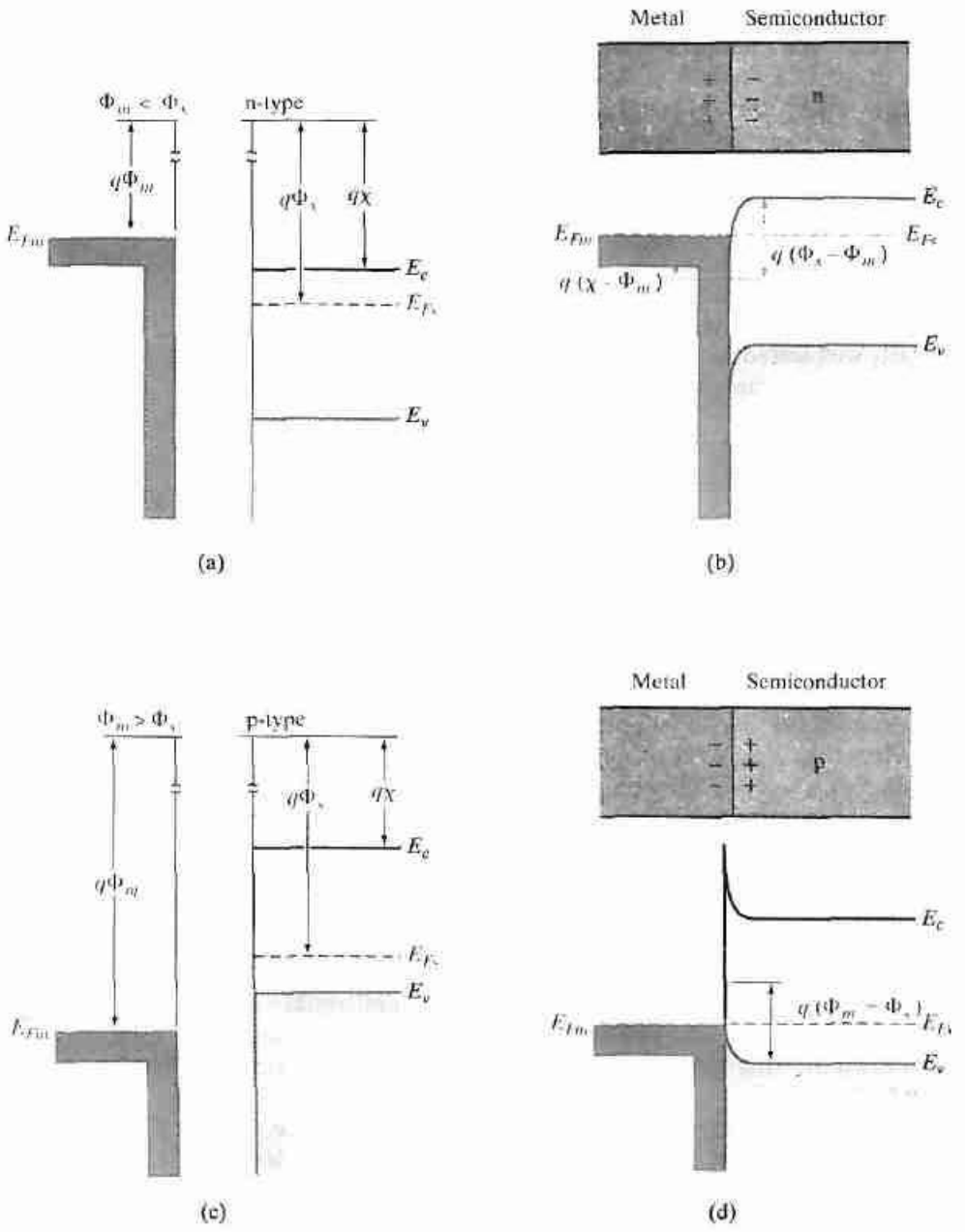


Figure 5–43
Ohmic metal–semiconductor contacts: (a) $\Phi_m < \Phi_s$ for an n-type semiconductor, and (b) the equilibrium band diagram for the junction; (c) $\Phi_m > \Phi_s$ for a p-type semiconductor, and (d) the junction at equilibrium.

5.7.4 Typical Schottky Barriers

The discussion of ideal metal-semiconductor contacts does not include certain effects of the junction between the two dissimilar materials. Unlike a p-n junction, which occurs within a single crystal, a Schottky barrier junction includes a termination of the semiconductor crystal. The semiconductor surface contains *surface states* due to incomplete covalent bonds and other effects, which can lead to charges at the metal-semiconductor interface. Furthermore, the contact is seldom an atomically sharp discontinuity between the semiconductor crystal and the metal. There is typically a thin interfacial layer, which is neither semiconductor nor metal. For example, silicon crystals are covered by a thin (10–20 Å) oxide layer even after etching or cleaving in atmospheric conditions. Therefore, deposition of a metal on such a Si surface leaves a glassy interfacial layer at the junction. Although electrons can tunnel through this thin layer, it does affect the barrier to current transport through the junction.

Because of surface states, the interfacial layer, microscopic clusters of metal–semiconductor phases, and other effects, it is difficult to fabricate junctions with barriers near the ideal values predicted from the work functions of the two isolated materials. Therefore, measured barrier heights are used in device design. In compound semiconductors the interfacial layer introduces states in the semiconductor band gap that pin the Fermi level at a fixed position, regardless of the metal used (Fig. 5–44). For example, a collection of interface states located 0.7 ~ 0.9 eV below the conduction band pins E_F at the surface of n-type GaAs, and the Schottky barrier height is determined from this pinning effect rather than by the work function of the metal. An interesting case is n-type InAs (Fig. 5–44b), in which E_F at the interface is pinned *above* the conduction band edge. As a result, ohmic contact to n-type

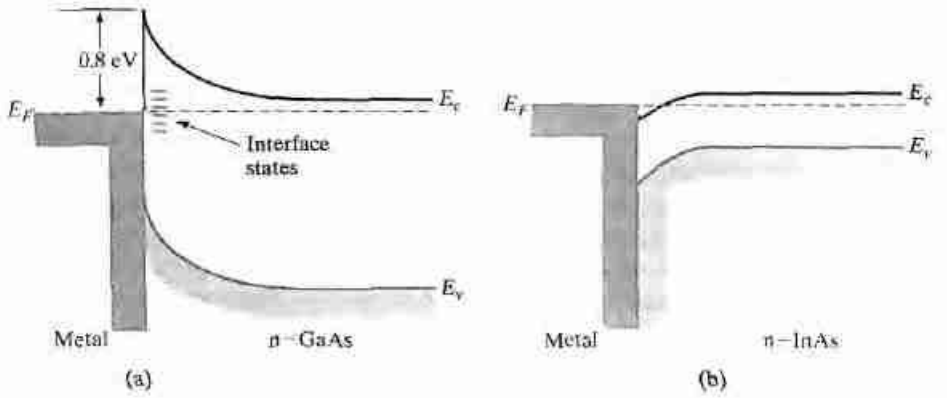


Figure 5–44
Fermi level pinning by interface states in compound semiconductors: (a) E_F is pinned near $E_C - 0.8$ eV in n-type GaAs, regardless of the choice of metal; (b) E_F is pinned above E_C in n-type InAs, providing an excellent ohmic contact.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 233

InAs can be made by depositing virtually any metal on the surface. For Si, good Schottky barriers are formed by various metals, such as Au or Pt. In the case of Pt, heat treatment results in a platinum silicide layer, which provides a reliable Schottky barrier with $\Phi_B \approx 0.85$ V on n-type Si.

A full treatment of Schottky barrier diodes results in a forward current equation of the form

$$I = ABT^2 e^{-q\Phi_B/kT} e^{qV/nkT} \tag{5-81}$$

where B is a constant containing parameters of the junction properties and n is a number between 1 and 2, similar to the ideality factor in Eq. (5-74) but arising from different reasons. The mathematics of this derivation is similar to that of *thermionic emission*, and the factor B corresponds to an effective Richardson constant in the thermionic problem.

5.8 HETERO-JUNCTIONS

Thus far we have discussed p-n junctions formed within a single semiconductor (*homojunctions*) and junctions between a metal and a semiconductor. The third important class of junctions consist of those between two lattice-matched semiconductors with different band gaps (*heterojunctions*). We discussed lattice-matching in Section 1.4.1. The interface between two such semiconductors may be virtually free of defects, and continuous crystals containing single or multiple heterojunctions can be formed. The availability of heterojunctions and multilayer structures in compound semiconductors opens a broad range of possibilities for device development. We will discuss many of these applications in later chapters, including heterojunction bipolar transistors, field-effect transistors, and semiconductor lasers.

When semiconductors of different band gaps, work functions, and electron affinities are brought together to form a junction, we expect discontinuities in the energy bands as the Fermi levels line up at equilibrium (Fig. 5-45). The discontinuities in the conduction band ΔE_c and the valence band ΔE_v accommodate the difference in band gap between the two semiconductors ΔE_g. In an ideal case, ΔE_c would be the difference in electron affinities $q(\chi_2 - \chi_1)$, and ΔE_v would be found from $\Delta E_g - \Delta E_c$. This is known as the Anderson affinity rule. In practice, the band discontinuities are found experimentally for particular semiconductor pairs. For example, in the commonly used system GaAs–AlGaAs (see Figs. 3-6 and 3-13), the direct band gap difference ΔE_g^{Γ} between the wider band gap AlGaAs and the narrower band gap GaAs is apportioned approximately $\frac{2}{3}$ in the conduction band and $\frac{1}{3}$ in the valence band for the heterojunction. The built-in contact potential is divided between the two semiconductors as required to align the Fermi levels at equilibrium. The resulting depletion region on each side of the heterojunction and the amount of built-in potential on each side (making up the contact potential V_0) are found by solving Poisson's equation with the boundary condition of continuous electric flux density, $\epsilon_1 \mathscr{E}_1 = \epsilon_2 \mathscr{E}_2$ at the junction. The barrier that electrons must overcome in moving from the n side to the p side may be quite different from

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 234

the barrier for holes moving from p to n. The depletion region on each side is analogous to that described in Eq. (5-23), except that we must account for the different dielectric constants in the two semiconductors.

To draw the band diagram for any semiconductor device involving homojunctions or heterojunctions, we need material parameters such as the band gap and the electron affinity which depend on the semiconductor material but not on the doping, and the work function which depends on the semiconductor as well as the doping. The electron affinity and work function are referenced to the vacuum level. The true vacuum level (or global vacuum level), E_{vac}, is the potential energy reference when an electron is taken out of the semiconductor to infinity, where it sees no forces. Hence, the true vacuum level is a constant (Fig. 5-45). That introduces an apparent

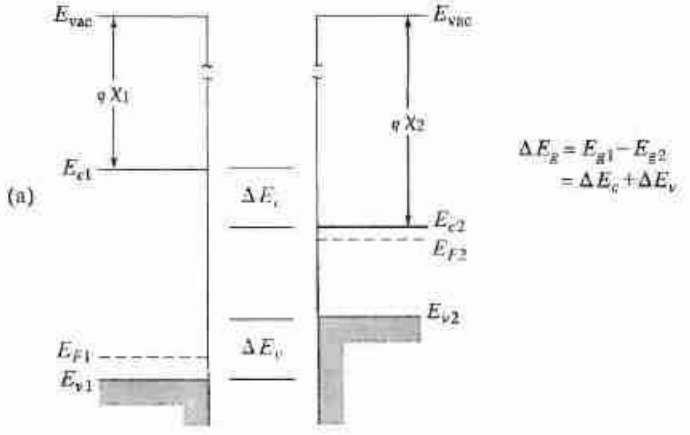
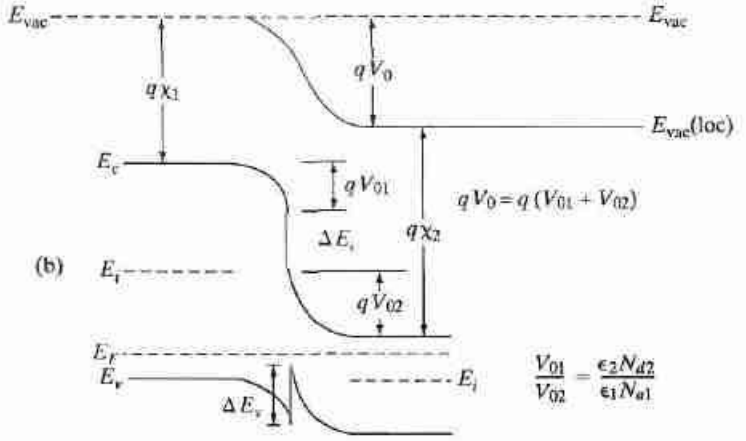


Figure 5–45
An ideal heterojunction between a p-type, wide band gap semiconductor and an n-type narrower band gap semiconductor: (a) band diagrams before joining; (b) band discontinuities and band bending at equilibrium.

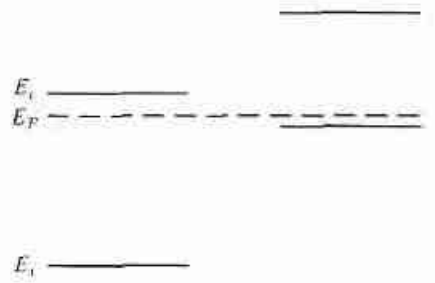


Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 235

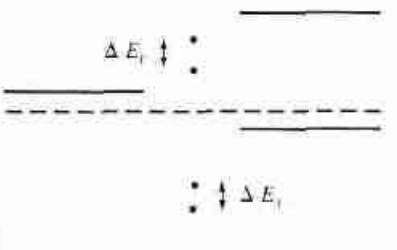
contradiction, however, because looking at the band bending in a semiconductor device, it seems to imply that the electron affinity in the semiconductor changes as a function of position, which is impossible because the electron affinity is a material parameter. Therefore, we need to introduce the new concept of the local vacuum level, E_{vac} (loc), which varies along with and parallel to the conduction band edge, thereby keeping the electron affinity constant. The local vacuum level tracks the potential energy of an electron if it is moved just outside of the semiconductor, but not far away. The difference between the local and global vacuum levels is due to the electrical work done against the fringing electric fields of the depletion region, and is equal to the potential energy qV_0 due to the built-in contact potential V_0 in equilibrium. This potential energy can, of course, be modified by an applied bias.

To draw the band diagram for a heterojunction accurately, we must not only use the proper values for the band discontinuities but also account for the band bending in the junction. To do this, we must solve Poisson's equation across the heterojunction, taking into account the details of doping and space charge, which generally requires a computer solution. We can, however, sketch an approximate diagram without a detailed calculation. Given the experimental band offsets ΔE_v and ΔE_c, we can proceed as follows:

1. Align the Fermi level with the two semiconductor bands separated. Leave space for the transition region.

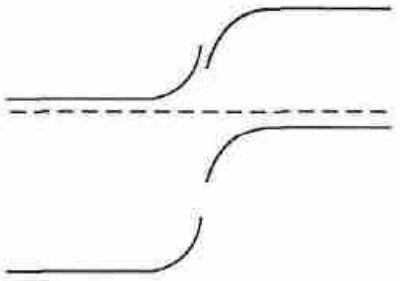


2. The metallurgical junction ($x = 0$) is located near the more heavily doped side. At $x = 0$ put ΔE_v and ΔE_c, separated by the appropriate band gaps.



Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 236

3. Connect the conduction band and valence band regions, keeping the band gap constant in each material.



Steps 2 and 3 of this procedure are where the exact band bending is important and must be obtained by solving Poisson's equation. In step 2 we must use the band offset values ΔE_c and ΔE_v for the specific pair of semiconductors in the heterojunction.

EXAMPLE 5–7

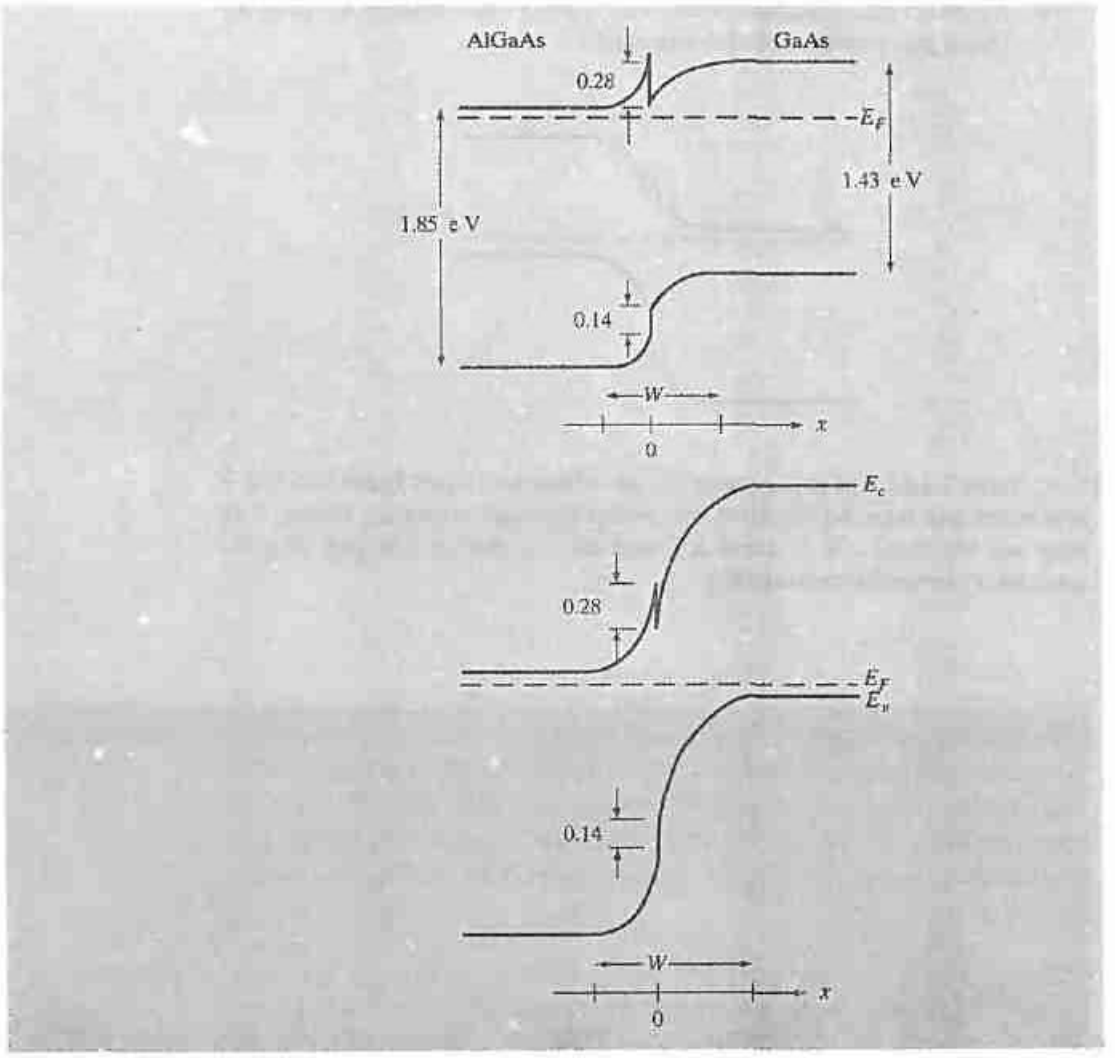
For heterojunctions in the GaAs–AlGaAs system, the direct (Γ) band gap difference ΔE_g^{Γ} is accommodated approximately $\frac{2}{3}$ in the conduction band and $\frac{1}{3}$ in the valence band. For an Al composition of 0.3, the AlGaAs is direct (see Fig. 3–6) with $\Delta E_g^{\Gamma} = 1.85$ eV. Sketch the band diagrams for two heterojunction cases: N^+-$Al_{0.3}Ga_{0.7}As$ on n-type GaAs, and N^+-$Al_{0.3}Ga_{0.7}As$ on p^+-GaAs.[18]

SOLUTION

Taking $\Delta E_g = 1.85 - 1.43 = 0.42$ eV, the band offsets are $\Delta E_c = 0.28$ eV and $\Delta E_v = 0.14$ eV. In each case we draw the equilibrium Fermi level, add the appropriate bands far from the junction, add the band offsets while estimating the relative amounts of band bending and position of $x = 0$ for the particular doping on the two sides, and finally sketch the band edges so that E_g is maintained in each separate semiconductor right up to the heterojunction at $x = 0$.

[18] In discussing heterojunctions, it is common to use a capital N or P to designate the wide band gap material.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 237



A particularly important example of a heterojunction is shown in Figure 5–46, in which heavily n-type AlGaAs is grown on lightly doped GaAs. In this example the discontinuity in the conduction band allows electrons to spill over from the N^+-AlGaAs into the GaAs, where they become trapped in the potential well. As a result, electrons collect on the GaAs side of the heterojunction and move the Fermi level above the conduction band in the GaAs near the interface. These electrons are confined in a narrow potential well in the GaAs conduction band. If we construct a device in which conduction occurs parallel to the interface, the electrons in such a potential well form a *two-dimensional electron gas* with very interesting device properties. As we shall see in Chapter 6, electron conduction in such a potential well can result in very high mobility electrons. This high mobility is due to the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 238

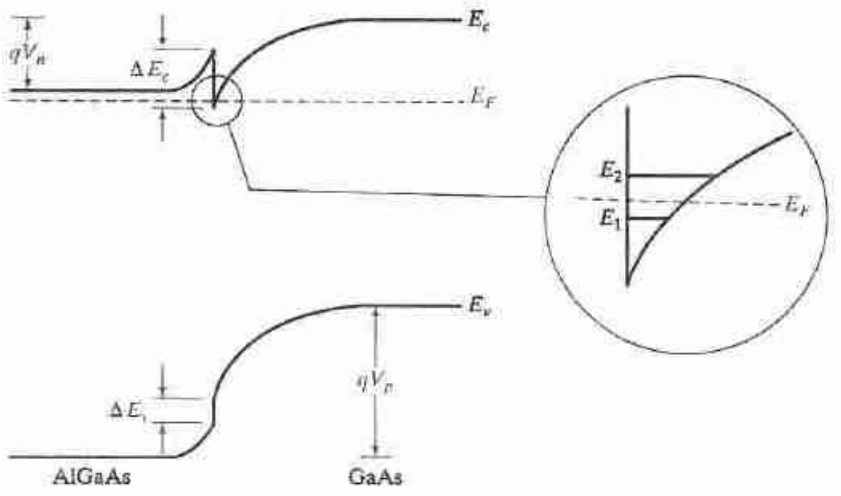


Figure 5–46
A heterojunction between N⁺-AlGaAs and lightly doped GaAs, illustrating the potential well for electrons formed in the GaAs conduction band. If this well is sufficiently thin, discrete states (such as E_1 and E_2) are formed, as discussed in Section 2.4.3.

fact that the electrons in this well come from the AlGaAs, and not from doping in the GaAs. As a result, there is negligible impurity scattering in the GaAs well, and the mobility is controlled almost entirely by lattice scattering (phonons). At low temperatures, where phonon scattering is low, the mobility in this region can be very high. If the band-bending in the GaAs conduction band is strong enough, the potential well may be extremely narrow, so that discrete states such as E_1 and E_2 in Fig. 5–46 are formed. We will return to this example in Chapter 6.

Another obvious feature of Fig. 5–46 is that the concept of a contact potential barrier qV_0 for both electrons and holes in a homojunction is no longer valid for the heterojunction. In Fig. 5–46 the barrier for electrons qV_n is smaller than the barrier for holes qV_p. This property of a heterojunction can be used to alter the relative injection of electrons and holes, as we shall see in Section 7.9.

SUMMARY

5.1 Diodes and other semiconductor devices are made by combinations of steps such as oxidation, selective doping (via implant or diffusion), and the deposition of various insulators or metals in concert with etching, using patterns formed by photolithography.

5.2 When we bring p- and n-type semiconductors into contact to make a p-n junction diode, carriers diffuse across the junction until we get a flat Fermi level in equilibrium. A *built-in junction potential barrier* is formed between the p and n sides, which reflects the voltage drop across the depletion region. This is a *dynamic* equilibrium, in which there is a continual diffusion of electrons from the n to the p side (and holes from p to n), but at a reduced rate over the potential barrier, and these fluxes are cancelled by opposing flows of minority carriers diffusing to the depletion edges and getting swept across the junction.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 239

5.3 The electrostatics of the depletion region are determined by solving the Poisson equation. For uniformly doped *abrupt* step junctions we get *linearly varying electric fields* that are highest at the metallurgical junction. We get a wider depletion region on the more lightly doped side and equal and opposite depletion charges on the two sides of the junction.

5.4 For an *ideal Shockley diode*, one assumes *negligible generation–recombination inside the depletion region*. In forward bias, the built-in potential barrier is lowered, making it exponentially easier for majority carriers, to diffuse across.

5.5 The opposing minority carrier fluxes are unaffected because they are limited by how *often* the few minority carriers diffuse to the depletion edges. Far from the junction, the current is carried by drift of the majority carriers, which are injected across the junction to the other side, where, as minority carriers they give rise to diffusion currents.

5.6 In reverse bias, for ideal diodes, the *voltage-independent* reverse current is due to the collection of thermally generated minority carriers on either side diffusing to the depletion region and then being swept across. The current flow from n to p is small and is the basis of diode rectifiers.

5.7 For high reverse biases, diodes undergo (reversible) breakdown due to quantum mechanical *tunneling* (*Zener* mechanism) across a narrow depletion region for very heavily doped junctions or due to *impact ionization* or *avalanche* multiplication of carriers in a wider depletion region for lightly doped junctions. For narrow diodes, there can also be punch-through from one contact to the other.

5.8 Varying the bias leads to diode switching; the diode transient behavior can be found by solving the continuity equation, for example, using Laplace transforms, with suitable initial and boundary conditions.

5.9 Small signal capacitance in a semiconductor device is caused by a *change* in the charge storage as a function of bias. There are two components of diode capacitance: *depletion capacitance*, due to exposed dopant charges in the depletion region (dominates in reverse bias); and *diffusion capacitance*, due to stored excess mobile carriers (dominates in forward bias).

5.10 *Real diodes* can deviate from Shockley "ideal" diodes, which have negligible generation–recombination in the depletion region. *Generation–recombination* in the depletion region increases the diode ideality factor $\mathbf{n}$ from 1 to 2 in forward bias and introduces a roughly square root voltage dependence of reverse leakage current.

5.11 *High-level carrier injection* for large forward bias, in which the injected minority carrier concentration is comparable to the background majority carrier concentration, also makes $\mathbf{n} = 2$. *Series resistance* effects play a role as well for higher currents.

5.12 Graded junctions, where the doping concentrations on either side are not constant, are qualitatively similar to abrupt junctions, but are harder to analyze. They have different C–V than abrupt junctions have.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 240

5.13 Metal–semiconductor junctions behave as *Schottky diodes* (if the Fermi level alignments are such that there is a depletion of majority carriers in the semiconductor) or *ohmic contacts* (if there is no depletion region formed in the semiconductor).

5.14 Junctions between dissimilar semiconductors are called *heterojunctions*. With respect to a vacuum reference level, one looks at the conduction band edges (electron affinities) and band gaps to determine the band offsets and at the Fermi levels (work functions) to determine the direction of carrier transfer. Electrons flow from the high to low Fermi level regions, and holes flow the opposite way.

PROBLEMS

5.1 A 900-nm oxide is grown on (100) Si in wet oxygen at 1100°C (see App. VI). How long does it take to grow the first 200 nm, the next 300 nm, and the final 400 nm?

A square window (1 mm × 1 mm) is etched in this oxide, and the wafer is reoxidized at 1150°C in wet oxygen such that the oxide thickness *outside* of the window region increases to 2000 nm. Draw a cross section of the wafer and mark off all the thicknesses, dimensions, and oxide–Si interfaces relative to the original Si surface. Calculate the step heights in Si and in the oxide at the edge of the window.

5.2 When impurities are diffused into a sample from an unlimited source such that the surface concentration N_0 is held constant, the impurity distribution (profile) is given by

$$N(x, t) = N_0 \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right)$$

where D is the diffusion coefficient for the impurity, t is the diffusion time, and erfc is the complementary error function.

If a certain number of impurities are placed in a thin layer on the surface before diffusion, and if no impurities are added and none escape during diffusion, a gaussian distribution is obtained:

$$N(x, t) = \frac{N_s}{\sqrt{\pi Dt}} e^{-(x/2\sqrt{Dt})^2}$$

where N_s is the quantity of impurity placed on the surface (atoms/cm^2) prior to $t = 0$. Notice that this expression differs from Eq. (4–44) by a factor of two. Why?

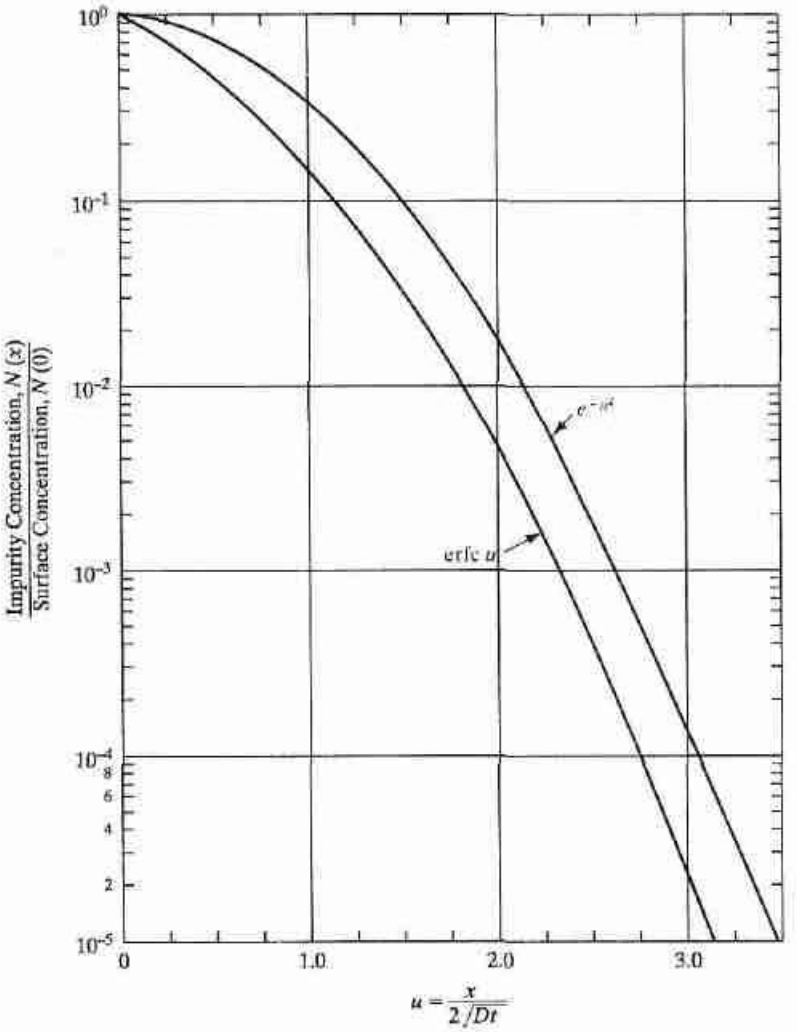
Figure P5-2 gives curves of the complementary error function and gaussian factors for the variable u, which in our case is $x/2\sqrt{Dt}$. Assume that boron is diffused into n-type Si (uniform $N_d = 5 \times 10^{16}$cm^{-3}) at 1000°C for 30 min. The diffusion coefficient for B in Si at this temperature is $D = 3 \times 10^{-14}$ cm^2/s.

(a) Plot $N_a(x)$ after the diffusion, assuming that the surface concentration is held constant at $N_0 = 5 \times 10^{20}$ cm^{-3}. Locate the position of the junction below the surface.

(b) Plot $N_a(x)$ after the diffusion, assuming that B is deposited in a thin layer on the surface prior to diffusion ($N_s = 5 \times 10^{13}$ cm^{-2}), and no additional B atoms are available during the diffusion. Locate the junction for this case.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 241

Figure P5–2



Hint: Plot the curves on five-cycle semilog paper, with an abscissa varying from zero to $\frac{1}{2}$ μm. In plotting $N_a(x)$, choose values of x that are simple multiples of $2\sqrt{Dt}$.

5.3 Assuming a constant (unlimited) source diffusion of P at 1000°C into p-type Si ($N_a = 2 \times 10^{16}$ cm^{-3}), calculate the time required to achieve a junction depth of 1 micron. See the equations in Prob. 5.2 and data in Appendices VII and VIII.

5.4 We wish to do an As implant into a Si wafer with a 0.1 μm oxide such that the peak lies at the oxide-silicon interface, with a peak value of 5×10^{19} cm^{-3}. What implant parameters (energy, dose, and beam current) would you choose? The scan area is 200 cm^2, and the desired implant time is 20 s. Assume similar range statistics in oxide and Si.

5.5 Silicon is implanted with P ions at 200 keV to a dose of 2.1×10^{14} cm^{-2}. Calculate and plot the P distribution on a semilog plot as in Fig. 5-4.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 242

5.6 We are interested in patterning the structure shown in Fig. P5-6. Design the mask aligner optics in terms of the numerical aperture of the lens and the wavelength of the source.

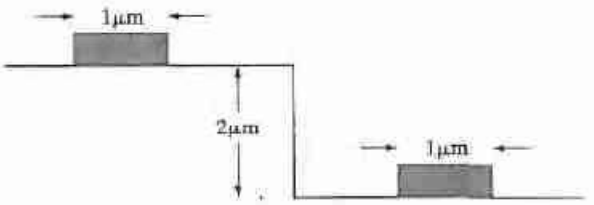


Figure P5-6

5.7 (a) Using Eq. (5-8), calculate the contact potential V_0 of a Si p-n junction operating at 300 K for $N_a = 10^{14}$ and 10^{19} cm^{-3}, with $N_d = 10^{14}$, 10^{15}, 10^{16}, 10^{17}, 10^{18}, and 10^{19} cm^{-3} in each case, and plot V_0 vs. N_d.

(b) Plot the maximum electric field E_0 vs. N_d for the junctions described in (a).

5.8 In a p$^+$-n junction the hole diffusion current in the neutral n material is given by Eq. (5-32). What are the electron diffusion and electron drift components of current at point x_n in the neutral n region?

5.9 An abrupt Si p-n junction has $N_a = 10^{17}$ cm^{-3} on the p side and $N_d = 10^{16}$ cm^{-3} on the n side. At 300 K, (a) calculate the Fermi levels, draw an equilibrium band diagram, and find V_0 from the diagram; (b) compare the result from (a) with V_0 calculated from Eq. (5-8).

5.10 Boron is implanted into an n-type Si sample ($N_d = 10^{16}$ cm^{-3}), forming an abrupt junction of square cross section with area $= 2 \times 10^{-3}$ cm^2. Assume that the acceptor concentration in the p-type region is $N_a = 4 \times 10^{18}$ cm^{-3}. Calculate V_0, x_{n0}, x_{p0}, Q_+, and E_0 for this junction at equilibrium (300 K). Sketch $\mathscr{E}$ and the charge density to scale, as in Fig. 5-12.

5.11 In a p$^+$-n junction, the n side has a donor concentration of 10^{16} cm^{-3}. If $n_i = 10^{10}$ cm^{-3}, the relative dielectric constant $\epsilon_r = 12$, $D_n = 50$ cm^2/s, $D_p = 20$ cm^2/s, and the electron and hole minority carriers have lifetimes $\tau = 100$ ns and 50 ns, respectively, and a forward bias of 0.6 V, calculate the *hole diffusion current density* 2 μm from the depletion edge on the n side. If we double the p$^+$ doping, what effect will it have on this hole diffusion current?

5.12 A Si p$^+$-n junction has a donor doping of 5×10^{16} cm^{-3} on the n side and a cross-sectional area of 10^{-3} cm^2. If $\tau_p = 1$ μs and $D_p = 10$ cm^2/s, calculate the current with a forward bias of 0.5 V at 300 K.

5.13 (a) Explain physically why the charge storage capacitance is unimportant for reverse-biased junctions.

(b) Assuming that a GaAs junction is doped to equal concentrations on the n and p sides, would you expect electron or hole injection to dominate in forward bias? Explain.

5.14 (a) A Si p$^+$-n junction 10^{-2} cm^2 in area has $N_d = 10^{15}$ cm^{-3} doping on the n side. Calculate the junction capacitance with a reverse bias of 10 V.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 243

(b) An abrupt p^+-n junction is formed in Si with a donor doping of $N_d = 10^{15}\ cm^{-3}$. What is the depletion region thickness W just prior to avalanche breakdown?

5.15 Using Eqs. (5-17) and (5-23), show that the peak electric field in the transition region is controlled by the doping on the more lightly doped side of the junction.

5.16 In a one-sided abrupt Si n^+-p junction (area = 0.0001 cm^2), we have the following parameters at 300 K:

n side	p side
N_d = Very high	$N_a = 10^{17} cm^{-3}$
$\tau_p = 10\ \mu s$	$\tau_n = 0.1\ \mu s$
$\mu_n = 100\ cm^2/V\text{-}s$	$\mu_n = 700\ cm^2/V\text{-}s$
$\mu_p = 450\ cm^2/V\text{-}s$	$\mu_p = 200\ cm^2/V\text{-}s$

Calculate the peak electric field at the junction and the depletion capacitance at a reverse bias of 100 V. (*Hint*: This is much higher than the contact potential, and remember that this is a one-sided junction!) For a certain forward bias, we have a current flow of 20 mA in this device. Calculate the *total excess stored electron charge* on the p side for this bias and the electric field far from the depletion region on the p side. (*Warning*: There is a lot of extraneous information here!)

5.17 The electron injection efficiency of a junction is I_n/I at $x_p = 0$.

(a) Assuming that the junction follows the simple diode equation, express I_n/I in terms of the diffusion constants, diffusion lengths, and equilibrium minority carrier concentrations.

(b) Show that I_n/I can be written as $[1 + L_n^p p_p \mu_p^n / L_p^n n_n \mu_n^p]^{-1}$, where the superscripts refer to the n and p regions. What should be done to increase the electron injection efficiency of a junction?

5.18 Assume that the doping concentration N_a on the p side of an abrupt junction is the same as N_d on the n side. Each side is many diffusion lengths long. Find the expression for the hole current I_p in the p-type material.

5.19 A Si p-n junction with cross-sectional area $A = 0.001\ cm^2$ is formed with $N_a = 10^{15}\ cm^{-3}$, $N_d = 10^{17}\ cm^{-3}$. Calculate:

(a) Contact potential, V_0.

(b) Space-charge width at equilibrium (zero bias).

(c) Current with a forward bias of 0.5 V. Assume that the current is diffusion dominated. Assume $\mu_n = 1500\ cm^2/V\text{-}s$, $\mu_p = 450\ cm^2/V\text{-}s$, $\tau_n = \tau_p = 2.5$ ms. Which carries most of the current, electrons or holes, and why? If you wanted to double the electron current, what should you do?

5.20 A Si n^+-p junction has an area of $5 \times 5\ \mu m$. Calculate the *total* junction capacitance associated with this junction at an applied reverse bias of 2 V. Assume that the n^+ region is doped $10^{20}\ cm^{-3}$ and the p doping is $1 \times 10^{16}\ cm^{-3}$. If we forward bias this junction 0.5 V, what is the electric field *far* from the junction on the p side,

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 244

assuming a hole mobility of 250 $cm^2/V\text{-}s$, an electron mobility of 100 $cm^2/V\text{-}s$, and a reverse saturation current density of 1 nA/cm^2 for this ideal diode?

5.21 In a p^+-n junction, the n-doping N_d is doubled. How do the following change if everything else remains unchanged? Indicate only increase or decrease.

(a) Junction capacitance

(b) Built-in potential

(c) Breakdown voltage

(d) Ohmic losses

5.22 We make a Si bar with a p-type doping of $2 \times 10^{16}\ cm^{-3}$ on the left half and a p-type doping of $10^{18}\ cm^{-3}$ on the right side. Sketch the equilibrium band diagram, with precise values marked off, *far* from the doping transition at 600 K when the intrinsic carrier concentration is $10^{16}\ cm^{-3}$. (*Note*: This is p type on both sides. Such a junction is known as a *high–low junction* rather than a p-n junction. Observe that the doping level is comparable to n_i on the left side! Do not worry about the exact details right near the doping transition!)

5.23 Assume that an abrupt Si p-n junction with area $10^{-4}\ cm^2$ has $N_a = 10^{17}\ cm^{-3}$ on the p side and $N_d = 10^{17}\ cm^{-3}$ on the n side. The diode has a forward bias of 0.7 V. Using mobility values from Fig. 3–23 and assuming that $\tau_n = \tau_p = 1\ \mu s$, plot I_p and I_n versus distance on a diagram such as Fig. 5–17, including both sides of the junction. Neglect recombination within W.

5.24 In a p^+-n junction reverse biased at 10 V, the capacitance is 10 pF. If the doping of the n side is doubled and the reverse bias changed to 80 V, what is the capacitance?

5.25 An abrupt p^+-n junction is formed in Si with a donor doping of $N_d = 10^{15}\ cm^{-3}$. What is the minimum thickness of the n region that will ensure avalanche breakdown rather than punch-through?

5.26 Calculate the capacitance for the following Si n^+-p junction.

$N_a = 10^{15}\ cm^{-3}$

Area = $0.001\ cm^2$

Reverse bias = 1, 5, and 10 V

Plot $1/C^2$ vs. V_R

Demonstrate that the slope yields N_a. Repeat calculations for $N_a = 10^{17}\ cm^{-3}$.

Since the doping is not specified on the n^+ side, use a suitable approximation.

5.27 We assumed in Section 5.2.3 that carriers are excluded within W and that the semiconductor is neutral outside W. This is known as the *depletion approximation*. Obviously, such a sharp transition is unrealistic. In fact, the space charge varies over a distance of several *Debye lengths*, given by

$$L_D = \left[\frac{\epsilon_s kT}{q^2 N_d}\right]^{1/2} \quad \text{on the n side.}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 245

Calculate the Debye length on the n side for Si junctions having $N_a = 10^{18}\ \text{cm}^{-3}$ on the p side and $N_d = 10^{14}, 10^{16}$, and $10^{18}\ \text{cm}^{-3}$ on the n side and compare with the size of W in each case.

5.28 We have a symmetric p-n silicon junction ($N_a = N_d = 10^{17}\ \text{cm}^{-3}$). If the peak electric field in the junction at breakdown is 5×10^5 V/cm, what is the reverse breakdown voltage in this junction?

5.29 A long p^+-n diode is forward biased with a current I flowing in it. The current is suddenly tripled at $t = 0$.

(a) What is the slope of the hole distribution at $x_n = 0$ just after the current is tripled?

(b) Assuming that the voltage is always $\gg kT/q$, relate the final junction voltage (at $t = \infty$) to the initial voltage (before $t = 0$).

5.30 A long p^+-n junction has its forward-bias current switched from I_{F1} to I_{F2} at $t = 0$. Find an expression for the stored charge Q_p as a function of time in the n region.

5.31 A p^+-n diode is switched from zero bias ($I = 0$) to a forward current I at $t = 0$.

(a) Find the expression for the excess hole charge $Q_p(t)$ during the turn-on transient. Sketch $Q_p(t)$ vs. t.

(b) Assuming that $\delta p(x_n)$ is always essentially exponential, find $\Delta p_n(t)$ and the voltate $v(t)$.

5.32 The diode of Fig. 5-23c is used in a simple half-wave rectifier circuit in which the diode is placed in series with a load resistor. Assume that the diode offset voltage E_0 is 0.4 V and that $R = dv/di = 400\ \Omega$. For a load resistor of 1 kΩ and a sinusoidal input of 2 sin ωt, sketch the output voltage (across the load resistor) over two cycles.

5.33 The ideality factor **n** can be used to describe the relative importance of recombination within the transition region and the neutral region. Calculate and plot the I–V characteristics of a diode, using Eq. (5–74) for values of the ideality factor of 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0.

$A = 100\ \mu\text{m}^2$, $N_a = 10^{19}\ \text{cm}^{-3}$, and

$N_d = 10^{19}\ \text{cm}^{-3}$, $\tau_n = \tau_p = \mu\text{s}$

5.34 Assume holes are injected from a p^+-n junction into a short n region of length l. If $\delta p(x_n)$ varies linearly from Δp_n at $x_n = 0$ to zero at the ohmic contact ($x_n = l$), find the steady state charge in the excess hole distribution Q_p and the current I.

5.35 Assume that a p^+-n diode is built with an n region width l smaller than a hole diffusion length ($l < L_p$). This is the so-called *narrow base diode*. Since for this case holes are injected into a short n region under forward bias, we cannot use the assumption $\delta p(x_n = \infty) = 0$ in Eq. (4–35). Instead, we must use as a boundary condition the fact that $\delta p = 0$ at $x_n = l$.

(a) Solve the diffusion equation to obtain

$$\delta p(x_n) = \frac{\Delta p_n \left[e^{(l - x_n)/L_p} - e^{(x_n - l)/L_p} \right]}{e^{l/L_p} - e^{-l/L_p}}$$

(b) Show that the current in the diode is

$$I = \left(\frac{qAD_p p_n}{L_p} \operatorname{ctnh} \frac{l}{L_p}\right)(e^{qV/kT} - 1)$$

5.36 Given the narrow base diode result (Prob. 5.35), (a) calculate the current due to recombination in the n region, and (b) show that the current due to recombination at the ohmic contact is

$$I(\text{ohmic contact}) = \left(\frac{qAD_p p_n}{L_p} \operatorname{csch} \frac{l}{L_p}\right)(e^{qV/kT} - 1)$$

5.37 Assume that a p$^+$-n junction is built with a graded n region in which the doping is described by $N_d(x) = Gx^m$. The depletion region ($W \cong x_{n0}$) extends from essentially the junction at $x = 0$ to a point W within the n region. The singularity at $x = 0$ for negative **m** can be neglected.

(a) Integrate Gauss's law across the depletion region to obtain the maximum value of the electric field $\mathscr{E}_0 = -qGW^{(m+1)}/\epsilon(\mathbf{m} + 1)$.

(b) Find the expression for $\mathscr{E}(x)$, and use the result to obtain $V_0 - V = qGW^{(m+2)}/\epsilon(\mathbf{m} + 2)$.

(c) Find the charge Q due to ionized donors in the depletion region; write Q explicitly in terms of $(V_0 - V)$.

(d) Using the results of (c), take the derivative $dQ/d(V_0 - V)$ to show that the capacitance is

$$C_j = A\left[\frac{qG\epsilon^{(m+1)}}{(m+2)(V_0 - V)}\right]^{1/(\mathbf{m}+2)}$$

5.38 We deposit a metal with a work function of 4.6 eV on Si (electron affinity of 4 eV) and acceptor doping level of 10^{18} cm^{-3}. Draw the equilibrium band diagram and mark off the Fermi level, the band edges, and the vacuum level. Is this a Schottky or ohmic contact, and why? By how much should the metal work function be altered to change the type of contact? Explain with reference to the band diagram.

5.39 Design an ohmic contact for n-type GaAs using InAs, with an intervening graded InGaAs region (see Fig. 5-44).

5.40 A Schottky barrier is formed between a metal having a work function of 4.3 eV and p-type Si (electron affinity = 4 eV). The acceptor doping in the Si is 10^{17} cm^{-3}.

(a) Draw the equilibrium band diagram, showing a numerical value for qV_0.

(b) Draw the band diagram with 0.3 V forward bias. Repeat for 2 V reverse bias.

READING LIST

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

Campbell, S. A. *The Science and Engineering of Microelectronic Fabrication*, 2d ed. NY: Oxford, 2001.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 247

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

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

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

Pierret, R. F. *Semiconductor Device Fundamentals.* Reading, MA: Addison-Wesley, 1996.

Plummer, J. D., M. D. Deal, and P. B. Griffin. *Silicon VLSI Technology.* Upper Saddle River, NJ: Prentice Hall, 2000.

Shockley, W. "The Theory of P-N Junctions in Semiconductors and P-N Junction Transistors." *Bell Syst. Tech. J.* 28 (1949), 435.

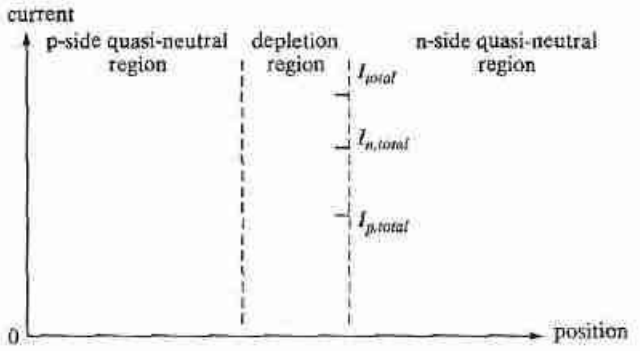
Wolf, S., and R. N. Tauber. *Silicon Processing for the VLSI Era.*, 2d ed. Sunset Beach, CA: Lattice Press, 2000.

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

SELF QUIZ

Question 1

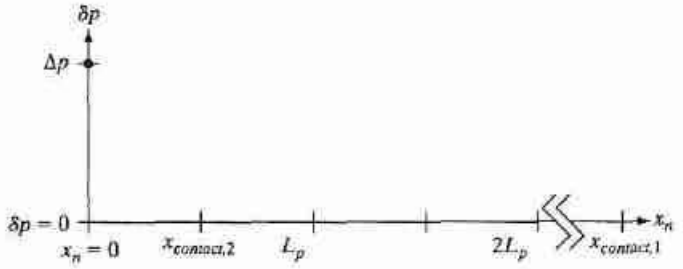
Consider a forward-biased *ideal* (abrupt junction, no recombination or generation in the depletion region) long p-n junction diode under forward bias. On the following graph, sketch and label the total current I_{total}, the total electron current $I_{n,total}$ and the total hole current $I_{p,total}$ as a function of position throughout the entire device. The value of each has been given at the n-side edge of the depletion region for reference. (*Hint*: the excess carrier concentrations are essentially pinned to zero at the contacts.)



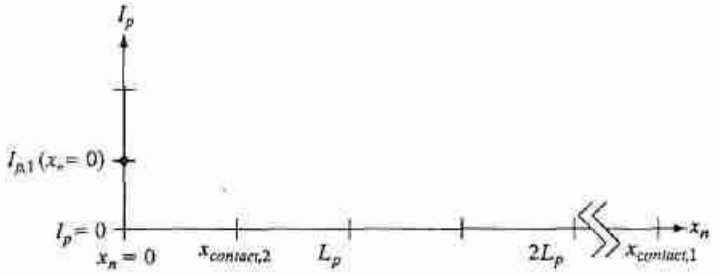
Question 2

(a) Consider (1) a long p-n junction diode with the ohmic contact on the n side, $x_{contact,1} >> L_p$, and (2) a short p-n junction diode with the ohmic contact on the n side well within a diffusion length of the depletion region, $x_{contact2} < L_p$. Given the indicated excess hole (equals electron) concentration at the n-side edge of

the depletion region Δp for both diodes, sketch the excess hole concentrations as a function of position $\delta p(x_n)$ within the region for the two cases on the following graph, and label them (1) and (2) accordingly.



(b) Consistent with your answers to part (a), sketch the hole (diffusion) currents within the region for the same two diodes on the following graph, and label them (1) and (2) again. The value of the hole current at the edge of the depletion region for the long diode $I_{p,1}(x_n = 0)$ is provided for reference.



Question 3

Consider two Si p-n junction diodes, one long and one short (contacts within a diffusion length of the depletion region) but otherwise identical. Under identical forward-bias voltage, which diode would have greater current flow?

Question 4

If the depletion capacitance of a p-n junction diode is $C_{d,o}$ at equilibrium and the contact potential is 0.5 V, how much reverse-bias voltage would have to be applied to reduce the depletion capacitance to 0.5 $C_{d,o}$?

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 249

Question 5

(a) What is the difference between depletion and diffusion capacitance in a diode? Which one dominates in forward bias and why? In reverse bias?

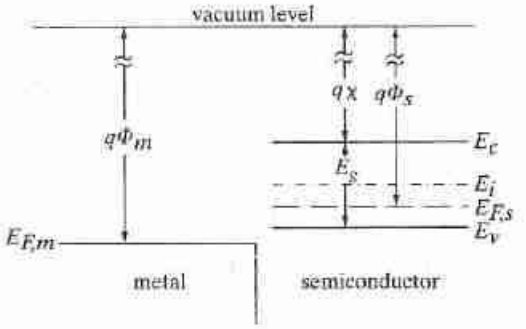
(b) Why is it meaningful to define small signal capacitance and conductance in semiconductor devices such as diodes? How are they defined?

Question 6

We grow a pseudomorphic heterostructure consisting of an epitaxial film with lattice constant of 6 Å and band gap of 2 eV on a thick substrate with lattice constant of 4 Å and band gap of 1 eV. Both the substrate and epitaxial layer have a cubic crystal structure in the unstrained state. If 60% of the band edge discontinuity is in the conduction band, sketch a simplified band diagram of this heterostructure. Also, qualitatively show a 2-D view of the crystal structure in relation to the band diagram.

Question 7

(a) In the space below, sketch the equilibrium band diagram resulting from bringing together the illustrated metal and lightly doped semiconductor indicating the Fermi level; the conduction band and valence band offsets from the Fermi level at the metal–semiconductor interface in terms of $q\Phi_m$, $q\Phi_s$, $q\chi$ and/or E_g; any band-bending in the semiconductor in terms of $q\Phi_m$, $q\Phi_s$, $q\chi$ and/or E_g; and (qualitatively) any charge depletion or accumulation layer. Assume no interface traps.



(b) Is this a Schottky contact or an ohmic contact?

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 250

Chapter 6
Field-Effect Transistors

OBJECTIVES

1. Derive I–V characteristics of JFETs, MESFETs, and HEMTs
2. Determine MOS C-V behavior and threshold voltage; study leakage characteristics of gate dielectrics
3. Understand MOSFET band diagrams, and linear and saturation behavior
4. Understand "effective" channel mobility, body effect, and subthreshold slopes
5. Analyze second-order effects—DIBL, GIDL, charge sharing, and V_T roll-off

The modern era of semiconductor electronics was ushered in by the invention of the bipolar transistor in 1948 by Bardeen, Brattain, and Shockley at the Bell Telephone Laboratories. This device, along with its field-effect counterpart, has had an enormous impact on virtually every area of modern life. In this chapter we will learn about the operation, applications, and fabrication of the field-effect transistor (FET).

The field-effect transistor comes in several forms. In a *junction* FET (called a *JFET*) the control (*gate*) voltage varies the depletion width of a reverse-biased p-n junction. A similar device results if the junction is replaced by a Schottky barrier (*metal–semiconductor* FET, called a *MESFET*). Alternatively, the metal gate electrode may be separated from the semiconductor by an insulator (*metal–insulator–semiconductor* FET, called a *MISFET*). A common special case of this type uses an oxide layer as the insulator (*MOSFET*).

In Chapter 5 we found that two dominant features of p-n junctions are the injection of minority carriers with forward bias and a variation of the depletion width W with reverse bias. These two p-n junction properties are used in two important types of transistors. The *bipolar junction transistor (BJT)* discussed in Chapter 7 uses the injection of minority carriers across a forward-biased junction, and the junction field-effect transistor discussed in this chapter depends on control of a junction depletion width under reverse bias. The FET is a majority carrier device, and is therefore often called a *unipolar* transistor. The BJT, on the other hand, operates by the injection and collection of *minority* carriers. Since the action of both electrons and holes is important in this device, it is called a *bipolar* transistor. Like its bipolar counterpart, the FET is a three-terminal device in which the current through two terminals is

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 251

controlled at the third. Unlike the BJT, however, field-effect devices are controlled by a voltage at the third terminal rather than by a current.

The history of BJTs and FETs is rather interesting. It was the FET that was proposed first in 1930 by Lilienfeld, but he never got it to work because he did not fully appreciate the role of surface defects or surface states. In the process of trying to demonstrate experimentally such a field-effect transistor, Bardeen and Brattain somewhat serendipitously invented the first bipolar transistor, the Ge point contact transistor. This major breakthrough was rapidly followed by Shockley's extension of the concept to the BJT. It was only much later, after the problem of surface states was resolved by growing an oxide insulator on Si, that the first MOSFET was demonstrated in 1960 by Kahng and Atalla. Although the BJT reigned supreme in the early days of semiconductor integrated electronics, it has been gradually supplanted in most applications by the Si MOSFET. The main reason is, unlike BJTs, the various types of FET are characterized by a high *input impedance*, since the control voltage is applied to a reverse-biased junction or Schottky barrier, or across an insulator. These devices are particularly well suited for controlled switching between a conducting state and a nonconducting state, and are therefore useful in digital circuits. They are also suitable for integration of many devices on a single chip, as we shall see in Chapter 9. In fact, millions of MOS transistors are commonly used together in semiconductor memory devices and microprocessors.

6.1 TRANSISTOR OPERATION

We begin this section with a general discussion of amplification and switching, the basic circuit functions performed by transistors. The transistor is a three-terminal device with the important feature that the current through two terminals can be controlled by small changes we make in the current or voltage at the third terminal. This control feature allows us to amplify small a-c signals or to switch the device from an *on* state to an *off* state and back. These two operations, *amplification* and *switching*, are the basis of a host of electronic functions. This section provides a brief introduction to these operations, as a foundation for understanding both bipolar and field-effect transistors.

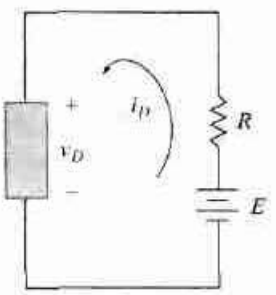
6.1.1 The Load Line

Consider a two-terminal device that has a nonlinear I–V characteristic, as in Fig. 6–1. We might determine this curve experimentally by measuring the current for various applied voltages, or by using an oscilloscope called a *curve tracer*, which varies I and V repetitively and displays the resulting curve. When such a device is biased with the simple battery–resistor combination shown in the figure, steady state values of I_D and V_D are attained. To find these values we begin by writing a loop equation around the circuit:[1]

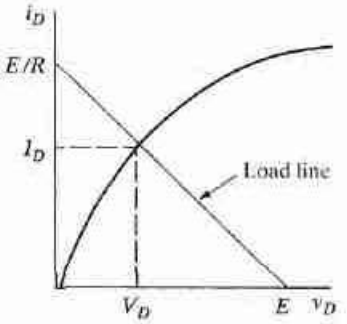
$$E = i_D R + v_D \tag{6-1}$$

[1]We use I_D to symbolize the total current i_D for the d-c value, and i_d for the a-c component. A similar scheme is used for other currents and voltages.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 252



(a)

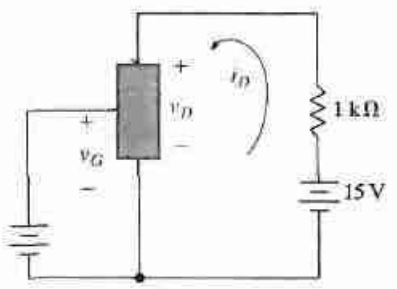


(b)

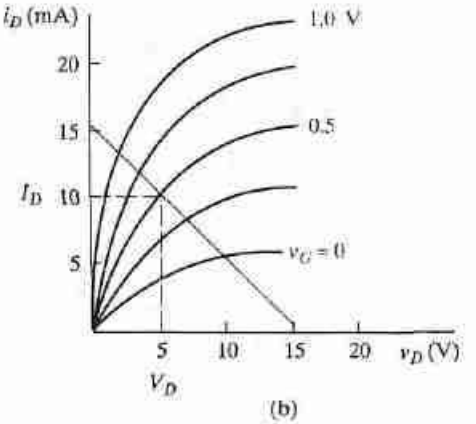
Figure 6–1
A two-terminal nonlinear device: (a) biasing circuit; (b) *I–V* characteristic and load line.

This gives us one equation describing the circuit, but it contains two unknowns (i_D and v_D). Fortunately, we have another equation of the form $i_D = f(v_D)$ in the curve of Fig. 6-1b, giving us two equations with two unknowns. The steady state current and voltage are found by a simultaneous solution of these two equations. However, since one equation is analytical and the other is graphical, we must first put them into the same form. It is easy to make the linear equation (6-1) graphical, so we plot it on Fig. 6-1b to find the simultaneous solution. The end points of the line described by Eq. (6-1) are at E when $i_D = 0$ and at E/R when $v_D = 0$. The two graphs cross at $v_D = V_D$ and $i_D = I_D$, the steady state values of current and voltage for the device with this biasing circuit.

Now let's add a third terminal which somehow controls the *I–V* characteristic of the device. For example, assume that the device current-voltage curve can be moved up the current axis by increasing the control voltage as in Fig. 6-2b. This results in a family of i_D-v_D curves, depending upon the choice of v_G. We can still write the loop equation (6-1) and draw it on the set of



(a)



(b)

Figure 6–2
A three-terminal nonlinear device that can be controlled by the voltage at the third terminal v_G: (a) biasing circuit; (b) *I–V* characteristic and load line. If $V_G = 0.5$ V, the d-c values of I_D and V_D are as shown by the dashed lines.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 253

curves, but now the simultaneous solution depends on the value of v_G. In the example of Fig. 6-2, V_G is 0.5 V and the d-c values of I_D and V_D are found at the intersection to be 10 mA and 5 V, respectively. Whatever the value of the control voltage v_G at the third terminal, values of I_D and V_D are obtained from points along the line representing Eq. (6-1). This is called the *load line*.

6.1.2 Amplification and Switching

If an a-c source is added to the control voltage, we can achieve large variations in i_D by making small changes in v_G. For example, as v_G varies about its d-c value by 0.25 V in Fig. 6-2, v_d varies about its d-c value V_D by 2 V. Thus the amplification of the a-c signal is 2/0.25 = 8. If the curves for equal changes in v_G are equally spaced on the i_D axis, a faithful amplified version of the small control signal can be obtained. This type of voltage-controlled amplification is typical of field-effect transistors. For bipolar transistors, a small control current is used to achieve large changes in the device current, achieving current amplification.

Another important circuit function of transistors is the controlled switching of the device off and on. In the example of Fig. 6-2, we can switch from the bottom of the load line ($i_D = 0$) to almost the top ($i_D \simeq E/R$) by appropriate changes in v_G. This type of switching with control at a third terminal is particularly useful in digital circuits.

6.2 THE JUNCTION FET

In a *junction FET (JFET)* the voltage-variable depletion region width of a junction is used to control the effective cross-sectional area of a conducting *channel*. In the device of Fig. 6-3, the current I_D flows through an n-type channel between two p^+ regions. A reverse bias between these p^+ regions and the channel causes the depletion regions to intrude into the n material, and therefore the effective width of the channel can be restricted. Since the resistivity of the channel region is fixed by its doping, the channel resistance varies with changes in the effective cross-sectional area. By analogy, the variable depletion regions serve as the two doors of a gate, which open and close on the conducting channel.

In Fig. 6-3 electrons in the n-type channel drift from left to right, opposite to current flow. The end of the channel from which electrons flow is called the *source*, and the end toward which they flow is called the *drain*. The p^+ regions are called *gates*. If the channel were p-type, holes would flow from the source to the drain, in the same direction as the current flow, and the gate regions would be n^+. It is common practice to connect the two gate regions electrically; therefore, the voltage V_G refers to the potential from each gate region G to the source S. Since the conductivity of the heavily doped p^+ regions is high, we can assume that the potential is uniform throughout each gate. In the lightly doped channel material, however, the potential varies with position (Fig. 6-3b). If the channel of Fig. 6-3 is considered as a distributed resistor carrying a current I_D

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 254

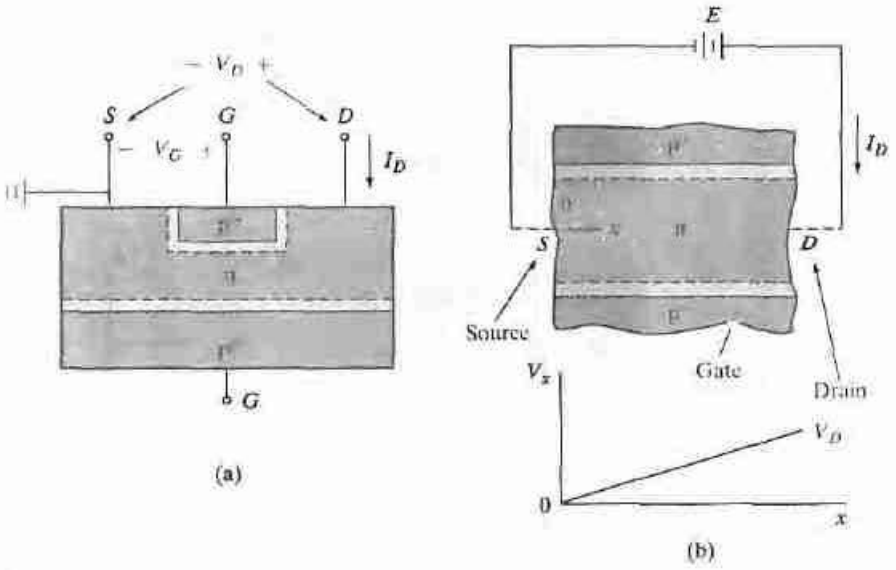


Figure 6–3
Simplified cross-sectional view of a junction FET: (a) transistor geometry; (b) detail of the channel and voltage variation along the channel with $V_G = 0$ and small I_D.

it is clear that the voltage from the drain end of the channel D to the source electrode S must be greater than the voltage from a point near the source end to S. For low values of current we can assume a linear variation of voltage V_x in the channel, varying from V_D at the drain end to zero at the source end (Fig. 6–3b).

6.2.1 Pinch-off and Saturation

In Figure 6–4 we consider the channel in a simplified way by neglecting voltage drops between the source and drain electrodes and the respective ends of the channel. For example, we assume that the potential at the drain end of the channel is the same as the potential at the electrode D. This is a good approximation if the source and drain regions are relatively large, so that there is little resistance between the ends of the channel and the electrodes. In Fig. 6–4 the gates are short circuited to the source ($V_G = 0$), such that the potential at $x = 0$ is the same as the potential everywhere in the gate regions. For very small currents, the widths of the depletion regions are close to the equilibrium values (Fig. 6–4a). As the current I_D is increased, however, it becomes important that V_x is large near the drain end and small near the source end of the channel. Since the reverse bias across each point in the gate-to-channel junction is simply V_x when V_G is zero, we can estimate the shape of the depletion regions as in Fig. 6–4b. The reverse bias is relatively large near the drain

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 255

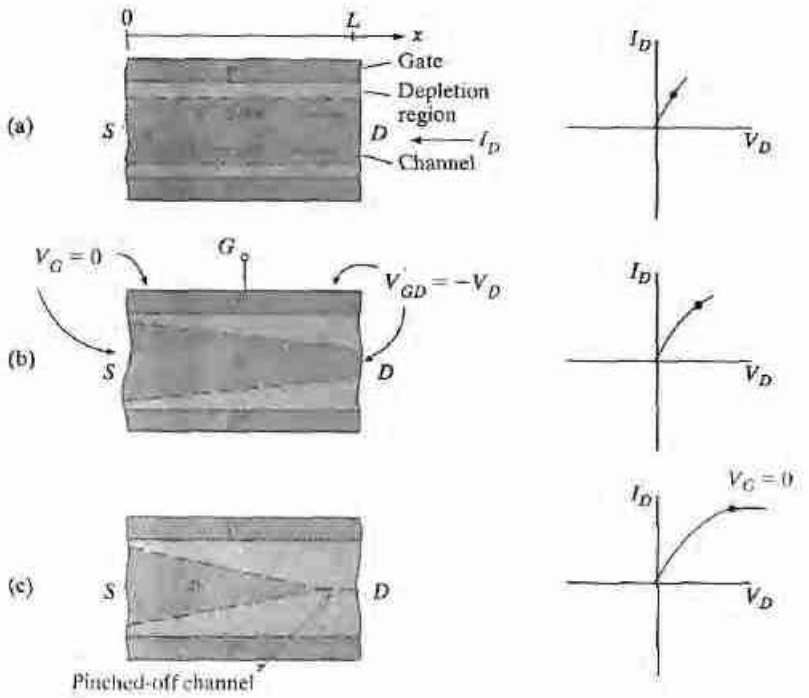


Figure 6–4
Depletion regions in the channel of a JFET with zero gate bias for several values of V_D: (a) linear range; (b) near pinch-off; (c) beyond pinch-off.

($V_{GD} = -V_D$) and decreases toward zero near the source. As a result, the depletion region intrudes into the channel near the drain, and the effective channel area is constricted.

Since the resistance of the constricted channel is higher, the I–V plot for the channel begins to depart from the straight line that was valid at low current levels. As the voltage V_D and current I_D are increased still further, the channel region near the drain becomes more constricted by the depletion regions and the channel resistance continues to increase. As V_D is increased, there must be some bias voltage at which the depletion regions meet near the drain and essentially *pinch off* the channel (Fig. 6–4c). When this happens, the current I_D cannot increase significantly with further increase in V_D. Beyond pinch-off the current is *saturated* approximately at its value at pinch-off.[2] Once electrons from the channel enter the electric field of the depletion region, they are swept through and ultimately flow to the positive drain contact. After the current saturates beyond pinch-off, the differential channel resistance dV_D/dI_D becomes very high. To a good approximation, we can

[2] *Saturation* is used by device engineers in more different contexts than any other word is used. We have discussed velocity saturation, the reverse saturation current of a junction, and, now, the saturation of FET characteristics. In Chapter 7, we will discuss the saturation of a BJT. The student has probably also reached saturation by now in trying to absorb these various meanings!

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 256

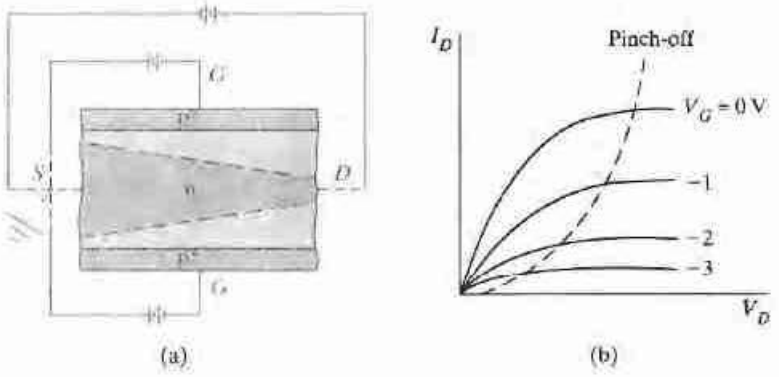


Figure 6–5
Effects of a negative gate bias: (a) increase of depletion region widths with V_G negative; (b) family of current–voltage curves for the channels as V_G is varied.

calculate the current at the critical pinch-off voltage and assume there is no further increase in current as V_D is increased.

6.2.2 Gate Control

The effect of a negative gate bias $-V_G$ is to increase the resistance of the channel and induce pinch-off at a lower value of current (Fig. 6–5). Since the depletion regions are larger with V_G negative, the effective channel width is smaller and its resistance is higher in the low-current range of the characteristic. Therefore, the slopes of the I_D vs. V_D curves below pinch-off become smaller as the gate voltage is made more negative (Fig. 6–5b). The pinch-off condition is reached at a lower drain-to-source voltage, and the saturation current is lower than for the case of zero gate bias. As V_G is varied, a family of curves is obtained for the I–V characteristic of the channel, as in Fig. 6–5b.

Beyond the pinch-off voltage the drain current I_D is controlled by V_G. By varying the gate bias we can obtain amplification of an a-c signal. Since the input control voltage V_G appears across the reverse-biased gate junctions, the input impedance of the device is high.

We can calculate the pinch-off voltage rather simply by representing the channel in the approximate form of Fig. 6–6. If the channel is symmetrical and the effects of the gates are the same in each half of the channel region, we can restrict our attention to the channel half-width $h(x)$, measured from the center line ($y = L$). The metallurgical half-width of the channel (i.e., neglecting the depletion region) is a. We can find the pinch-off voltage by calculating the reverse bias between the n channel and the p^+ gate at the drain end of the channel ($x = L$). For simplicity we shall assume that the channel width at the drain decreases uniformly as the reverse bias increases to pinch-off. If the reverse bias between the gate and the drain is $-V_{GD}$, the width of the depletion region at $x = L$ can be found from Eq. (5–57):

$$W(x = L) = \left[\frac{2\epsilon(-V_{GD})}{qN_d}\right]^{1/2} \quad (V_{GD} \text{ negative}) \tag{6–2}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 257

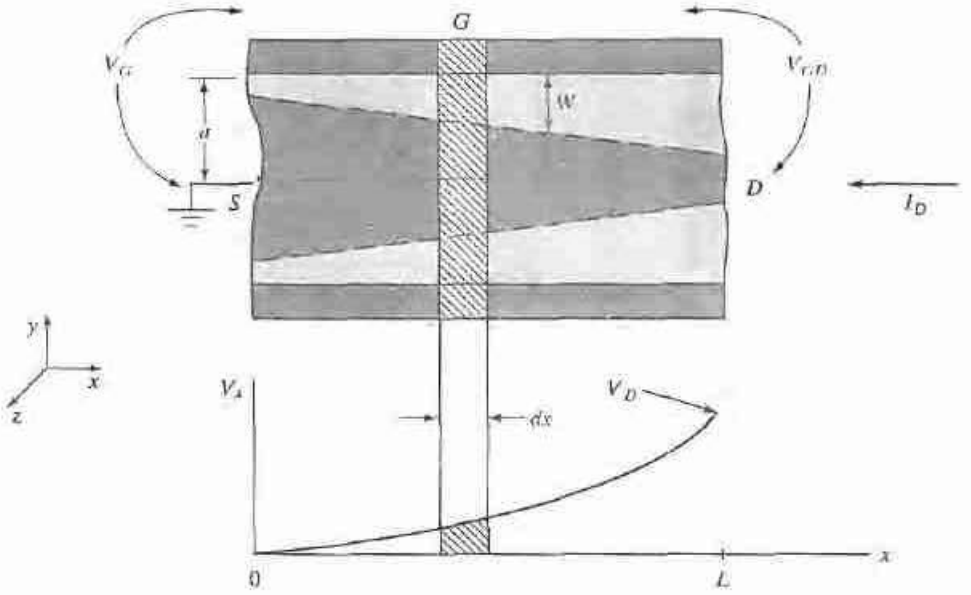


Figure 6–6
Simplified diagram of the channel with definitions of dimensions and differential volume for calculations.

In this expression we assume the equilibrium contact potential V_0 is negligible compared with V_{GD} and the depletion region extends primarily into the channel for the p⁺-n junction. Including V_0 is left for Prob. 6.1.

Pinch-off occurs at the drain end of the channel when

$$h(x = L) = a - W(x = L) = 0 \tag{6-3}$$

that is, when $W(x = L) = a$. If we define the value of $-V_{GD}$ at pinch-off as V_p, we have

$$\left[\frac{2\epsilon V_P}{qN_d}\right]^{1/2} = a$$

$$\boxed{V_P = \frac{qa^2N_d}{2\epsilon}} \tag{6-4}$$

The pinch-off voltage V_p is a positive number; its relation to V_D and V_G is

$$V_P = -V_{GD}(\text{pinch-off}) = -V_G + V_D \tag{6-5}$$

where V_G is zero or negative for proper device operation. A forward bias on the gate would cause hole injection from the p⁺ regions into the channel, eliminating the field-effect control of the device. From Eq. (6-5) it is clear that

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 258

pinch-off results from a combination of gate-to-source voltage and drain-to-source voltage. Pinch-off is reached at a lower value of V_D (and therefore a lower I_D) when a negative gate bias is applied, in agreement with Fig. 6-5b.

6.2.3 Current–Voltage Characteristics

Calculation of the exact channel current is complicated, although the mathematics is relatively straightforward below pinch-off. The approach we shall take is to find the expression for I_D just at pinch-off, and then assume the saturation current beyond pinch-off remains fairly constant at this value.

In the coordinate system defined in Fig. 6-6, the center of the channel at the source end is taken as the origin. The length of the channel in the x-direction is L, and the depth of channel in the z-direction is Z. We shall call the resistivity of the n-type channel material ρ (valid only in the neutral n material, outside the depletion regions). If we consider the differential volume of neutral channel material $Z2h(x)dx$, the resistance of the volume element is $\rho dx/Z2h(x)$ [see Eq. (3-44)]. Since the current does not change with distance along the channel, I_D is related to the differential voltage change in the element dV_x by the conductance of the element:

$$I_D = \frac{Z2h(x)}{\rho}\frac{dV_x}{dx} \tag{6-6}$$

The term $2h(x)$ is the channel width at x.

The half-width of the channel at point x depends on the local reverse bias between gate and channel $-V_{Gx}$:

$$h(x) = a - W(x) = a - \left[\frac{2\epsilon(-V_{Gx})}{qN_d}\right]^{1/2} = a\left[1 - \left(\frac{V_x - V_G}{V_p}\right)^{1/2}\right] \tag{6-7}$$

since $V_{Gx} = V_G - V_x$ and $V_p = qa^2N_d/2\epsilon$. Implicit in Eq. (6-7) is the assumption that the expression for $W(x)$ can be obtained by a simple extension of Eq. (6-2) to point x in the channel. This is called the *gradual channel approximation*; it is valid if $h(x)$ does not vary abruptly in any element dx.

The voltage V_{Gx} will be negative since the gate voltage V_G is chosen zero or negative for proper operation. Substituting Eq. (6-7) into Eq. (6-6), we have

$$\frac{2Za}{\rho}\left[1 - \left(\frac{V_x - V_G}{V_p}\right)^{1/2}\right]dV_x = I_D dx \tag{6-8}$$

We can solve this equation to obtain

$$\boxed{I_D = G_0 V_p\left[\frac{V_D}{V_p} + \frac{2}{3}\left(-\frac{V_G}{V_p}\right)^{3/2} - \frac{2}{3}\left(\frac{V_D - V_G}{V_p}\right)^{3/2}\right]} \tag{6-9}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 259

where V_G is negative and $G_0 \equiv 2aZ/\rho L$ is the conductance of the channel for negligible $W(x)$, i.e., with no gate voltage and low values of I_D. This equation is valid only up to pinch-off, where $V_D - V_G = V_p$. If we assume the saturation current remains essentially constant at its value at pinch-off, we have

$$I_D(\text{sat.}) = G_0 V_P \left[\frac{V_D}{V_P} + \frac{2}{3}\left(-\frac{V_G}{V_P} \right)^{3/2} - \frac{2}{3} \right]$$

$$= G_0 V_P \left[\frac{V_G}{V_P} + \frac{2}{3}\left(-\frac{V_G}{V_P} \right)^{3/2} + \frac{1}{3} \right] \qquad (6\text{-}10)$$

where

$$\frac{V_D}{V_P} = 1 + \frac{V_G}{V_P}$$

The resulting family of I-V curves for the channel agrees with the results we predicted qualitatively (Fig. 6-5b). The saturation current is greatest when V_G is zero and becomes smaller as V_G is made negative.

We can represent the device biased in the saturation region by an equivalent circuit where changes in drain current are related to gate voltage changes by

$$g_m(\text{sat.}) = \frac{\partial I_D(sat.)}{\partial V_G} = G_0 \left[1 - \left(-\frac{V_G}{V_P} \right)^{1/2} \right] \qquad (6\text{-}11)$$

The quantity g_m is the *mutual transconductance*, with units (A/V) called siemens (S), sometimes called mhos. As a figure of merit for FET devices it is common to describe the transconductance per unit channel width Z. This quantity g_m/Z is usually given in units of millisiemens per millimeter.

It is found experimentally that a square-law characteristic closely approximates the drain current in saturation:

$$I_D(\text{sat.}) = I_{DSS}\left(1 + \frac{V_G}{V_P} \right)^2, \quad (V_G \textit{ negative}) \qquad (6\text{-}12)$$

where I_{DSS} is the saturated drain current with $V_G = 0$.

The appearance of a constant value of channel resistivity (in the G_0 term) in Eqs. (6-9)–(6-11) implies that the electron mobility is constant. As mentioned in Section 3.4.4, electron velocity saturation at high fields may make this assumption invalid. This is particularly likely for very short channels, where even moderate drain voltage can result in a high field along the channel. Another departure from the ideal model results from the fact that the effective channel length decreases as the drain voltage is increased beyond pinch-off, as Fig. 6-4c suggests. In short-channel devices this effect can cause I_D to increase beyond pinch-off, since L appears in the denominator

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 260

of Eq. (6–10), in G_0. Therefore, the assumption of constant saturation current is not valid for very short-channel devices.

6.3 THE METAL–SEMICONDUCTOR FET

The depletion of the channel discussed above for a JFET can be accomplished by the use of a reverse-biased Schottky barrier instead of a p-n junction. The resulting device is called a MESFET, indicating that a metal–semiconductor junction is used. This device is useful in high-speed digital or microwave circuits, where the simplicity of Schottky barriers allows fabrication to close geometrical tolerances. There are particular speed advantages for MESFET devices in III–V compounds such as GaAs or InP, which have higher mobilities and carrier drift velocities than Si.

6.3.1 The GaAs MESFET

Figure 6–7 shows schematically a simple MESFET in GaAs. The substrate is undoped or doped with chromium, which has an energy level near the center of the GaAs band gap. In either case the Fermi level is near the center of the gap, resulting in very high resistivity material ($\sim 10^8$ Ω-cm), generally called *semi-insulating* GaAs. On this nonconducting substrate a thin layer of lightly-doped n-type GaAs is grown epitaxially, to form the channel region of the FET.[3] The photolithographic processing consists of defining patterns in the metal layers for source and drain ohmic contacts (e.g., Au–Ge) and for the Schottky barrier gate (e.g., Al). By reverse biasing the Schottky gate, the channel can be depleted to the semi-insulating substrate, and the resulting I-V characteristics are similar to the JFET device.

By using GaAs instead of Si, a higher electron mobility is available (see Appendix III), and furthermore GaAs can be operated at higher temperatures (and therefore higher power levels). Since no diffusions are involved in Fig. 6–7, close geometrical tolerances can be achieved and the MESFET can be made very small. Gate lengths $L \leq 0.25$ μm are common in these devices. This is important at high frequencies, since drift time and capacitances must be kept to a minimum.

It is possible to avoid the epitaxial growth of the n-type layer and the etched isolation in Fig. 6–7 by using ion implantation. Starting with a semi-insulating GaAs substrate, a thin n-type layer at the surface of each transistor region can be formed by implanting Si or a column VI donor impurity such as Se. This implantation requires an anneal to remove the radiation damage, but the epitaxial growth step is eliminated. In either the fully implanted device or the epitaxial device of Fig. 6–7, the source and drain contacts may be improved by further n^+ implantation in these regions. Because

[3] In many cases, a high-resistivity GaAs epitaxial layer (called a *buffer layer*) is grown between the two layers shown in Fig. 6–7.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 261

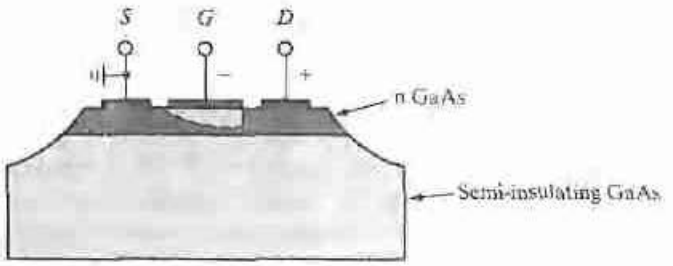


Figure 6–7 GaAs MESFET formed on an n-type GaAs layer grown epitaxially on a semi-insulating substrate. Common metals for the Schottky gate in GaAs are Al or alloys of Ti, W, and Au. The ohmic source and drain contacts may be an alloy of Au and Ge. In this example the device is isolated from others on the same chip by etching through the n region to the semi-insulating substrate.

of the relative simplicity of implanted GaAs MESFETs and the isolation between devices provided by the semi-insulating substrate, these structures are commonly used in GaAs integrated circuits.

6.3.2 The High Electron Mobility Transistor (HEMT)

Since the metal-semiconductor field effect transistor (MESFET) is compatible with the use of III–V compounds, it is possible to exploit the band gap engineering available with heterojunctions in these materials. In order to maintain high transconductance in a MESFET, the channel conductivity must be as high as possible. Obviously, the conductivity can be increased by increasing the doping in the channel and thus the carrier concentration. However, increased doping also causes increased scattering by the ionized impurities, which leads to a degradation of mobility (see Fig. 3-23). What is needed is a way of creating a high electron concentration in the channel of a MESFET by some means other than doping. A clever approach to this requirement is to grow a thin undoped well (e.g., GaAs) bounded by wider band gap, doped barriers (e.g., AlGaAs). This configuration, called *modulation doping*, results in conductive GaAs when electrons from the doped AlGaAs barriers fall into the well and become trapped there, as shown in Fig. 6-8a. Since the donors are in the AlGaAs rather than the GaAs, there is no impurity scattering of electrons in the well. If a MESFET is constructed with the channel along the GaAs well (perpendicular to the page in Fig. 6-8), we can take advantage of this reduced scattering and resulting higher mobility. The effect is especially strong at low temperature where lattice (phonon) scattering is also low. This device is called a *modulation doped field-effect transistor (MODFET)* and is also called a *high electron mobility transistor (HEMT)*.

In Fig. 6-8a we have left out the band-bending expected at the AlGaAs/GaAs interfaces. Based upon the discussion in Section 5.8, we expect the electrons to accumulate at the corners of the well due to band-bending at the heterojunction. In fact, only one heterojunction is required to trap electrons, as shown in Fig. 6-8b. Generally, the donors in the AlGaAs layer are purposely separated from the interface by ~100 Å. Using this configuration, we can achieve a high electron concentration in the channel while retaining high mobility, since the GaAs channel region is spatially separated from the ionized impurities which provide the free carriers.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 262

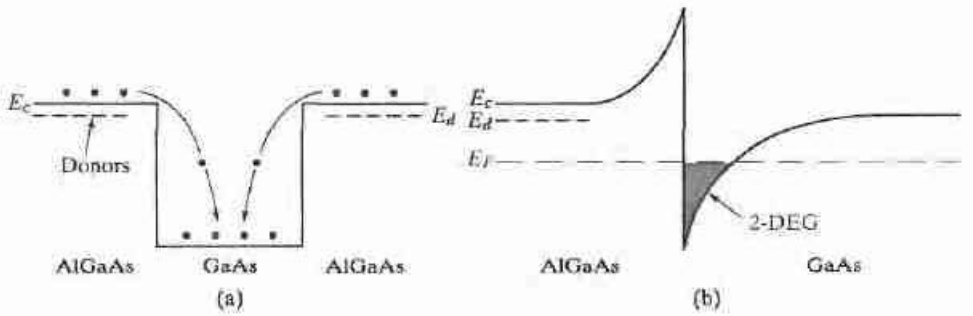


Figure 6–8
(a) Simplified view of modulation doping, showing only the conduction band. Electrons in the donor-doped AlGaAs fall into the GaAs potential well and become trapped. As a result, the undoped GaAs becomes n-type, without the scattering by ionized donors which is typical of bulk n-type material. (b) Use of a single AlGaAs/GaAs heterojunction to trap electrons in the undoped GaAs. The thin sheet of charge due to free electrons at the interface forms a two-dimensional electron gas (2-DEG), which can be exploited in HEMT devices.

In Fig. 6–8b, mobile electrons generated by the donors in the AlGaAs diffuse into the small band gap GaAs layer, and they are prevented from returning to the AlGaAs by the potential barrier at the AlGaAs/GaAs interface. The electrons in the (almost) triangular well form a two-dimensional electron gas (sometimes abbreviated 2-DEG). Sheet carrier densities as high as 10^{12} cm^{-2} can be obtained at a single interface such as that shown in Fig. 6–8b. Ionized impurity scattering is greatly reduced simply by separating the electrons from the donors. Also, screening effects due to the extremely high density of the two-dimensional electron gas can reduce ionized impurity scattering further. In properly designed structures, the electron transport approaches that of bulk GaAs with no impurities, so that mobility is limited by lattice scattering. As a result, mobilities above 250,000 cm^2/V-s at 77 K and 2,000,000 cm^2/V-s at 4 K can be achieved.

The advantages of a HEMT are its ability to locate a large electron density (~10^{12} cm^{-2}) in a very thin layer (<100 Å thick) very close to the gate while simultaneously eliminating ionized impurity scattering. The AlGaAs layer in a HEMT is fully depleted under normal operating conditions, and since the electrons are confined to the heterojunction, device behavior closely resembles that of a MOSFET. The advantages of the HEMT over the Si MOSFET are the higher mobility and maximum electron velocity in GaAs compared with Si, and the smoother interfaces possible with an AlGaAs/GaAs heterojunction compared with the Si/SiO_2 interface. The high performance of the HEMT translates into an extremely high cutoff frequency, and devices with fast access times.

Although we have discussed the HEMT in terms of the AlGaAs/GaAs heterojunction, other materials are also promising, such as the InGaAsP/InP

system. A motivation for avoiding $Al_xGa_{1-x}As$ is the presence of a deep-level defect called the DX center for $x > 0.2$, which traps electrons and impairs the HEMT operation. Since very thin layers are involved, materials with slight lattice mismatch can be grown to form *pseudomorphic* HEMTs. An example of such a system is the use of a thin layer of InGaAs grown pseudomorphically on GaAs, followed by AlGaAs. An advantage of this system is that a useful band discontinuity can be achieved using AlGaAs of low enough Al composition to avoid the DX center problem.

The HEMT, or MODFET, is also referred to as a *two-dimensional electron gas FET (2-DEG FET,* or *TEGFET)* to emphasize the fact that conduction along the channel occurs in a thin sheet of charge. The device has also been called a *separately doped FET (SEDFET),* to emphasize the fact that the doping occurs in a separate region from the channel.

6.3.3 Short Channel Effects

As mentioned in Section 6.2.3, a variety of modifications to the simple theory of JFET and MESFET operation must be made when the channel length is small (typically < 1 μm). In the past, these short-channel effects would be considered unusual, but now it is common to encounter FET devices in which these effects dominate the *I–V* characteristics. For example, high-field effects occur when 1 V appears across a channel length of 1 μm (10^{-4} cm), giving an electric field of 10 kV/cm.

A simple piecewise-linear approximation to the velocity-field curve assumes a constant mobility (linear) dependence up to some critical field $\mathscr{E}_c$ and a constant saturation velocity v_s for higher fields. For Si a better approximation is

$$v_d = \frac{\mu\mathscr{E}}{1 + \mu\mathscr{E}/v_s} \tag{6-13}$$

where μ is the low-field mobility. These two approximations are shown in Fig. 6–9a. If we assume that the electrons passing through the channel drift with a constant saturation velocity v_s, the current takes a simple form

$$I_D = qnv_sA = qN_dv_sZh \tag{6-14}$$

where h is a slow function of V_G. In this case the saturated current follows the velocity saturation, and does not require a true pinch-off in the sense of depletion regions meeting at some point in the channel. In the saturated velocity case, the transconductance g_m is essentially constant, in contrast with the constant mobility case described by Eq. (6–11). As shown in Fig. 6–9b, the $I_D - V_D$ curves are more evenly spaced if constant saturation velocity dominates, compared with the V_G-dependent spacing shown in Fig. 6–5b for the long-channel constant-mobility case.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 264

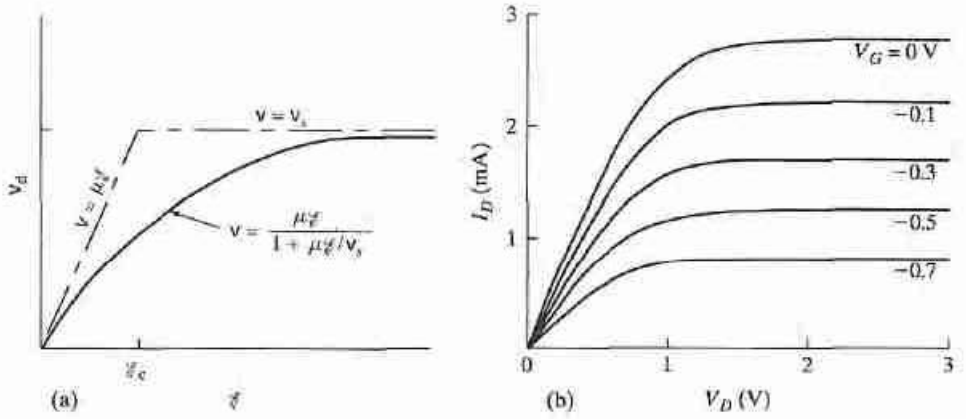


Figure 6–9
Effects of electron velocity saturation at high electric fields: (a) approximations to the saturation of drift velocity with increasing field; (b) drain current–voltage characteristics for the saturated velocity case, showing almost equally spaced curves with increasing gate voltage.

Most devices operate with characteristics intermediate between the constant mobility and the constant velocity regimes. Depending on the details of the field distribution, it is possible to divide up the channel into regions dominated by the two extreme cases, or to use an approximation such as Eq. (6–13).

Another important short-channel effect, described in Section 6.2.3, is the reduction in effective channel length after pinch-off as the drain voltage is increased. This effect is not significant in long-channel devices, since the change in L due to intrusion of the depletion region is a minor fraction of the total channel length. In short-channel devices, however, the effective channel length can be substantially shortened, leading to a slope in the saturated I–V characteristic that is analogous to the Early (base-width narrowing) effect in bipolar transistors discussed in Section 7.7.2.

6.4 THE METAL–INSULATOR–SEMICONDUCTOR FET

One of the most widely used electronic devices, particularly in digital integrated circuits, is the *metal–insulator–semiconductor (MIS) transistor*. In this device the channel current is controlled by a voltage applied at a gate electrode that is isolated from the channel by an insulator. The resulting device may be referred to generically as an insulated-gate field-effect transistor (IGFET). However, since most such devices are made using silicon for the semiconductor, SiO_2 for the insulator, and metal or heavily doped polysilicon for the gate electrode, the term *MOS field-effect transistor* (MOSFET) is commonly used.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 265

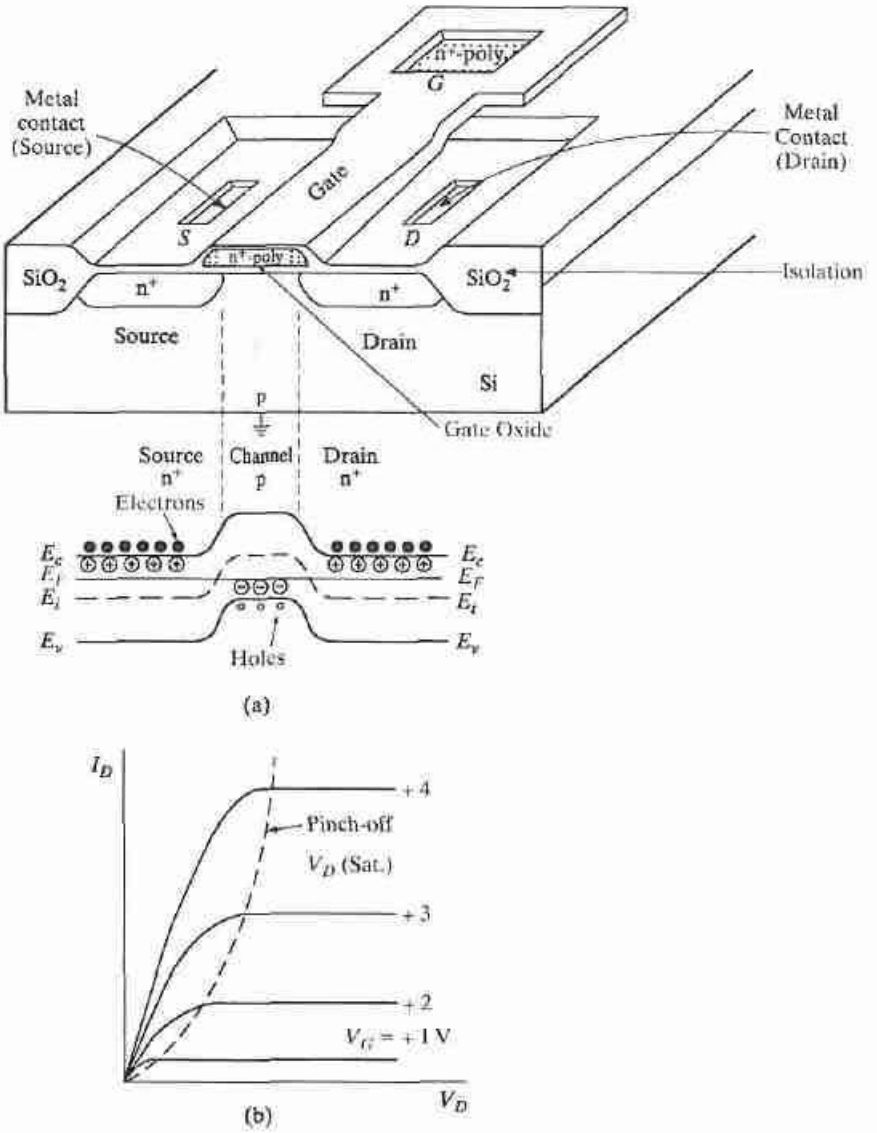


Figure 6–10 An enhancement-type n-channel MOSFET: (a) isometric view of device and equilibrium band diagram along channel; (b) drain current–voltage output characteristics as a function of gate voltage.

6.4.1 Basic Operation and Fabrication

The basic MOS transistor is illustrated in Fig. 6-10a for the case of an enhancement-mode n-channel device formed on a p-type Si substrate. The n^+ source and drain regions are diffused or implanted into a relatively lightly doped p-type substrate, and a thin oxide layer separates the conducting gate from the Si surface. No current flows from drain to source without a conducting n channel between them. This can be understood clearly by

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 266

looking at the band diagram of the MOSFET in equilibrium along the channel (Fig. 6–10a). The Fermi level is flat in equilibrium. The conduction band is close to the Fermi level in the n^+ source/drain, while the valence band is closer to the Fermi level in the p-type material. Hence, there is a potential barrier for an electron to go from the source to the drain, corresponding to the built-in potential of the back-to-back p-n junctions between the source and drain.

When a positive voltage is applied to the gate relative to the substrate (which is connected to the source in this case), positive charges are in effect deposited on the gate metal. In response, negative charges are induced in the underlying Si, by the formation of a depletion region and a thin surface region containing mobile electrons. These induced electrons form the channel of the FET, and allow current to flow from drain to source. As Fig. 6–10b suggests, the effect of the gate voltage is to vary the conductance of this induced channel for low drain-to-source voltage, analogous to the JFET case. Since electrons are electrostatically induced in the p-type channel region, the channel becomes less p-type, and therefore the valence band moves down, farther away from the Fermi level. This obviously reduces the barrier for electrons between the source, the channel, and the drain. If the barrier is reduced sufficiently by applying a gate voltage in excess of what is known as the *threshold voltage*, V_T, there is significant current flow from the source to the drain. Thus, one view of a MOSFET is that it is a gate-controlled potential barrier. It is very important to have high-quality, low-leakage p-n junctions in order to ensure a low off-state leakage in the MOSFET. For a given value of V_G there will be some drain voltage V_D for which the current becomes saturated, after which it remains essentially constant.

The threshold voltage V_T is the minimum gate voltage required to induce the channel. In general, the positive gate voltage of an n-channel device (such as that shown in Fig. 6–11) must be larger than some value V_T before a conducting channel is induced. Similarly, a p-channel device (made on an n-type substrate with p-type source and drain implants or diffusions) requires a gate voltage more negative than some threshold value to induce the required positive charge (mobile holes) in the channel. There are exceptions to this general rule, however, as we shall see. For example, some n-channel devices have a channel already with zero gate voltage, and in fact a negative gate voltage is required to turn the device off. Such a "normally on" device is called a *depletion-mode* transistor, since gate voltage is used to deplete a channel which exists at equilibrium. The more common MOS transistor is "normally off" with zero gate voltage, and operates in the *enhancement mode* by applying gate voltage large enough to induce a conducting channel.

An alternative view of a MOSFET is that it is a gate-controlled resistor. If the (positive) gate voltage exceeds the threshold voltage in an n-channel device, electrons are induced in the p-type substrate. Since this channel is connected to the n^+ source and drain regions, the structure looks electrically like an induced n-type resistor. As the gate voltage increases, more electron

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 267

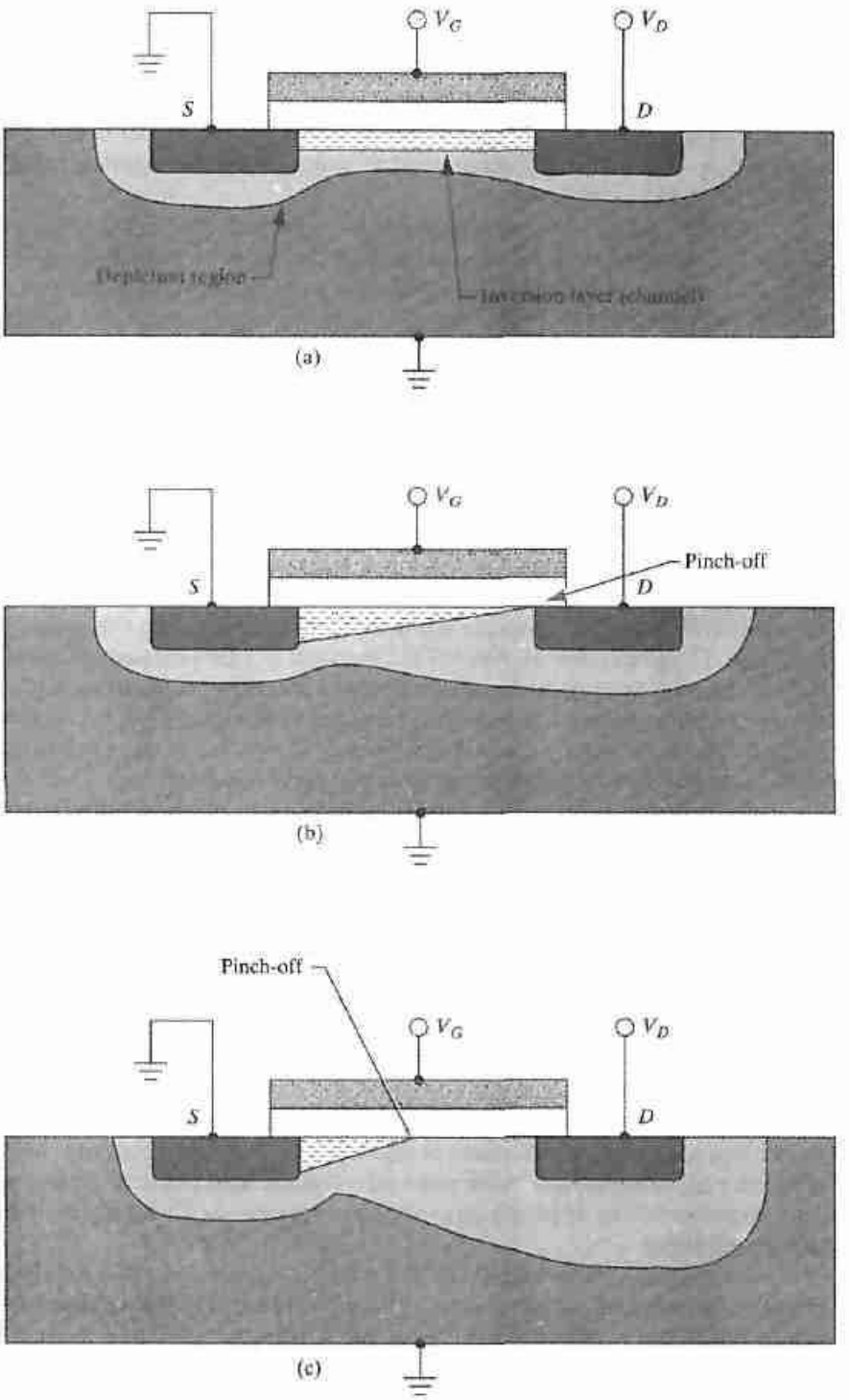


Figure 6–11 n-channel MOSFET cross-sections under different operating conditions: (a) linear region for $V_G > V_T$ and $V_D < (V_G - V_T)$; (b) onset of saturation at pinch-off, $V_G > V_T$ and $V_D = (V_G - V_T)$; (c) strong saturation, $V_G > V_T$ and $V_D > (V_G - V_T)$.

charge is induced in the channel and, therefore, the channel becomes more conducting. The drain current initially increases linearly with the drain bias (the *linear* regime) (Fig. 6-10b). As more drain current flows in the channel, however, there is more ohmic voltage drop along the channel such that the channel potential varies from zero near the grounded source to whatever the applied drain potential is near the drain end of the channel. Hence, the voltage difference between the gate and the channel reduces from V_G near the source to $(V_G - V_D)$ near the drain end. Once the drain bias is increased to the point that $(V_G - V_D) = V_T$, threshold is barely maintained near the drain end, and the channel is said to be pinched off. Increasing the drain bias beyond this point (V_D(sat)) causes the point at which the channel gets pinched off to move more and more into the channel, closer to the source end (Fig. 6-11c). Electrons in the channel are pulled into the pinch-off region and travel at the saturation drift velocity because of the very high longitudinal electric field along the channel. Now, the drain current is said to be in the *saturation* region because it does not increase with drain bias significantly (Fig. 6-10b). Actually, there is a slight increase of drain current with drain bias due to various effects such as channel length modulation and drain-induced barrier lowering (DIBL) that will be discussed in Section 6.5.10.

The MOS transistor is particularly useful in digital circuits, in which it is switched from the "off" state (no conducting channel) to the "on" state. The control of drain current is obtained at a gate electrode which is insulated from the source and drain by the oxide. Thus the d-c input impedance of an MOS circuit can be very large.

Both n-channel and p-channel MOS transistors are in common usage. The n-channel type illustrated in Fig. 6-10 is generally preferred because it takes advantage of the fact that the electron mobility in Si is larger than the mobility of holes. In much of the discussion to follow we will use the n-channel (p-type substrate) example, although the p-channel case will be kept in mind also.

Let us give a very simplified description of how such an n-channel MOSFET can be fabricated. A much more detailed discussion is given in Section 9.3.1. An ultra-thin (~5–10 nm) dry thermal silicon dioxide is grown on the p-type substrate. This serves as the gate insulator between the conducting gate and the channel. We immediately cover it with LPCVD of polysilicon, which is doped very heavily n^+ using P diffusion in order to make it behave electrically like a metal electrode. The doped polysilicon layer is then patterned to form the gates, and etched anisotropically by RIE to achieve vertical walls (Section 5.1.7). The gate itself is used as an implant mask for an n^+ implant which forms the source/drain junctions abutted to the gate edges, but is blocked from the channel region. Such a scheme is called a *self-aligned process* because we did not have to use a separate lithography step for the source/drain formation. Self-alignment is simple and is very useful because we thereby guarantee that there will be some overlap of the gate with the source/drain but not too much overlap. The advantages of this are discussed in Section 6.5.8. The implanted dopants must be annealed for reasons discussed

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 269

in Section 5.1.4. Finally, the MOSFETs have to be properly interconnected according to the circuit layout, using metallization. This involves LPCVD of an oxide dielectric, etching contact holes by RIE, sputter depositing a suitable metal such as Al, patterning and etching it.

As shown in Fig. 6-10a, the MOSFET is surrounded on all sides by a thick SiO_2 layer. This layer provides critical electrical isolation between adjacent transistors on an integrated circuit. We shall see in Section 9.3.1 that such *isolation* or *field* regions can be formed in several ways, such as *LOCal Oxidation of Silicon (LOCOS)*. Briefly, it involves depositing a LPCVD Si_3N_4 layer over the entire substrate before the fabrication of the MOSFETs, patterning and etching it so that it is removed only in the isolation regions, but not in the *active* regions where the MOSFETs will be formed subsequently. A boron *channel stop* implant is then done in the isolation regions. Exploiting the useful property of Si_3N_4 that it blocks thermal oxidation, a thick LOCOS oxide is selectively grown by wet oxidation only in the isolation regions. The reason why a thick field oxide layer and a boron channel stop implant leads to electrical isolation is discussed in Section 6.5.5.

6.4.2 The Ideal MOS Capacitor

The surface effects that arise in an apparently simple MOS structure are actually quite complicated. Although many of these effects are beyond the scope of this discussion, we will be able to identify those which control typical MOS transistor operation. We begin by considering an uncomplicated idealized case, and then include effects encountered in real surfaces in the next section.

Some important definitions are made in the energy band diagram of Fig. 6-12a. The work function characteristic of the metal (see Section 2.2.1) can be defined in terms of the energy required to move an electron from the Fermi level to outside the metal. In MOS work it is more convenient to use a *modified work function* $q\Phi_m$ for the metal–oxide interface. The energy $q\Phi_m$ is measured from the metal Fermi level to the conduction band of the oxide.[4] Similarly, $q\Phi_s$ is the modified work function at the semiconductor–oxide interface. In this idealized case we assume that $\Phi_m = \Phi_s$, so there is no difference in the two work functions. Another quantity that will be useful in later discussions is $q\phi_{fv}$ which measures the position of the Fermi level below the intrinsic level E_i for the semiconductor. This quantity indicates how strongly p-type the semiconductor is [see Eq. (3-25)].

The MOS structure of Fig. 6-12a is essentially a capacitor in which one plate is a semiconductor. If we apply a negative voltage between the metal and the semiconductor (Fig. 6-12b), we effectively deposit a negative charge on the metal. In response, we expect an equal net positive charge to accumulate

[4]On the MOS band diagrams of this section, we show a break in the electron energy scale leading to the insulator conduction band, since the band gap of SiO_2 (or other typical insulators) is much greater than that of the Si.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 270

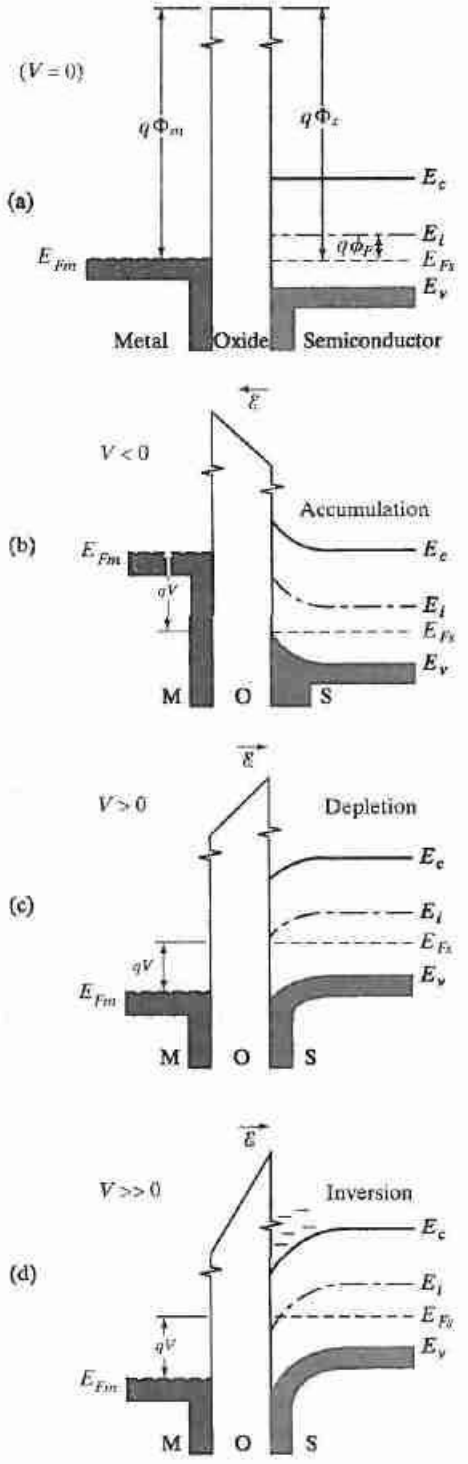


Figure 6–12
Band diagram for the ideal MOS structure at: (a) equilibrium; (b) negative voltage causes hole accumulation in the p-type semiconductor; (c) positive voltage depletes holes from the semiconductor surface; (d) a larger positive voltage causes inversion—an "n-type" layer at the semiconductor surface.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 271

at the surface of the semiconductor. In the case of a p-type substrate this occurs by *hole accumulation* at the semiconductor-oxide interface.

Since the applied negative voltage *depresses* the electrostatic potential of the metal relative to the semiconductor, the electron energies are *raised* in the metal relative to the semiconductor.[5] As a result, the Fermi level for the metal E_{Fm} lies above its equilibrium position by qV, where V is the applied voltage.

Since Φ_m and Φ_s do not change with applied voltage, moving E_{Fm} up in energy relative to E_{Fs} causes a tilt in the oxide conduction band. We expect such a tilt since an electric field causes a gradient in E_i (and similarly in E_v and E_c) as described in Section 4.4.2:

$$\mathscr{E}(x) = \frac{1}{q}\frac{dE_i}{dx} \qquad \text{(see 4-26)}$$

The energy bands of the semiconductor bend near the interface to accommodate the accumulation of holes. Since

$$p = n_i e^{(E_i - E_F)/kT} \qquad \text{(see 3-25)}$$

it is clear that an increase in hole concentration implies an increase in $E_i - E_F$ at the semiconductor surface.

Since no current passes through the MOS structure, there can be no variation in the Fermi level within the semiconductor. Therefore, if $E_i - E_F$ is to increase, it must occur by E_i moving up in energy near the surface. The result is a bending of the semiconductor bands near the interface. We notice in Fig. 6-12b that the Fermi level near the interface lies closer to the valence band, indicating a larger hole concentration than that arising from the doping of the p-type semiconductor.

In Fig. 6-12c we apply a positive voltage from the metal to the semiconductor. This raises the potential of the metal, lowering the metal Fermi level by qV relative to its equilibrium position. As a result, the oxide conduction band is again tilted. We notice that the slope of this band, obtained by simply moving the metal side down relative to the semiconductor side, is in the proper direction for the applied field, according to Eq. (4-26).

The positive voltage deposits positive charge on the metal and calls for a corresponding net negative charge at the surface of the semiconductor. Such a negative charge in p-type material arises from *depletion* of holes from the region near the surface, leaving behind uncompensated ionized acceptors. This is analogous to the depletion region at a p-n junction discussed in Section 5.2.3. In the depleted region the hole concentration decreases, moving E_i closer to E_F, and bending the bands down near the semiconductor surface.

If we continue to increase the positive voltage, the bands at the semiconductor surface bend down more strongly. In fact, a sufficiently large voltage can

[5]Recall that an electrostatic potential diagram is drawn for positive test charges, in contrast to an electron energy diagram, which is drawn for negative charges.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 272

bend E_i *below* E_F (Fig. 6–12d). This is a particularly interesting case, since $E_F \gg E_i$ implies a large electron concentration in the conduction band.

The region near the semiconductor surface in this case has conduction properties typical of n-type material, with an electron concentration given by Eq. (3–25a). This n-type surface layer is formed not by doping, but instead by *inversion* of the originally p-type semiconductor due to the applied voltage. This inverted layer, separated from the underlying p-type material by a depletion region, is the key to MOS transistor operation.

We should take a closer look at the inversion region, since it becomes the conducting channel in the FET. In Fig. 6–13 we define a potential ϕ at any point x, measured relative to the equilibrium position of E_i. The energy $q\phi$ tells us the extent of band-bending at x, and $q\phi_s$ represents the band-bending at the surface. We notice that $\phi_s = 0$ is the *flat band* condition for this ideal MOS case (i.e., the bands look like Fig. 6–12a). When $\phi_s < 0$, the bands bend up at the surface, and we have hole accumulation (Fig. 6–12b). Similarly, when $\phi_s > 0$, we have depletion (Fig. 6–12c). Finally, when ϕ_s is positive and larger than ϕ_F, the bands at the surface are bent down such that $E_i(x = 0)$ lies below E_F, and inversion is obtained (Fig. 6–12d).

While it is true that the surface is inverted whenever ϕ_s is larger than ϕ_F, a practical criterion is needed to tell us whether a true n-type conducting channel exists at the surface. The best criterion for *strong inversion* is that the surface should be as strongly n-type as the substrate is p-type. That is, E_i should lie as far below E_F at the surface as it is above E_F far from the surface. This condition occurs when

$$\phi_s(\text{inv.}) = 2\phi_F = 2\frac{kT}{q}\ln\frac{N_a}{n_i} \tag{6-15}$$

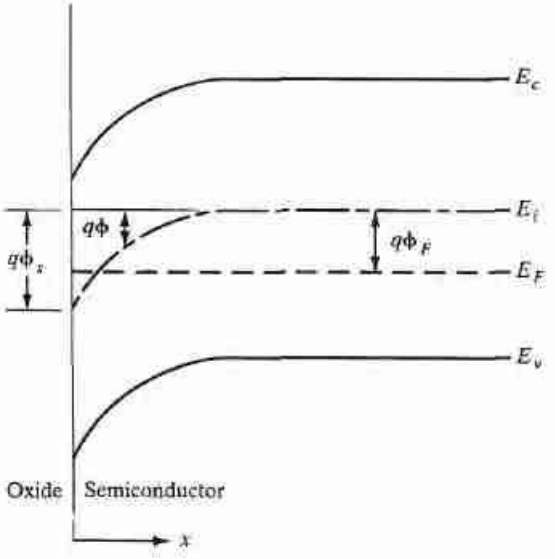


Figure 6–13
Bending of the semiconductor bands at the onset of strong inversion: the surface potential ϕ_s is twice the value of ϕ_F in the neutral p material.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 273

A surface potential of ϕ_F is required to bend the bands down to the intrinsic condition at the surface ($E_i = E_F$), and E_i must then be depressed another $q\phi_F$ at the surface to obtain the condition we call strong inversion.

The electron and hole concentrations are related to the potential $\phi(x)$ defined in Fig. 6-13. Since the equilibrium electron concentration is

$$n_0 = n_i e^{(E_F - E_i)/kT} = n_i e^{-q\phi_F/kT} \tag{6-16}$$

we can easily relate the electron concentration at any x to this value:

$$n = n_i e^{-q(\phi_F - \phi)/kT} = n_0 e^{q\phi/kT} \tag{6-17}$$

and similarly for holes:

$$p_0 = n_i e^{q\phi_F/kT} \tag{6-18a}$$

$$p = p_0 e^{-q\phi/kT} \tag{6-18b}$$

at any x. We could combine these equations with Poisson's equation (6-19) and the usual charge density expression (6-20) to solve for $\phi(x)$:

$$\frac{\partial^2 \phi}{\partial x^2} = -\frac{\rho(x)}{\epsilon_s} \tag{6-19}$$

$$\rho(x) = q(N_d^+ - N_a^- + p - n) \tag{6-20}$$

Let us solve this equation to determine the total integrated charge per unit area, Q_s, as a function of the surface potential, ϕ_s. Substituting Eqs. (6-16), (6-17), and (6-18) for the electron and hole concentrations in Eqs. (6-19) and (6-20), we get

$$\frac{\partial^2 \phi}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial \phi}{\partial x}\right) = -\frac{q}{\epsilon_s}\left[p_0\left(e^{\frac{-q\phi}{kT}} - 1\right) - n_0\left(e^{\frac{q\phi}{kT}} - 1\right)\right] \tag{6-21}$$

It should be kept in mind that

$$\frac{-\partial \phi}{\partial x}$$

is the electric field, $\mathscr{E}$, at a depth x.

Integrating Eq. (6-21) from the bulk (where the bands are flat, the electric fields are zero, and the carrier concentrations are determined solely by the doping), towards the surface, we get

$$\int_0^{\frac{\partial \phi}{\partial x}} \left(\frac{\partial \phi}{\partial x}\right) d\left(\frac{\partial \phi}{\partial x}\right) = -\frac{q}{\epsilon_s}\int_0^{\phi}\left[p_0\left(e^{\frac{-q\phi}{kT}} - 1\right) - n_0\left(e^{\frac{q\phi}{kT}} - 1\right)\right] d\phi \tag{6-22}$$

After integration, we then get

$$\mathscr{E}^2 = \left(\frac{2kT\, p_0}{\epsilon_s}\right)\left[\left(e^{\frac{q\phi}{kT}} + \frac{q\phi}{kT} - 1\right) + \frac{n_0}{p_0}\left(e^{\frac{q\phi}{kT}} - \frac{q\phi}{kT} - 1\right)\right] \tag{6-23}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 274

A particularly important case is at the surface ($x = 0$) where the surface perpendicular electric field, $\mathscr{E}_s$, becomes

$$\mathscr{E}_s = \frac{\sqrt{2}kT}{qL_D}\left[\left(e^{-\frac{q\phi_s}{kT}} + \frac{q\phi_s}{kT} - 1\right) + \frac{n_0}{p_0}\left(e^{\frac{q\phi_s}{kT}} - \frac{q\phi_s}{kT} - 1\right)\right]^{\frac{1}{2}} \quad (6\text{-}24)$$

where we have introduced a new term, the *Debye screening length*,

$$L_D = \sqrt{\frac{\epsilon_s kT}{q^2 p_0}} \quad (6\text{-}25)$$

The Debye length is a very important concept in semiconductors. It gives us an idea of the distance scale in which charge imbalances are screened or smeared out. For example, if we think of inserting a positively charged sphere in an n-type semiconductor we know that the mobile electrons will crowd around the sphere. If we move away from the sphere by several Debye lengths, the positively charged sphere and the negative electron cloud will look like a neutral entity. Not surprisingly, L_D depends inversely on doping because the higher the carrier concentration, the more easily screening takes place. For n-type material we should use n_0 in Eq. (6-25).

By using Gauss's law at the surface, we can relate the integrated space charge per unit area to the electric displacement, keeping in mind that the electric field or displacement deep in the substrate is zero.

$$Q_s = -\epsilon_s \mathscr{E}_s \quad (6\text{-}26)$$

The space charge density per unit area Q_s in Eq. (6-26) is plotted as a function of the surface potential ϕ_s in Fig. 6-14. We see from Eq. (6-24) and Fig. 6-14 that when the surface potential is zero (flat band conditions), the net space charge is zero. This is because the fixed dopant charges are cancelled by the mobile carrier charges at flat band. When the surface potential is negative, it attracts and forms an accumulation layer of the majority carrier holes at the surface. The first term in Eq. (6-24) is the dominant one, and the accumulation space charge increases very strongly (exponentially) with negative surface potential. It is easy to see why by looking at Eq. (6-18), which gives the surface hole concentration in a p-type semiconductor as a function of surface potential. Since the band-bending decreases as a function of depth, the integrated accumulation charge involves averaging over depth and introduces a factor of 2 in the exponent. Mathematically, this is due to the square root in Eq. (6-24). It must be noted that since this charge is due to the mobile majority carriers (holes in this case), the charge piles up near the oxide-silicon interface, since typical accumulation layer thicknesses are ~20 nm. Also, because of the exponential dependence of accumulation charge on surface potential, the band-bending is generally small or is said to be *pinned* to nearly zero.

On the other hand, for a positive surface potential, we see from Eq. (6-24) that initially the second (linear) term is the dominant one. Although

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 275

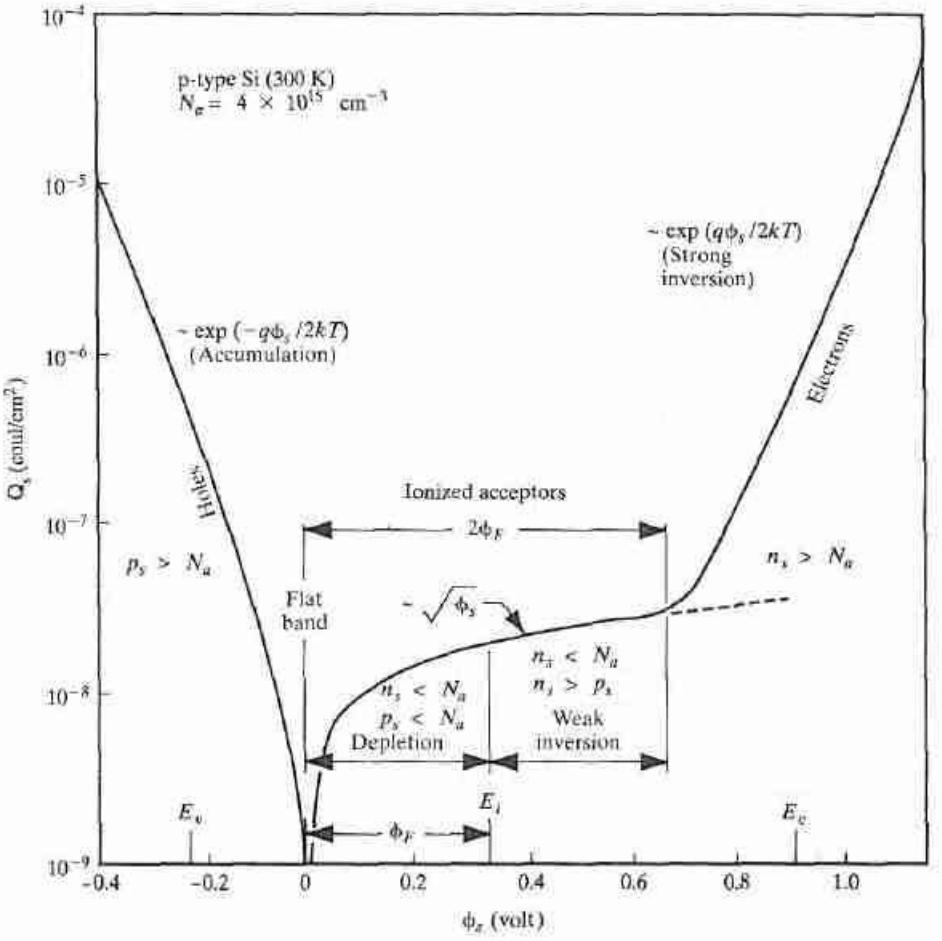


Figure 6–14
Variation of space-charge density in the semiconductor as a function of the surface potential ϕ_s for p-type silicon with $N_a = 4 \times 10^{15}$ cm^{-3} at room temperature. p_s and n_s are the hole and electron concentrations at the surface, ϕ_F is the potential difference between the Fermi level and the intrinsic level of the bulk. (Garrett and Brattain, Phys. Rev., 99, 376 (1955).)

the exponential term $\exp(q\phi_s/kT)$ is very large, it is multiplied by the ratio of the minority to majority carrier concentration which is very small, and is initially negligible. Hence, the space charge for small positive surface potentials increases as $\sim\sqrt{\phi_s}$ as shown in Fig. 6-14. As discussed in detail later in this section, this corresponds to the depletion region charge due to the exposed, fixed immobile dopants (acceptors in this case). The depletion width typically extends over several hundred nm. At some point, the band-bending is twice the Fermi potential ϕ_F, which is enough for the onset of strong inversion. Now, the exponential term $\exp(q\phi_s(\text{inv.})/kT)$ multiplied by the minority carrier concentration n_0 is equal to the majority carrier concentration

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 276

p_0. Hence, for band bending beyond this point, it becomes the dominant term. As in the case of accumulation, the mobile inversion charge now increases very strongly with bias, as indicated by Eq. (6-17) and shown in Fig. 6-14. The typical inversion layer thicknesses are ~5 nm, and the surface potential now is essentially pinned at $2\phi_F$.

It may be pointed out that in accumulation, and especially in inversion, the carriers are confined in the x-direction in narrow, essentially triangular potential wells, causing quantum mechanical particle-in-a-box states or subbands, similar to those discussed in Chapter 2. However, the carriers are free in the other directions (parallel to the oxide-silicon interface). This leads to a two-dimensional electron gas (2DEG) or hole gas, with a "staircase" constant density of states, as discussed in Appendix IV. The detailed analysis of these effects is, unfortunately, beyond the scope of our discussion here.

The charge distribution, electric field, and electrostatic potential for the inverted surface are sketched in Fig. 6-15. For simplicity we use the depletion approximation of Chapter 5 in this figure, assuming complete depletion for $0 < x < W$, and neutral material for $x > W$. In this approximation the charge per unit area[6] due to uncompensated acceptors in the depletion region is $-qN_aW$. The positive charge Q_m on the metal is balanced by the negative charge Q_s in the semiconductor, which is the depletion layer charge plus the charge due to the inversion region Q_n:

$$Q_m = -Q_s = qN_aW - Q_n \tag{6-27}$$

The width of the inversion region is exaggerated in Fig. 6-15 for illustrative purposes. Actually, the width of this region is generally less than 100 Å. Thus we have neglected it in sketching the electric field and potential distribution. In the potential distribution diagram we see that an applied voltage V appears partially across the insulator (V_i) and partially across the depletion region of the semiconductor (ϕ_s):

$$V = V_i + \phi_s \tag{6-28}$$

The voltage across the insulator is obviously related to the charge on either side, divided by the capacitance:

$$V_i = \frac{-Q_s d}{\epsilon_i} = \frac{-Q_s}{C_i} \tag{6-29}$$

where ϵ_i is permittivity of the insulator and C_i is the insulator capacitance per unit area. The charge Q_s will be negative for the n channel, giving a positive V_i.

Using the depletion approximation, we can solve for W as a function of ϕ_s (Prob. 6.7). The result is the same as would be obtained for an n^+-p junction in Chapter 5, for which the depletion region extends almost entirely into the p region:

[6]In this chapter, we will use charge per unit area (Q) and capacitance per unit area (C) to avoid repeating A throughout the discussion.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 277

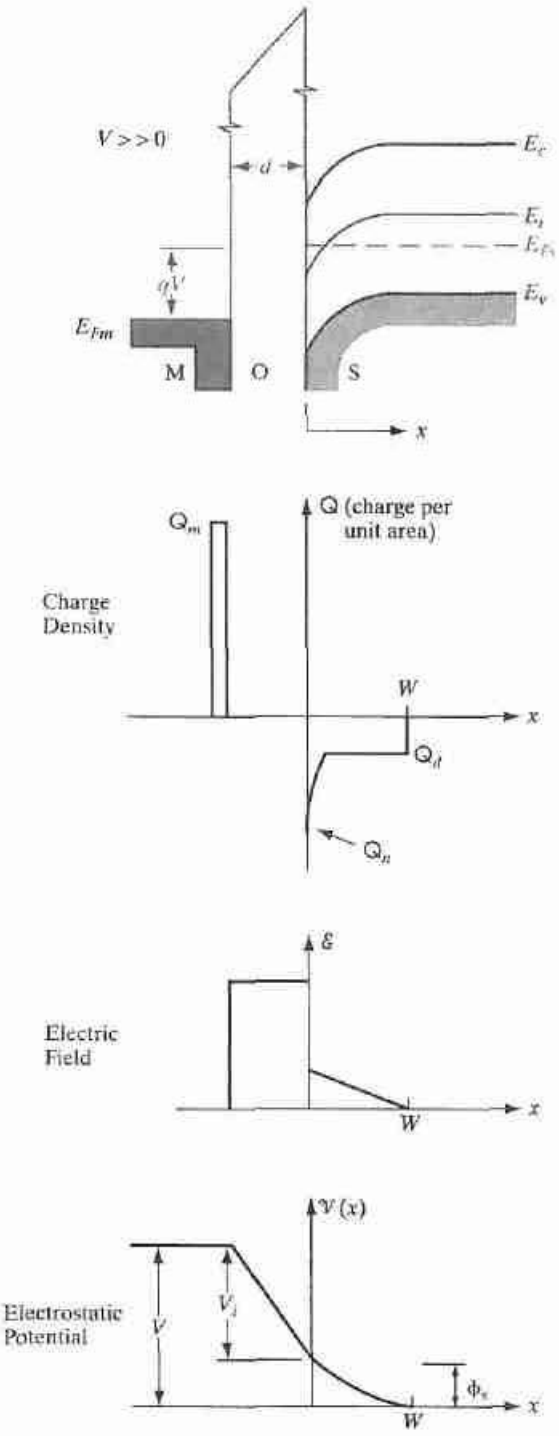


Figure 6–15
Approximate distributions of charge, electric field, and electrostatic potential in the ideal MOS capacitor in inversion. The relative width of the inverted region is exaggerated for illustrative purposes, but is neglected in the field and potential diagrams.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 278

$$W = \left[\frac{2\epsilon_s \phi_s}{qN_a}\right]^{1/2} \tag{6-30}$$

This depletion region grows with increased voltage across the capacitor until strong inversion is reached. After that, further increases in voltage result in stronger inversion rather than in more depletion. Thus the maximum value of the depletion width is

$$W_m = \left[\frac{2\epsilon_s \phi_s(\text{inv.})}{qN_a}\right]^{1/2} = 2\left[\frac{\epsilon_s kT \ln(N_a/n_i)}{q^2 N_a}\right]^{1/2} \tag{6-31}$$

using Eq. (6-15). We know the quantities in this expression, so W_m can be calculated.

The charge per unit area in the depletion region Q_d at strong inversion is[7]

$$Q_d = -qN_a W_m = -2(\epsilon_s q N_a \phi_F)^{1/2} \tag{6-32}$$

The applied voltage must be large enough to create this depletion charge plus the surface potential ϕ_s(inv.). The *threshold* voltage required for strong inversion, using Eqs. (6-15), (6-28), and (6-29), is

$$V_T = -\frac{Q_d}{C_i} + 2\phi_F \quad (\textit{ideal case}) \tag{6-33}$$

This assumes the negative charge at the semiconductor surface Q_s at inversion is mostly due to the depletion charge Q_d. The threshold voltage represents the minimum voltage required to achieve strong inversion, and is an extremely important quantity for MOS transistors. We will see in the next section that other terms must be added to this expression for real MOS structures.

The capacitance–voltage characteristics of this ideal MOS structure (Fig. 6-16) vary depending on whether the semiconductor surface is in accumulation, depletion, or inversion.

Since the capacitance for MOSFETs is voltage dependent, we must use the more general expression in Eq. (5-55) for the voltage-dependent semiconductor capacitance,

$$C_s = \frac{dQ}{dV} = \frac{dQ_s}{d\phi_s} \tag{6-34}$$

Actually, if one looks at the electrical equivalent circuit of a MOS capacitor or MOSFET, it is the series combination of a fixed, voltage-independent gate oxide (insulator) capacitance, and a voltage-dependent semiconductor capacitance [defined according to Eq. (6-34)], such that the overall MOS capacitance becomes voltage dependent. The semiconductor capacitance itself can be determined from the slope of the Q_s versus ϕ_s plot (Fig. 6-14). It is

[7] In the p-channel (n-type substrate) case, for which ϕ_F is negative, we use $Q_d = +qN_a W_m = 2(\epsilon_s q N_a |\phi_F|)^{1/2}$.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 279

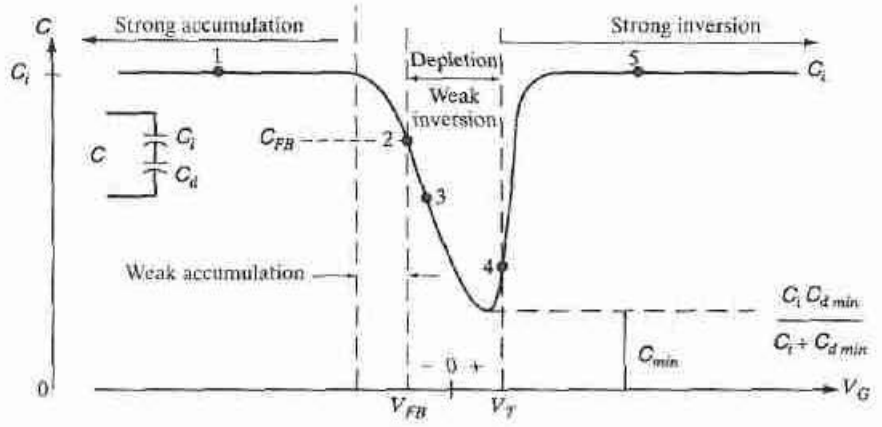


Figure 6–16
Capacitance–voltage relation for an n-channel (p-substrate) MOS capacitor. The dashed curve for $V > V_T$ is observed at high measurement frequencies. The flat band voltage V_{FB} will be discussed in Section 6.4.3. When the semiconductor is in depletion, the semiconductor capacitance C_s is denoted as C_d.

clear that the semiconductor capacitance in accumulation is very high because the slope is so steep; i.e., the accumulation charge changes a lot with surface potential. Hence, the series capacitance in accumulation is basically the insulator capacitance, C_i. Since, for negative voltage, holes are accumulated at the surface (Fig. 6-12b), the MOS structure appears almost like a parallel-plate capacitor, dominated by the insulator properties $C_i = \epsilon_i/d$ (point 1 in Fig. 6-16). As the voltage becomes less negative, the semiconductor surface is depleted. Thus a depletion-layer capacitance C_d is added in series with C_i:

$$C_i = \frac{\epsilon_s}{W} \tag{6-35}$$

where ϵ_s is the semiconductor permittivity and W is the width of the depletion layer from Eq. (6-30). The total capacitance is

$$C = \frac{C_i C_d}{C_i + C_d} \tag{6-36}$$

The capacitance decreases as W grows from flatband (point 2), past weak inversion (point 3), until finally strong inversion is reached at V_T (point 4). In the depletion region, the small signal semiconductor capacitance is given by the same formula [Eq. (6-34)] which gives the variation of the (depletion) space charge with surface potential. Since the charge increases as $-\sqrt{\phi_s}$, the depletion capacitance will obviously decrease as $1/\sqrt{\phi_s}$, exactly as for the depletion capacitance of a p-n junction [see Eq. (5-63)].

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 280

After inversion is reached, the small signal capacitance depends on whether the measurements are made at high (typically ~1 MHz) or low (typically ~1–100 Hz) frequency, where "high" and "low" are with respect to the generation-recombination rate of the minority carriers in the inversion layer. If the gate voltage is varied rapidly, the charge in the inversion layer cannot change in response, and thus does not contribute to the small signal a-c capacitance. Hence, the semiconductor capacitance is at a minimum, corresponding to a maximum depletion width.

On the other hand, if the gate bias is changed slowly, there is time for minority carriers to be generated in the bulk, drift across the depletion region to the inversion layer, or go back to the substrate and recombine. Now, the semiconductor capacitance, using the same Eq. (6-34), is very large because we saw in Fig. 6–14 that the inversion charge increases exponentially with ϕ_s. Hence, the low frequency MOS series capacitance in strong inversion is basically C_i once again (point 5).

What is the frequency dependence of the capacitance in accumulation (Fig. 6–12a)? We get a very high capacitance both at low and high frequencies because the majority carriers in the accumulation layer can respond much faster than minority carriers. While minority carriers respond on the time scale of generation-recombination times (typically hundreds of microseconds in Si), majority carriers respond on the time scale of the dielectric relaxation time, $\tau_D = \rho\epsilon$, where ρ is the resistivity and ϵ is the permittivity. τ_D is analogous to the RC time constant of a system, and is small for the majority carriers ($\sim 10^{-13}$ s). As an interesting aside, it may be pointed out that in inversion, although the high-frequency capacitance for MOS capacitors is low, it is high ($= C_i$) for MOSFETs because now the inversion charge can flow in readily and very fast ($\sim\tau_D$) from the source/drain regions rather than having to be created by generation-recombination in the bulk.

6.4.3 Effects of Real Surfaces

When MOS devices are made using typical materials (e.g., n^+ polysilicon-SiO_2-Si), departures from the ideal case described in the previous section can strongly affect V_T and other properties. First, there is a work function difference between the doped polysilicon gate and substrate, which depends on the substrate doping. Here, the heavily doped polysilicon acts as a metal electrode. Second, there are inevitably charges at the Si-SiO_2 interface and within the oxide which must be taken into account.

Work Function Difference. We expect Φ_s to vary depending on the doping of the semiconductor. Figure 6–17 illustrates the work function potential difference $\Phi_{ms} = \Phi_m - \Phi_s$ for n^+ polysilicon on Si as the doping is varied. We note that Φ_{ms} is always negative for this case, and is most negative for heavily doped p-type Si (i.e., for E_F close to the valence band).

If we try to construct an equilibrium diagram with Φ_{ms} negative (Fig. 6–18a), we find that in aligning E_F we must include a tilt in the oxide conduction band

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 281

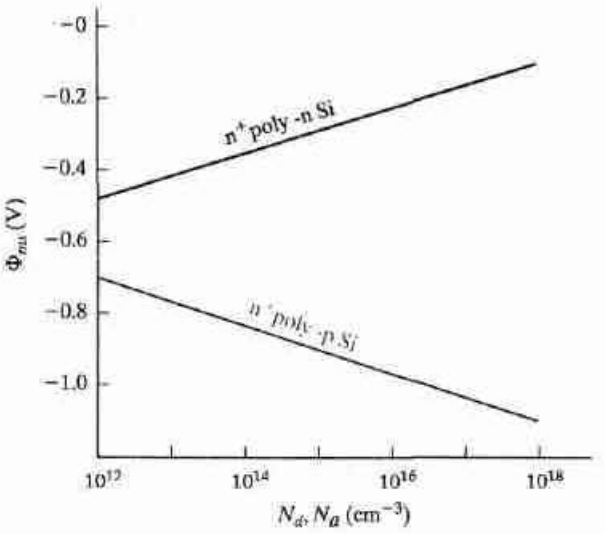


Figure 6–17 Variation of the metal–semiconductor work function potential difference Φ_{ms} with substrate doping concentration, for n^+ poly-Si.

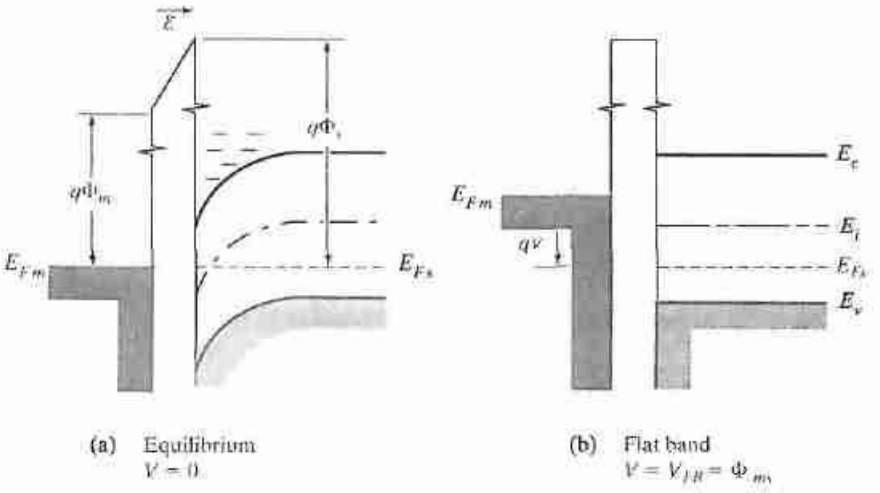


Figure 6–18 Effect of a negative work function difference ($\Phi_{ms} < 0$): (a) band bending and formation of negative charge at the semiconductor surface; (b) achievement of the flat band condition by application of a negative voltage.

(implying an electric field). Thus the metal is positively charged and the semiconductor surface is negatively charged at equilibrium, to accommodate the work function difference. As a result, the bands bend down near the semiconductor surface. In fact, if Φ_{ms} is sufficiently negative, an inversion region can exist with no external voltage applied. To obtain the *flat band* condition pictured in Fig. 6–18b, we must apply a negative voltage to the metal ($V_{FB} = \Phi_{ms}$).

Interface Charge. In addition to the work function difference, the equilibrium MOS structure is affected by charges in the insulator and at the semiconductor–oxide interface (Fig. 6–19). For example, alkali metal ions (particularly Na^+) can be incorporated inadvertently in the oxide during

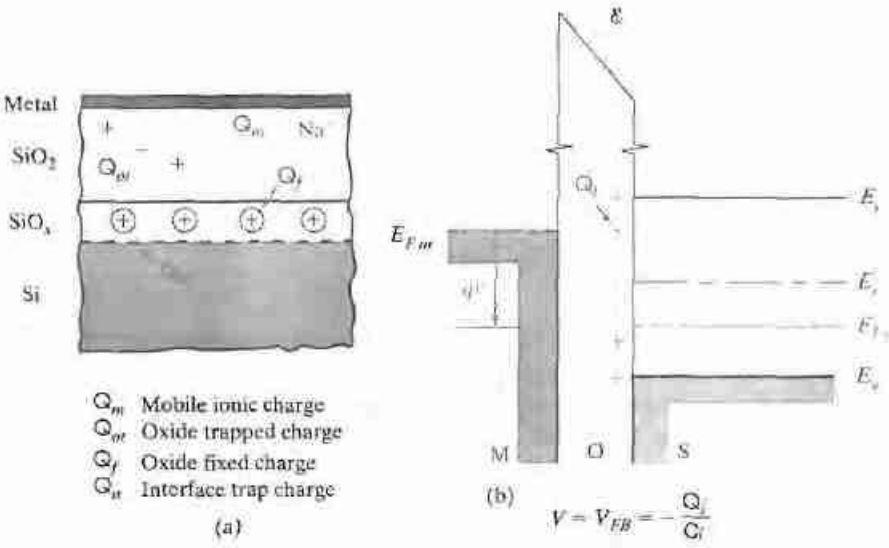


Figure 6–19
Effects of charges in the oxide and at the interface: (a) definitions of charge densities (C/cm^2) due to various sources; (b) representing these charges as an equivalent sheet of positive charge Q_f at the oxide–semiconductor interface. This positive charge induces an equivalent negative charge in the semiconductor, which requires a negative gate voltage to achieve the flat band condition.

growth or subsequent processing steps. Since sodium is a common contaminant, it is necessary to use extremely clean chemicals, water, gases, and processing environment to minimize its effect on dielectric layers. Sodium ions introduce positive charges (Q_m) in the oxide, which in turn induce negative charges in the semiconductor. The effect of such positive ionic charges in the oxide depends upon the number of ions involved and their distance from the semiconductor surface (Prob. 6.12). The negative charge induced in the semiconductor is greater if the Na^+ ions are near the interface than if they are farther away. The effect of this ionic charge on threshold voltage is complicated by the fact that Na^+ ions are relatively mobile in SiO_2, particularly at elevated temperatures, and can thus drift in an applied electric field. Obviously, a device with V_T dependent on its past history of voltage bias is unacceptable. Fortunately, Na contamination of the oxide can be reduced to tolerable levels by proper care in processing. The oxide also contains trapped charges (Q_{ot}) due to imperfections in the SiO_2.

In addition to oxide charges, a set of positive charges arises from *interface states* at the Si-SiO_2 interface. These charges, which we will call Q_{it}, result from the sudden termination of the semiconductor crystal lattice at the oxide interface. Near the interface is a transition layer (SiO_x) containing fixed charges (Q_f). As oxidation takes place in forming the SiO_2 layer, Si is removed from the surface and reacts with the oxygen. When the oxidation is stopped, some ionic Si is left near the interface. These ions, along with uncompleted Si bonds at the surface, result in a sheet of positive charge Q_f near the interface. This charge depends on oxidation rate and subsequent heat treatment, and also on crystal orientation. For carefully treated Si-SiO_2 interfaces, typical charge densities due to Q_{it} and Q_f are about 10^{10} charges/cm^2 for samples with [100] surfaces. The interface charge density is about a factor

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 283

of ten higher on {111} surfaces. That is why MOS devices are generally made on {100} Si.

For simplicity we will include the various oxide and interface charges in an *effective* positive charge at the interface Q_i (C/cm^2). The effect of this charge is to induce an equivalent negative charge in the semiconductor. Thus an additional component must be added to the flat band voltage:

$$V_{FB} = \Phi_{ms} - \frac{Q_i}{C_i} \tag{6-37}$$

Since the difference in work function and the positive interface charge both tend to bend the bands down at the semiconductor surface, a negative voltage must be applied to the metal relative to the semiconductor to achieve the flat band condition of Fig. 6-19b.

6.4.4 Threshold Voltage

The voltage required to achieve flat band should be added to the threshold voltage equation (6-33) obtained for the ideal MOS structure (for which we assumed a zero flat band voltage):

$$\boxed{V_T = \Phi_{ms} - \frac{Q_i}{C_i} - \frac{Q_d}{C_i} + 2\phi_F} \tag{6-38}$$

Thus the voltage required to create strong inversion must be large enough to first achieve the flat band condition (Φ_{ms} and Q_i/C_i terms), then accommodate the charge in the depletion region (Q_d/C_i), and finally to induce the inverted region ($2\phi_F$). This equation accounts for the dominant threshold voltage effects in typical MOS devices. It can be used for both n-type and p-type substrates[8] if appropriate signs are included for each term (Fig. 6-20). Typically Φ_{ms} is negative, although its value varies as in Fig. 6-17. The interface charge is positive, so the contribution of the $-Q_i/C_i$ term is negative for either substrate type. On the other hand, the charge in the depletion region is negative for ionized acceptors (p-type substrate, n-channel device) and is positive for ionized donors (n-type substrate, p channel). Also, the term ϕ_F, which is defined as $(E_i - E_F)/q$ in the neutral substrate, can be positive or negative, depending on the conductivity type of the substrate. Considering the signs in Fig. 6-20, we see that all four terms give negative contributions in the p-channel case. Thus we expect negative threshold voltages for typical p-channel devices. On the other hand, n-channel devices may have either positive or negative threshold voltages, depending on the relative values of terms in Eq. (6-38).

[8]It is important to remember that n-channel devices are made on p-type substrates and that p-channel devices have n-type substrates.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 284

(a)

$V_T =$	Φ_{ms}	$-\frac{Q_i}{C_i}$	$-\frac{Q_d}{C_i}$	$+2\phi_F$
	(−)	(−)	(+) n channel (−) p channel	(+) n channel (−) p channel

(b)

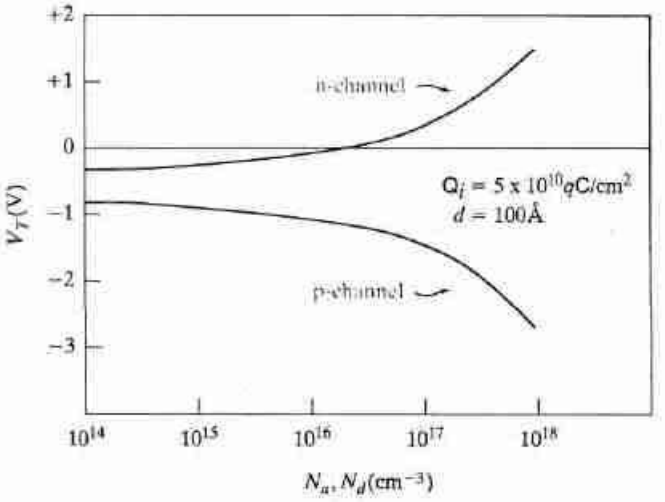


Figure 6–20
Influence of materials parameters on threshold voltage: (a) the threshold voltage equation indicating signs of the various contributions; (b) variation of V_T with substrate doping for n-channel and p-channel n^+ poly-SiO_2–Si devices.

All terms in Eq. (6–38) except Q_i/C_i depend on the doping in the substrate. The terms Φ_{ms} and ϕ_F have relatively small variations as E_F is moved up or down by the doping. Large changes can occur in Q_d, which varies with the square root of the doping impurity concentration as in Eq. (6–32). We illustrate the variation of threshold voltage with substrate doping in Fig. 6–20. As expected from Eq. (6–38), V_T is always negative for the p-channel case. In the n-channel case, the negative flat band voltage terms can dominate for lightly doped p-type substrates, resulting in a negative threshold voltage. However, for more heavily doped substrates, the increasing contribution of N_a to the Q_d term dominates, and V_T becomes positive.

We should pause here and consider what positive or negative V_T means for the two cases. In a p-channel device we expect to apply a negative voltage from metal to semiconductor in order to induce the positive charges in the channel. In this case a negative threshold voltage means that the negative voltage we apply must be larger than V_T in order to achieve strong inversion. In the n-channel case we expect to apply a positive voltage to the metal to induce the channel. Thus a positive value for V_T means the applied voltage must be larger than this threshold value to obtain strong inversion and a conducting n channel. On the other hand, a negative V_T in this case means

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 285

that a channel exists at $V = 0$ due to the Φ_{ms} and Q_i effects (Figs. 6-18 and 6-19), and we must apply a negative voltage V_T to turn the device off. Since lightly doped substrates are desirable to maintain a high breakdown voltage for the drain junction, Fig. 6-20 suggests that V_T will be negative for n-channel devices made by standard processing. This tendency for the formation of depletion mode (normally on) n-channel transistors is a problem which must be dealt with by special fabrication methods to be described in Section 6.5.5.

EXAMPLE 6-1 An n^+-polysilicon-gate n-channel MOS transistor is made on a p-type Si substrate with $N_a = 5 \times 10^{15}\text{cm}^{-3}$. The SiO_2 thickness is 100 Å in the gate region, and the effective interface charge Q_i is $4 \times 10^{10}\, q\text{C/cm}^2$. Find C_i and C_{min} on the C–V characteristics, and find W_m, V_{FB}, and V_T.

SOLUTION

$$\phi_F = \frac{kT}{q}\ln\frac{N_a}{n_i} = 0.0259\ln\frac{5\times10^{15}}{1.5\times10^{10}} = 0.329\text{ eV}$$

$$W_m = 2\left[\frac{\epsilon_s\phi_F}{qN_a}\right]^{1/2} = 2\left[\frac{11.8\times8.85\times10^{-14}\times0.329}{1.6\times10^{-19}\times5\times10^{15}}\right]^{1/2}$$

$$= 4.15\times10^{-5}\text{cm} = \mathbf{0.415\ \mu m}$$

From Fig. 6-17, $\Phi_{ms} \approx -0.95$ V, and we have

$$Q_i = 4\times10^{10}\times1.6\times10^{-19} = 6.4\times10^{-9}\text{ C/cm}^2$$

$$C_i = \frac{\epsilon_i}{d} = \frac{3.9\times8.85\times10^{-14}}{0.1\times10^{-5}} = 3.45\times10^{-7}\text{ F/cm}^2$$

$$V_{FB} = \Phi_{ms} - Q_i/C_i = -0.95 - 6.4\times10^{-9}/3.45\times10^{-7} = \mathbf{-0.969\text{ V}}$$

$$Q_d = -qN_aW_m = -1.6\times10^{-19}\times5\times10^{15}\times4.15\times10^{-5}$$

$$= -3.32\times10^{-8}\text{C/cm}^2$$

$$V_T = V_{FB} - \frac{Q_d}{C_i} + 2\phi_F = -0.969 + \frac{3.32\times10^{-8}}{3.45\times10^{-7}} + 0.658 = \mathbf{-0.215\text{ V}}$$

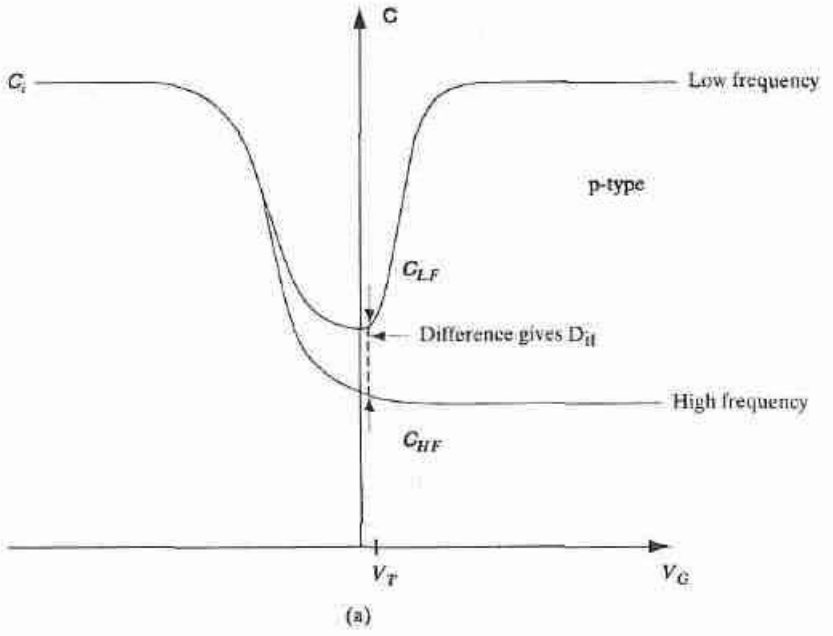
$$C_d = \frac{\epsilon_s}{W_m} = \frac{11.8\times8.85\times10^{-14}}{4.15\times10^{-5}} = 2.5\times10^{-8}\text{ F/cm}$$

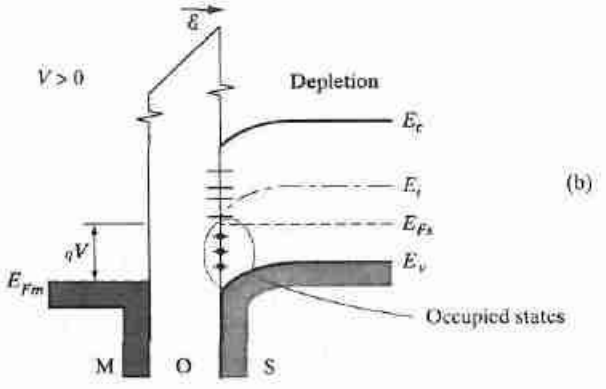
$$C_{min} = \frac{C_iC_d}{C_i + C_d} = \frac{3.45\times10^{-7}\times2.5\times10^{-8}}{3.45\times10^{-7} + 2.5\times10^{-8}} = \mathbf{2.33\times10^{-8}\text{ F/cm}^2}$$

6.4.5 MOS Capacitance–Voltage Analysis

Let us see how the various parameters of a MOS device such as insulator thickness, substrate doping, and V_T can be determined from the C–V characteristics (Fig. 6-21). First, the shape of the C–V curve depends upon the type of substrate

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 286





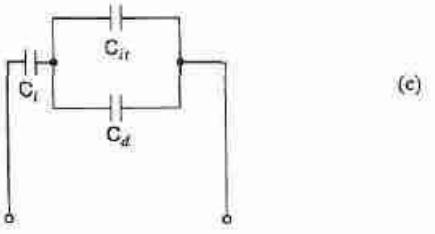


Figure 6–21
Fast interface state determination: (a) High-frequency and low-frequency C–V curves showing impact of fast interface states; (b) energy levels in the band gap due to fast interface states; (c) equivalent circuit of MOS structure showing capacitance components due to gate oxide (C_i), depletion layer in the channel (C_d), and fast interface states (C_{it}).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 287

doping. If the high-frequency capacitance is large for negative gate biases and small for positive biases, it is a p-type substrate, and vice versa. From the low-frequency C–V curve for p-type material, as the gate bias is made more positive (or less negative), the capacitance goes down slowly in depletion and then rises rapidly in inversion. As a result, the low-frequency C–V is not quite symmetric in shape. For n-type substrates, the C–V curves would be the mirror image of Fig. 6-21.

The capacitance $C_i = \epsilon_i/d$ in accumulation or strong inversion (at low frequencies) gives us the insulator thickness, d. The minimum MOS capacitance, C_{min}, is the series combination of C_i and the minimum depletion capacitance, $C_{dmin} = \epsilon_s/W_m$, corresponding to the maximum depletion width. We can in principle use the measurement of C_{min} to determine the substrate doping. However, from Eq. (6-31) we see that the dependence of W_m on N_a is complicated, and we get a transcendental equation which can only be solved numerically. Actually, an approximate, iterative solution exists which gives us N_a in terms of the minimum depletion capacitance, C_{dmin}.

$$N_a = 10^{[30.388 + 1.683 \log C_{dmin} - 0.03177(\log C_{dmin})^2]} \tag{6-39}$$

where C_{dmin} is in F/cm^2.

Once the substrate doping is obtained, we can determine the flat band capacitance from it. It can be shown that the semiconductor capacitance at flat band C_{FB} (point 2 in Fig. 6-16) is determined from the Debye length capacitance

$$C_{debye} = \frac{\epsilon_s}{L_D} \tag{6-40}$$

where the Debye length is dependent on doping as described in Eq. (6-25). The overall MOS flat band capacitance, C_{FB}, is the series combination of C_{debye} and C_i. We can thus determine V_{FB} corresponding to the C_{FB}. Once C_i, V_{FB}, and substrate doping are obtained, all terms in the V_T expression (Eq. 6-38) are known. Interestingly, the threshold voltage V_T does not correspond to exactly the minimum of the C–V characteristics, C_{min}, but a slightly higher capacitance marked as point 4 in Fig. 6-16. In fact, it corresponds to the series combination of C_i and $2C_{dmin}$, rather than the series combination of C_i and C_{dmin}. The reason for this is that when we change the gate bias around strong inversion, the change of charge in the semiconductor is the sum of the change in depletion charge and the mobile inversion charge, where the two are equal in magnitude at the onset of strong inversion.

We can also determine MOS parameters such as the *fast interface state* density, D_{it}, and mobile ion charges, Q_m, from C–V measurements (Figs. 6-21 and 6-22). The term fast interface state refers to the fact that these defects can change their charge state relatively fast in response to changes of the gate bias. As the surface potential in a MOS device is varied, the fast interface

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 288

states or traps in the band gap can move above or below the Fermi level in response to the bias, because their positions relative to the band edges are fixed (Fig. 6-21b). Keeping in mind the property of the Fermi-Dirac distribution that energy levels below the Fermi level have a high probability of occupancy by electrons, while levels above the Fermi level tend to be empty, we see that a fast interface state moving above the Fermi level would tend to give up its trapped electron to the semiconductor (or equivalently capture a hole). Conversely, the same fast interface state below the Fermi level captures an electron (or gives up a hole). It obviously makes sense to talk in terms of electrons or holes, depending on which is the majority carrier in the semiconductor. Since charge storage results in capacitance, the fast interface states give rise to a capacitance which is in parallel with the depletion capacitance in the channel (and hence is additive), and this combination is in series with the insulator capacitance C_i. The fast interface states can keep pace with low frequency variations of the gate bias (~1-1000 Hz), but not at extremely high frequencies (~1 MHz). So the fast interface states contribute to the low frequency capacitance C_{LF}, but not the high frequency capacitance C_{HF}. Clearly, from the difference between the two, we ought to be able to compute the fast interface state density. Although we will not do the detailed derivation here, it can be shown that

$$D_{it} = \frac{1}{q}\left(\frac{C_i C_{LF}}{C_i - C_{LF}} - \frac{C_i C_{HF}}{C_i - C_{HF}}\right) \text{cm}^{-2}\text{eV}^{-1} \qquad (6\text{-}41)$$

While the fast interface states can respond quickly to voltage changes, the fixed oxide charges Q_f, as the name implies, do not change their charge state regardless of the gate bias or surface potential. As mentioned above, the effect of these charges on the flat band and threshold voltage depends not only on the number of charges but also their location relative to the oxide-silicon interface (Fig. 6-22). Hence, we must take a weighted sum of these charges, counting charges closer to the oxide-silicon interface more heavily than those that are farther away. This position dependence is the basis of what is called the *bias-temperature stress test* for measuring the mobile ion content, Q_m. We heat up the MOS device to ~200-300°C (to make the ions more mobile) and apply a positive gate bias to generate a field of ~1 MV/cm within the oxide. After cooling the capacitor to room temperature, the C-V characteristics are measured. We have seen how V_{FB} can be determined from the C-V curve, using Eq. (6-40) and $C_i V_{FB}$ is also given by Eq. (6-37). The positive bias repels positive mobile ions such as Na^+ to the oxide-silicon interface so that they contribute fully to a flat band voltage we can call V_{FB}^+. Next, the capacitor is heated up again, a negative bias is applied so that the ions drift to the gate electrode, and another C-V measurement is made. Now, the mobile ions are too far away to affect the semiconductor band-bending, but induce an equal and opposite charge on the gate electrode. From the resulting C-V, the new flat band, V_{FB}^-, is determined. From

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 289

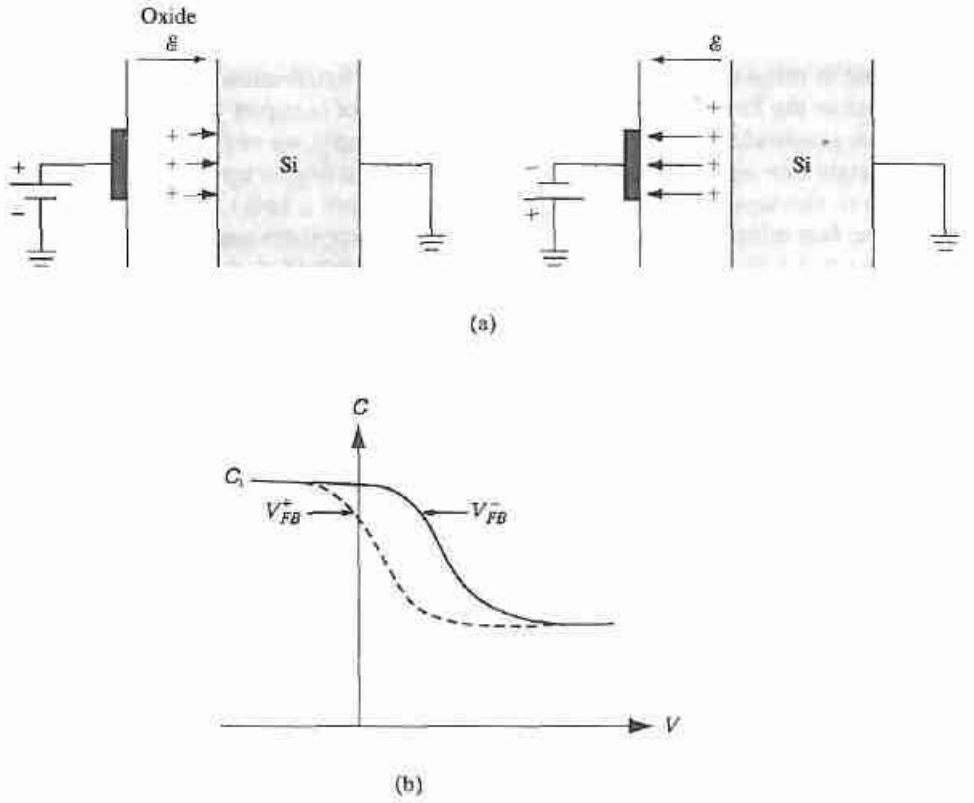


Figure 6–22
Mobile ion determination: (a) Movement of mobile ions due to positive and negative bias-temperature stress; (b) *C–V* characteristics under positive (dashed line) and negative (solid line) bias-temperature.

the difference of the two flat band voltages, we can determine the mobile ion content using

$$Q_m = C_i(V_{FB}^- - V_{FB}^+) \tag{6-42}$$

6.4.6 Time-Dependent Capacitance Measurements

During *C–V* measurements, if the gate bias is varied rapidly from accumulation to inversion, the depletion width can momentarily become greater than the theoretical maximum for gate biases beyond V_T. This phenomenon is known as *deep depletion*, and causes the MOS capacitance to drop below the theoretical minimum, C_{min}, for a transient period. After a time period characteristic of the minority carrier lifetime, which determines

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 290

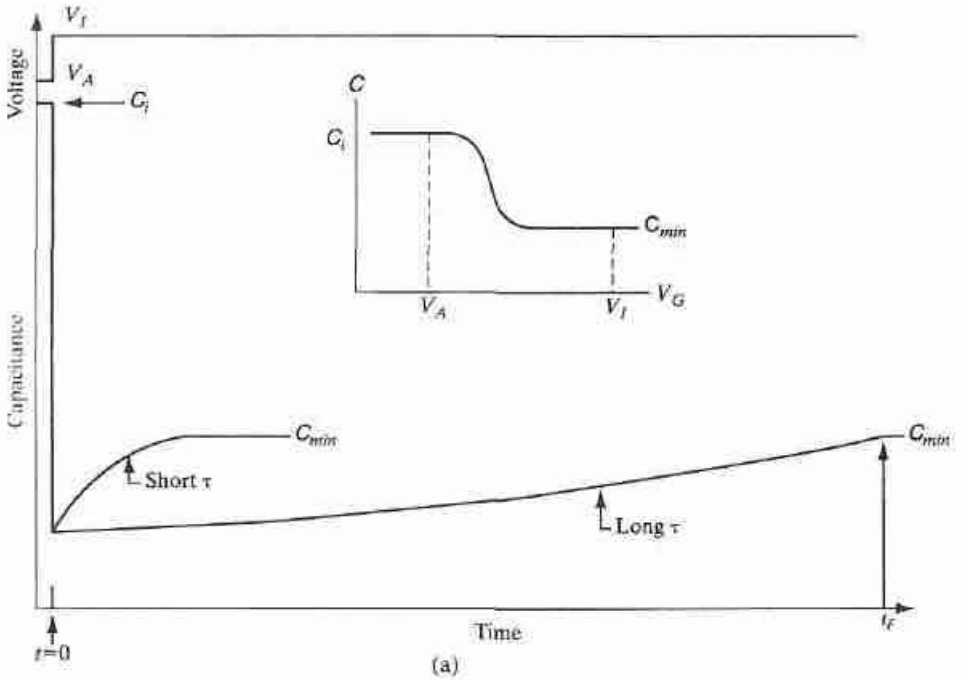


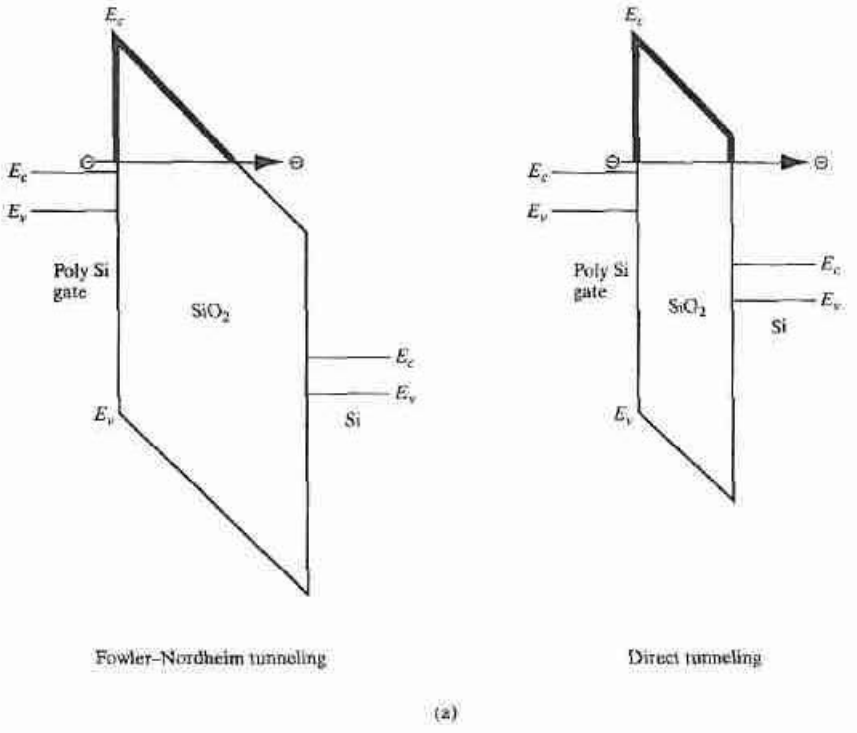
Figure 6–23
Time-dependent MOS capacitance ($C_{t\#}$) due to the application of a step voltage V_A (which puts the capacitor in accumulation) to V_I (which puts the capacitor in inversion).

the rate of generation of the minority carriers in the MOS device, the depletion width collapses back to the theoretical maximum and the capacitance recovers to C_{min} (Fig. 6-23). This capacitance transient, C-t, forms the basis of a powerful technique to measure the lifetime, known as the Zerbst technique.

6.4.7 Current–Voltage Characteristics of MOS Gate Oxides

An ideal gate insulator does not conduct any current, but for real insulators there can be some leakage current which varies with the voltage or electric field across the gate oxide. By looking at the band diagram of the MOS system perpendicular to the oxide-silicon interface (Fig. 6-24), we see that for electrons in the conduction band, there is a barrier, ΔE_c (= 3.1 eV). Although electrons with energy less than this barrier cannot go through the oxide classically, it was discussed in Chapter 2 that quantum mechanically electrons can tunnel through a barrier, especially if the barrier thickness is sufficiently small. The detailed calculation of the *Fowler–Nordheim tunneling* current for electrons going from the Si conduction band to the conduction band of SiO_2, and then having the electrons "hop" along in the oxide to the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 291



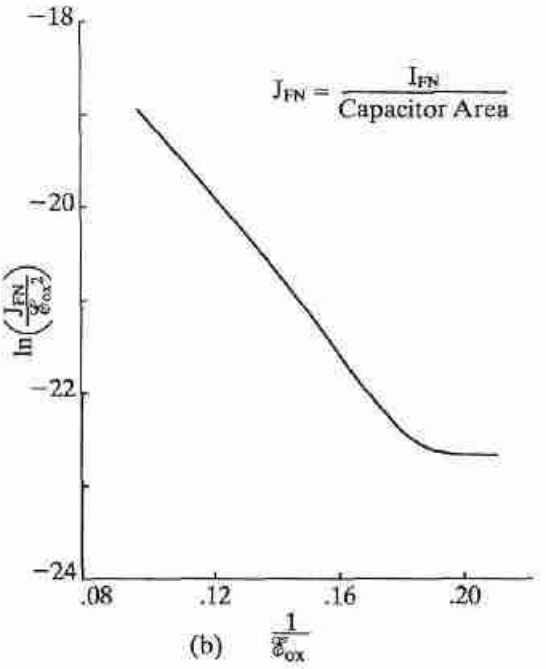


Figure 6–24 Current–voltage characteristics of gate oxides: (a) Fowler–Nordheim and direct tunneling through thin gate oxides; (b) plot of Fowler–Nordheim tunneling leakage current as a function of electric field across the oxide.

gate electrode, involves solving the Schrödinger equation for the electron wave function. The Fowler-Nordheim tunneling current I_{FN} can be expressed as a function of the electric field in the gate oxide:

$$I_{FN} \propto \mathscr{E}_{ox}^2 \exp\left(\frac{-B}{\mathscr{E}_{ox}}\right) \tag{6-43}$$

where B is a constant depending on m_n^* and the barrier height.

As gate oxides are made thinner in successive generations of MOSFETs, the tunneling barrier in the gate oxide becomes so thin that the electrons in the conduction band of Si can tunnel through the gate oxide and emerge in the gate, without having to go via the conduction band of the gate oxide. This is known as *direct tunneling* rather than Fowler-Nordheim tunneling. The overall physics is similar, but some of the details are different. For instance, Fowler-Nordheim tunneling involves a triangular barrier, while direct tunneling is through a trapezoidal barrier (Fig. 6-24a). Such tunneling currents are becoming a major problem in modern devices because the useful feature of high input impedance for MOS devices is degraded.

It is necessary to increase the gate capacitance in successive generations of MOSFETs in order to increase the drain current, as discussed later in Section 6.5. One can achieve higher C_i by using insulators with a dielectric constant higher than SiO_2 instead of reducing the gate oxide thickness, d, as can be seen from Eq. (6-29). The advantage of not reducing d too much is that it keeps the tunneling barrier wide, and the gate oxide field low, thereby keeping the gate tunneling leakage current low (Eq. 6-43). Such high dielectric constant insulators such as hafnium dioxide, HfO_2, are also known as high-k gate dielectrics because k is also sometimes used as a symbol for dielectric constant.

Prolonged charge transport through gate oxides can ultimately cause catastrophic electrical breakdown of the oxides. This is known as *Time-Dependent Dielectric Breakdown (TDDB)*. One of the popular models that explains this degradation involves electrons tunneling into the conduction band of the gate oxide from the negative electrode (cathode), then gaining energy from the electric field, thus becoming "hot" electrons in the gate oxide. If they gain sufficient energy, they can cause impact ionization within the oxide and create electron-hole pairs. The electrons are accelerated toward the (positive) Si substrate, while the holes travel toward the gate. However, electron and hole mobilities are extremely small in SiO_2. Hole mobilities are particularly low (~0.01 cm^2/Vs). Hence, there is a great propensity for these impact-generated holes to be trapped at defect sites within the oxide, near the cathode. The resulting band diagram (Fig. 6-25) is altered by this sheet of trapped positive charge, which causes the internal electric field between this point and the gate to increase. A similar distortion of the electric field near the Si anode is created by the trapped impact-generated electrons. However, the steepest slope in Fig. 6-25, and therefore the highest field, is near the gate. As a result, the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 293

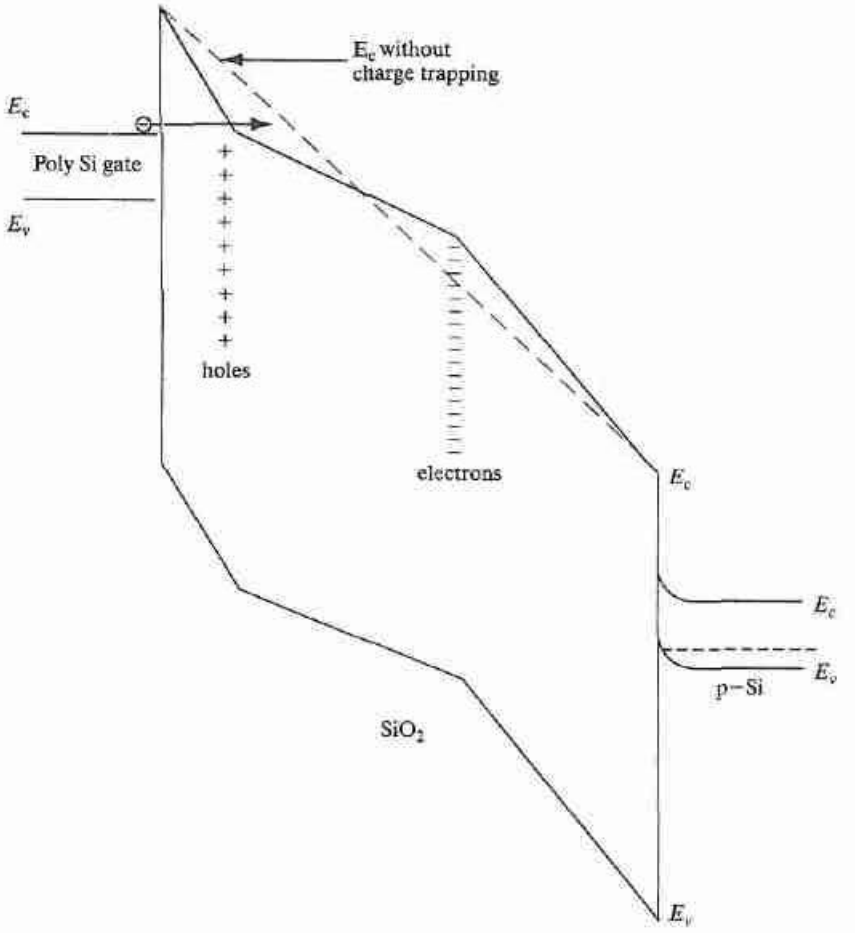


Figure 6–25
Time-dependent dielectric breakdown of oxides: Band diagram of a MOS device showing the band edges in the polysilicon gate, oxide, and Si substrate. Trapped holes and electrons in the oxide distort the band edges, and increase the electric field in the oxide near the gate. The tunneling barrier width is seen to be less than if there were no charge trapping (dashed line).

barrier for electron tunneling from the gate into the oxide is reduced. More electrons can tunnel into the oxide, and cause more impact ionization. We get a positive feedback effect that can lead to a runaway TDDB process.

6.5 THE MOS FIELD-EFFECT TRANSISTOR

The MOS transistor is also called a surface field-effect transistor, since it depends on control of current through a thin channel at the surface of the semiconductor (Fig. 6–10). When an inversion region is formed under the gate, current can flow from drain to source (for an n-channel device). In this section

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 294

we analyze the conductance of this channel and find the $I_D - V_D$ characteristics as a function of gate voltage V_G. As in the JFET case, we will find these characteristics below saturation and then assume I_D remains essentially constant above saturation.

6.5.1 Output Characteristics

The applied gate voltage V_G is accounted for by Eq. (6-28) plus the voltage required to achieve flat band:

$$V_G = V_{FB} - \frac{Q_s}{C_i} + \phi_s \tag{6-44}$$

The induced charge Q_s in the semiconductor is composed of mobile charge Q_n and fixed charge in the depletion region Q_d. Substituting $Q_n + Q_d$ for Q_s, we can solve for the mobile charge:

$$Q_n = -C_i\left[V_G - \left(V_{FB} + \phi_s - \frac{Q_d}{C_i}\right)\right] \tag{6-45}$$

At threshold the term in brackets can be written $V_G - V_T$ from Eq. (6-38).

With a voltage V_D applied, there is a voltage rise V_x from the source to each point x in the channel. Thus the potential $\phi_s(x)$ is that required to achieve strong inversion ($2\phi_F$) plus the voltage V_x:

$$Q_n = -C_i\left[V_G - V_{FB} - 2\phi_F - V_x - \frac{1}{C_i}\sqrt{2q\epsilon_s N_a(2\phi_F + V_x)}\right] \tag{6-46}$$

If we neglect the variation of $Q_d(x)$ with bias V_x, Eq. (6-46) can be simplified to

$$Q_n(x) = -C_i(V_G - V_T - V_x) \tag{6-47}$$

This equation describes the mobile charge in the channel at point x (Fig. 6-26). The conductance of the differential element dx is $\bar{\mu}_n Q_n(x) Z/dx$, where Z is the width of the channel and $\bar{\mu}_n$ is a *surface* electron mobility (indicating the mobility in a thin region near the surface is not the same as in the bulk material). At point x we have

$$I_D dx = \bar{\mu}_n Z|Q_n(x)|dV_x \tag{6-48}$$

Integrating from source to drain,

$$\int_0^L I_D dx = \bar{\mu}_n Z C_i \int_0^{V_D} (V_G - V_T - V_x)dV_x$$

$$I_D = \frac{\bar{\mu}_n Z C_i}{L}[(V_G - V_T)V_D - \tfrac{1}{2}V_D^2] \tag{6-49}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 295

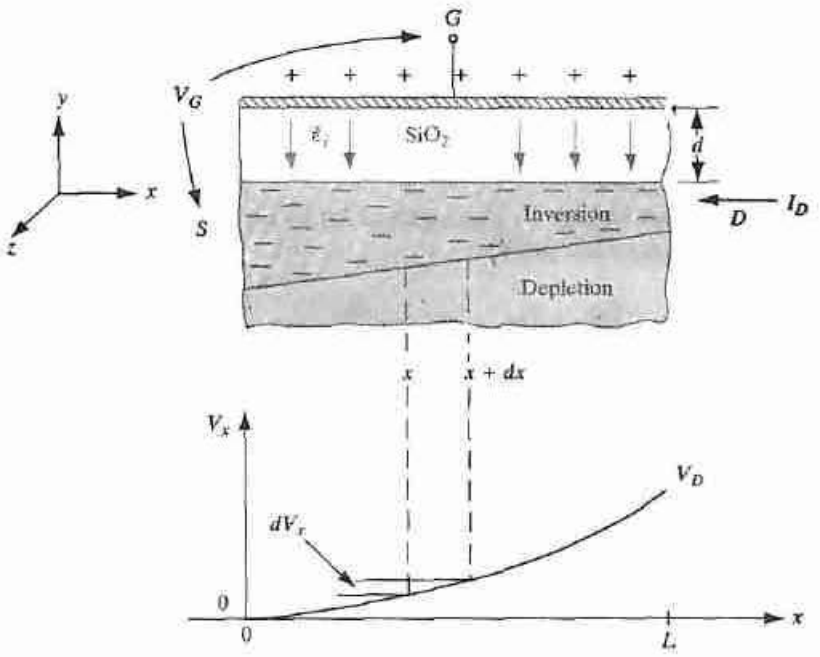


Figure 6–26
Schematic view of the n-channel region of a MOS transistor under bias below pinch-off, and the variation of voltage V_x along the conducting channel.

where

$$\frac{\overline{\mu}_n Z C_i}{L} = k_N$$

determines the conductance and transconductance of the *n*-channel MOSFET [see Eqs. (6–51) and (6–54)].

In this analysis the depletion charge Q_d in the threshold voltage V_T is simply the value with no drain current. This is an approximation, since $Q_d(x)$ varies considerably when V_D is applied, to reflect the variation in V_x (see Fig. 6–26b). However, Eq. (6–49) is a fairly accurate description of drain current for low values of V_D, and is often used in approximate design calculations because of its simplicity. A more accurate and general expression is obtained by including the variation of $Q_d(x)$. Performing the integration of Eq. (6–48) using Eq. (6–46) for $Q_n(x)$, one obtains

$$I_D = \frac{\overline{\mu}_n Z C_i}{L} \times \left\{ (V_G - V_{FB} - 2\phi_F - \tfrac{1}{2}V_D)V_D - \frac{2}{3}\frac{\sqrt{2\epsilon_s q N_a}}{C_i}\left[(V_D + 2\phi_F)^{3/2} - (2\phi_F)^{3/2}\right] \right\} \quad (6\text{–}50)$$

The drain characteristics that result from these questions are shown in Fig. 6–10c. If the gate voltage is above threshold ($V_G > V_T$), the drain current

is described by Eq. (6–50) or approximately by Eq. (6–49) for low V_D. Initially the channel appears as an essentially linear resistor, dependent on V_G. The conductance of the channel in this linear region can be obtained from Eq. (6–49) with $V_D \ll (V_G - V_T)$:

$$g = \frac{\partial I_D}{\partial V_D} \simeq \frac{Z}{L}\bar{\mu}_n C_i (V_G - V_T) \tag{6-51}$$

where $V_G > V_T$ for a channel to exist.

As the drain voltage is increased, the voltage across the oxide decreases near the drain, and Q_n becomes smaller there. As a result the channel becomes pinched off at the drain end, and the current saturates. The saturation condition is approximately given by

$$V_D(\text{sat.}) \simeq V_G - V_T \tag{6-52}$$

The drain current at saturation remains essentially constant for larger values of drain voltage. Substituting Eq. (6–52) into Eq. (6–49), we obtain

$$I_D(\text{sat.}) \simeq \tfrac{1}{2}\bar{\mu}_n C_i \frac{Z}{L}(V_G - V_T)^2 = \frac{Z}{2L}\bar{\mu}_n C_i V_D^2(\text{sat.}) \tag{6-53}$$

for the approximate value of drain current at saturation.

The transconductance in the saturation range can be obtained approximately by differentiating Eq. (6–53) with respect to the gate voltage:

$$g_m(\text{sat.}) = \frac{\partial I_D(\text{sat.})}{\partial V_G} \simeq \frac{Z}{L}\bar{\mu}_n C_i (V_G - V_T) \tag{6-54}$$

The derivations presented here are based on the n-channel device. For the p-channel enhancement transistor the voltages V_D, V_G, and V_T are negative, and current flows from source to drain (Fig. 6–27).

6.5.2 Transfer Characteristics

The output characteristics plot the drain current as a function of the drain bias, with gate bias as a parameter (Fig. 6–27). On the other hand, the *transfer* characteristics plot the output drain current as a function of the input gate bias, for fixed drain bias (Fig. 6–28a). Clearly, in the linear region, I_D versus V_G should be a straight line from Eq. (6–49). The intercept of this line on the V_G axis is the linear region threshold voltage, V_T (lin.) and the slope (divided by the applied V_D) gives us the linear value of k_N, k_N(lin.), of the n-channel MOSFET. If we look at actual data, however, we see that while the characteristics are approximately linear at low gate bias, at high gate biases the drain current increases sub-linearly. The transconductance, g_m (lin.), in the linear region can be obtained by differentiating the right-hand side of Eq. (6–49) with respect to gate bias.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 297

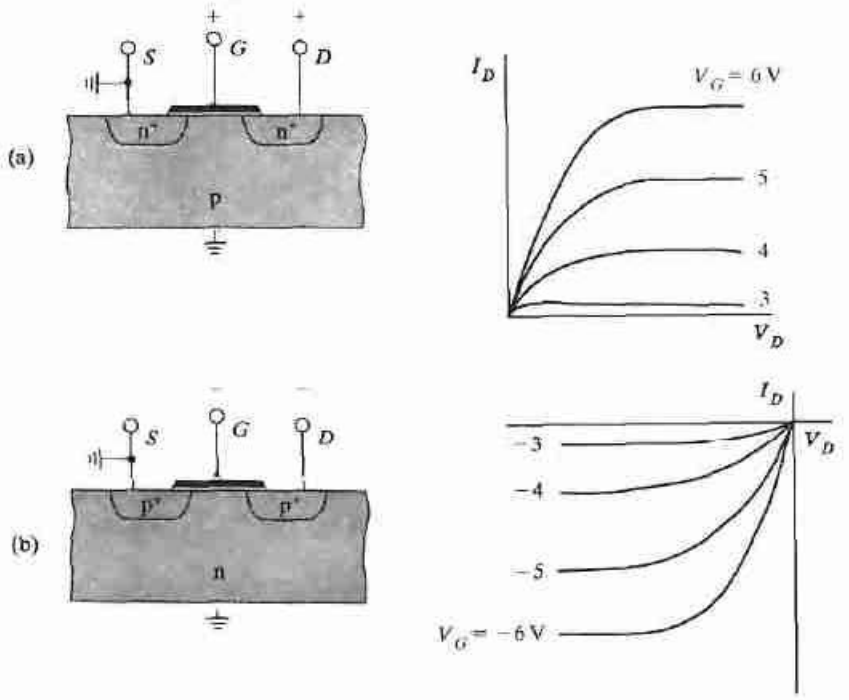


Figure 6–27 Drain current–voltage characteristics for enhancement transistors: (a) for n-channel V_D, V_G, V_T, and I_D are positive; (b) for p-channel all these quantities are negative.

The g_m (lin.) is plotted as a function of V_G in Fig. 6–28b. It may be noted that the transconductance is zero below V_T because there is little drain current. It goes through a maximum at the point of inflection of the I_D–V_G curve, and then decreases. This decrease is due to two factors that will be discussed in Sections 6.5.3 and 6.5.8: degradation of the effective channel mobility as a function of increasing transverse electric field across the gate oxide, and source/drain series resistance.

For the transfer characteristics in the saturation region, since Eq. (6–53) shows a quadratic dependence of I_D on V_G, we get a linear behavior by plotting not the drain current, but rather the square root of I_D, as a function of V_G (Fig. 6–29). In this case the intercept gives us the threshold voltage in the saturation region, V_T(sat.). We shall see in Section 6.5.10 that due to effects such as drain-induced barrier lowering (DIBL), for short channel length MOSFETs the V_T(sat.) can be lower than V_T(lin.), while the long channel values are similar. Similarly, the slope of the transfer characteristics can be used to determine the value of k_N in the saturation region, k_N(sat.) for the n-channel MOSFET, which can be different from k_N(lin.) for short channel devices.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 298

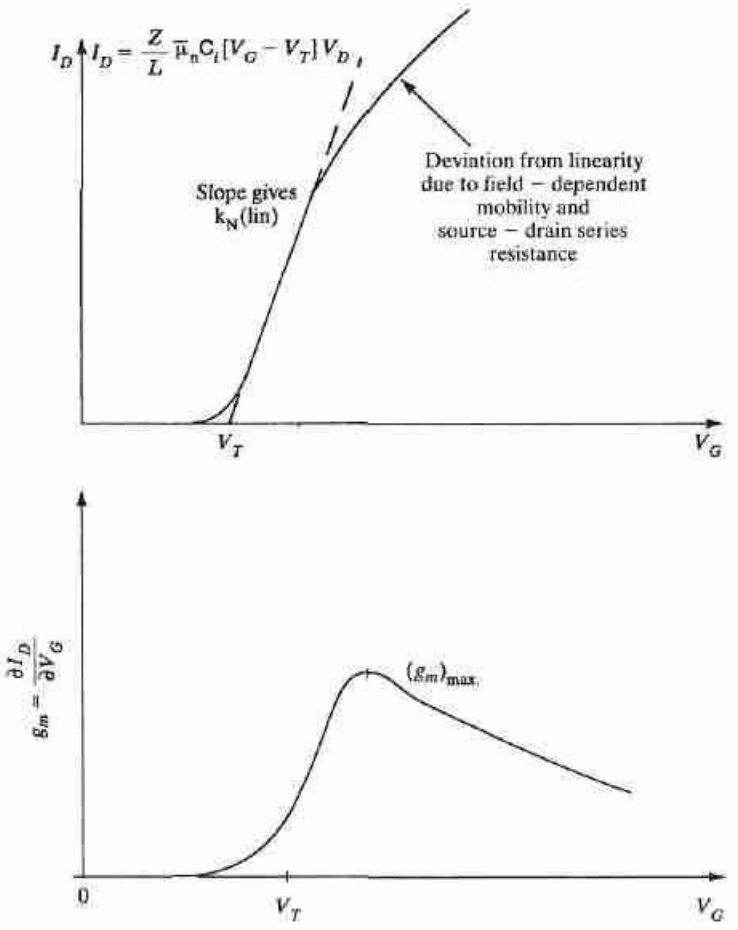


Figure 6–28
Linear region transfer characteristics: (a) plot of drain current versus gate voltage for MOSFETs in the linear region; (b) transconductance as a function of gate bias.

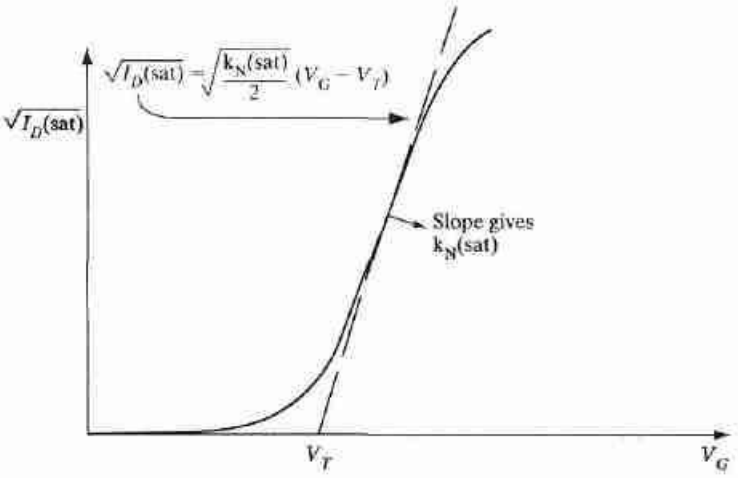


Figure 6–29
Saturation region transfer characteristics: plot of square root of the drain current versus gate voltage for MOSFETs.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 299

EXAMPLE 6–2

For an n-channel MOSFET with a gate oxide thickness of 10 nm, $V_T = 0.6$ V, and $Z = 25\ \mu$m, $L = 1\ \mu$m. Calculate the drain current at $V_G = 5$ V and $V_D = 0.1$ V. Repeat for $V_G = 3$ V and $V_D = 5$ V. Discuss what happens for $V_D = 7$ V. Assume an electron channel mobility of $\bar{\mu}_n = 200\ \text{cm}^2/\text{V-s}$.

SOLUTION

$$C_i = \frac{\epsilon_i}{d} = \frac{(3.9)(8.85 \times 10^{-14})}{10^{-6}} = 3.45 \times 10^{-7}\ \text{F/cm}^2$$

For $V_G = 5$ V and it follows that $V_D = 0.1$ V, and $V_T = 0.6$ V, $V_D < (V_G - V_T)$, we are in the linear region. Thus,

$$I_D = \frac{Z}{L}\bar{\mu}_n C_i\left[(V_G - V_T)V_D - \frac{1}{2}V_D^2\right]$$

$$= \frac{25}{1}(200)(3.45 \times 10^{-7})\left[(5 - 0.6) \times 0.1 - \frac{1}{2}(0.1)^2\right] = \mathbf{7.51 \times 10^{-4}\ A}$$

For $V_G = 3$ V, $V_D = 5$ V, $V_D(\text{sat.}) = V_G - V_T = 3 - 0.6 = 2.4$ V

$$I_D = \frac{Z}{L}\bar{\mu}_n C_i\left[(V_G - V_T)V_D(\text{sat.}) - \frac{1}{2}V_D^2(\text{sat.})\right]$$

$$= \frac{25}{1}(200)(3.45 \times 10^{-7})\left[(2.4)^2 - \frac{1}{2}(2.4)^2\right] = \mathbf{4.97 \times 10^{-3}\ A}$$

For $V_D = 7$ V, I_D will not increase, because we are in the saturation region.

6.5.3 Mobility Models

The mobility of carriers in the channel of a MOSFET is lower than in bulk semiconductors because there are additional scattering mechanisms. Since carriers in the channel are very close to the semiconductor-oxide interface, they are scattered by surface roughness and by coulombic interaction with fixed charges in the gate oxide. When the carriers travel in the inversion layer from the source to the drain, they encounter microscopic roughness on an atomistic scale at the oxide-silicon interface and undergo scattering because, as discussed in Section 3.4.1, any deviation from a perfectly periodic crystal potential results in scattering. This mobility degradation increases with the gate bias because a higher gate bias draws the carriers closer to the oxide-silicon interface, where they are more influenced by the interfacial roughness.

It is very interesting to note that if we plot the effective carrier mobility in the MOSFET as a function of the average transverse electric field in the middle of the inversion layer, we get what is known as a "universal"

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 300

mobility degradation curve for any MOSFET, which is independent of the technology or device structural parameters such as oxide thickness and channel doping (Fig. 6-30). We can apply Gauss's law to the region marked by the colored box in Fig. 6-31, which encloses all the depletion charge and half of the inversion charge in the channel. We see that the average transverse field in the middle of the inversion region is given by

$$\mathscr{E}_{eff} = \frac{1}{\epsilon_s}\left(Q_d + \frac{1}{2}Q_n\right) \tag{6-55a}$$

While this model works quite well for electrons, for reasons that are not clearly understood at present, it has to be modified slightly for holes in the sense that the average transverse field must now be defined as

$$\mathscr{E}_{eff} = \frac{1}{\epsilon_s}\left(Q_d + \frac{1}{3}Q_n\right) \tag{6-55b}$$

This degradation of mobility with gate bias is often compactly described by writing the drain current expression as

$$I_D = \frac{\bar{\mu}_n Z C_i}{L\{1 + \theta(V_G - V_T)\}}\left[(V_G - V_T)V_D - \frac{1}{2}V_D^2\right] \tag{6-56}$$

where θ is called the *mobility degradation parameter*. Because of the additional $(V_G - V_T)$ term in the denominator, the drain current increases sub-linearly with gate bias for high gate voltages.

In addition to this dependence of the channel mobility on gate bias or transverse electric field, there is also a strong dependence on drain bias or the

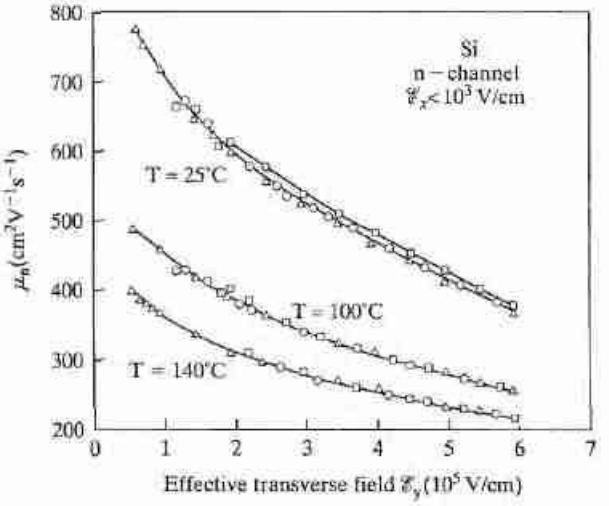


Figure 6–30
Inversion layer electron mobility versus effective transverse field, at various temperatures. The triangles, circles, and squares refer to different MOSFETs with different gate oxide thicknesses and channel dopings. (After Sabnis and Clemens, *IEEE IEDM*, 1979.)

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 301

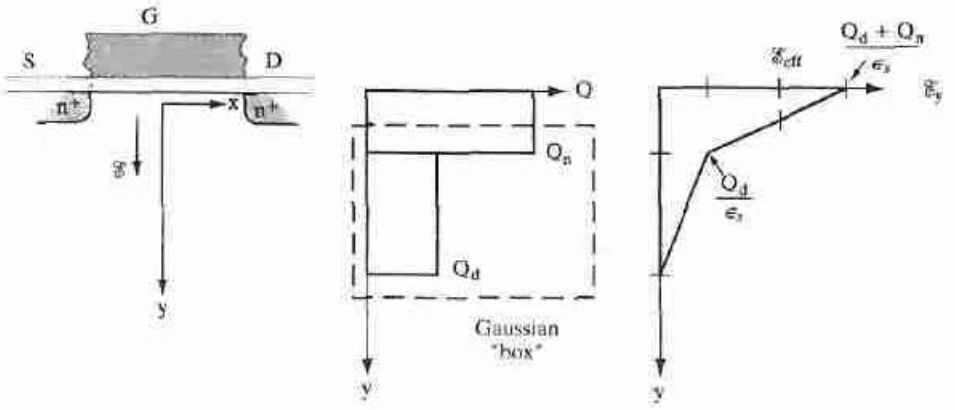


Figure 6–31
Determination of effective transverse field. Idealized charge distribution and transverse electric field in the inversion layer and depletion layer, as a function of depth in the channel of a MOSFET. The region to which we apply Gauss's law is shown in color.

longitudinal electric field. As shown in Fig. 3-24, the carrier drift velocity increases linearly with electric field (ohmic behavior) until the field reaches $\mathcal{E}_{sat}$; in other words, the mobility is constant up to $\mathcal{E}_{sat}$. After this, the velocity saturates at v_s, and it can no longer be described in terms of mobility. These effects can be described as:

$$v = \mu\mathcal{E} \text{ for } \mathcal{E} < \mathcal{E}_{sat} \tag{6–57}$$

$$\text{and } v = v_s \text{ for } \mathcal{E} > \mathcal{E}_{sat} \tag{6–58}$$

The maximum longitudinal electric field near the drain end of the channel is approximately given by the voltage drop along the pinch-off region ($V_D - V_D(\text{sat.})$), divided by the length of the pinch-off region, ΔL:

$$\mathcal{E}_{max} = \left(\frac{V_D - V_D(\text{sat.})}{\Delta L}\right) \tag{6–59}$$

From a two-dimensional solution of the Poisson equation near the drain end, one can show that the pinch-off region ΔL shown in Fig. 6-11c is approximately equal to $\sqrt{(3dx_j)}$, where d is the gate oxide thickness and x_j is the source/drain junction depth. The factor of 3 is due to the ratio of the dielectric constant for Si to that of SiO_2.

6.5.4 Short Channel MOSFET I–V Characteristics

In short channel devices, the analysis has to be somewhat modified. As mentioned in the previous section, the effective channel mobility decreases with increasing transverse electric field perpendicular to the gate oxide (i.e., the

gate bias). Furthermore, for very high longitudinal electric fields in the pinch-off region, the carrier velocity saturates (Fig. 3-24). For short channel lengths, the carriers travel at the saturation velocity over most of the channel. In that case, the drain current is given by the width times the channel charge per unit area times the saturation velocity:

$$I_D(\text{sat.}) \approx ZC_i(V_G - V_T)v_s \tag{6-60}$$

As a result, the saturation drain current does not increase quadratically with $(V_G - V_T)$ as shown in Eq. (6-53), but rather shows a linear dependence (note the equal spacing of curves in Fig. 6-32). Due to the advances in Si device processing, particularly photolithography, MOSFETs used in modern integrated circuits tend to have short channels, and are commonly described by Eq. (6-60) rather than Eq. (6-53).

6.5.5 Control of Threshold Voltage

Since the threshold voltage determines the requirements for turning the MOS transistor on or off, it is very important to be able to adjust V_T in designing the

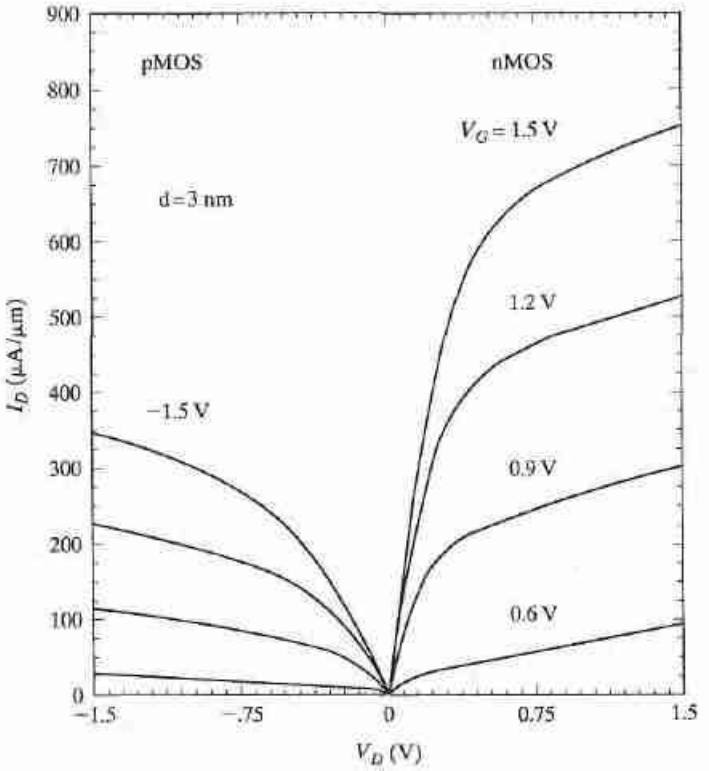


Figure 6–32 Experimental output characteristics of n-channel and p-channel MOSFETs with 0.1-μm channel lengths. The curves exhibit almost equal spacing, indicating a linear dependence of I_D on V_G, rather than a quadratic dependence. We also see that I_D is not constant but increases somewhat with V_D in the saturation region. The p-channel devices have lower currents because hole mobilities are lower than electron mobilities.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 303

device. For example, if the transistor is to be used in a circuit driven by a 3-V battery, it is clear that a 4-V threshold voltage is unacceptable. Some applications require not only a low value of V_T, but also a precisely controlled value to match other devices in the circuit.

All of the terms in Eq. (6-38) can be controlled to some extent. The work function potential difference Φ_{ms} is determined by choice of the gate conductor material; ϕ_F depends on the substrate doping; Q_i can be reduced by proper oxidation methods and by using Si grown in the (100) orientation; Q_d can be adjusted by doping of the substrate; and C_i depends on the thickness and dielectric constant of the insulator. We shall discuss here several methods of controlling these quantities in device fabrication.

Choice of Gate Electrode. Since V_T depends on Φ_{ms}, the choice of the gate electrode material (i.e., the gate electrode work function) has an impact on the threshold voltage. When MOSFETs were first made in the 1960s, they used Al gates. However, since Al has a low melting point, it precluded the use of a self-aligned source/drain technology because that required a high temperature source/drain implant anneal after the gate formation. Hence, Al was supplanted by n^+ doped LPCVD polysilicon *refractory* (high melting point) gates, where the Fermi level lines up with the conduction band edge in Si. While this works quite well for n-channel MOSFETs, we shall see in Section 9.3.1 that it can create problems for p-channel MOSFETs. Therefore, sometimes, a p^+ doped polysilicon gate is used for p-channel devices. Refractory metal gates with suitable work functions are also being researched as possible replacements for doped polysilicon. One attractive candidate is tungsten,whose work function is such that the Fermi level happens to lie near the midgap of Si.

Control of C_i. Since a low value of V_T and a high drive current is usually desired, a thin oxide layer is used in the gate region to increase $C_i = \epsilon_i/d$ in Eq. (6-38). From Fig. 6-20 we see that increasing C_i makes V_T less negative for p-channel devices and less positive for n-channel with $-Q_d > Q_i$. For practical considerations, the gate oxide thickness is generally 20-100 Å (2-10 nm) in modern devices having submicron gate length. An example of such a device is shown in Figure 6-33. The gate oxide, easily observable in this micrograph, is 40 Å thick. The interfacial layer between the crystalline silicon and the amorphous SiO_2 is also observable.

Although a low threshold voltage is desirable in the gate region of a transistor, a large value of V_T is needed between devices. For example, if a number of transistors are interconnected on a single Si chip, we do not want inversion layers to be formed inadvertently between devices (generally called the *field*). One way to avoid such parasitic channels is to increase V_T in the field by using a very thick oxide. Figure 6-34 illustrates a transistor with a gate oxide 10 nm thick and a field oxide of 0.5 μm.

The value of C_i can also be controlled by varying ϵ_i. A SiO_2 layer which has some N incorporated in it, leading to the formation of a silicon

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 304

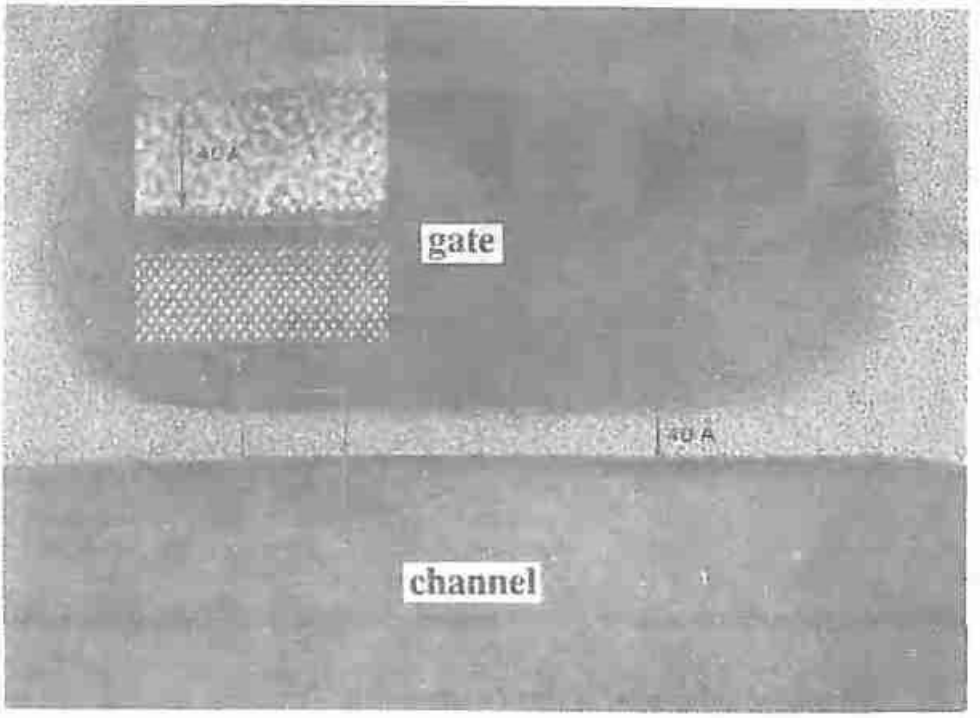


Figure 6–33
Cross section of a MOSFET. This high resolution transmission electron micrograph of a silicon Metal–Oxide Semiconductor Field Effect Transistor shows the silicon channel and metal gate separated by a thin (40 Å, 4 nm) silicon–dioxide insulator. The inset shows a magnified view of the three regions, in which individual rows of atoms in the crystalline silicon can be distinguished. (Photograph courtesy of AT&T Bell Laboratories.)

oxynitride, is often used. Such silicon oxynitrides have slightly higher ϵ_i and C_i than SiO_2, with excellent interface properties. Other high dielectric constant materials such as Ta_2O_5, ZrO_2 and ferroelectrics (e.g., barium–strontium–titanate) are also being investigated as replacements for SiO_2 as the gate dielectric in MOSFETs in order to increase $C_i = \epsilon_i/d$ and, therefore, the drive current of the MOSFET. Generally speaking, we cannot use these high dielectric constant materials directly on the Si substrate; a very thin (~0.5 nm) interfacial SiO_2 layer is needed to achieve a low fast interface state density. It is clear from the expression for C_i that for these high dielectric constant materials, a physically thicker layer, d, can be used than for SiO_2 and still achieve a certain C_i. This is very useful for reducing the tunneling leakage current through the gate dielectric, discussed in Section 6.4.7. A physically thicker layer implies a wider tunneling barrier with a reduced tunneling probability.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 305

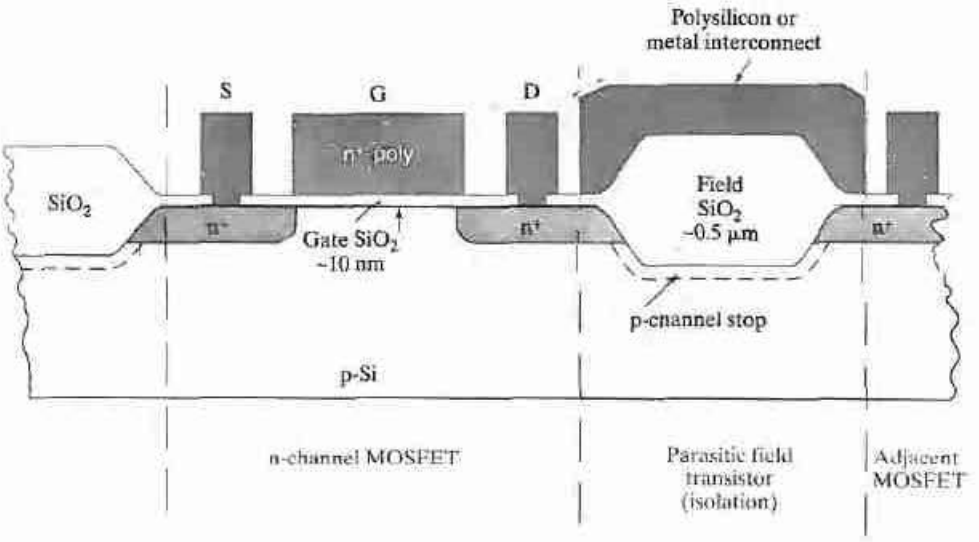


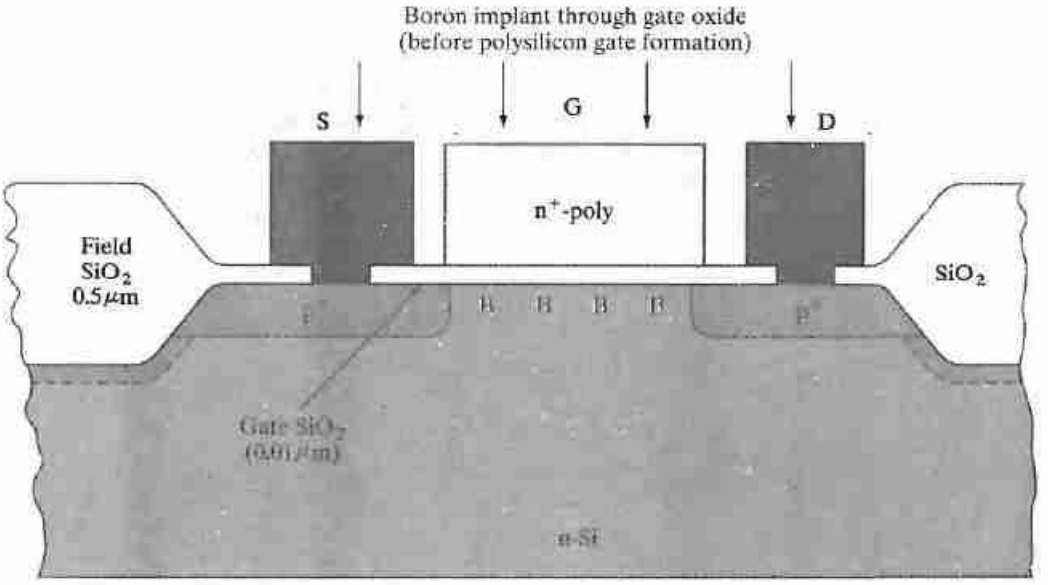
Figure 6–34
Thin oxide in the gate region and thick oxide in the field between transistors for V_T control (not to scale).

Threshold Adjustment by Ion Implantation. The most valuable tool for controlling threshold voltage is ion implantation (Section 5.1.4). Since very precise quantities of impurity can be introduced by this method, it is possible to maintain close control of V_T. For example, Fig. 6–35 illustrates a boron implantation through the gate oxide of a p-channel device such that the implanted peak occurs just below the Si surface. The negatively charged boron acceptors serve to reduce the effects of the positive depletion charge Q_d. As a result, V_T becomes less negative. Similarly, a shallow boron implant into the p-type substrate of an n-channel transistor can make V_T positive, as required for an enhancement device.

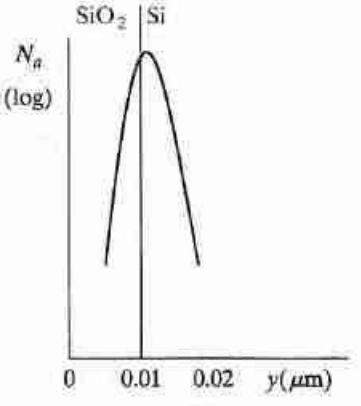
If the implantation is performed at higher energy, or into the bare Si instead of through an oxide layer, the impurity distribution lies deeper below the surface. In such cases the essentially gaussian impurity concentration profile cannot be approximated by a spike at the Si surface. Therefore, effects of distributed charge on the Q_d term of Eq. (6–38) must be considered. Calculations of the effects on V_T in this case are more complicated, and the shift of threshold voltage with implantation dose is often obtained empirically instead.

The implantation energy required for shallow V_T adjustment implants is low (50–100 keV), and relatively low doses are needed. A typical V_T adjustment requires only about 10 s of implantation for each wafer, and therefore this procedure is compatible with large-scale production requirements.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 306



(a)



(b)

Figure 6–35
Adjustment of V_T in a p-channel transistor by boron implantation: (a) boron ions are implanted through the thin gate oxide but are absorbed within the thick oxide regions; (b) variation of implanted boron concentration in the gate region—here the peak of the boron distribution lies just below the Si surface.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 307

EXAMPLE 6–3 For a p-channel transistor with a gate oxide thickness of 10 nm, calculate the boron ion dose F_B(B^+ ions/cm^2) required to reduce V_T from – 1.1 V to – 0.5 V. Assume that the implanted acceptors form a sheet of negative charge just below the Si surface. If, instead of a shallow B implant, it was a much broader distribution, how would the V_T calculation change? Assuming a boron ion beam current of 10^{-5} A, and supposing that the area scanned by the ion beam is 650 cm^2, how long does this implant take?

SOLUTION Since the B negative ions are assumed to form a sheet of charge just below the gate oxide, they affect the fixed oxide charge term in V_{FB}. Thus,

$$C_i = \frac{\epsilon_i}{d} = \frac{(3.9)(8.85 \times 10^{-14})}{10^{-6}} = 3.45 \times 10^{-7}\ \text{F/cm}^2$$

$$-0.5 = -1.1 + \frac{qF_B}{C_i}$$

$$F_B = \frac{3.45 \times 10^{-7}}{1.6 \times 10^{-19}}(0.6) = 1.3 \times 10^{12}\ \text{cm}^{-2}$$

If the B distribution is deeper, we cannot assume it to be a sheet charge in V_{FB}. If it is approximately *constant over the maximum depletion width*, we should instead change the substrate doping term in the expression for V_T. Otherwise, the calculation has to be done by numerically solving Poisson's equation in the depletion region, with a varying doping concentration corresponding to the B distribution, to obtain the voltage drop in the semiconductor.

For a beam current of 10 μA scanned over a 650-cm^2 target area,

$$\frac{10^{-5}(\text{C/s})}{650\ \text{cm}^2}\, t(s) = 1.3 \times 10^{12}\ (\text{ions/cm}^2) \times 1.6 \times 10^{-19}(\text{C/ion})$$

The implant time is $t = 13.5$ s.

If the implantation is continued to higher doses, V_T can be moved past zero to the *depletion-mode* condition (Fig. 6–36). This capability provides considerable flexibility to the integrated-circuit designer, by allowing enhancement- and depletion-mode devices to be incorporated on the same chip. For example, a depletion-mode transistor can be used instead of a resistor as a load element for the enhancement device. Thus an array of MOS transistors can be fabricated in an IC layout, with some adjusted by implantation to have the desired enhancement mode V_T and others implanted to become depletion loads.

As mentioned, V_T control is important not only in the MOSFETs but also in the isolation or field regions. In addition to using a thick field oxide, we can do a *channel stop implant* (so called because it stops turning on a parasitic channel in the isolation regions) selectively in the isolation regions under the

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 308

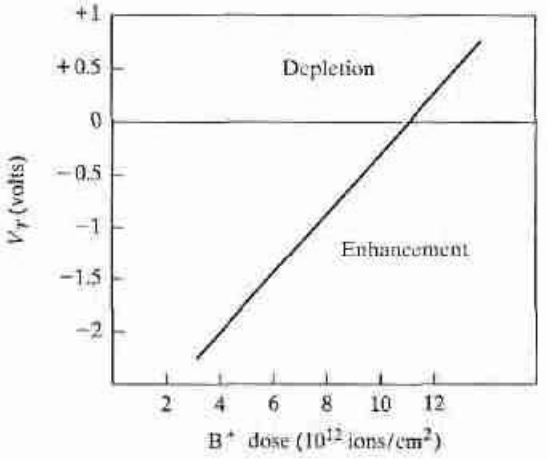


Figure 6–36
Typical variation of V_T for a p-channel device with increased implanted boron dose. The originally enhancement p-channel transistor becomes a depletion-mode device ($V_T > 0$) by sufficient B implantation.

field oxide (Fig. 6–34). Generally, a B channel stop implant is used for n-channel devices. (It must be noted that such an acceptor implant will raise the field thresholds for n-channel MOSFETs made in a p-substrate, but will *decrease* the field thresholds for p-MOSFETs made in an n-substrate).

6.5.6 Substrate Bias Effects

In the derivation of Eq. (6–49) for current along the channel, we assumed that the source S was connected to the substrate B (Fig. 6–27). In fact, it is possible to apply a voltage between S and B (Fig. 6–37). With a reverse bias between the substrate and the source (V_B negative for an n-channel device), the depletion region is widened and the threshold gate voltage required to achieve inversion must be increased to accommodate the larger Q_d'. A simplified view of the result is that W is widened uniformly along the channel, so that Eq. (6–32) should be changed to

$$Q_d' = -[2\epsilon_s q N_a(2\phi_F - V_B)]^{1/2} \tag{6-61}$$

The change in threshold voltage due to the substrate bias is

$$\Delta V_T = \frac{\sqrt{2\epsilon_s q N_a}}{C_i}[(2\phi_F - V_B)^{1/2} - (2\phi_F)^{1/2}] \tag{6-62}$$

If the substrate bias V_B is much larger than $2\phi_F$ (typically ~0.6 V), the threshold voltage is dominated by V_B and

$$\Delta V_T \approx \frac{\sqrt{2\epsilon_s q N_a}}{C_i}(-V_B)^{1/2} \quad \text{(n channel)} \tag{6-63}$$

where V_B will be negative for the n-channel case. As the substrate bias is increased, the threshold voltage becomes more positive. The effect of this bias becomes more dramatic as the substrate doping is increased, since ΔV_T is also proportional to $\sqrt{N_a}$. For a p-channel device the bulk-to-source voltage

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 309

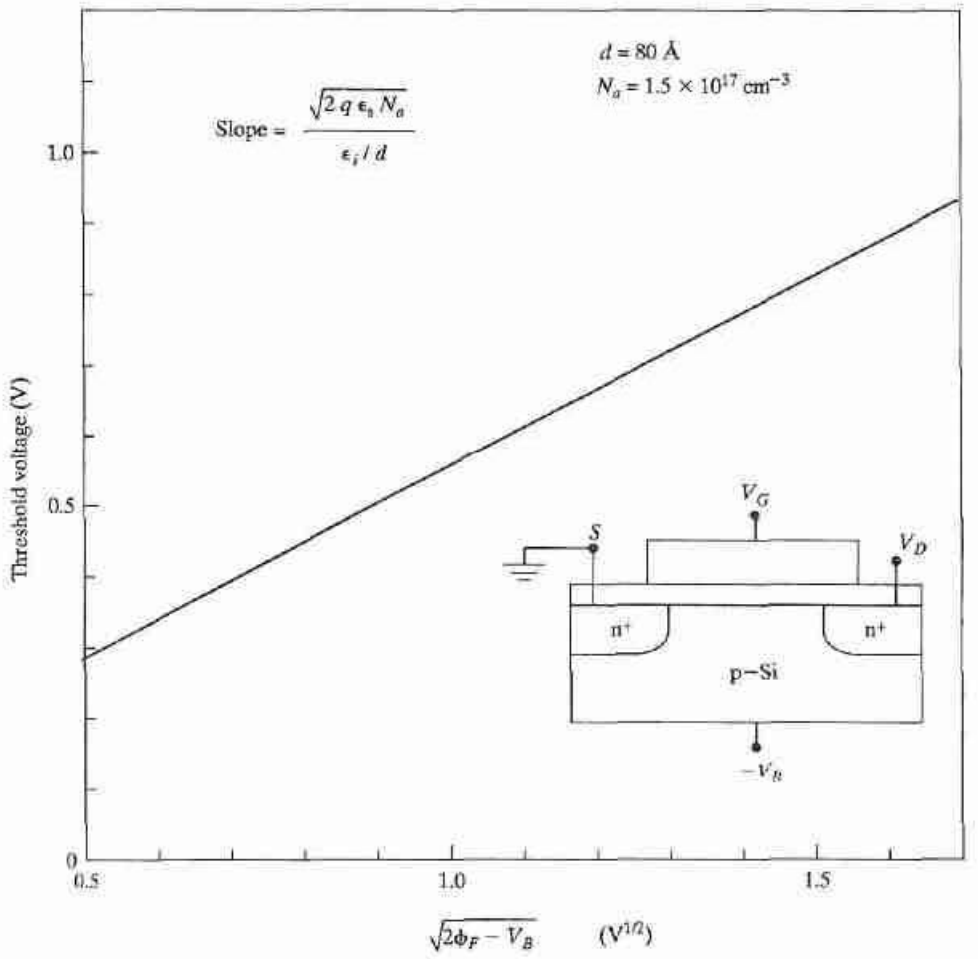


Figure 6–37
Threshold voltage dependence on substrate bias resulting from application of a voltage V_B from the substrate (i.e., bulk) to the source. For n channel, V_B must be zero or negative to avoid forward bias of the source junction. For p channel, V_B must be zero or positive.

V_B is positive to achieve a reverse bias, and the approximate change ΔV_T for $V_B \gg 2\phi_F$ is

$$\Delta V_T \simeq -\frac{\sqrt{2\epsilon_s q N_d}}{C_i} V_B^{1/2} \quad \text{(p channel)} \tag{6-64}$$

Thus the p-channel threshold voltage becomes more negative with substrate bias.

The substrate bias effect (also called the *body effect*) increases V_T for either type of device. This effect can be used to raise the threshold voltage

of a marginally enhancement device ($V_T \simeq 0$) to a somewhat larger and more manageable value. This can be an asset for n-channel devices particularly (see Fig. 6-20). The effect can present problems, however, in MOS integrated circuits for which it is impractical to connect each source region to the substrate. In these cases, possible V_T shifts due to the body effect must be taken into account in the circuit design.

6.5.7 Subthreshold Characteristics

If we look at the drain current expression (Eq. 6-53), it appears that the current abruptly goes to zero as soon as V_G is reduced to V_T. In reality, there is still some drain conduction below threshold, and this is known as *subthreshold conduction*. This current is due to weak inversion in the channel between flat band and threshold (for band-bending between zero and $2\phi_F$), which leads to a diffusion current from source to drain. The drain current in the subthreshold region is equal to

$$I_D = \mu(C_d + C_{it})\frac{Z}{L}\left(\frac{kT}{q}\right)^2\left(1 - e^{\frac{-qV_D}{kT}}\right)\left(e^{\frac{q(V_G - V_T)}{c_r kT}}\right) \qquad (6\text{-}65)$$

where

$$c_r = \left[1 + \frac{C_d + C_{it}}{C_i}\right]$$

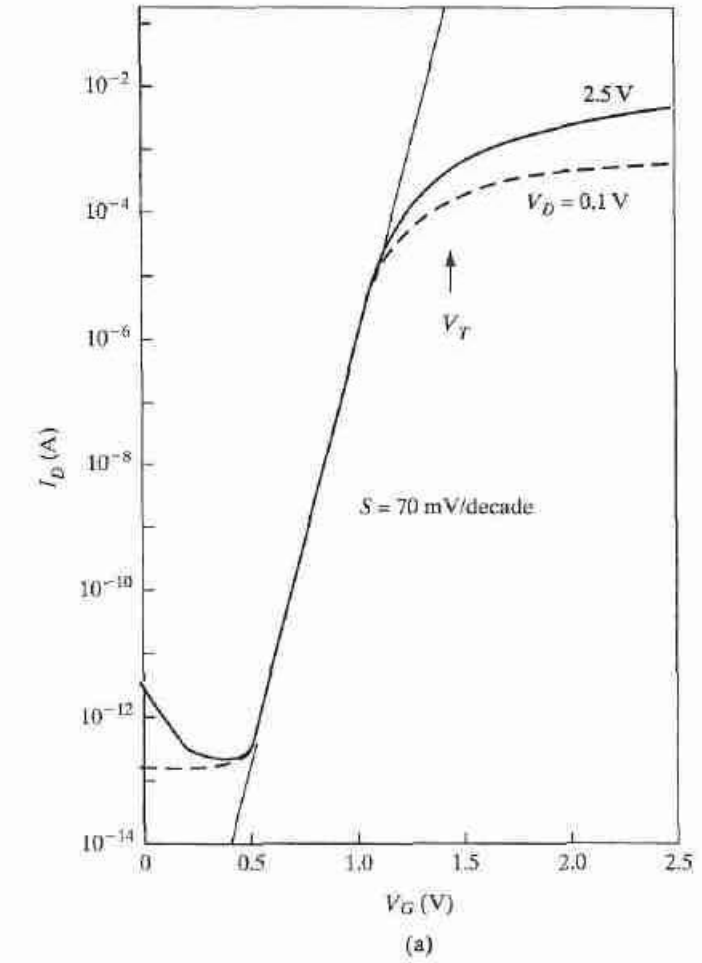
It can be seen that I_D depends exponentially on gate bias, V_G. However, V_D has little influence once V_D exceeds a few kT/q. Obviously, if we plot ln I_D as a function of gate bias V_G, we should get a linear behavior in the subthreshold regime, as shown in Fig. 6-38a. The slope of this line (or more precisely the reciprocal of the slope) is known as the subthreshold slope, S, which has typical values of ~70 mV/decade at room temperature for state-of-the-art MOSFETs. This means that a change in the input V_G of 70 mV will change the output I_D by an order of magnitude. Clearly, the smaller the value of S, the better the transistor is as a switch. A small value of S means a small change in the input bias can modulate the output current considerably.

It can be shown that the expression for S is given by

$$S = \frac{dV_G}{d(\log I_D)} = \ln 10 \frac{dV_G}{d(\ln I_D)} = 2.3\frac{kT}{q}\left[1 + \frac{C_d + C_{it}}{C_i}\right] \qquad (6\text{-}66)$$

Here, the factor ln 10 (= 2.3) is introduced to change from $\log_{10}$ to natural logarithm, ln. This equation can be understood by looking at the electrical equivalent circuit of the MOSFET in terms of the capacitors (Fig. 6-38b). Between the gate and the substrate, we find the gate capacitance, C_i, in series with the parallel combination of the depletion capacitance in the channel, C_d, and the fast interface state capacitance, $C_{it} = qD_{it}$. The expression in brackets in Eq. (6-66) is simply the capacitor divider ratio which tells us what

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 311



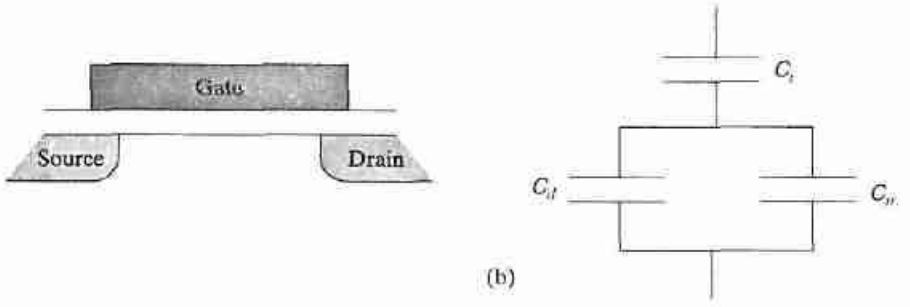


Figure 6–38
Subthreshold conduction in MOSFETs: (a) Semilog plot of I_D versus V_G; (b) equivalent circuit showing capacitor divider which determines subthreshold slope.

fraction of the applied gate bias, V_G, appears at the Si-SiO_2 interface as the surface potential. Ultimately, it is the surface potential that is responsible for modulating the barrier between source and drain, and therefore the drain current. Hence, S is a measure of the efficacy of the gate potential in modulating I_D. From Eq. (6-66), we observe that S is improved by reducing the gate oxide thickness, which is reasonable because if the gate electrode is closer to the channel, the gate control is obviously better. The value of S is higher for heavy channel doping (which increases the depletion capacitance) or if the silicon–oxide interface has many fast interface states.

For a very small gate voltage, the subthreshold current is reduced to the leakage current of the source/drain junctions. This determines the off-state leakage current, and therefore the standby power dissipation in many complementary MOS (CMOS) circuits involving both n-channel and p-channel MOSFETs. It also underlines the importance of having high quality source/drain junctions. From the subthreshold characteristics, it can be seen that if the V_t of a MOSFET is too low, it cannot be turned off fully at $V_G = 0$. Also, unavoidable statistical variations of V_T cause drastic variations of the subthreshold leakage current. On the other hand, if V_T is too high, one sacrifices drive current, which depends on the difference between the power supply voltage and V_T. For these reasons, the V_T of MOSFETs has historically been designed to be ~0.7 V. However, with the recent advent of various types of low-voltage, low-power portable electronics, there are new challenges in device and circuit design to optimize speed and power dissipation.

6.5.8 Equivalent Circuit for the MOSFET

When we attempt to draw an equivalent circuit of a MOSFET, we find that in addition to the intrinsic MOSFET itself, there are a variety of parasitic elements associated with it. An important addition to the gate capacitance is the so-called *Miller overlap capacitance* due to the overlap between the gate and the drain region (Fig. 6–39). This capacitance is particularly problematic because it represents a feedback path between the output drain terminal and the input gate terminal. One can measure the Miller capacitance at high frequency by holding the gate at ground ($V_G = 0$) so that an inversion layer is not formed in the channel. Thereby, most of the measured capacitance between gate and drain is due to the Miller capacitance, rather than the gate capacitance C_i. It is possible to minimize this capacitance by using a so-called *self-aligned gate*. In this process, the gate itself is used to mask the source/drain implants, thereby achieving alignment. Even in this design, however, there is still a certain amount of overlap because of the lateral straggle or spread of the implanted dopants underneath the gate, further exacerbated by the lateral diffusion which occurs during high temperature annealing. This spread of the source/drain junctions under the gate edge determines what is called the channel length reduction, ΔL_R (Fig. 6–40). Hence, we get the electrical or "effective" channel length, L_{eff}, in terms of the physical gate length, L as

$$L_{eff} = L - \Delta L_R \tag{6-67}$$

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 313

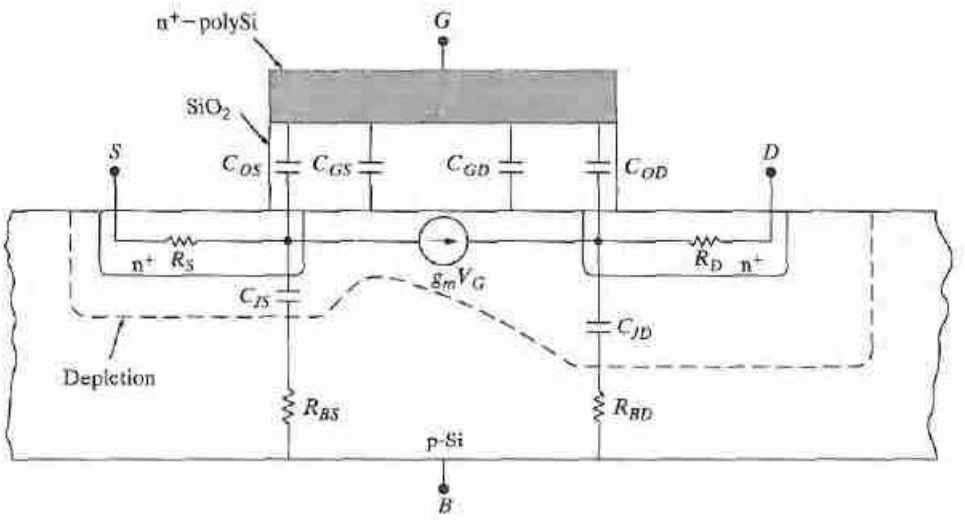


Figure 6–39
Equivalent circuit of a MOSFET, showing the passive capacitive and resistive components. The gate capacitance C_i is the sum of the distributed capacitances from the gate to the source end of the channel (C_{GS}) and the drain end (C_{GD}). In addition, we have an overlap capacitance (where the gate electrode overlaps the source/drain junctions) from the gate-to-source (C_{OS}) and gate-to-drain (C_{OD}). C_{OD} is also known as the Miller overlap capacitance. We also have p-n junction depletion capacitances associated with the source (C_{JS}) and drain (C_{JD}). The parasitic resistances include the source/drain series resistances (R_S and R_D), and the resistances in the substrate between the bulk contact and the source and drain (R_{BS} and R_{BD}). The drain current can be modeled as a (gate) voltage-controlled constant-current source.

There can also be a width reduction, ΔZ, which changes the effective width, Z_{eff}, from the physical width Z of the MOSFET. The width reduction results from the electrical isolation regions that are formed around all transistors, generally by LOCOS. The LOCOS isolation technique is discussed in Section 9.3.1.

Another very important parameter in the equivalent circuit is the source/drain series resistance, $R_{SD} = (R_S + R_D)$, because it degrades the drain current and transconductance. For a certain applied drain bias to the source/drain terminals, part of the applied voltage is "wasted" as an ohmic voltage drop across these resistances, depending on the drain current (or gate bias). Hence, the actual drain voltage applied to the intrinsic MOSFET itself is less; this causes I_D to increase sub-linearly with V_G.

We can determine R_{SD}, along with ΔL_R, from the overall resistance of the MOSFET in the linear region,

$$\left(\frac{V_D}{I_D}\right)$$

This corresponds to the intrinsic channel impedance R_{Ch}, plus the source-drain resistance R_{SD}. Modifying Eq. (6-51) we get

$$\frac{V_D}{I_D} = R_{Ch} + R_{SD} = \frac{L - \Delta L_R}{Z - \Delta Z} \frac{1}{\overline{\mu_n C_i (V_G - V_T)}} + R_{SD} \qquad (6\text{-}68)$$

We can measure V_D/I_D in the linear range for various MOSFETs having the same width, but different channel lengths, as a function of substrate bias. Varying the substrate bias changes the V_T through the body effect, and therefore the slope of the straight lines that result from plotting the overall resistance as a function of L. The lines pass through a point, having values which correspond to ΔL_R and R_{SD}, as shown in Fig. 6-40.

6.5.9 MOSFET Scaling and Hot Electron Effects

Much of the progress in semiconductor integrated circuit technology can be attributed to the ability to shrink or *scale* the devices. Scaling down MOSFETs has a multitude of benefits. From Table 6-1, we see the benefits of scaling in terms of the improvement of packing density, speed and power dissipation. A key concept in scaling, first due to Dennard at IBM, is that the various structural parameters of the MOSFET should be scaled in concert if the device is to keep functioning properly. In other words, if lateral dimensions such as the channel length and width are reduced by a factor of K, so should the vertical dimensions such as source/drain junction depths (x_j) and gate insulator thickness (Table 6-1). Scaling of depletion widths is achieved indirectly by scaling up doping concentrations. However, if we simply reduced the dimensions of the device and kept the power supply voltages the same, the internal electric fields in the device would increase. For ideal scaling, power supply voltages should also be reduced to keep the internal electric fields reasonably constant from one technology generation to the next. Unfortunately, in practice, power supply voltages are not scaled hand-in-hand with the device dimensions, partly because of other system-related constraints. The longitudinal electric fields in the pinch-off region, and the transverse electric fields across the gate oxide, increase with MOSFET scaling. A variety of problems then arise which are generically known as hot electron effects and short channel effects (Fig. 6-41).

When an electron travels from the source to the drain along the channel, it gains kinetic energy at the expense of electrostatic potential energy in the pinch-off region, and becomes a "hot" electron. At the conduction band edge, the electron has potential energy only; as it gains more kinetic energy, it moves higher up in the conduction band. A few of the electrons can become energetic enough to surmount the 3.1-eV potential barrier between the Si channel and the gate oxide (Fig. 6-25). Some of these injected hot electrons can go through the gate oxide and be collected as gate current, thereby reducing the input impedance. More importantly, some of these electrons can

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 315

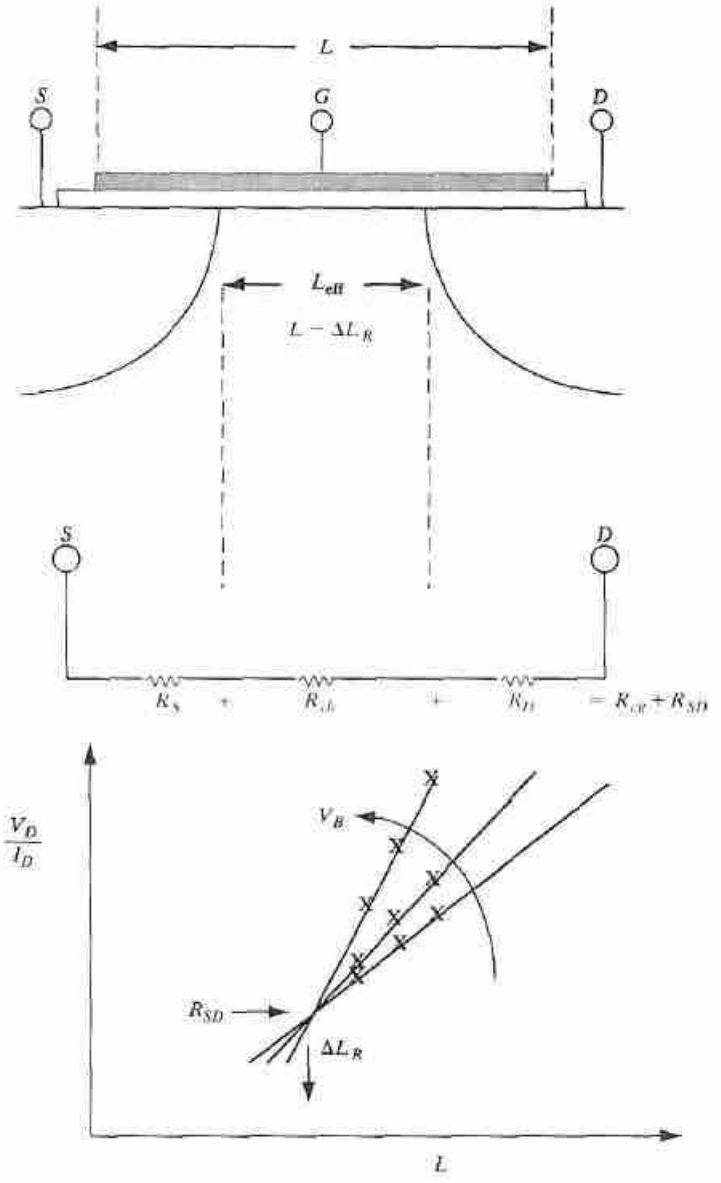


Figure 6–40
Determination of length reduction and source/drain series resistance in a MOSFET. The overall resistance of a MOSFET in the linear region is plotted as a function of channel length, for various substrate biases. The x's mark data points for three different physical gate lengths L.

be trapped in the gate oxide as fixed oxide charges. According to Eq. (6-37), this increases the flat band voltage, and therefore the V_T. In addition, these energetic hot carriers can rupture Si-H bonds that exist at the Si-SiO_2 interface, creating fast interface states that degrade MOSFET parameters such as transconductance and subthreshold slope, with stress. The results of such hot carrier degradation are shown in Fig. 6-42, where we see the increase of V_T and decrease of slope, and therefore transconductance, with stress. The solution to this problem is to use what is known as a *lightly doped drain* (LDD).

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 316

Table 6–1 Scaling rules for MOSFETs according to a constant factor K. The horizontal and vertical dimensions are scaled by the same factor. The voltages are also scaled to keep the internal electric fields more or less constant, and the hot carrier effects manageable.

	Scaling factor
Surface dimensions (L,Z)	1/K
Vertical dimensions (d,x_j)	1/K
Impurity Concentrations	K
Current, Voltages	1/K
Current Density	K
Capacitance (per unit area)	K
Transconductance	1
Circuit Delay Time	1/K
Power Dissipation	$1/K^2$
Power Density	1
Power-Delay Product	$1/K^3$

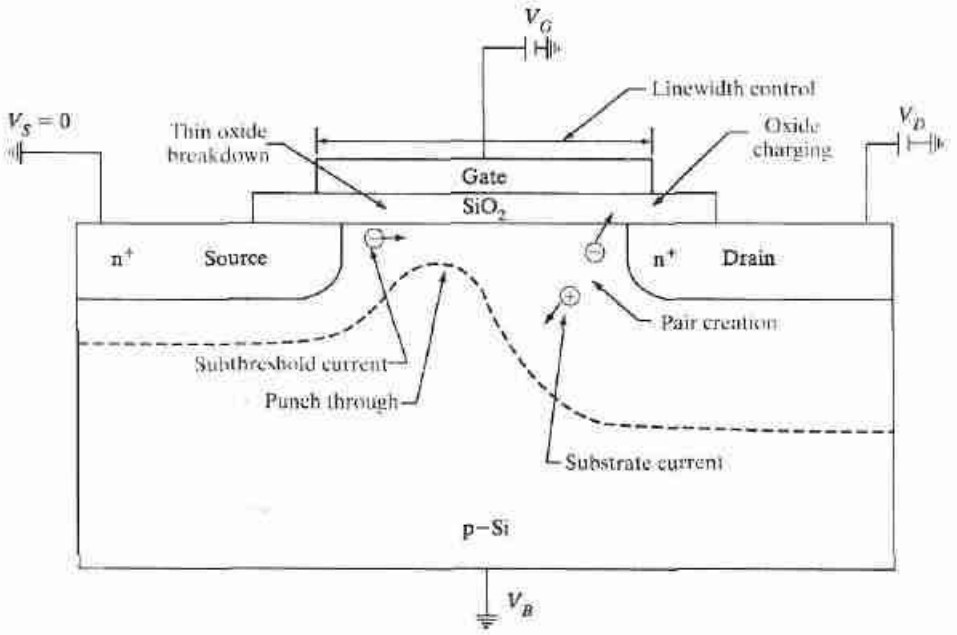


Figure 6–41
Short channel effects in MOSFETs. As MOSFETs are scaled down, potential problems due to short channel effects include hot carrier generation (electron–hole pair creation) in the pinch-off region, punch-through breakdown between source and drain, and thin gate oxide breakdown.

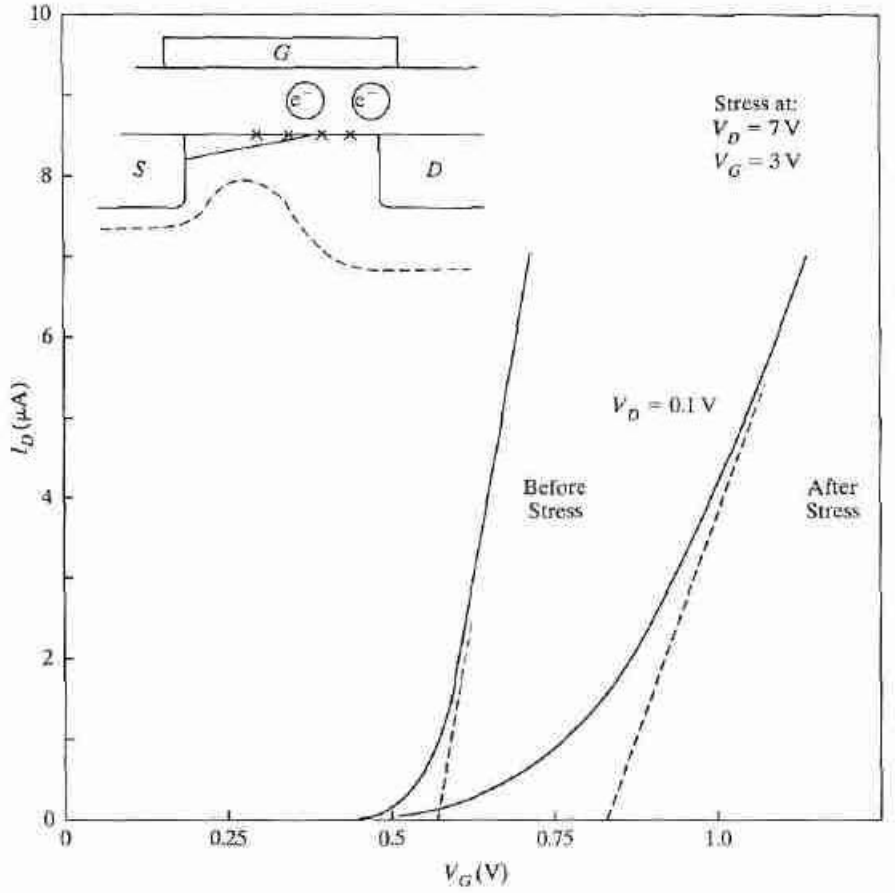


Figure 6–42
Hot carrier degradation in MOSFETs. The linear region transfer characteristics before and after hot carrier stress indicate an increase of V_T and decrease of transconductance (or channel mobility) due to hot electron damage. The damage can be due to hot electron injection into the gate oxide which increases the fixed oxide charge, and increasing fast interface state densities at the oxide-silicon interface (indicated by x).

As discussed in more detail in Section 9.3.1, by reducing the doping concentration in the source/drain, the depletion width at the reverse-biased drain-channel junction is increased and the electric field is reduced.

Hot carrier effects are less problematic for holes in p-channel MOSFETs than for electrons in n-channel devices for two reasons. The channel mobility of holes is approximately half that of electrons; hence, for the same electric field, there are fewer hot holes than hot electrons. Unfortunately, the lower hole mobility is also responsible for lower drive currents in p-channel than in n-channel. Also, the barrier for hole injection in the valence band between Si and SiO_2 is higher (5 eV) than for electrons in the conduction band (3.1 eV), as

shown in Fig. 6–25. Hence, while LDD is mandatory for n-channel, it is often not used for p-channel devices.

One "signature" for hot electron effects is substrate current (Fig. 6–43). As the electrons travel towards the drain and become hot, they can create secondary electron-hole pairs by impact ionization (Fig. 6–41). The secondary electrons are collected at the drain, and cause the drain current in saturation to increase with drain bias at high voltages, thereby leading to a decrease of the output impedance. The secondary holes are collected at the substrate as substrate current. This current can create circuit problems such as noise or latchup in CMOS circuits (Section 9.3.1). It can also be used as a monitor for hot electron effects. As shown in Fig. 6–43, substrate current initially increases with gate bias (for a fixed, high drain bias), goes through a peak and then

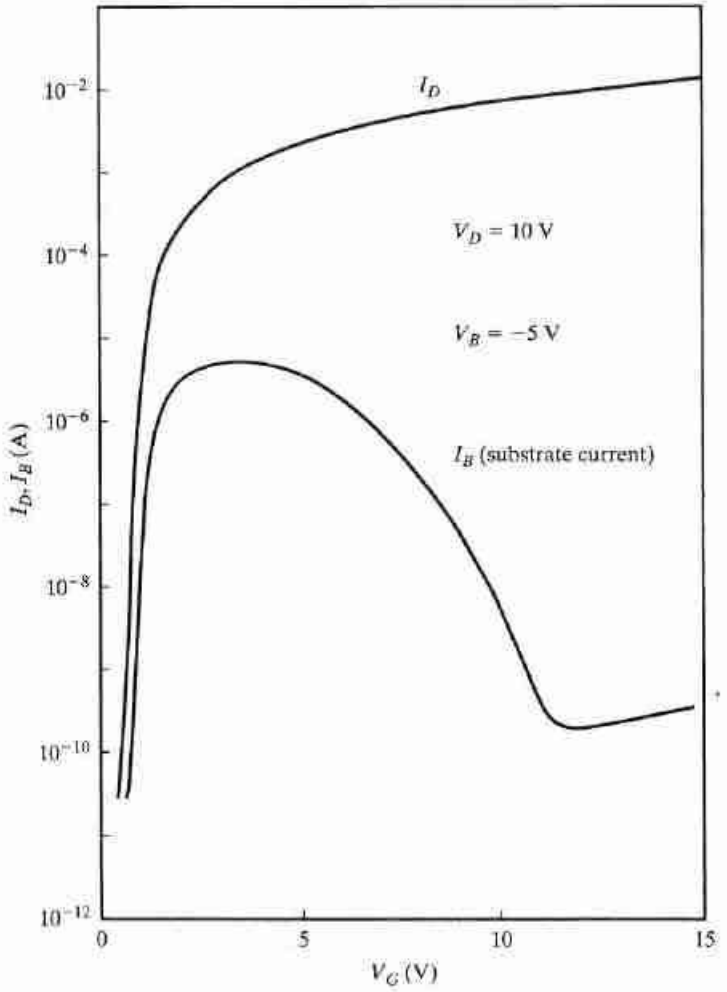


Figure 6–43
Substrate current in a MOSFET. The substrate current in n-channel MOSFETs due to impact-generated holes in the pinch-off region, as a function of gate bias. The substrate current initially increases with V_G because of the corresponding increase of I_D. However, for even higher V_G, the MOSFET goes from the saturation to the linear region, and the high electric fields in the pinch-off region decrease, causing less impact ionization. (After Kamata, et. al., *Jpn. J. Appl. Phys.*, 15 (1976), 1127.)

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 319

decreases. The reason for this behavior is that initially, as the gate bias increases, the drain current increases and thereby provides more primary carriers into the pinch-off region for impact ionization. However, for even higher gate bias, the MOSFET goes from the saturation region into the linear region when the fixed V_D drops below $V_D(\text{sat.}) = (V_G - V_T)$. The longitudinal electric field in the pinch-off region drops, thereby reducing the impact ionization rates. Hot electron reliability studies are done under "worst case" conditions of peak substrate current. These are generally done under accelerated conditions of higher-than-normal operating voltages so that if there are any potential problems, they show up in a reasonable time period. The degradation data is then extrapolated to the actual operating conditions.

6.5.10 Drain-Induced Barrier Lowering

If small channel length MOSFETs are not scaled properly, and the source/drain junctions are too deep or the channel doping is too low, there can be unintended electrostatic interactions between the source and the drain known as *Drain-Induced Barrier Lowering (DIBL)*. This leads to punch-through leakage or breakdown between the source and the drain, and loss of gate control. The phenomenon can be understood from Fig. 6–44, where we have schematically plotted the surface potential along the channel for a long channel device and a short device. We see that as the drain bias is increased, the conduction band edge (which reflects the electron energies) in the drain is pulled down, and the drain-channel depletion width expands. For a long channel MOSFET, the drain bias does not affect the source-to-channel potential barrier, which corresponds to the built-in potential of the source-channel p-n junction. Hence, unless the gate bias is increased to lower this potential barrier, there is little drain current. On the other hand, for a short channel MOSFET, as the drain bias is raised and the conduction band edge in the drain is pulled down (with a concomitant increase of the drain depletion width), the source-channel potential barrier is lowered due to DIBL. This can be shown numerically by a solution of the two-dimensional Poisson equation in the channel region. Simplistically, the onset of DIBL is sometimes considered to correspond to the drain depletion region expanding and merging with the source depletion region, and causing punch-through breakdown between source and drain. However, it must be kept in mind that DIBL is ultimately caused by the lowering of the source-junction potential barrier below the built-in potential. Hence, if we get DIBL in a MOSFET for a grounded substrate, the problem can be mitigated by applying a substrate reverse bias, because that raises the potential barrier at the source end. This works in spite of the fact that the drain depletion region interacts even more with the source depletion region under such back bias. Once the source-channel barrier is lowered by DIBL, there can be significant drain leakage current, with the gate being unable to shut it off.

What are the solutions to this problem? The source/drain junctions must be made sufficiently shallow (i.e., scaled properly) as the channel lengths are reduced, to prevent DIBL. Secondly, the channel doping must be made

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 320

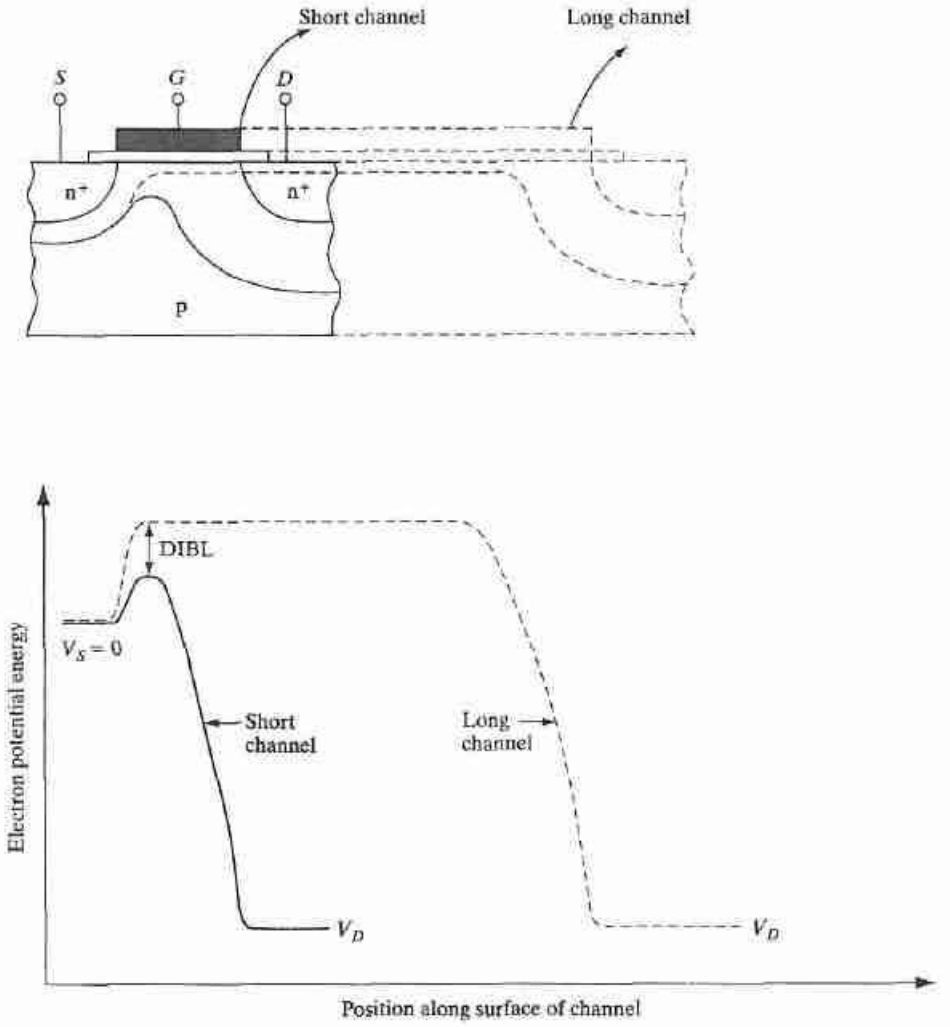


Figure 6–44
Drain-induced barrier lowering in MOSFETs. Cross sections and potential distribution along the channel for a long channel and short channel MOSFET.

sufficiently high to prevent the drain from being able to control the source junction. This is achieved by performing what is known as an *anti-punch-through* implant in the channel. Sometimes, instead of such an implant throughout the channel (which can have undesirable consequences such as raising the V_T or the body effect), a localized implant is done only near the source/drains. These are known as *halo* or *pocket* implants. The higher doping reduces the source/drain depletion widths and prevents their interaction.

For short channel MOSFETs, DIBL is related to the electrical modulation of the channel length in the pinch-off region, ΔL. Since the drain current

is inversely proportional to the electrical channel length, we get

$$I_D \propto \frac{1}{L - \Delta L} = \frac{1}{L}\left(1 + \frac{\Delta L}{L}\right) \tag{6-69}$$

for small pinch-off regions, ΔL. We assume that the fractional change in the channel length is proportional to the drain bias,

$$\frac{\Delta L}{L} = \lambda V_D \tag{6-70}$$

where λ is the *channel length modulation parameter*. Hence, in the saturation region, the expression for the drain current becomes

$$I_D = \frac{Z}{2L}\bar{\mu}_n C_i (V_G - V_T)^2 (1 + \lambda V_D) \tag{6-71}$$

This leads to a slope in the output characteristics, or a lowering of the output impedance (Fig. 6–32).

6.5.11 Short Channel Effect and Narrow Width Effect

If we plot the threshold voltage as a function of channel length in MOSFETs, we find that V_T decreases with L for very small geometries. This effect is called the *short channel effect (SCE)*, and is somewhat similar to DIBL. The mechanism is due to something called *charge sharing* between the source/drain and the gate (Fig. 6-45)[9]. From the equation for the threshold voltage (6–38), we notice that one of the terms is the depletion charge under the gate.

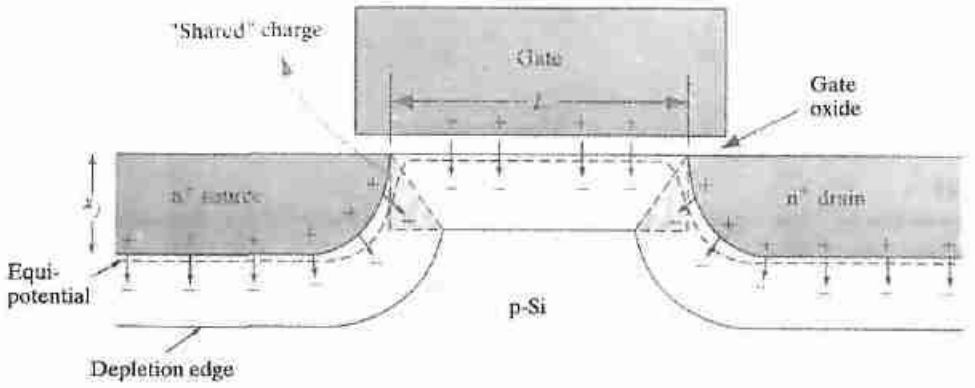


Figure 6–45
Short channel effect in a MOSFET. Cross-sectional view of MOSFET along the length showing depletion charge sharing (colored regions) between the gate, source and drain.

[9] L. Yau, "A simple theory to predict the threshold voltage of short-channel IGFETs," Solid-State Electronics, 17 (1974): 1059.

The equipotential lines in Fig. 6-45 designating the depletion regions curve around the contours of the source/drain junctions. Keeping in mind that the electric field lines are perpendicular to the equipotential contours, we see that the depletion charges that are physically underneath the gate in the approximately triangular regions near the source/drains have their field lines terminate not on the gate, but instead on the source/drains. Hence, electrically these depletion charges are "shared" with the source and drain regions and should not be counted in the V_T expression, Eq. (6-38). We can deal with this effect by replacing the orginal Q_d in the rectangular region underneath the gate by a lower Q_d in the trapezoidal region in Fig. 6-45. Clearly, for a long channel device, the triangular depletion charge regions near the source and drain are a very small fraction of the total depletion charge underneath the gate. However, as the channel lengths are reduced, the shared charge becomes a larger fraction of the total, and this results in a V_T *roll-off* as a function of L (Fig. 6-46). This is important because it is hard to control the channel lengths precisely in manufacturing. The channel length variations then lead to problems with V_T control.

In the last several years, another effect has been observed in n-channel MOSFETs with decreasing L. The V_T initially goes up before it goes down due to the short channel effect. This phenomenon has been dubbed the *reverse short channel effect (RSCE)*, and is due to interactions between Si point defects that are created during the source/drain implant and the B doping in the channel, causing the B to pile up near the source and drains, and thus raise the V_T.

Another related effect in MOSFETs is the *narrow width effect*, where the V_T goes up as the channel width Z is reduced for very narrow devices (Fig. 6-46). This can be understood from Fig. 6-47, where some of the depletion charges under the LOCOS isolation regions have field lines electrically terminating on the gate. Unlike the SCE, where the effective depletion charge is reduced due to charge sharing with the source/drain, here the depletion charge belonging to the gate is increased. The effect is not important for very wide devices, but becomes quite important as the widths are reduced below 1μm.

6.5.12 Gate-Induced Drain Leakage

If we examine the subthreshold characteristics shown in Fig. 6-38, we find that as the gate voltage is reduced below V_T, the subthreshold current drops and

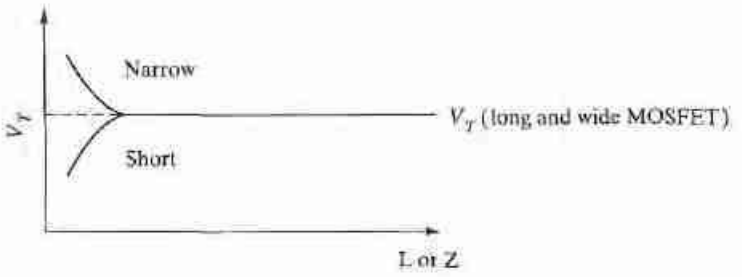


Figure 6-46
Roll-off of V_T with decreasing channel length, and increase of V_T with decreasing width.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 323

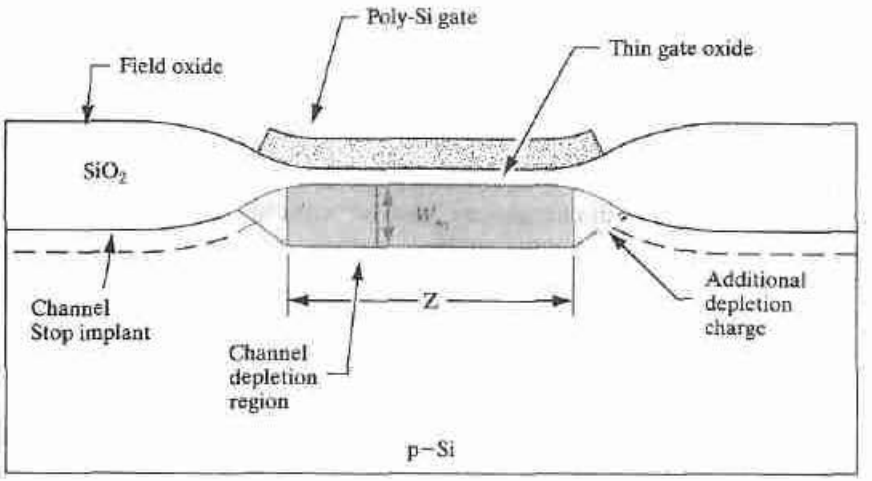


Figure 6–47 Narrow width effect in a MOSFET. Cross-sectional view of MOSFET along the width, showing additional depletion charge (colored regions) underneath the field or the LOCOS isolation regions.

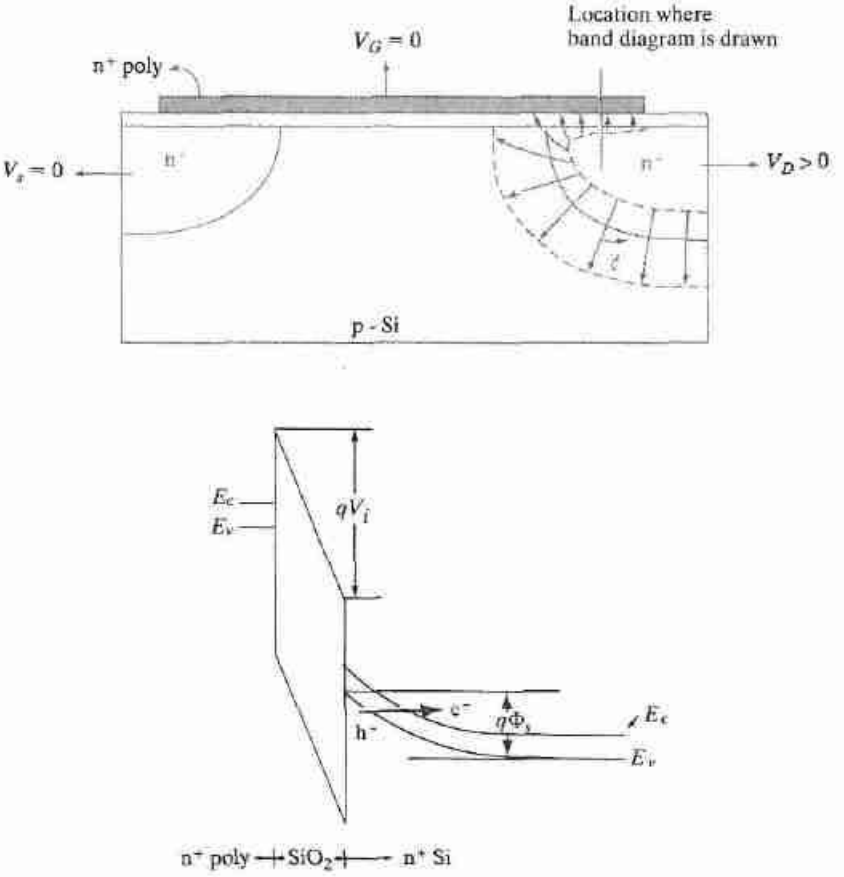


Figure 6–48 Gate-induced drain leakage in MOSFETs. The band diagram for the location shown in color is plotted as a function of depth in the gate–drain overlap region, indicating band-to-band tunneling and creation of electron–hole pairs in the drain region in the Si substrate.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 324

then bottoms out at a level determined by the source/drain diode leakage. However, for even more negative gate biases we find that the off-state leakage current actually goes up as we try to turn off the MOSFET more for high V_D; this is known as *gate-induced drain leakage (GIDL)*. The same effect is seen at a fixed gate bias of near zero, for increasing drain bias. The reason for GIDL can be understood from Fig. 6-48, where we show the band diagram as a function of depth in the region where the gate overlaps the drain junction. As the gate is made more negative (or alternatively, for a fixed gate bias, the drain is made more positive), a depletion region forms in the n-type drain. Since the drain doping is high, the depletion widths tend to be narrow. If the band-bending is more than the band gap E_g across a narrow depletion region, the conditions are conducive to band-to-band tunneling in this region, thereby creating electron–hole pairs. The electrons then go to the drain as GIDL. It must be emphasized that this tunneling is not through the gate oxide (Section 6.4.7), but entirely in the Si drain region. For GIDL to occur, the drain doping level should be moderate ($\sim 10^{18}$ cm^{-3}). If it is much lower than this, the depletion widths and tunneling barriers are too wide. On the other hand, if the doping in the drain is very high, most of the voltage drops in the gate oxide, and the band-bending in the Si drain region drops below the value E_g. GIDL is an important factor in limiting the off-state leakage current in state-of-the-art MOSFETs.

SUMMARY

6.1 *Three-terminal active* devices capable of *power gain* are key to the switching and amplification of electronic signals. The *input impedance* is (desirably) high for voltage-controlled field-effect transistors and low for current-controlled bipolar junction transistors. The power gain of the a-c signal comes from the d-c power supply. It is also desirable to have a high output impedance and a linear dependence of output current with respect to input voltage (or current).

6.2 Field effect devices such as the JFETs (or MESFETs and HEMTs) use a reverse-biased p-n junction (or Schottky diode) as a *gate* to control the undepleted part of a *channel* and thereby modulate the current flow between a *source* (S) of carriers and a *drain* (D) of carriers.

6.3 The *MOSFET* uses an oxide-isolated gate (MOS capacitor) that has a very high input impedance to modulate the barrier between S and D and to change the conductivity of the channel.

6.4 An enhancement-mode NMOSFET has a p-type channel between the n$^+$ source and the drain regions. At the *flat band conditions* (V_{FB}), there is no band-bending as a function of depth in the channel. A gate bias more negative than V_{FB} causes exponentially more holes to be accumulated near the surface with band-bending.

6.5 More positive gate voltages initially repel the majority carrier holes and create a *depletion* region. Then, above the *threshold voltage* (V_T), they attract minority carrier electrons near the surface and *invert* the surface, thereby forming a conductive channel between source and drain. PMOSFETs have the opposite polarities.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 325

6.6 The V_T can be increased by increasing the channel doping and the gate oxide thickness (lower C_i). It is also affected by V_{FB}, which depends on the gate work function and oxide charges, and increases with substrate reverse bias (*body effect*) as well.

6.7 The overall MOS capacitance is the series combination of a fixed gate capacitance C_i and the voltage-dependent semiconductor capacitance in the channel. The inversion charge per unit area is approximately $C_i(V_G - V_T)$.

6.8 For long channel MOSFETs, initially I_D increases *linearly* with V_D, and then it *saturates* once the inversion channel is *pinched off* near the drain end.

6.9 There is a quadratic dependence of I_D on V_G for long channel MOSFETs. For *short channel* devices that operate in the *velocity-saturated region* or *source-injection regime*, the dependence of I_D on V_G is approximately linear.

6.10 For short L, I_{DSat} often increases slightly with V_D due to various short channel effects, leading to a decrease of output impedance. Short channel effects include *drain-induced barrier lowering* (*DIBL*) and *charge sharing*, which leads to V_T *roll-off* with L.

6.11 There is some conduction below V_T, characterized by a subthreshold slope, that is a measure of the effectiveness of the MOSFET as a switch.

6.12 Conduction above V_T in the channel inversion region reflects an *effective channel mobility* lower than that in bulk due to *surface roughness scattering*, and increasingly so at higher V_G.

6.13 In highly scaled MOSFETs, there can be *gate-tunneling* leakage currents (mitigated by high-k dielectrics) and *hot electron* (and, to a lesser extent, hole) effects, mitigated by lower power supply voltages and lightly doped drains (*LDDs*). The resulting high-S/D parasitic series resistance is minimized by Self-Aligned siLICIDEs (*SALICIDE*), described in Chapter 9.

PROBLEMS

6.1 Assume the JFET shown in Fig. 6-6 is Si and has p$^+$ regions doped with 10^{18} acceptors/cm^3 and a channel with 10^{16} donors/cm^3. If the channel half-width a is 1 μm, compare V_P with V_0. What voltage V_{GD} is required to cause pinch-off when V_0 is included? With $V_G = -3$ V, at what value of V_D does the current saturate?

6.2 If the ratio $Z/L = 10$ for the JFET of Prob. 6.1, and $\mu_n = 1000$ cm^2/V-s, calculate I_D(sat.) for $V_G = 0, -2, -4$, and -6 V. Plot I_D(sat.) vs. V_D(sat.).

6.3 For the JFET of Prob. 6.2, plot I_D vs. V_D for the same four values of V_G. Terminate each plot at the point of saturation.

6.4 Use Eqs. (6-9) and (6-10) to calculate and plot $I_D(V_D, V_G)$ at 300 K for a Si JFET with $a = 1000$ Å, $N_d = 7 \times 10^{17}$ cm^{-3}, Z = 100 μm, and L = 5 μm. Allow V_D to range from 0 to 5 V, and allow V_G to take on the values 0, –1, –2, –3, –4, and –5 V.

6.5 Use the field-dependent mobility expression to calculate and plot $I_D(V_D, V_G)$ for the Si JFET described in Prob. 6.4 for gate lengths of 0.25, 0.50, 1.0, 2.0,

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 326

and 5.0 μm and a gate voltage of 0 V. The field-dependent mobility model of Eq. (8-12) takes the form

$$\mu(\mathscr{E}) = \frac{v_d}{\mathscr{E}} = \frac{\mu_0}{1 + (\mu\mathscr{E}/v_s)}$$

where μ_0 is the low-field mobility (as given in Fig. 3-23) and v_s is the saturation velocity (10^7 cm/s). Assume that $\mathscr{E} = V_D/L$, and assume that the drain current saturates when it reaches a maximum. Discuss the advantages of having a short gate length device.

6.6 The current I_D varies almost linearly with V_D in a JFET for low values of V_D.

(a) Use the binomial expansion with $V_D/(-V_G) < 1$ to rewrite Eq. (6-9) as an approximation to this case.

(b) Show that the expression for the channel conductance I_D/V_D in the linear range is the same as g_m(sat.) given by Eq. (6-11).

(c) What value of gate voltage V_G turns the device off such that the channel conductance goes to zero?

6.7 Show that the width of the depletion region in Fig. 6-15 is given by Eq. (6-30). Assume the carriers are completely swept out within W, as was done in Section 5.2.3.

6.8 Find the maximum depletion width, minimum capacitance C_i, and threshold voltage for an *ideal* MOS capacitor with a 10-nm gate oxide (SiO_2) on p-type Si with $N_a = 10^{16}$ cm^{-3}. Next, include the effects of flat band voltage, assuming an n^+ polysilicon gate and fixed oxide charge of $5 \times 10^{10} q$ (C/cm^2).

6.9 An Al-gate p-channel MOS transistor is made on an n-type Si substrate with $N_d = 5 \times 10^{17}$ cm^{-3}. The SiO_2 thickness is 100 Å in the gate region, and the effective interface charge Q_i is $5 \times 10^{10} q$ C/cm^2. Find W_m, V_{FB}, and V_T. Sketch the C–V curve for this device and give important numbers for the scale.

6.10 Use Eq. (6-50) to calculate and plot $I_D(V_D, V_G)$ at 300 K for an n-channel Si MOSFET with an oxide thickness $d = 200$ Å, a channel mobility $\overline{\mu}_n = 1000$ cm^2/V-s, $Z = 100$ μm, $L = 5$ μm, and N_a of 10^{14}, 10^{15}, 10^{16}, and 10^{17} cm^{-3}. Allow V_D to range from 0 to 5 V and allow V_G to take on values of 0, 1, 2, 3, 4, and 5 V. Assume that $Q_i = 5 \times 10^{11} q$ C/cm^2.

6.11 Calculate the V_T of a Si-P-channel MOS transistor for an n^+-polysilicon gate with silicon oxide thickness = 50 Å, $N_d = 1 \times 10^{18}$ cm^{-3} and a fixed charge of $2 \times 10^{10} q$ C/cm^2. Is it an enhancement- or depletion-mode device? What B dose is required to change the V_T to 0 V? Assume a shallow B implant.

6.12 (a) Find the voltage V_{FB} required to reduce to zero the negative charge induced at the semiconductor surface by a sheet of positive charge Q_{ox} located x' below the metal.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 327

(b) In the case of an arbitrary distribution of charge $\rho(x')$ in the oxide, show that

$$V_{FB} = -\frac{1}{C_i}\int_0^d \frac{x'}{d}\rho(x')dx'$$

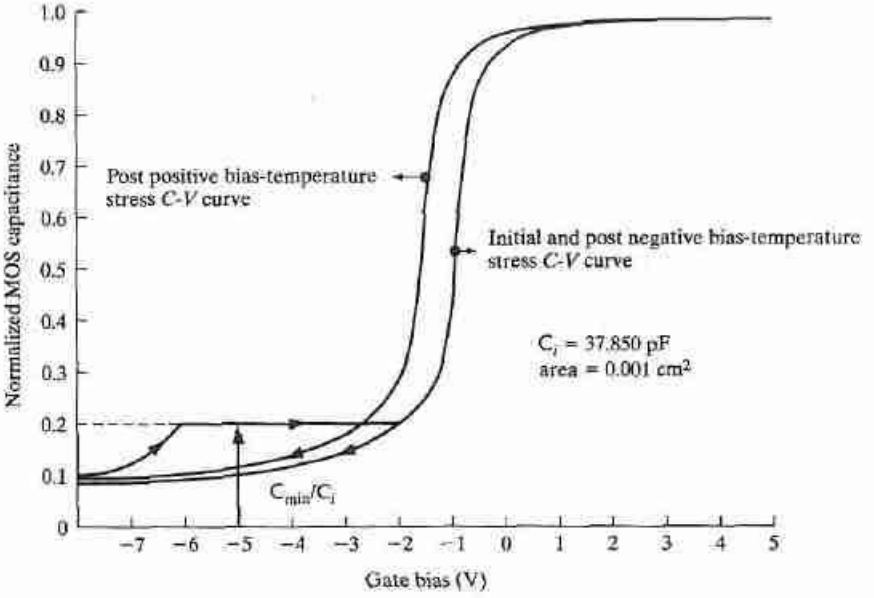
6.13 Redraw Figs. 6–12, 6–13, and 6–15 for the p-channel case (n-type substrate).

6.14 Calculate the V_T of an MOS capacitor where we deposit a high-k gate dielectric, HfO_2, whose relative dielectric constant is 25, on a *novel* p-type semiconductor whose electron affinity is 4 eV, band gap is 1.5 eV, relative dielectric constant is 10, and intrinsic carrier concentration is 10^{12} cm^{-3}. The gate is made of a metal whose work function is 5 eV, gate oxide thickness is 100 Å, and N_A is 10^{18} cm^{-3} and that has a fixed oxide charge of 5×10^{10} q C/cm^2. At V_T, what are the electron and hole concentrations at the oxide–semiconductor interface and deep in the substrate? Sketch a labeled band diagram normal to the surface at V_T, and mark off the relevant values on the basis of the numbers given.

Sketch the *low- and high*-frequency C–V characteristics of this capacitor, and explain their differences. How would the characteristics change *at large negative gate bias* if we doubled the oxide thickness? How about if we doubled the substrate doping?

6.15 A Si MOS capacitor has the high frequency C–V curve shown in Fig. P6–15 normalized to the capacitance in strong accumulation. Determine the oxide thickness and substrate doping assuming a gate-to-substrate work function difference of −0.35 V.

Figure P6–15



Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 328

6.16 For the capacitor in Prob. 6.15, determine the initial flat band voltage.

6.17 For the capacitor in Prob. 6.15, determine the fixed oxide charge, Q_i, and the mobile ion content.

6.18 For a long channel n-MOSFET with $V_T = 1$ V, calculate the V_G required for an I_D(sat.) of 0.1 mA and V_D(sat.) of 5 V. Calculate the small-signal output conductance g and the transconductance g_m(sat.) at $V_D = 10$ V. Sketch the cross section of this MOSFET, and schematically show the inversion charge and depletion charge distributions for $V_D = 1$ V, 5 V, and 10 V.

Recalculate the new I_D for $V_G - V_T = 3$ V and $V_D = 4$ V.

6.19 Calculate the V_T of a Si n-channel MOSFET for a gate-to-substrate work function difference $\phi_{ms} = -1.5$ eV, gate oxide thickness = 100 Å, $N_A = 10^{18}$ cm^{-3}, and fixed oxide charge of $5 \times 10^{10}\, q$ C/cm^2, for a substrate bias of –2.5 V. At V_T, what are the electron and hole concentrations at the oxide–Si interface and deep in the substrate? Sketch a labeled band diagram normal to the surface at V_T, showing the Fermi potential.

6.20 Calculate the V_T of a Si n-channel MOSFET for an n$^+$-polysilicon gate with gate oxide thickness = 100 Å, $N_a = 10^{18}$ cm^{-3}, and a fixed oxide charge of $5 \times 10^{10}\, q$ C/cm^2.

6.21 For a MOSFET with $V_T = 1$ V and $Z = 50$ μm, $L = 2$ μm, calculate the drain current at $V_G = 5$ V, $V_D = 0.1$ V. Repeat for $V_G = 3$ V, $V_D = 5$ V. Assume an electron channel mobility $\bar{\mu}_n = 200$ cm^2/V-s, and the substrate is connected to the source.

6.22 For the MOSFET characteristics shown in Fig. P6-22, calculate:

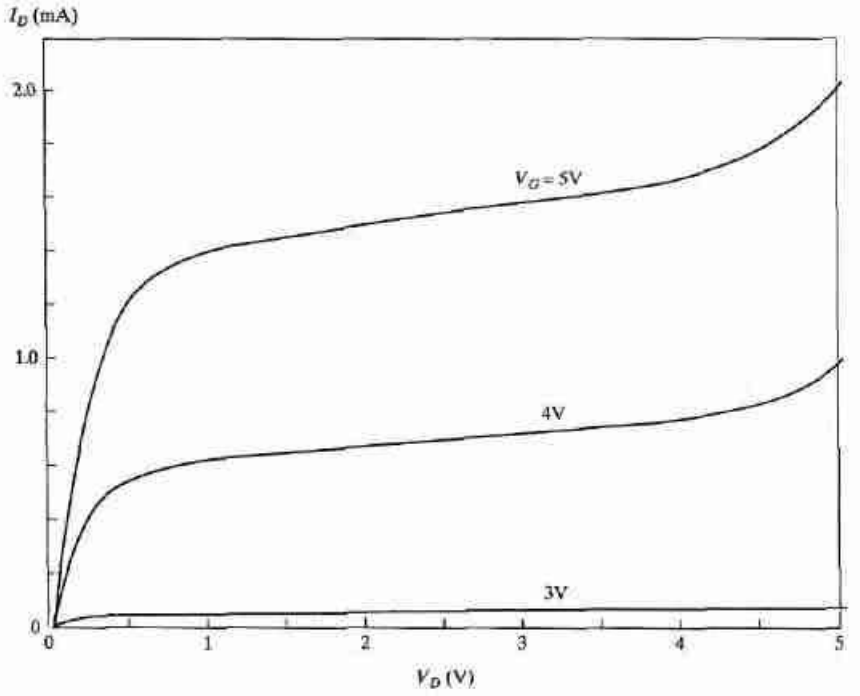


Figure P6–22

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 329

1. Linear V_T and k_N
2. Saturation V_T and k_N

Assume channel mobility, $\bar{\mu}_n = 500$ cm^2/V-s and $V_{FB} = 0$.

6.23 For Prob. 6.22, calculate the gate oxide thickness and substrate doping, either graphically or iteratively.

6.24 Assume that the inversion layer in a Si MOSFET can be treated as a 2-D electron gas trapped in an infinite rectangular potential well of width 100 Å. (In reality, it looks more like a triangular well.) Calculate the inversion charge per unit area assuming that the Fermi level lies midway between the second and third subbands. Assume $T = 77$ K, and effective mass $= 0.2\, m_0$. Assume also that the Fermi function can be treated as a rectangular function. Also sketch (E, k) for the first three subbands. Refer to Appendix IV.

6.25 An n-channel MOSFET with a 400-Å gate oxide requires its V_T to be lowered by 2 V. Using a 50-keV implant of a singly charged species, and assuming that the implant distribution is peaked at the oxide–Si interface and can be regarded as a sheet charge at that interface, what implant parameters (species, energy, dose, and beam current) would you choose? The scan area is 200 cm^2 and the desired implant time is 20 s. Assume similar range statistics in oxide and Si.

6.26 Plot I_D vs. V_D with several values of V_G for a thin-oxide p-channel transistor with a 10-μm oxide and $V_T = -1.1$ V. Assume that I_D(sat.) remains constant beyond pinch-off. Assume that $\bar{\mu}_p = 200$ cm^2/V-s, and $Z = 10L$.

6.27 A typical figure of merit for high-frequency operation of MOS transistors is the cutoff frequency $f_c = g_m/2\pi C_G LZ$, where the gate capacitance C_G is essentially C_i over most of the voltage range. Express f_c above pinch-off in terms of materials parameters and device dimensions, and calculate f_c for the transistor of Prob. 6.26, with $L = 1$ μm.

6.28 When an MOS transistor is biased with $V_D > V_D$ (sat.), the effective channel length is reduced by ΔL and the current I'_D is larger than I_D(sat.), as shown in Fig. 6-32. Assuming that the depleted region ΔL is described by an expression similar to Eq. (6-30) with $V_D - V_D$(sat.) for the voltage across ΔL, show that the conductance beyond saturation is

$$g'_D = \frac{\partial I'_D}{\partial V_D} = I_D(\text{sat.}) \frac{\partial}{\partial V_D}\left(\frac{L}{L - \Delta L}\right)$$

and find the expression for g'_D in terms of V_D.

6.29 From Fig. 6-44, it is clear that the depletion regions of the source and drain junctions can meet for short channels, a condition called *punch-through*. Assume that the source and drain regions of an n-channel Si MOSFET are doped with 10^{20} donors/cm^3 and the 1-μm-long channel is doped with 10^{16} acceptors/cm^3. If the source and substrate are grounded, what drain voltage will cause punch-through?

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 330

READING LIST

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

Information about MOS devices used in integrated circuits can be found at http://public.itrs.net/

Kahng, D. "A Historical Perspective on the Development of MOS Transistors and Related Devices," *IEEE Trans. Elec. Dev.*, ED-23 (1976): 655.

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.

Pierret, R. F. *Field Effect Devices*. Reading, MA: Addison-Wesley, 1990.

Sah, C. T. "Characteristics of the Metal–Oxide Semiconductor Transistors." *IEEE Trans. Elec. Dev.* ED-11 (1964): 324.

Sah, C. T. "Evolution of the MOS Transistor—From Conception to VLSI." *Proceedings of the IEEE* 76 (October 1988): 1280–1326.

Schroder, D. K. *Modular Series on Solid State Devices: Advanced MOS Devices*. Reading, MA: Addison-Wesley, 1987.

Shockley, W., and G. Pearson. "Modulation of Conductance of Thin Films of Semiconductors by Surface Charges." *Phys. Rev.* 74 (1948): 232.

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

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

Tsividis, Y. *Operation and Modeling of the MOS Transistor*. Boston: McGraw-Hill College, 1998.

SELF QUIZ

Question 1

(a) What is the main distinction between an active device and a passive device?

(b) If a device has power gain, where is the higher energy of the a-c signal at the output coming from?

(c) What is the distinction between a current-controlled versus voltage-controlled three-terminal active device? Which is preferable?

Question 2

For parts (a) through (c) below, consider the following low-frequency gate capacitance (per unit area) vs. gate voltage characteristic for a metal gate n-channel MOSFET.

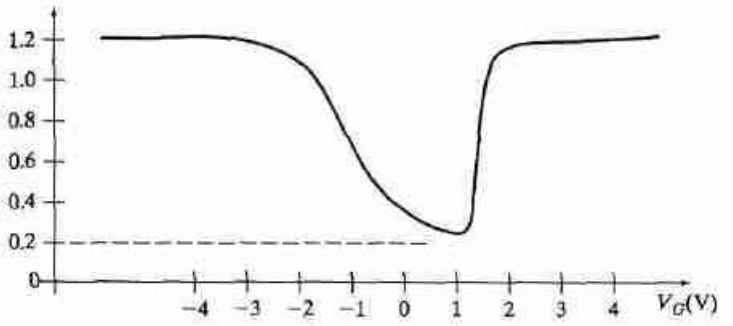
(a) On the below figure label the *approximate* regions or points of

- weak inversion (1)
- flat band (2)
- strong inversion (3)
- accumulation (4)
- threshold (5)
- depletion (6)

(b) What is the oxide capacitance per unit area?_____ $\mu F/cm^2$

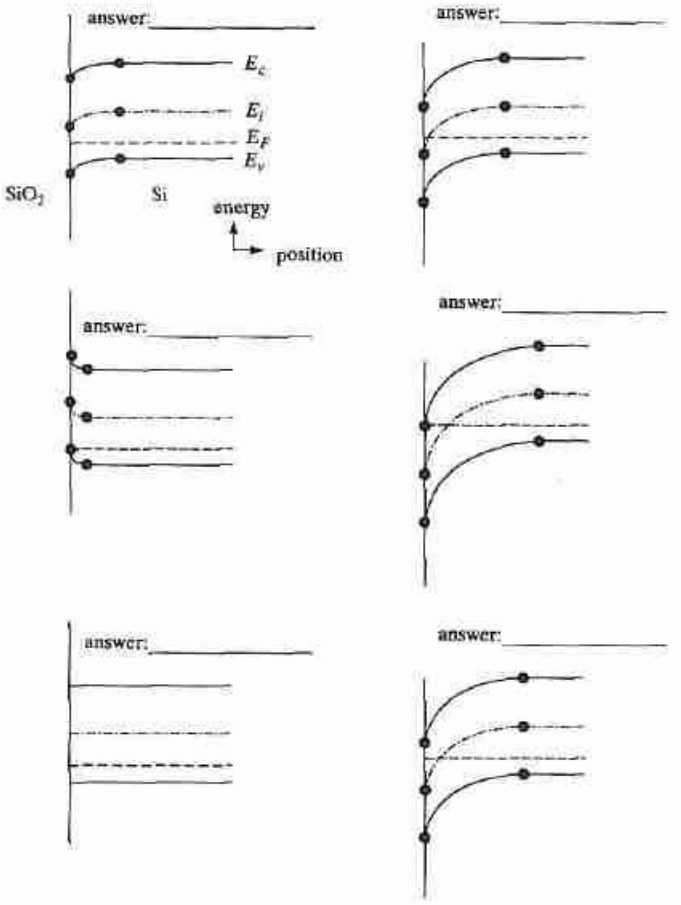
(c) On the curve below, sketch the high-frequency curves for this MOSFET with grounded source/drain.

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 331



Question 3

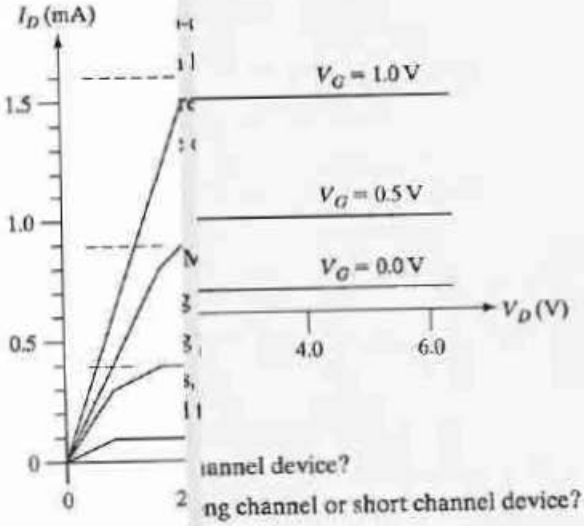
Label—in the "answer" space provided *above* the figure—the following MOS capacitor band diagrams as corresponding to accumulation, weak inversion, depletion, strong inversion, flat band or threshold. Use each possibility only once.



characteristic.

Field-Effect Transisto

Question 4
Consider the following MOSFE



annel device?

ng channel or short channel device?

hold voltage V_T?

enhancement-mode MOSFET?

(a) Is this an n-channel or p-

(b) Does this appear to be a

(c) What is the apparent thre

(d) Is this a depletion-mode

OSFET decreased (improved) by:

e oxide thickness?

e substrate doping?

Question 5
Is the subthreshold slope of a M

me one thing/factor/effect that limits the extent to reduce subthreshold swing.

(a) Increasing or decreasing

(b) Increasing or decreasing

For each of these two methods, which the method can be used

erature increase, decrease, or leave unchanged (cir-

nt of a p-n diode?

Question 6
Would increasing the device ten
cle the correct answers):

/ decrease / unchanged

-drain leakage current of MOSFET?

(a) the reverse saturation cu

/ decrease / unchanged

increase

(b) the subthreshold source-

increase

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 333

Question 7

Assuming no interface charge due to defects and/or traps, would decreasing the oxide thickness/increasing the oxide capacitance of an n-channel MOSFET *increase*, *decrease*, or leave essentially *unchanged* the following parameters (circle the correct answers):

(a) the flat band voltage, V_{FB}?

increase / decrease / unchanged

(b) the threshold voltage, V_T?

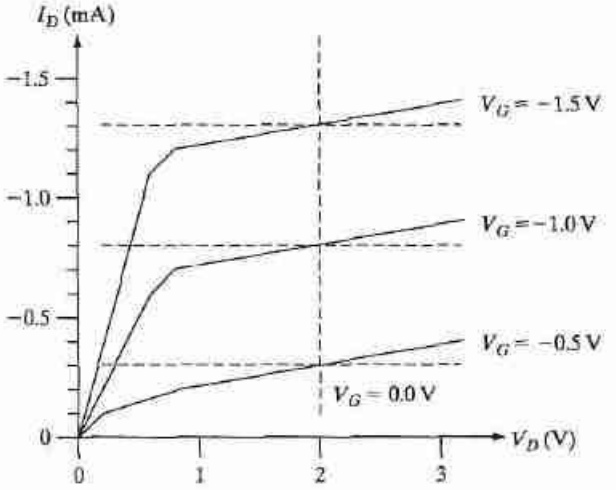
increase / decrease / unchanged

(c) the subthreshold slope?

increase / decrease / unchanged

Question 8

Consider the following MOSFET characteristic.



(a) What is the transconductance g_m at $V_D = -1.0$ V (in units of mhos)?

(b) What is the apparent threshold voltage at $V_D = -1.0$ V?

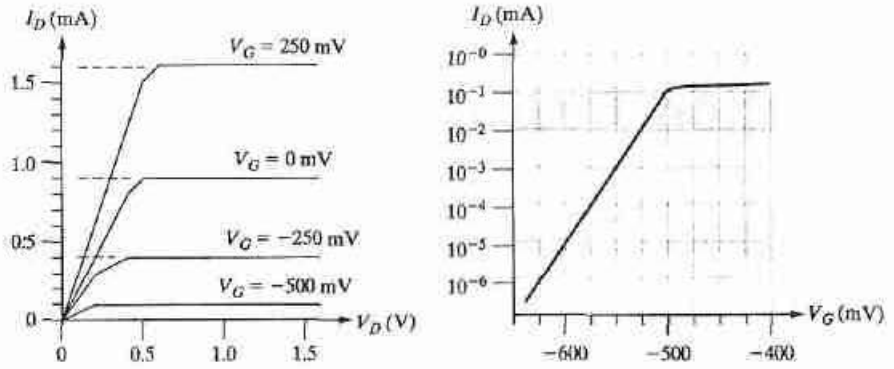
(c) Does this MOSFET appear to be a long channel or short channel device?

(d) Is this an n-channel or p-channel device?

(e) Is this a depletion-mode or enhancement-mode MOSFET?

Question 9

A senior in electrical and computer engineering in a device fabrication course presented the following characteristics as those of an n-channel MOSFET that he had fabricated and characterized at a temperature of 300 K.



(a) If we were to believe this student, would this be *normally on/depletion mode* or a *normally off/enhancement mode* MOSFET? Circle one choice below.

normally off / normally on

(b) If we were to believe this student, would this MOSFET appear to be a *long channel* or *short channel* device? Circle one.

long channel / short channel

(c) If we were to believe this student, what would be the *subthreshold slope/swing*, S, of this MOSFET?

(d) This student, however, was subsequently expelled from the university for falsifying this data, leaving him with nothing to show for his years at school but huge college loan debts. What *clearly* physically unrealistic aspect of these MOSFET characteristics (besides less than smooth curves) should have drawn suspicion?

Petitioner Canadian Solar (USA) Inc., Ex. 1026, p. 335