Chapter 5

Junctions

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 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 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

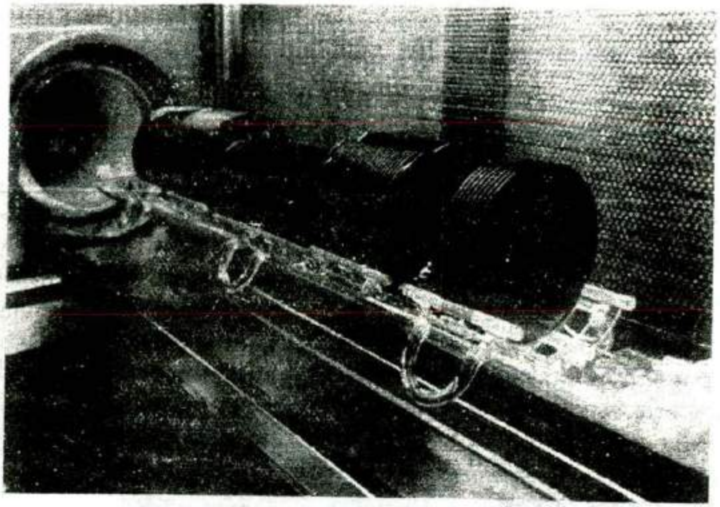


Figure 5–1a Silicon wafers being loaded into a furnace. For 8-inch and larger wafers, this type of horizontal loading is often replaced by a vertical furnace.

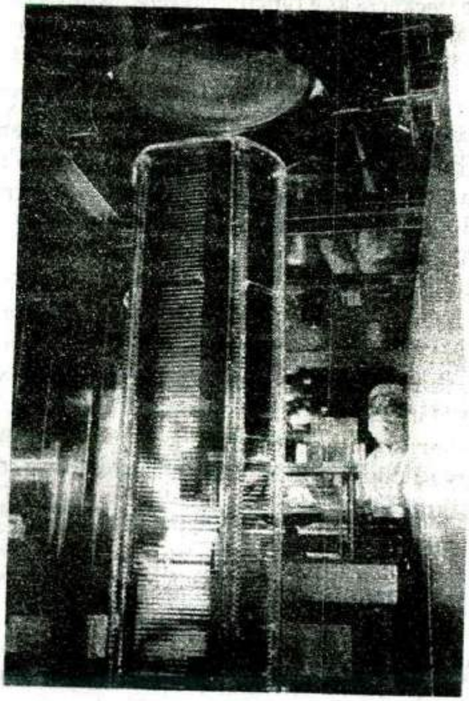


Figure 5–1b Vertical furnace for large Si wafers. The silica wafer holder is loaded with eight-inch Si wafers and moved into the furnace above for oxidation, diffusion, or deposition operations. (Photograph courtesy of Tokyo Electron Ltd.)

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:

$$\mathrm{Si} + \mathrm{O}_2 \rightarrow \mathrm{SiO}_2 \text{ (dry oxidation)}$$

$$\mathrm{Si} + 2\mathrm{H}_2\mathrm{O} \rightarrow \mathrm{SiO}_2 + 2\mathrm{H}_2 \text{ (wet oxidation)}$$

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 5.17, 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 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

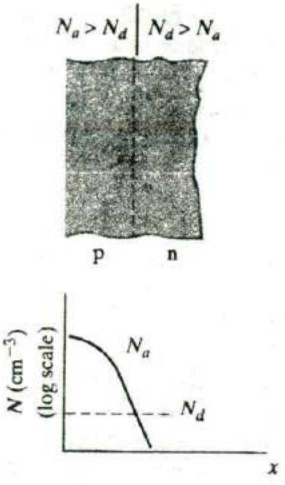


Figure 5–2
Impurity concentration profile for fabricating a p-n junction by diffusion.

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).

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 is 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) tung-

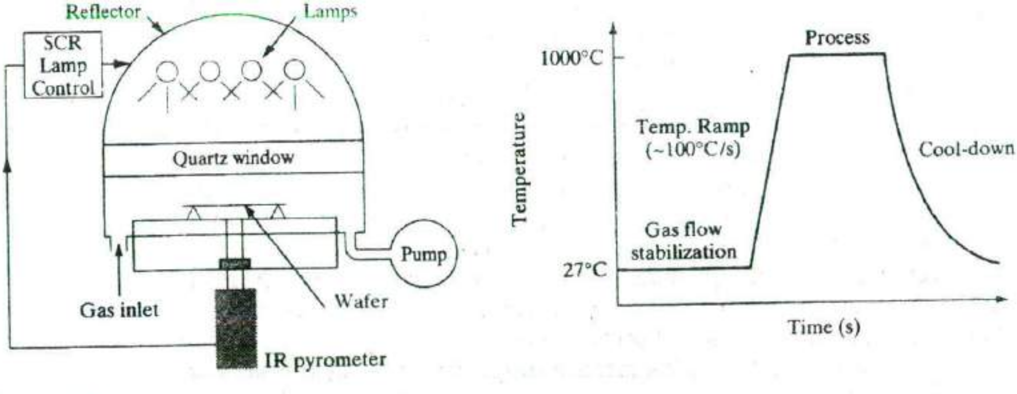


Figure 5–3
Schematic diagram of a rapid thermal processor, and typical time-temperature profile.

sten-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 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).

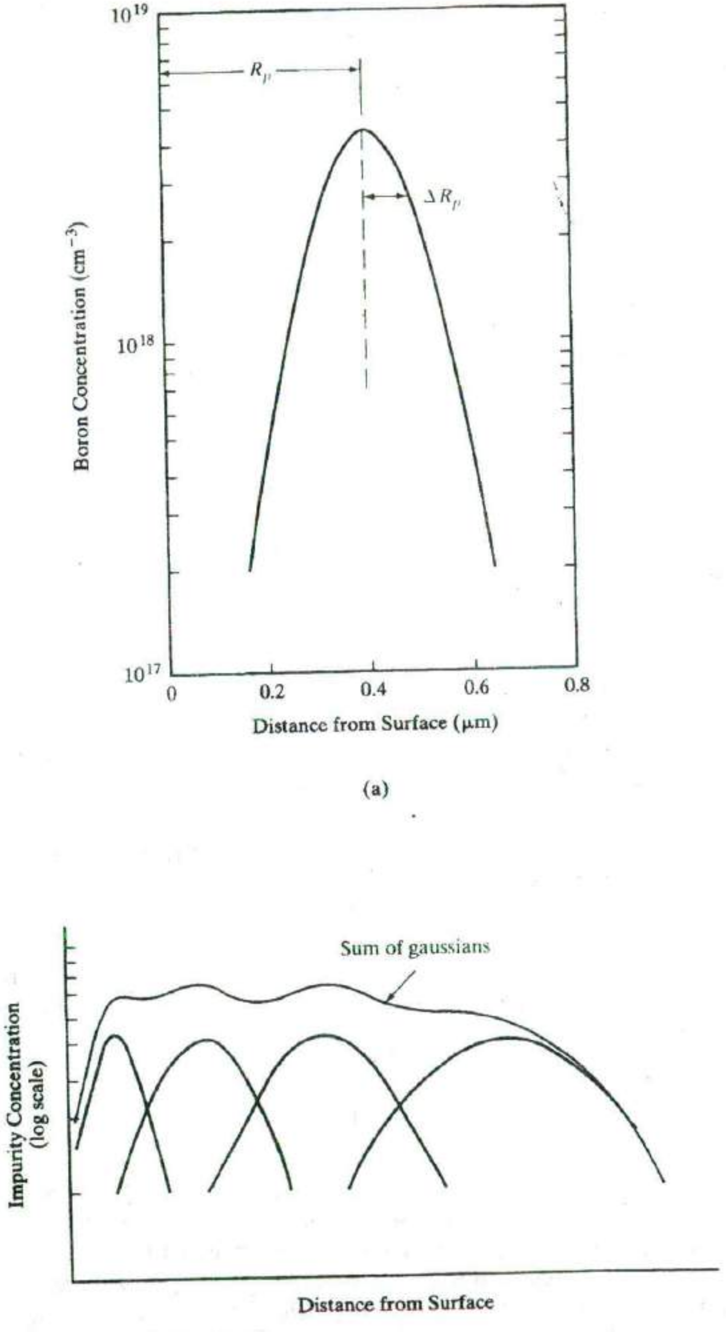
5.1.4 Ion Implantation

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. 5–4a. An implanted dose of ϕ ions/cm^2 is distributed approximately by a gaussian formula

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

where ΔR_p, called the *straggle*, measures the half-width of the distribution at $e^{-1/2}$ of the peak Fig. (5–4a). 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. By performing several implantations at different energies, it is possible to synthesize a desired impurity distribution, such as the uniformly doped region in Fig. 5–4b.

Figure 5–4 Distributions of implanted impurities: (a) 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); (b) a relatively flat distribution obtained by summing four gaussians implanted at selected energies and doses.



(a)

(b)

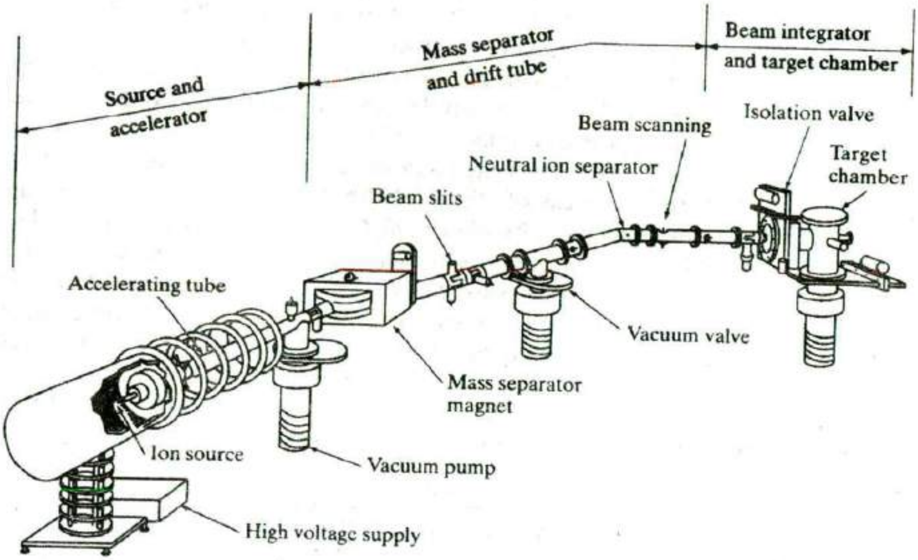


Figure 5–5
Schematic diagram of an ion implantation system.

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.

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

[1] In many ion implanters the mass separation occurs before the

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 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

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

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

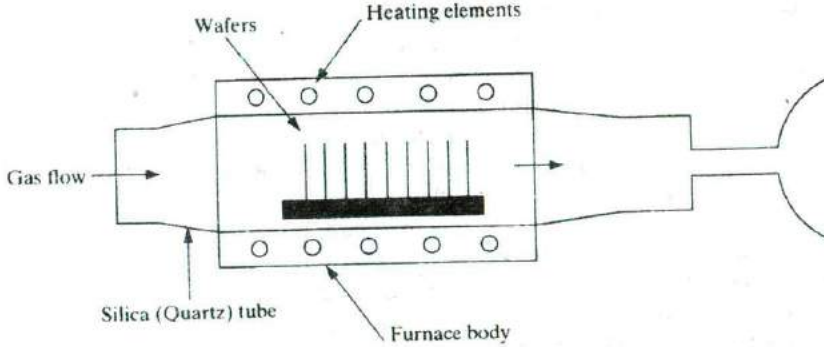


Figure 5–6
Low pressure chemical vapor deposition (LPCVD) reactor.

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

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 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

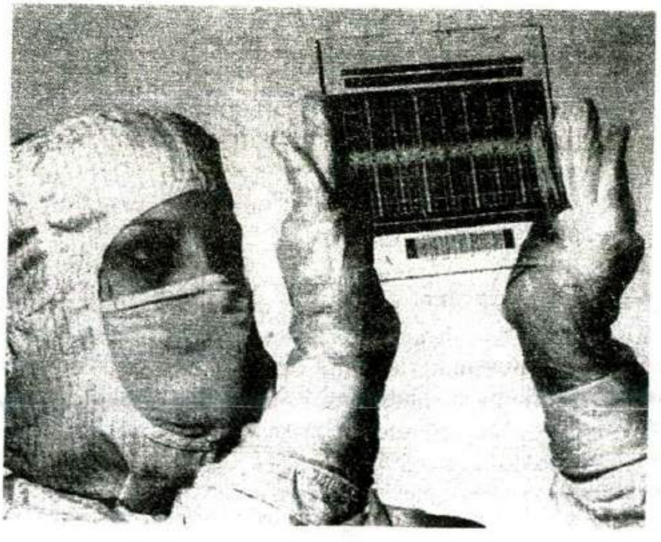


Figure 5–7a
A photolithographic reticle used for one step in the processing of a 16 Mb dynamic random access memory (DRAM). In a "stepper" projection exposure system, ultraviolet light shines through the glass plate and the image is projected onto the wafer to expose photoresist for one die in the array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

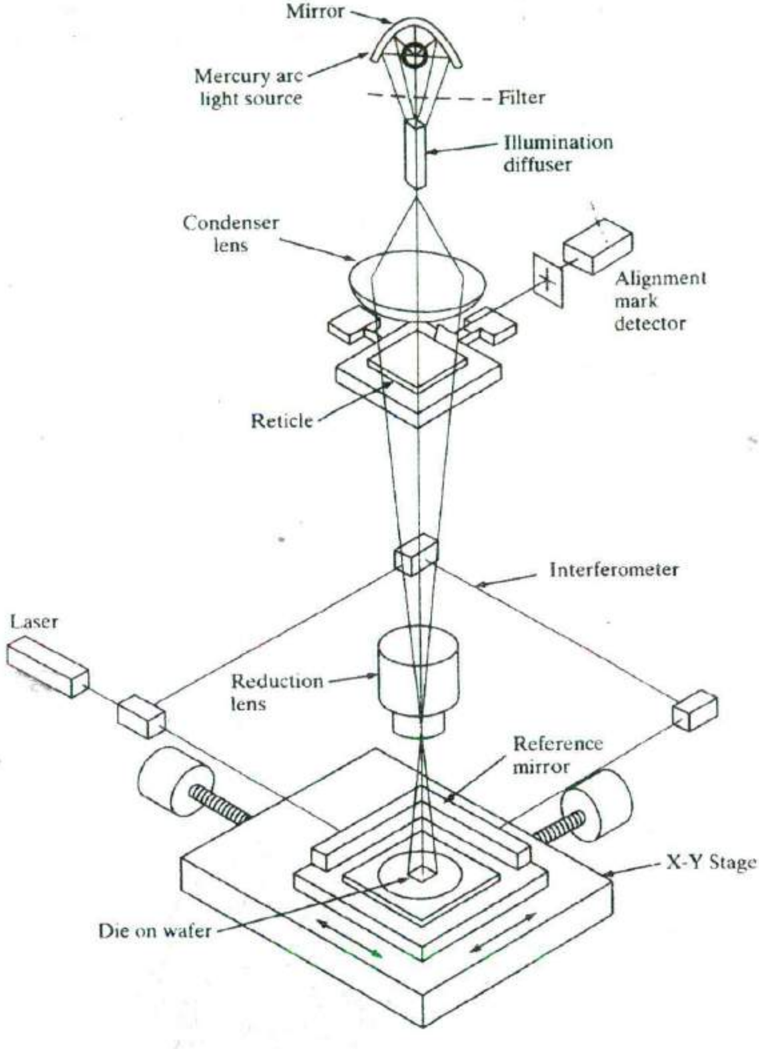


Figure 5–7b Schematic diagram of an optical stepper.

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 to ~0.25 μm using ultraviolet light from mercury lamps (wavelength ~0.365 μm). 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 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 re-focusing 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

$$l_{\min} = 0.8\,\lambda/NA \qquad (5\text{–}2a)$$

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) with argon fluoride (ArF) excimer lasers (0.193 μm), or extreme ultraviolet (EUV) sources (0.154 μm). Novel exposure techniques employing phase shifting and Fourier optics allow resolutions near the dimension of the wavelength being used. For a common ultraviolet wavelength of 0.365 μm, for example, one may achieve linewidths of about 0.25 μm. An ArF laser can be used for 0.15 μm linewidths.

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

$$\mathrm{DOF} = \frac{\lambda}{2\,(NA)^2} \qquad (5\text{–}2b)$$

The DOF tells us the range of distances around the focal plane where the image quality is sharp. Unfortunately, this expression implies that exposure 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 iarger 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 so much interest in X-ray and electron beam lithography. The de Broglie theorem states that the wavelength of a particle varies inversely with its momentum:

$$\lambda = \frac{h}{p} \qquad (5\text{–}2c)$$

Thus more massive particles or energetic photons should be considered to achieve shorter wavelengths. Viable candidates for this application are electrons, ions, or X-rays. For example, electron beams are easily generated, focused, and deflected. The basic technology for this process has been developed over many years in scanning electron microscopy. 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), but to expose the wafer photoresist using photons. In addition to the advances in deep ultraviolet sources mentioned previously, X-rays offer the promise of even smaller dimensions. For example, if a heavy metal is used in the mask, X-rays ($\lambda \sim 1$ Å) can be used to expose the wafer with 0.1 μm resolution. A particularly high flux of X-rays can be obtained from the synchrotron radiation emitted by electrons accelerated in a storage ring or synchrotron.

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

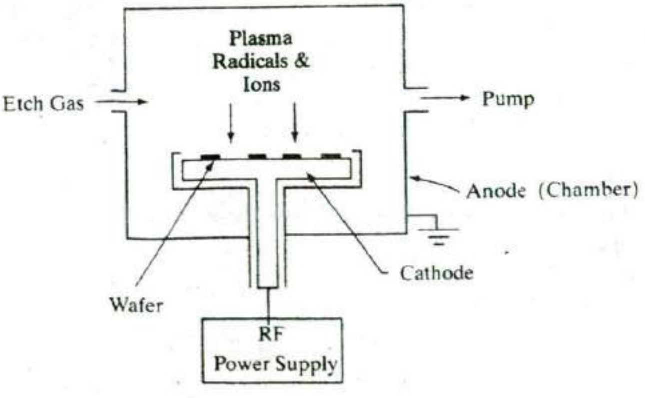


Figure 5–8
Reactive ion etcher. Single or multiple wafers are placed on the rf powered cathode to maximize the ion bombardment. Shown in the figure is a simple *diode* etcher in which we have just two electrodes. We can also use a third electrode to supply rf power separately to the etch gases in a *triode* 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.

into the chamber at reduced pressure (~1–100 mTorr), and a plasma is struck 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 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.

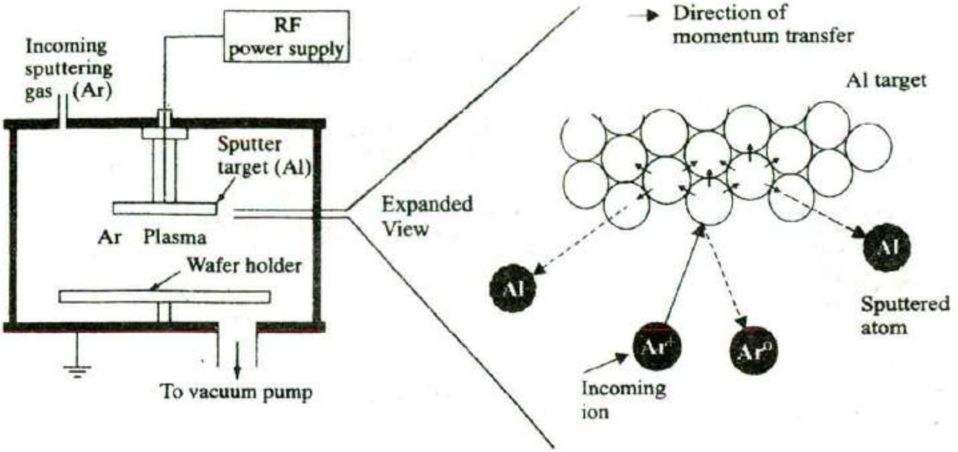


Figure 5–9
Aluminum sputtering by Ar^+ ions. The Ar^+ ions with energies of ~1–3 keV physically dislodge Al atoms which end up depositing on the Si wafers held in close proximity. The chamber pressures are kept low such that the mean free path of the ejected Al atoms is long compared to the target-to-wafer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

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 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.

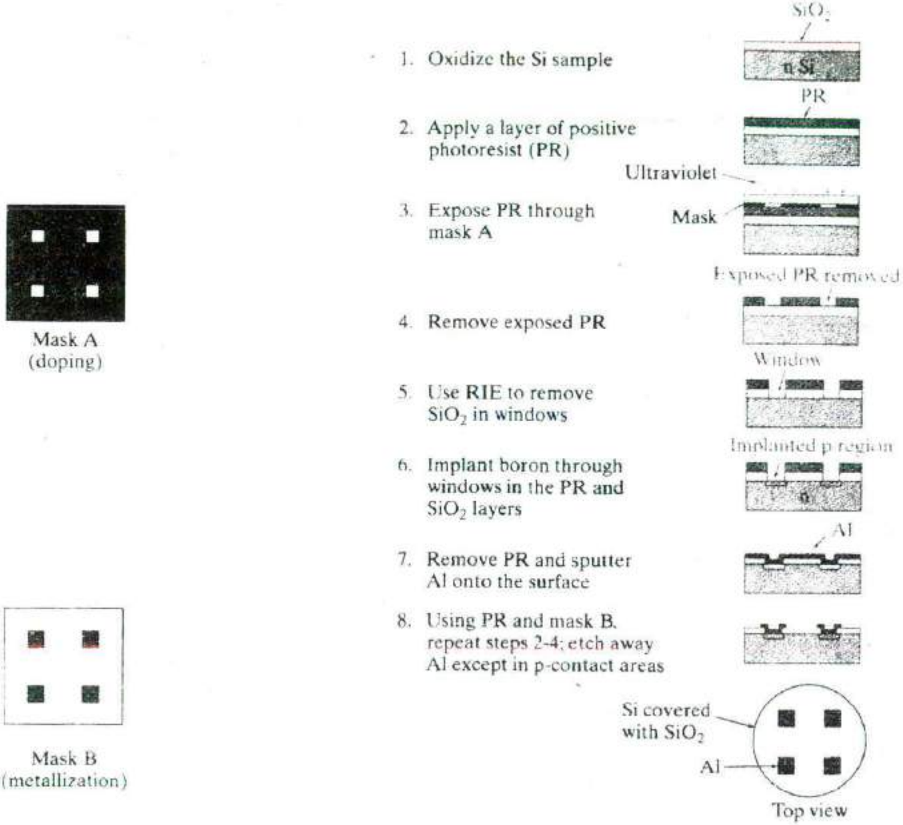

SiO_2
1. Oxidize the Si sample
n Si
PR
2. Apply a layer of positive photoresist (PR)
Ultraviolet
3. Expose PR through mask A
Mask
Exposed PR removed
4. Remove exposed PR
Window
5. Use RIE to remove SiO_2 in windows
Implanted p region
6. Implant boron through windows in the PR and SiO_2 layers
n
Al
7. Remove PR and sputter Al onto the surface
8. Using PR and mask B, repeat steps 2-4; etch away Al except in p-contact areas
Si covered with SiO_2
Al
Top view
Mask A (doping)
Mask B (metallization)

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.

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.

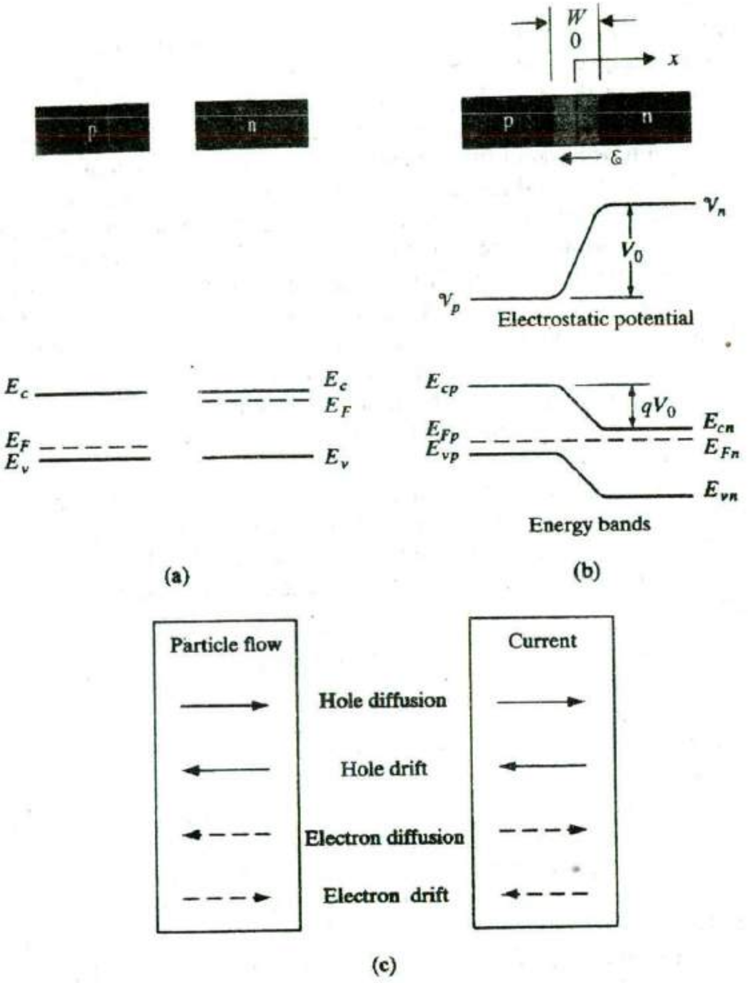


Figure 5–11 Properties of an equilibrium p-n junction: (a) isolated, neutral regions of p-type and n-type material and energy bands for the isolated regions; (b) junction, showing space charge in the transition region W, the resulting electric field $\mathscr{E}$ and contact potential V_0, and the separation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

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 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 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

[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.

[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.

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:

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

This equation can be rearranged to obtain

$$\frac{\mu_p}{D_p}\mathscr{E}(x) = \frac{1}{p(x)}\,\frac{dp(x)}{dx} \qquad (5\text{–}4b)$$

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, $\mathscr{E}(x) = -d\mathcal{V}(x)/dx$, so that Eq. (5–4b) becomes

$$-\frac{q}{kT}\,\frac{d\mathcal{V}(x)}{dx} = \frac{1}{p(x)}\,\frac{dp(x)}{dx} \qquad (5\text{–}5)$$

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, $\mathcal{V}_p$ and $\mathcal{V}_n$, 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 $\mathcal{V}$ can be taken reasonably as functions of x only. Integration of Eq. (5–5) gives

[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 $\mathcal{V}_n$ is a higher potential than $\mathcal{V}_p$ by the amount V_0, the electron energies on the n side are *lower* than those on the p side by qV_0.

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

$$-\frac{q}{kT}(\mathcal{V}_n - \mathcal{V}_p) = \ln p_n - \ln p_p = \ln\frac{p_n}{p_p} \qquad (5\text{–}6)$$

The potential difference $\mathcal{V}_n - \mathcal{V}_p$ 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:

$$V_0 = \frac{kT}{q} \ln \frac{p_p}{p_n} \qquad (5\text{–}7)$$

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

$$V_0 = \frac{kT}{q} \ln \frac{N_a}{n_i^2/N_d} = \frac{kT}{q} \ln \frac{N_a N_d}{n_i^2} \qquad (5\text{–}8)$$

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

Another useful form of Eq. (5–7) is

$$\frac{p_p}{p_n} = e^{qV_0/kT} \qquad (5\text{–}9)$$

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:

$$\boxed{\frac{p_p}{p_n} = \frac{n_n}{n_p} = e^{qV_0/kT}} \qquad (5\text{–}10)$$

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

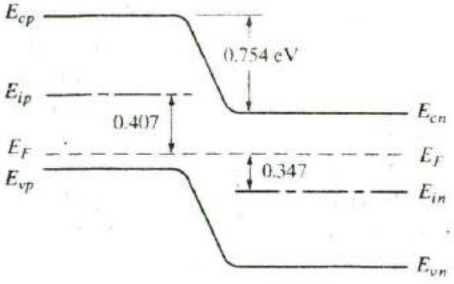
EXAMPLE 5–1 An abrupt Si p-n junction has $N_a = 10^{17}$ cm^{-3} on the p side and $N_d = 10^{16}$ cm^{-3} on the n side. At 300 K, (a) calculate the Fermi levels, draw an equilibrium band diagram and find V_0 from the diagram; (b) compare the result from (a) with V_0 calculated from Eq. (5–8).

SOLUTION (a) Find E_F on each side

$$E_{ip} - E_F = kT \ln\frac{p_p}{n_i} = 0.0259 \ln \frac{10^{17}}{(1.5 \times 10^{10})} = \mathbf{0.407\ eV}$$

$$E_F - E_{in} = kT\ln\frac{n_n}{n_i} = 0.0259\ln\frac{10^{16}}{(1.5\times10^{10})} = \mathbf{0.347\ eV}$$

$$qV_0 = 0.407 + 0.347 = \mathbf{0.754\ eV}$$



(b) Find V_0 from Eq. (5–8)

$$qV_0 = kT\ln\frac{N_aN_d}{n_i^2} = 0.0259\ln\frac{10^{33}}{2.25\times10^{20}} = \mathbf{0.754\ eV}$$

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

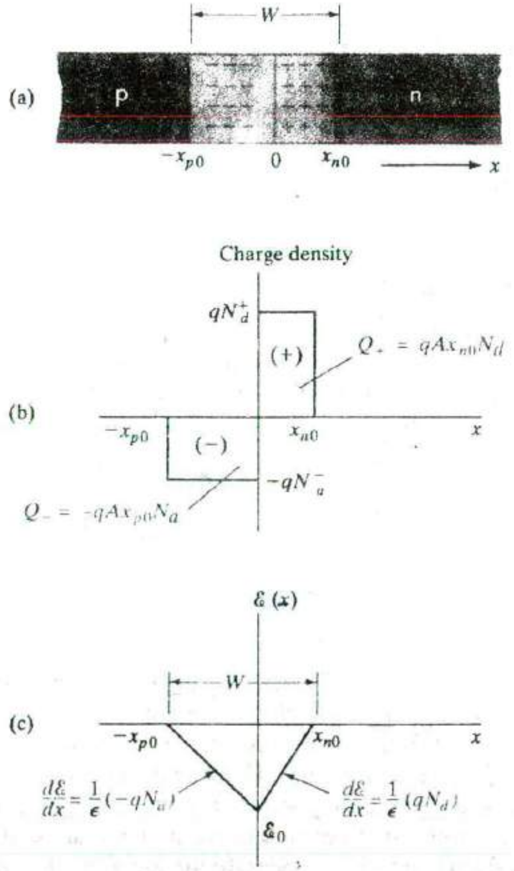


Figure 5–12 Space charge and electric field distribution within the transition region of a p-n junction with $N_d > N_a$: (a) the transition region, with $x = 0$ defined at the metallurgical junction; (b) charge density within the transition region, neglecting the free carriers; (c) the electric field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

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, 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:

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

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:

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

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

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

[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 become equal.

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:

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

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.

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

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):

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

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

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

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

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 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

Boron is implanted into an n-type Si sample ($N_d = 10^{16}\ \text{cm}^{-3}$), forming an abrupt junction of square cross section, with area $= 2 \times 10^{-3}\ \text{cm}^2$. Assume that the acceptor concentration in the p-type region is $N_a = 4 \times 10^{18}\ \text{cm}^{-3}$. Calculate V_0, 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

From Eq. (5–8),

$$V_0 = \frac{kT}{q}\ln\frac{N_a N_d}{n_i^2} = 0.0259\ln\frac{4 \times 10^{34}}{2.25 \times 10^{20}}$$

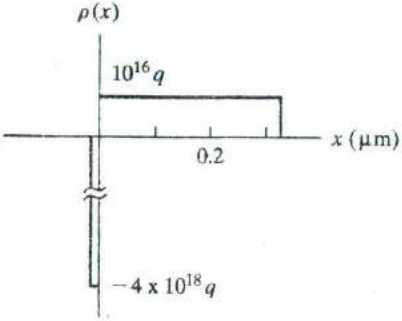
$$= 0.0259\ln(1.78 \times 10^{14}) = 0.85\ V$$

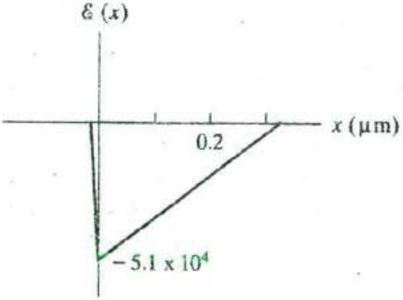
From Eq. (5–21),

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

$$= \left[\frac{2(11.8 \times 8.85 \times 10^{-14})(0.85)}{1.6 \times 10^{-19}}(0.25 \times 10^{-18} + 10^{-16})\right]^{1/2}$$

$$= 3.34 \times 10^{-5}\text{ cm} = 0.334\ \mu\text{m}$$





From Eq. (5–23),

$$x_{n0} = \frac{3.34 \times 10^{-5}}{1 + 0.0025} \simeq 0.333\ \mu\text{m}$$

$$x_{p0} = \frac{3.34 \times 10^{-5}}{1 + 400} \simeq 8.3 \times 10^{-8}\text{ cm} = 8.3\text{ Å}$$

Note that $x_{n0} \simeq W$.

$$Q_+ = -Q_- = qAx_{n0}N_d = (1.6 \times 10^{-19})(2 \times 10^{-3})(3.33 \times 10^{-5})(10^{16})$$

$$= 1.07 \times 10^{-10}\ \mathrm{C}$$

$$\mathscr{E}_0 = \frac{-qN_d x_{n0}}{\epsilon} = \frac{-(1.6 \times 10^{-19})(10^{16})(3.3 \times 10^{-5})}{(11.8)(8.85 \times 10^{-14})}$$

$$= -5.1 \times 10^4\ \mathrm{V/cm}$$

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.

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 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).*

[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.

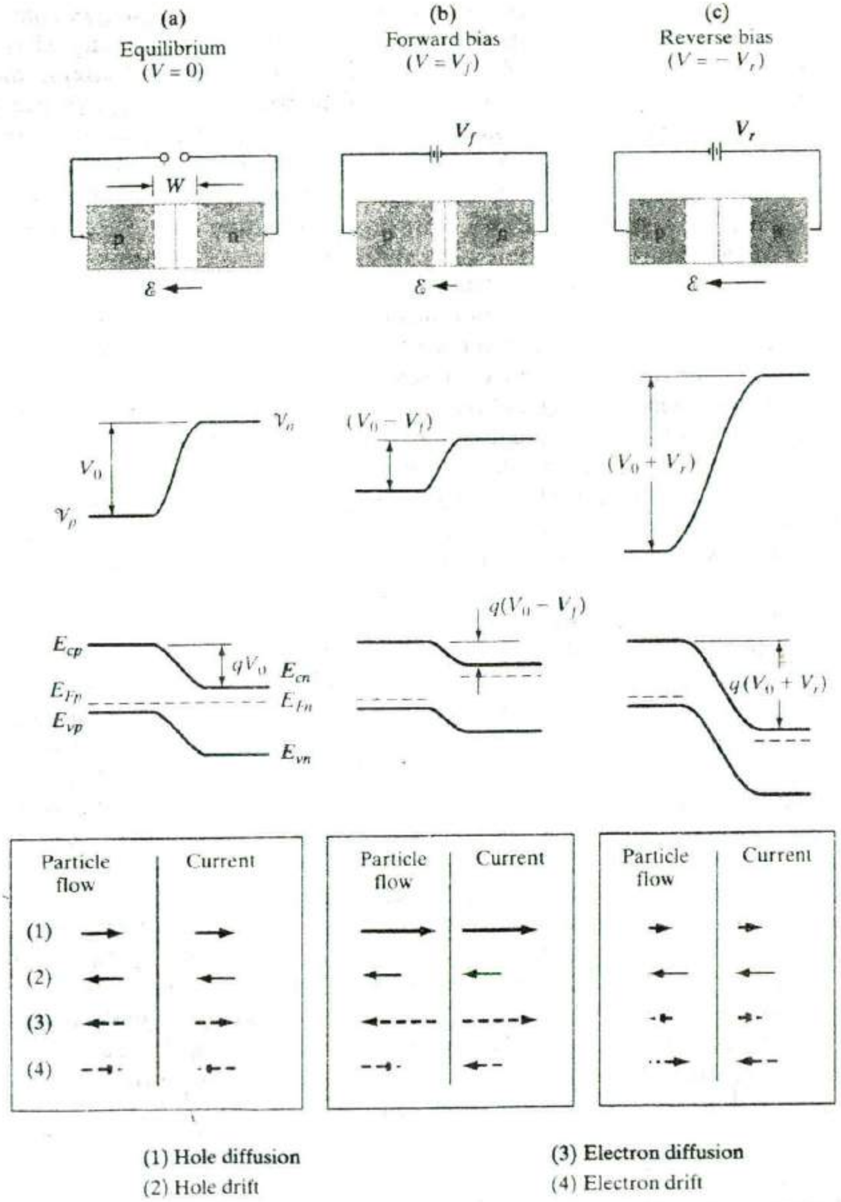


Figure 5–13 Effects of a bias at a p-n junction; transition region width and electric field, electrostatic potential, energy band diagram, and particle flow and current directions within W for (a) equilibrium, (b) forward bias, and (c) reverse bias.

The *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 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

[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).

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]

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

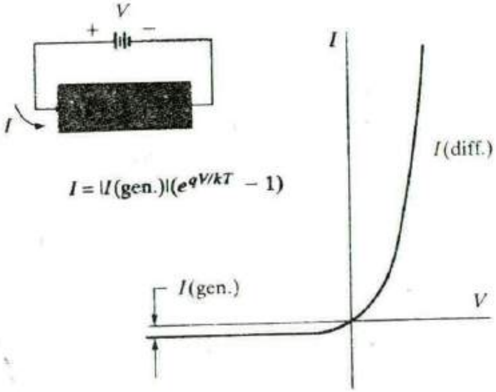


Figure 5–14
I–V characteristic of a p-n junction.

[10] The total current I is the sum of the generation and diffusion components. However, these components are oppositely directed, 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.

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

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

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 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:

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

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

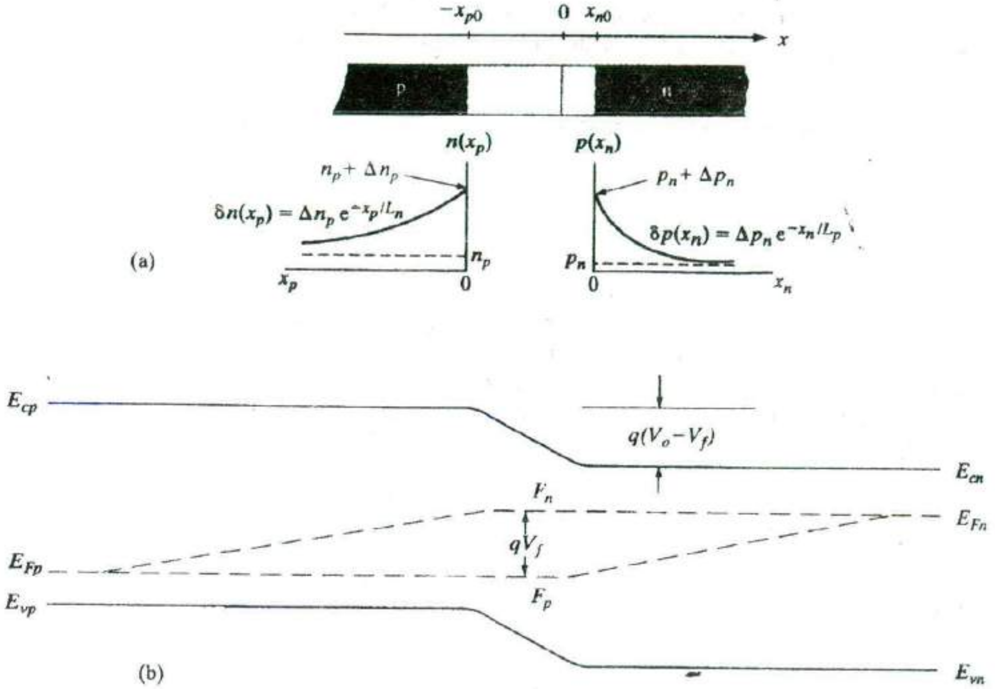


Figure 5–15
Forward-biased junction: (a) minority carrier distributions on the two sides of the transition region and definitions of distances x_n and x_p measured from the transition region edges; (b) variation of the quasi-Fermi levels with position.

The 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 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}:

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

[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 tha spatial variations of the carrier concentrations.

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 $V = -V_r$:

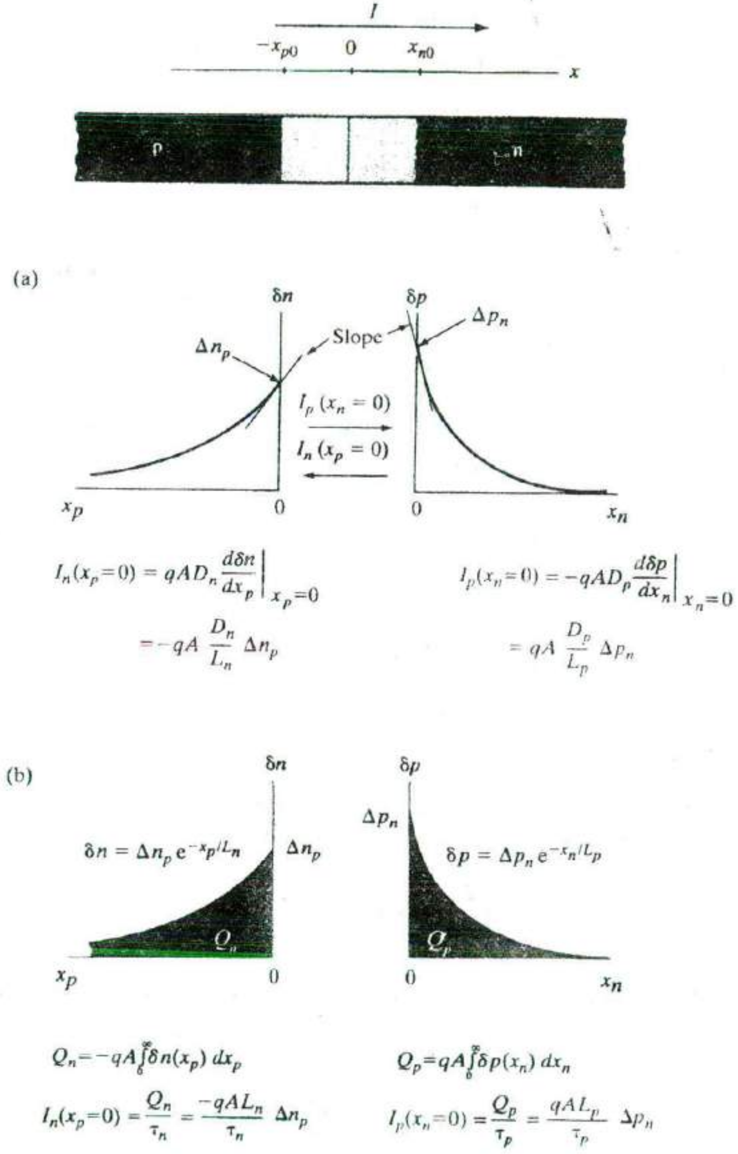
$$I = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)(e^{-qV_r/kT} - 1) \qquad (5\text{–}37a)$$

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

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

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

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.



$$I_n(x_p=0) = qAD_n \frac{d\delta n}{dx_p}\bigg|_{x_p=0} = -qA\,\frac{D_n}{L_n}\,\Delta n_p$$

$$I_p(x_n=0) = -qAD_p \frac{d\delta p}{dx_n}\bigg|_{x_n=0} = qA\,\frac{D_p}{L_p}\,\Delta p_n$$

$$Q_n = -qA\int_0^\infty \delta n(x_p)\,dx_p$$

$$I_n(x_p=0) = \frac{Q_n}{\tau_n} = \frac{-qAL_n}{\tau_n}\,\Delta n_p$$

$$Q_p = qA\int_0^\infty \delta p(x_n)\,dx_n$$

$$I_p(x_n=0) = \frac{Q_p}{\tau_p} = \frac{qAL_p}{\tau_p}\,\Delta p_n$$

(c)

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

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

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. We see that the quasi-Fermi levels are more or less flat within the depletion region, which appears to be inconsistent with what we learned in Section 4.4.6 about the current flow being proportional to the gradient of the quasi-Fermi levels. Keeping in mind that for an ideal diode, the electron (and hole) current is constant across the depletion region, we see that within the depletion region the product of the gradient of the quasi-Fermi level and the carrier concentration must be independent of position. For a given current, the gradient in the quasi-Fermi level must be significant for minority carriers, since the carrier concentration is small (see Eq. 4–52). On the other hand, for majority carriers, very little gradient is needed in the quasi-Fermi level. Within W there is an intermediate situation, where the carrier concentration is changing from majority on one side to minority on the other. Although there is some variation in F_n and F_p within W, it doesn't show up on the scale used in Fig. 5–15. A homework problem with typical values should help clarify the concept (Prob. 5.21). 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, $I_p(x_n = 0)$ must supply enough holes 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:

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

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* 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

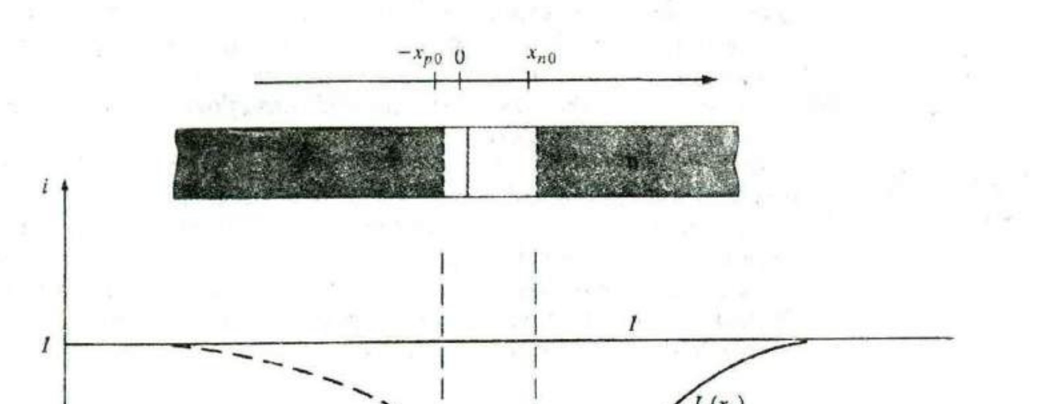


Figure 5–17
Electron and hole components of current in a forward-biased p-n junction. In this example, we have a higher injected minority hole current on the n-side than electron current on the p side because we have a lower n doping than p doping.

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

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

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

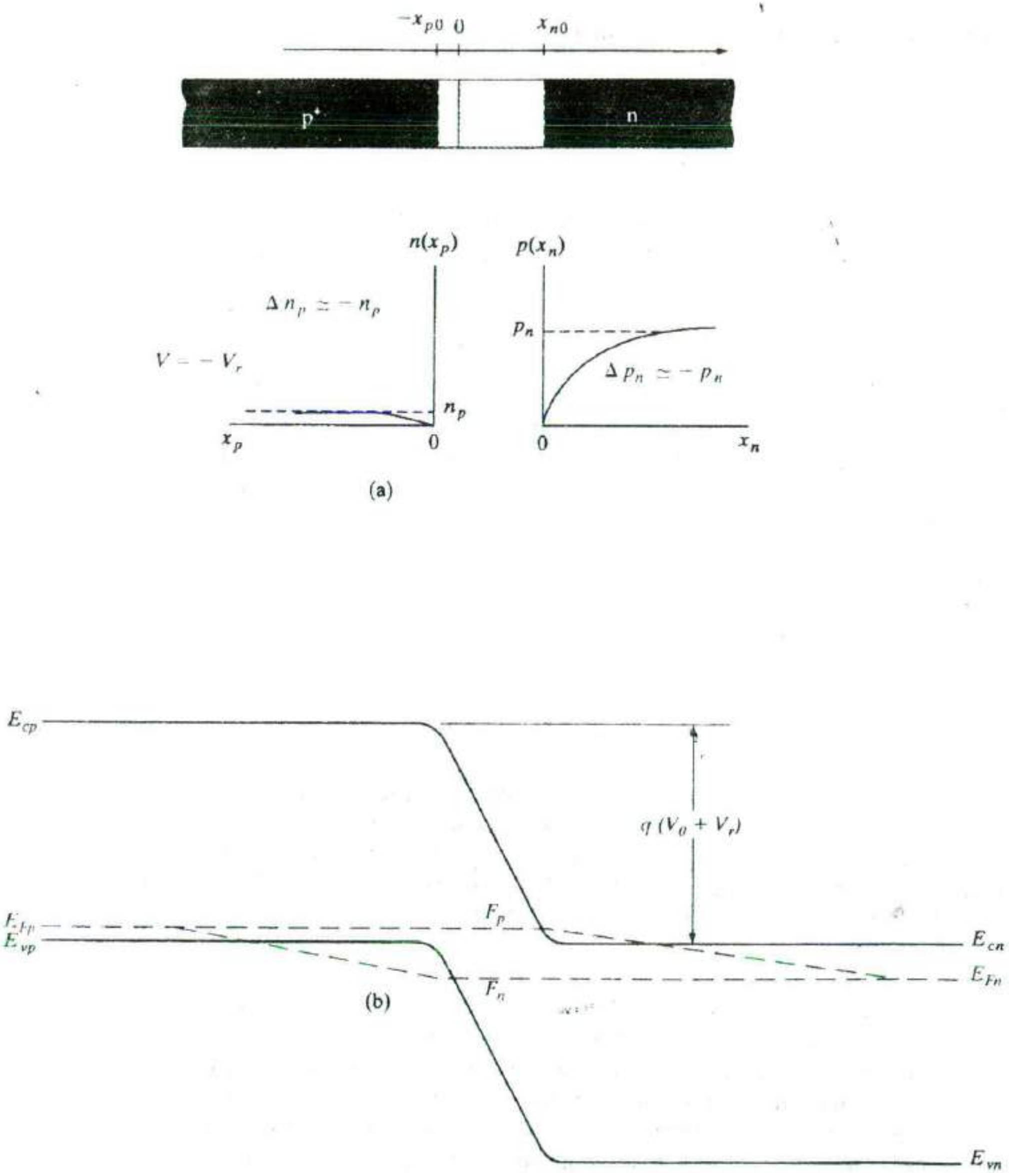


Figure 5–18
Reverse-biased p-n junction: (a) minority carrier distributions near the reverse-biased junction; (b) variation of the quasi-Fermi levels.

current, Eq. (5–38), represents the rate at which carriers are generated thermally within a diffusion length of each side of the transition region.

EXAMPLE 5–4

Consider a volume of n-type material of area A, with a length of one hole diffusion length L_p. The rate of thermal generation of holes within the volume is

$$AL_p\frac{p_n}{\tau_p} \quad \text{since } g_{\text{th}} = \alpha_r n_i^2 = \alpha_r n_n p_n = \frac{p_n}{\tau_p}$$

Assume that each thermally generated hole diffuses out of the volume before it can recombine. The resulting hole current is $I = qAL_p p_n/\tau_p$, which is the same as the saturation current for a p$^+$-n junction. We conclude that saturation current is due to the collection of minority carriers thermally generated within a diffusion length of the junction.

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.

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.

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

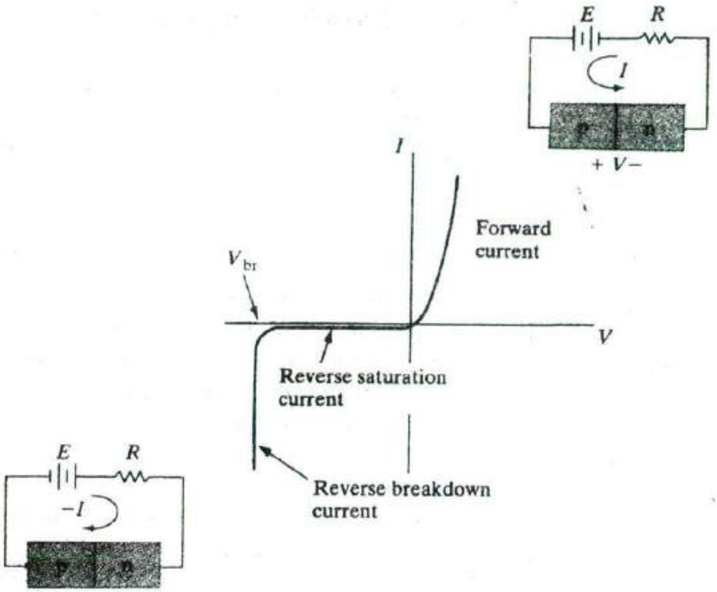


Figure 5–19 Reverse breakdown in a p-n junction.

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

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

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.*

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.

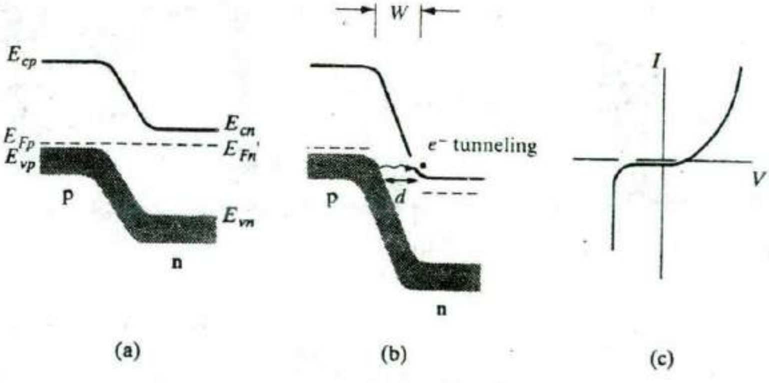


Figure 5–20
The Zener effect: (a) heavily doped junction at equilibrium; (b) reverse bias with electron tunneling from p to n; (c) I–V characteristic.

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

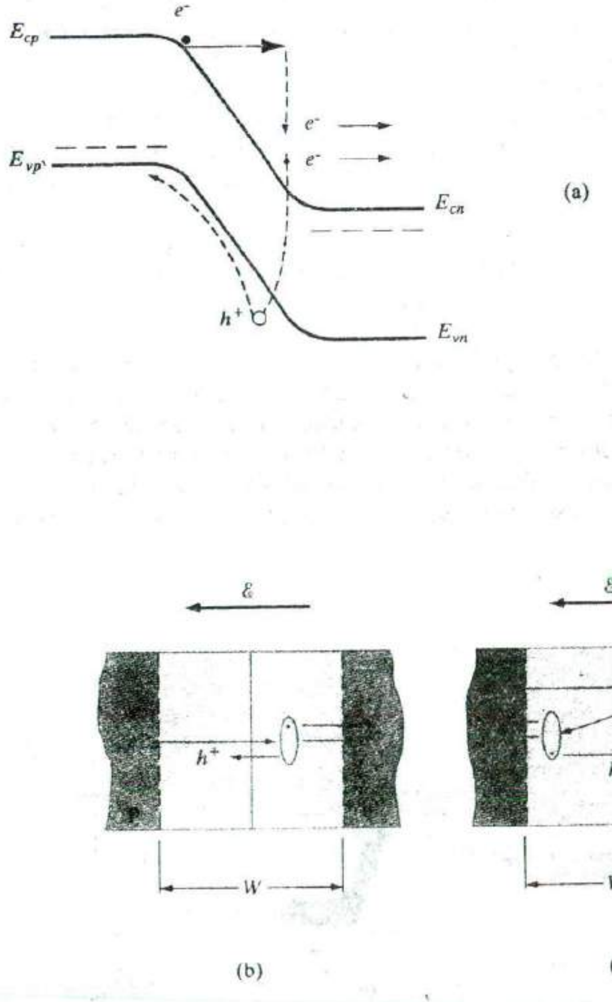


Figure 5–21 Electron-hole pairs created by impact ionization: (a) band diagram of a p-n junciton in reverse bias showing (primary) electron gaining kinetic energy in the field of the depletion region, and creating a (secondary) electron-hole pair by impact ionization, the primary electron losing most 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.

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. After the Pn_{in} collisions by the primary electrons, we have the primary plus the secondary electrons, $n_{\text{in}}(1 + P)$. 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 $n_{\text{in}}P$ secondary pairs there will be $(n_{\text{in}}P)P$ ionizing collisions and $n_{\text{in}}P^2$ tertiary pairs. Summing up the total number of electrons out of the region at n after many collisions, we have

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

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

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

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

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

where the exponent $\mathbf{n}$ varies from about 3 to 6, depending on the type of material used for the junction.

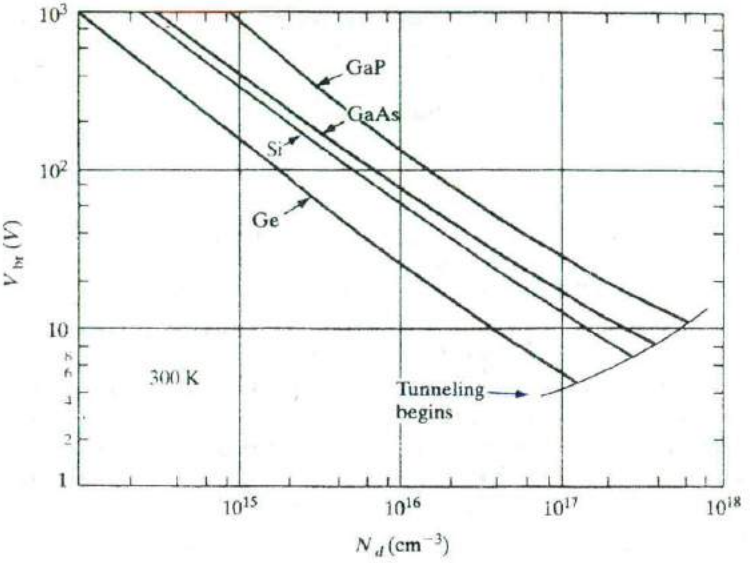


Figure 5–22 Variation of avalanche breakdown voltage in abrupt p^+-n junctions, as a function of donor concentration on the n side, for several semiconductors. [After S.M. Sze and G. Gibbons, *Applied Physics Letters*, vol. 8, p. 111 (1966).]

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

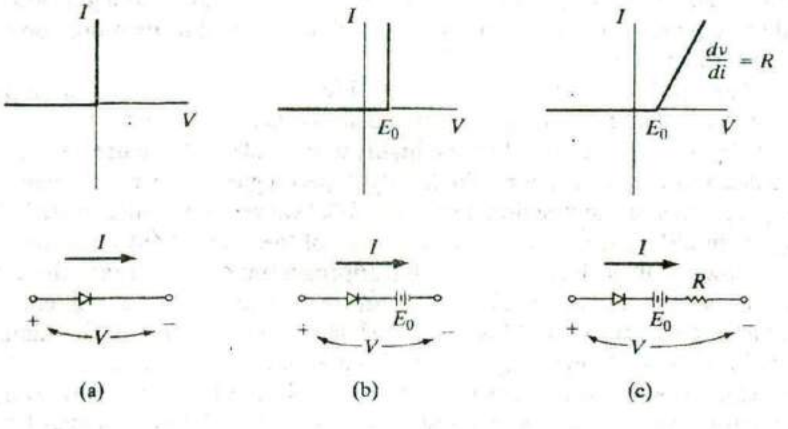


Figure 5–23 Piecewise-linear approximations of junction diode characteristics: (a) the ideal diode; (b) ideal diode with an offset voltage; (c) ideal diode with an offset voltage and a resistance to account for slope in the forward characteristic.

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.

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. 5.33). The result of punch-through is a breakdown below the value of V_{br} expected from Fig. 5–22.

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.

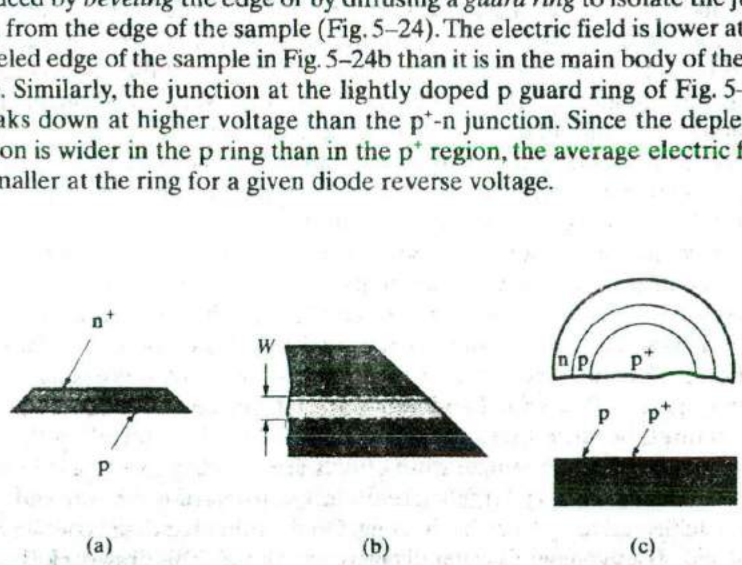


Figure 5–24 Beveled edge and guard ring to prevent edge breakdown under reverse bias: (a) diode with beveled edge; (b) closeup view of edge, showing reduction of depletion region near the bevel; (c) guard ring.

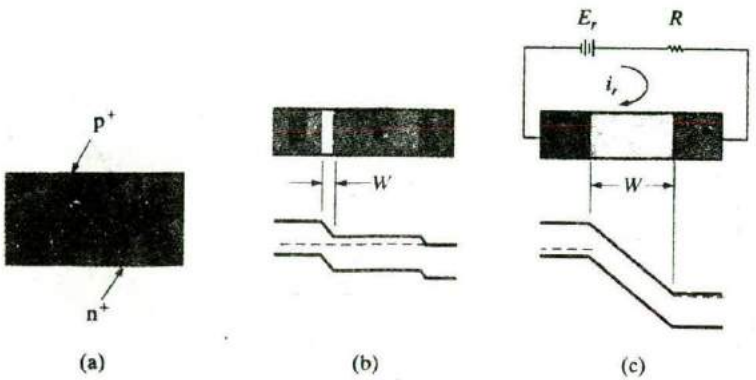


Figure 5–25
A p^+-n-n^+ junction diode: (a) device configuration; (b) zero-bias condition; (c) reverse-biased to punch-through.

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, 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

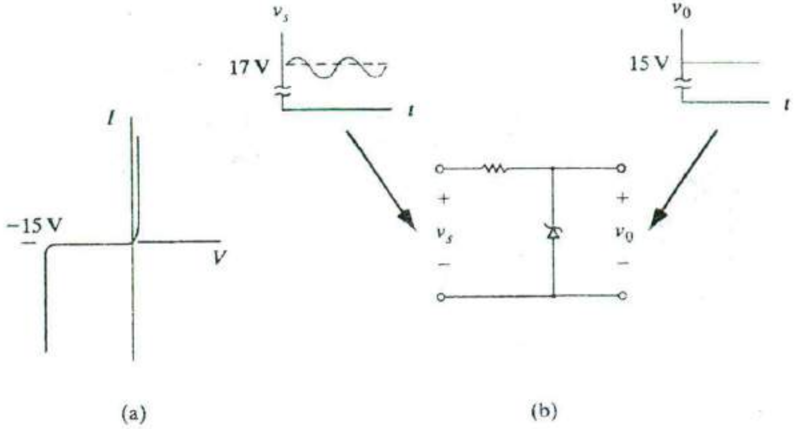


Figure 5–26 A breakdown diode: (a) *I–V* characteristic; (b) application as a voltage regulator.

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.

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

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

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 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.

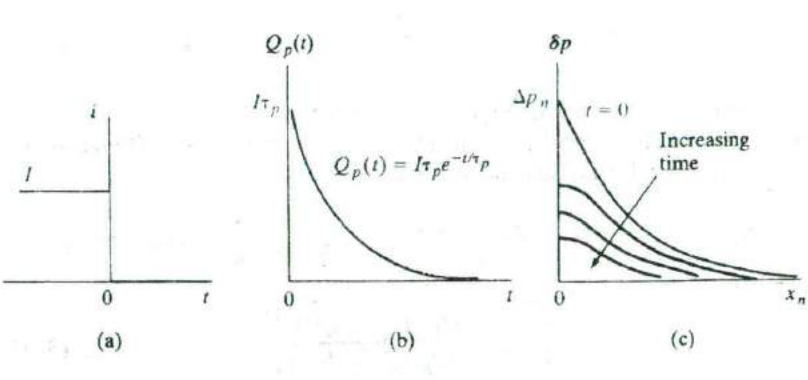


Figure 5–27 Effects of a step turn-off transient in a p$^+$-n diode: (a) current through the diode; (b) decay of stored charge in the n-region; (c) excess hole distribution in the n-region as a function of time during the transient.

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

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

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

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

we have for the stored charge at any instant

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

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

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

[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 $t = 0$.

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 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). 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 be-

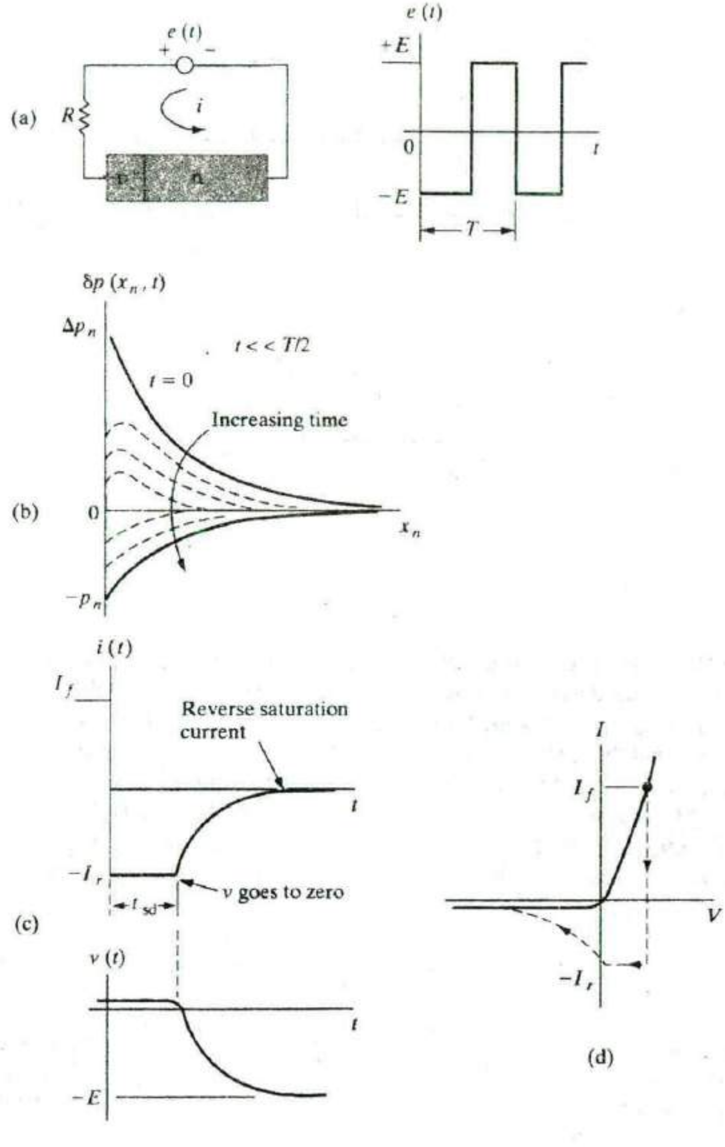


Figure 5–28 Storage delay time in a p$^+$-n diode: (a) circuit and input square wave; (b) hole distribution in the n-region as a function of time during the transient; (c) variation of current and voltage with time; (d) sketch of transient current and voltage on the device I–V characteristic.

comes 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

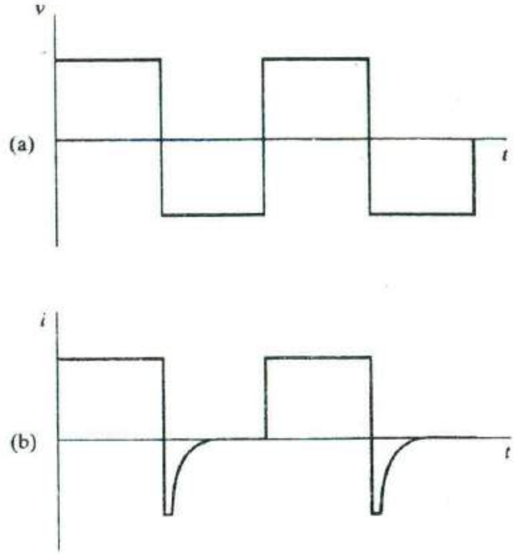


Figure 5–29 Effects of storage delay time on switching signal: (a) switching voltage; (b) diode current.

(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

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

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.

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),

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

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 straight-forward 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 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 *transition region capacitance* or *depletion layer capacitance*; (2) is often called the *diffusion capacitance*.

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

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

and therefore the charge on each side of the dipole is

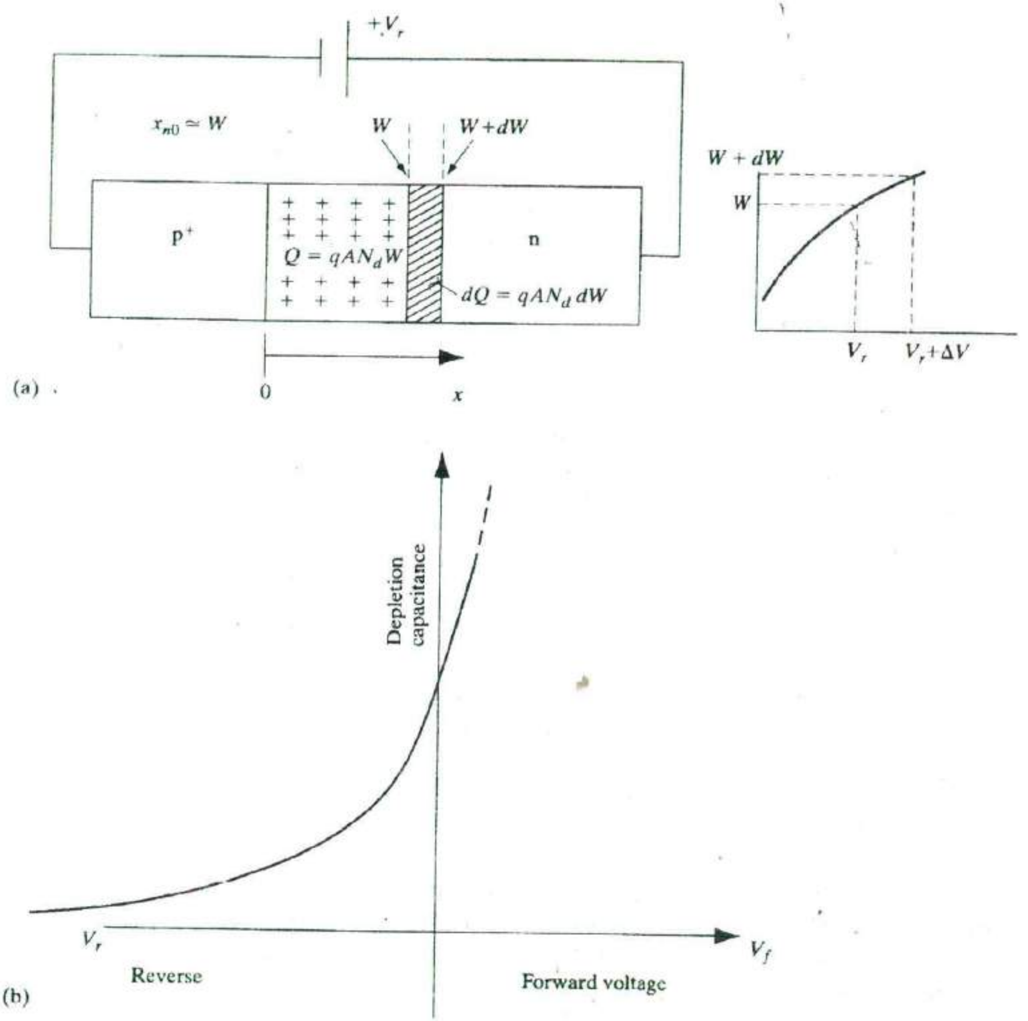


Figure 5–30
Depletion capacitance of a junction: (a) p⁺-n junction showing variation of depletion edge on n side with reverse bias. Electrically, the structure looks like a parallel plate capacitor whose dielectric is the depletion region, and the plates are the space charge neutral regions; (b) variation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

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

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:

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

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):

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

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)

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

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. 5.38).

The junction capacitance dominates the reactance of a p-n junction under reverse bias; for forward bias, however, the charge storage, or diffusion 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 where the stored charges are extracted, and where the relevant voltage drops occur.

[15] S. Loux 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), p. 396 (Feb 1999).

To illustrate the calculation let us look at the simplified case of a symmetric, abrupt p-n junction where the doping levels N_a and N_d are equal.

We will consider two cases. For the long diode, which we have been dealing with so far, the diffusion lengths are assumed to be small compared to the lengths of the p and n regions. In this case, the injected minority carriers on either side of the depletion region decay exponentially to their equilibrium value long before they reach the ohmic contacts (Fig. 5–31a). On the other hand, the diffusion lengths in a short diode are assumed to be long compared to the length of the p and n regions (Fig. 5–31b). In the short diode, the injected excess minority carrier concentrations decrease almost linearly to zero at the ohmic contacts designated $x = -a$ and $x = c$ in Fig. 5–31b. Minority carrier distributions are discussed in Probs. 5.34-5.36 and Section 5.3.2. We will discuss the almost linear excess carrier distribution in a narrow region in Section 7.4.1.

The total current in the diode is the sum of the particle currents and the displacement current evaluated at any suitable location (chosen here at $x = 0$).

$$J_t = J_n(0) + J_p(0) + J_d(0) \tag{5–64}$$

For the general case of time-dependent voltage and currents, we need to solve the hole and electron current continuity equations (Eq. 4–31a and 4-31b) for J_n and J_p and also take the time-derivative of Poisson's equation (Eq. 5–14) to obtain the displacement current:

$$J_d(x) = \frac{\partial}{\partial t}[\epsilon \mathscr{E}(x)] \tag{5–65a}$$

We can integrate Poisson's equation between 0 and c, and take the derivative with respect to time to get

$$J_d(0) = q\frac{\partial}{\partial t}\int_0^c (n - p)dx + J_d(c) \tag{5–65b}$$

We notice that the dopant charges do not appear here because they are time independent. Laux and Hess show in their paper that for most practical cases $J_d(c) = 0$, and that the displacement current $J_d(0)$ originates from a time-varying voltage across the depletion capacitance that was discussed earlier in this section.

Integrating the electron continuity equation (Eq. 4–31b) from $-a$ to 0 in the p-region, and the hole continuity equation (Eq. 4–31a) from 0 to c in the n-region, we can get the sum of the electron and hole particle current densities, at $x = 0$.

$$J_n(0) + J_p(0) = q\int_{-a}^{c} R\,dx + q\int_{-a}^{0} \frac{\partial n}{\partial t}dx + q\int_{0}^{c}\frac{\partial p}{\partial t}dx + J_n(-a) + J_p(c) \tag{5–65c}$$

where R is the net recombination rate at each point x.

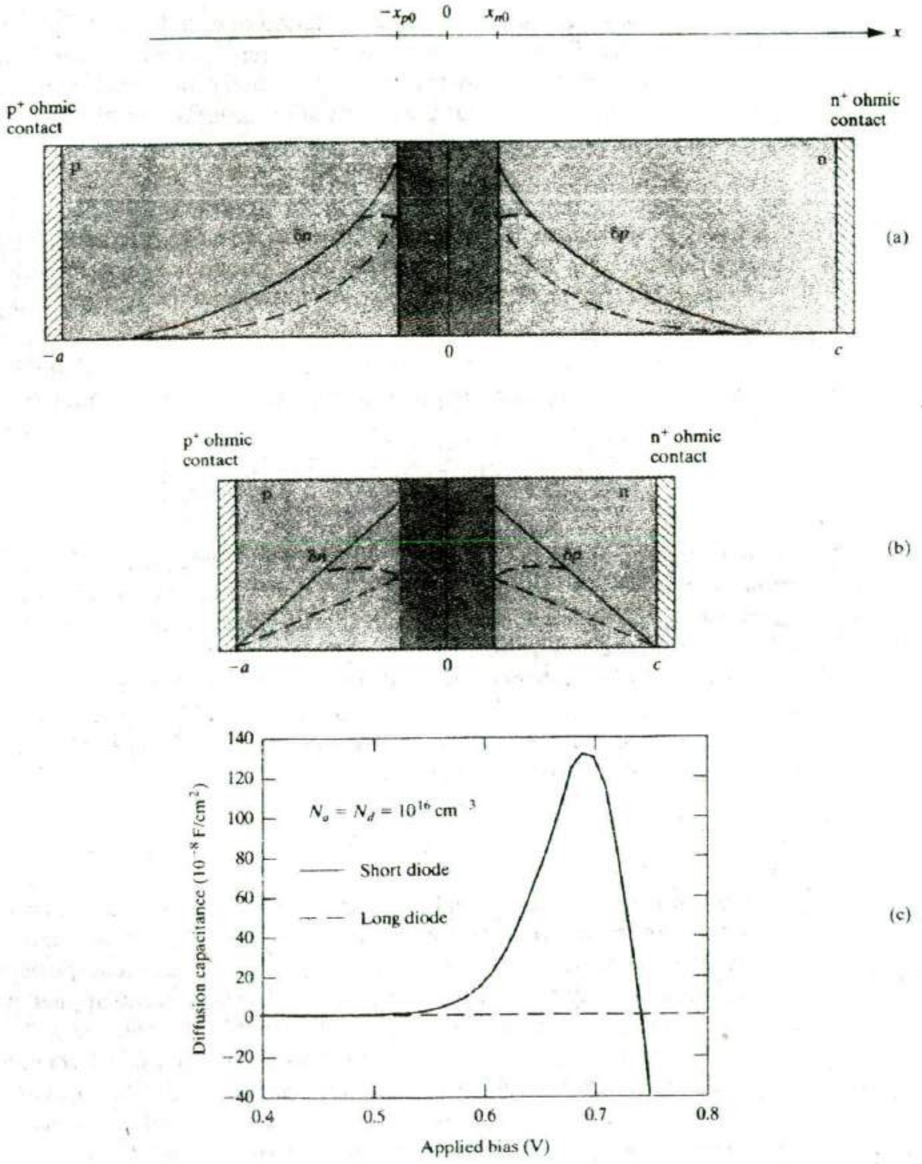


Figure 5–31

Diffusion capacitance in p-n junctions. (a) Steady-state minority carrier distribution for a forward bias, V (colored lines), and reduced forward bias, V-ΔV (dashed colored lines) in a long diode. The transient case when the current is reduced suddenly is shown by the black, dashed lines. Although the carrier distributions can change quickly near the junctions, they stay close to the original steady-state distributions far from the junctions at first. Gradually, the carrier distributions approach the new steady-state distributions for V-ΔV (dashed colored lines); (b) minority carrier distributions in a short diode; (c) diffusion capacitance as a function of forward bias in long and short diodes.

Let us see what the physical interpretation of each term in Eq (5–65c) is. For a long base diode, since the minority carrier concentrations reduce to zero before we reach the ohmic contacts at $-a$ and c, the terms $J_n(-a)$ and $J_p(c)$ are zero. Furthermore, for the steady state case, the terms

$$q\int_{-a}^{0}\frac{\partial n}{\partial t}\,dx + q\int_{0}^{c}\frac{\partial p}{\partial t}\,dx$$

are also zero because the time derivatives are zero. We see then that the dc current is given by the first term, which is the integrated carrier recombination rate. This is exactly the charge control model of the diode that we discussed in regard to Fig. 5–16.

For the time-dependent case, the integrands in the second and third terms in Eq. (5–65c) can be expressed, for example for the holes, by the chain rule as

$$q\int_{0}^{c}\frac{\partial p}{\partial t}dx = q\int_{0}^{c}\frac{\partial p}{\partial V}\cdot\frac{\partial V}{\partial t}dx = q\frac{\partial}{\partial V}\left\{\int_{0}^{c}pdx\right\}\frac{\partial V}{\partial t} \qquad (5\text{–}66)$$

We have interchanged the order of integration with respect to x and differentiation with respect to t. Equation (5–66) is in the form of a current, with a (diffusion) capacitance times the voltage ramp rate. There is a similar contribution from the electrons.

In conventional theories of the diffusion capacitance due to stored minority carriers, the second and third terms in Eq. (5–65c) are erroneously considered to be the only contributors. Furthermore, the stored charge, for example for holes, is approximately set equal to

$$q\int_{x_{n0}}^{c} pdx$$

in the neutral region, rather than the correct expression in Eq. (5–65c) which considers both the neutral and the depletion regions. Also, in conventional theories, we assume that all the applied voltage is dropped across the depletion region. In reality, there can be a significant fraction of the applied bias dropped across the neutral region from x_{n0} to c and from $-a$ to $-x_{p0}$.

More importantly, Laux and Hess have shown that the first term in Eq. (5–65c) due to carrier recombination cancels most of the diffusion capacitance in long base diodes. Physically, the reason for this cancellation of the capacitance effect is that if the injected minority carriers (holes) recombine on the n side between 0 and c, they cannot be fully "*reclaimed*" at the injecting ohmic contact at $-a$ where the external voltage is changed, and similarly for electron injection.

For steady state, holes lost due to EHP recombination in the diode must be replenished at $-a$ (and electrons at c). For capacitive effects to be manifested, however, we must consider the transient case. We must determine the transfer of charge through the external terminals, as a function of the applied voltage variation at those terminals. To understand why the reclaimable charge is less than the total stored minority carrier charge, let us consider the transient conditions in a p$^+$-n diode, as discussed in Section 5.5.2. As shown in Fig. 5–28, when the forward bias is reduced, the minority carrier hole concentration at the edge of the depletion region is reduced, and therefore the slope of the hole distribution changes near the junction. This reduction occurs by some holes near $x_n = 0$ moving to the left towards the p$^+$ ohmic contact. The arrival of holes at the p$^+$ contact is referred to as reclaimed charge. Not all of the reduction in the hole distribution (shown in color in Fig. 5–31a) occurs by reclaiming holes at the p$^+$ contact, however. From the shape of the hole distribution within the n region, there obviously continues to be a diffusion of holes to the right also, toward the n$^+$ ohmic contact. In a long diode, these holes do not make it all the way to the n$^+$ ohmic contact because they recombine with electrons on the way. These recombined electrons have to be replenished by the n$^+$ ohmic contact. The key point is that because some of the holes are diffusing to the right, not all the holes in the stored distribution can be extracted (reclaimed) at the p$^+$ ohmic contact at the left, when the forward bias is reduced by a small amount in the transient case. The resulting capacitance–voltage behavior (Fig. 5–31c) for long base diodes shows almost zero diffusion capacitance.

The situation is somewhat different for narrow or short-base diodes. Since the minority carrier diffusion lengths are much longer than the length of the diode, there is negligible carrier recombination within the charge distribution, and the term

$$q\int_{-a}^{c} R dx$$

in Eq. (5–65c) is small. On the other hand, since most of the injected minority carriers now reach the ohmic contacts, the fourth and fifth terms in Eq. (5–65c) are large, unlike for the long base case. To understand physically why the reclaimed hole charge at $-a$ is less than the total stored charge in the short diode, we must recognize once again that for capacitive effects to be manifested, we need to consider the transient case. When the current is reduced suddenly, the slope of the hole distribution at $x_n = 0$ reduces, but the slope at $x = c$ does not (Fig. 5–31b). Because most holes reach the n$^+$ contact (at c), there is a reduction in the "reclaimable" hole charge at the p^+-ohmic contact (at $-a$). Hence, the net charge that is driven through the external circuit is reduced, and the diffusion capacitance due to minority carrier storage is reduced for the short diode, although not as drastically as for the long diode case. An exact solution of the continuity equation in this case shows that the reclaimable charge is 2/3 of the total stored charge.

There is an exponentially increasing diffusion capacitance with applied forward bias for the short diode (Fig. 5–31c). For a triangular minority carrier charge distribution (Fig. 5–31b), the stored hole charge on the n side is given by half the product of the height times the base of the triangle (Prob. 5.34).

$$Q_p = \frac{1}{2}qA(c - x_{no})(\Delta p_n) = \frac{1}{2}qA(c - x_{no})p_n\left(e^{\frac{qV}{kT}} - 1\right) \quad (5\text{–}67a)$$

Since the reclaimable charge is 2/3 of this, the diffusion capacitance is:

$$C_s = \frac{dQ_p}{dV} = \frac{1}{3}\,\frac{q^2}{kT}A(c - x_{no})p_n e^{\frac{qV}{kT}} \tag{5–67b}$$

There is a similar contribution from the stored electrons on the p-side. Laux and Hess show in their paper that because of the voltage drop in the neutral regions, and the possibility of conductivity modulation occurring there due to high carrier concentration at large forward biases, the diffusion capacitance becomes negative around the built-in voltage, V_0. Most Si p-n junctions in practice behave like short-base diodes, while laser diodes made in direct bandgap (short lifetime) semiconductors often correspond to the long base case.

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

$$G_s = \frac{dI}{dV} = \frac{qAL_p p_n}{\tau_p}\,\frac{d}{dV}(e^{qV/kT}) = \frac{q}{kT}I \tag{5–67c}$$

5.5.5 The Varactor Diode

The term *varactor* is a shortened form of *var*iable *reactor*, 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

$$C_j \propto V_r^{-\mathbf{n}} \quad \text{for } V_r \gg V_0 \tag{5–68a}$$

For example, in a linearly graded junction the exponent **n** is one-third (Prob. 5.38). 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. (5–68a) 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.

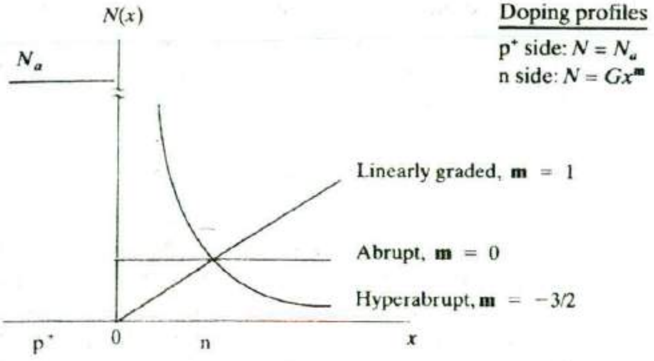


Figure 5–32 Graded junction profiles: linearly graded, abrupt, hyperabrupt.

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 $\mathbf{m}$ is 0, 1, or $-\frac{3}{2}$. We can show (Prob. 5.37) that the exponent $\mathbf{n}$ in Eq. (5–68a) is $1/(\mathbf{m} + 2)$ for the p$^+$-n junction. Thus for the profiles of Fig. 5–32, $\mathbf{n}$ is $\frac{1}{2}$ for the abrupt junction and $\frac{1}{3}$ for the linearly graded junction. The 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.

$$\omega_r = \frac{1}{\sqrt{LC}} \propto \frac{1}{\sqrt{V_r^{-n}}} \propto V_r, \quad \text{for } \mathbf{n} = 2 \qquad (5\text{–}68b)$$

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

[16] It is clear that $N_d(x)$ 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.

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.

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

$$I = \frac{qAD_p}{L_p} p_n e^{qV/kT} = \frac{qAD_p}{L_p} N_v e^{[qV-(E_{Fn}-E_{vn})]/kT} \qquad (5\text{–}69)$$

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

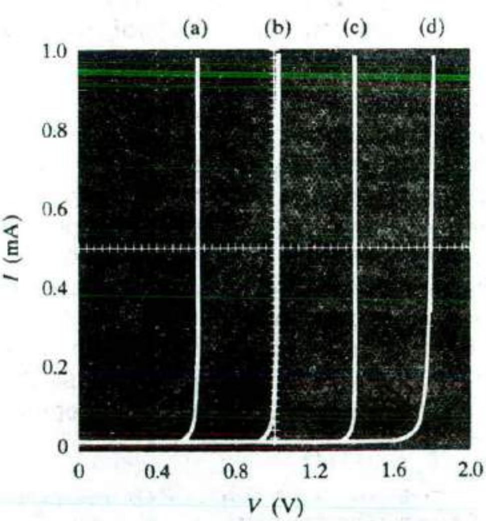


Figure 5–33 *I–V* characteristics of heavily doped p-n junction diodes at 77 K, illustrating the effects of contact potential on the forward current: (a) Ge, $E_g \simeq$ 0.7 eV; (b) Si, $E_g \simeq$ 1.4 eV; (c) GaAs, $E_g \simeq$ 1.4 eV; (d) GaAsP, $E_g \simeq$ 1.9 eV.

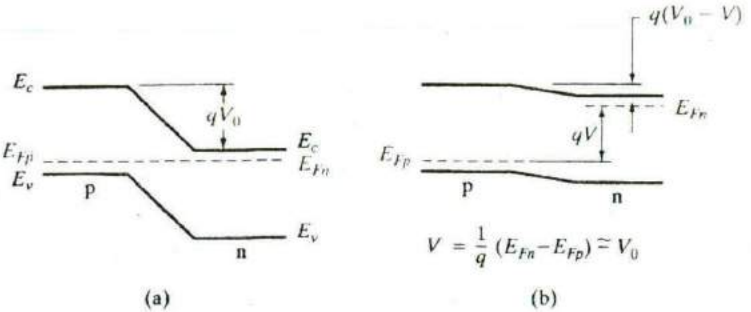


Figure 5–34 Examples of contact potential for a heavily doped p-n junction: (a) at equilibrium; (b) approaching the maximum forward bias $V = V_0$.

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).

The limiting forward bias across a p-n junction is equal to the contact potential, as in Fig. 5–34(b). 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

$$\frac{p(-x_{p0})}{p(x_{n0})} = \frac{p_p + \Delta p_p}{p_n + \Delta p_n} = e^{q(V_0 - V)/kT} = \frac{n_n + \Delta n_n}{n_p + \Delta n_p} \qquad (5\text{–}70)$$

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

$$pn = p(-x_{p0})n(-x_{p0}) = p(x_{n0})n(x_{n0}) = n_i^2 e^{\frac{F_n - F_p}{kT}} = n_i^2 e^{qV/kT} \qquad (5\text{–}71\text{a})$$

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

$$(p_p + \Delta p_p)(n_p + \Delta n_p) = n_i^2 e^{qV/kT} \qquad (5\text{–}71\text{b})$$

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

$$\Delta n_p = n_i e^{qV/2kT} \qquad (5\text{–}72)$$

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

$$I \propto e^{qV/2kT} \qquad (5\text{–}73)$$

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**:

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

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

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

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 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

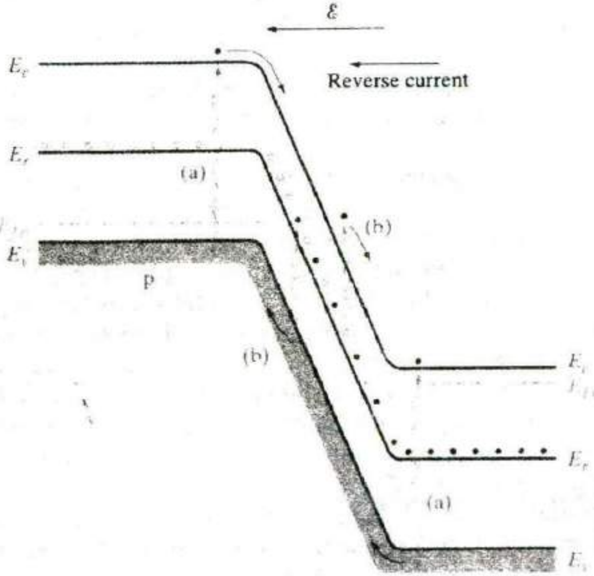
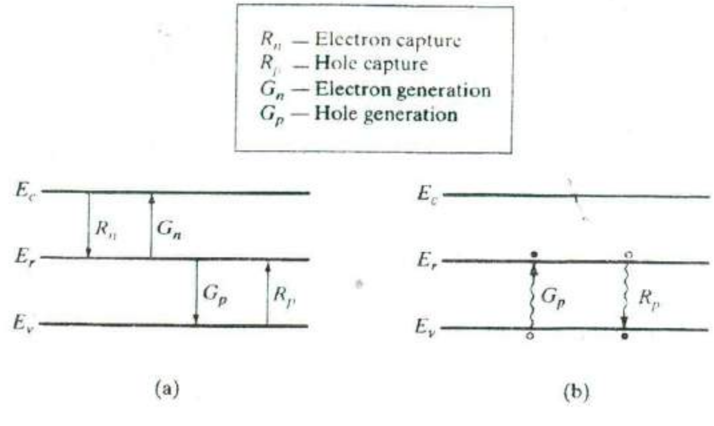


Figure 5–35
Current in a reverse-biased p-n junction due to thermal generation of carriers by (a) band-to-band EHP generation, and (b) generation from a recombination level.

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_r (hole generation) and electron deexcitation from E_r to E_v (hole capture by E_r).



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.

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. 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 semi-log scale, both for an ideal Shockley diode as well as for non-ideal devices. For an ideal forward-based diode, we get a straight line on a semi-log 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 (**n** = 2). For moderate currents, we get ideal low-level injection and diffusion-limited current (**n** = 1). At higher currents, we get high level injection and **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 effects.

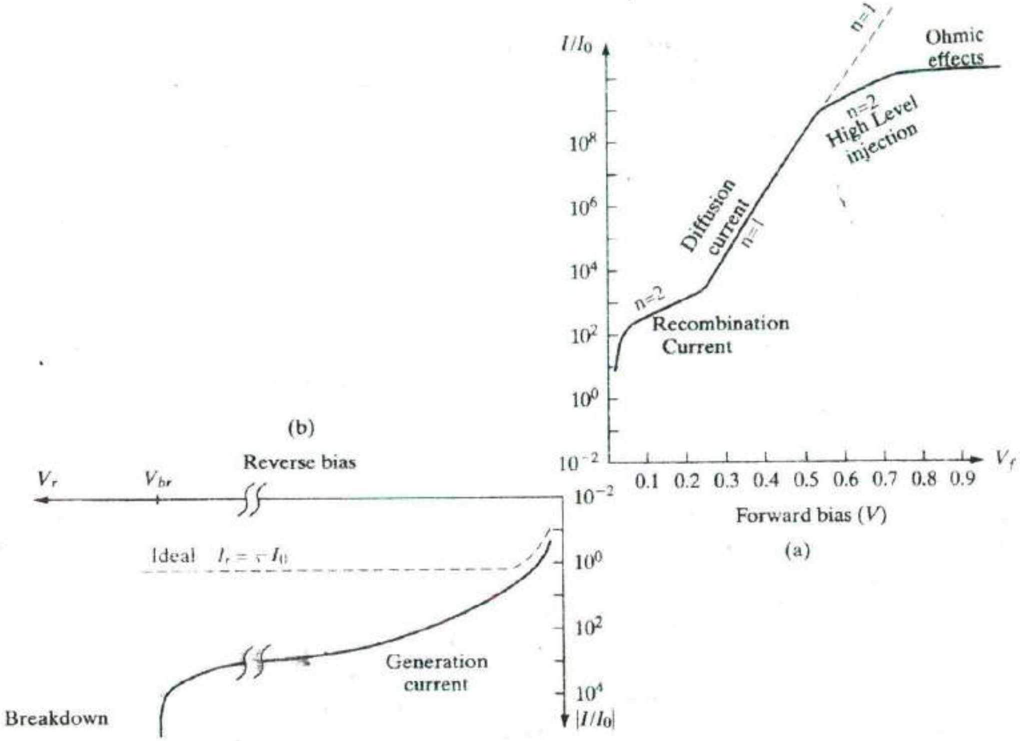


Figure 5–37
Forward and reverse current-voltage characteristics plotted on semi-log scales, with current normalized with respect to saturation current, I_0; (a) the ideal forward characteristic is an exponential with an ideality factor, **n** = 1 (dashed straight line on log-linear plot). The actual forward characteristics of a typical diode (colored line) have four regimes of operation; (b) ideal reverse characteristic (dashed line) is a voltage-independent current = $-I_0$. Actual leakage characteristics (colored line) are higher due to generation in the depletion region, and also show breakdown at high voltages.

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

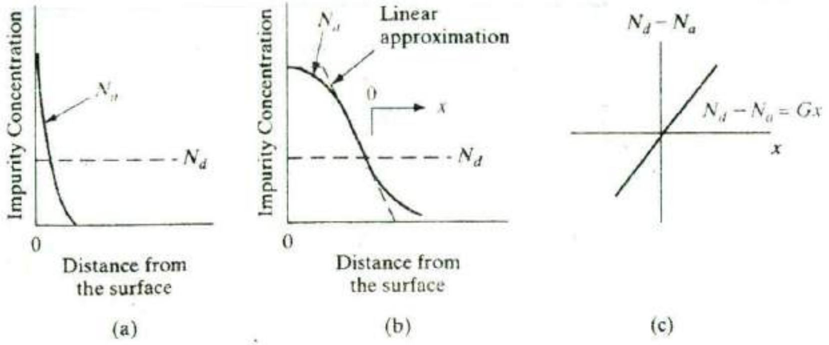


Figure 5–38 Approximations to diffused junctions: (a) shallow diffusion (abrupt); (b) deep drive-in diffusion with source removed (graded); (c) linear approximation to the graded junction.

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 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. 5.38), 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 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.

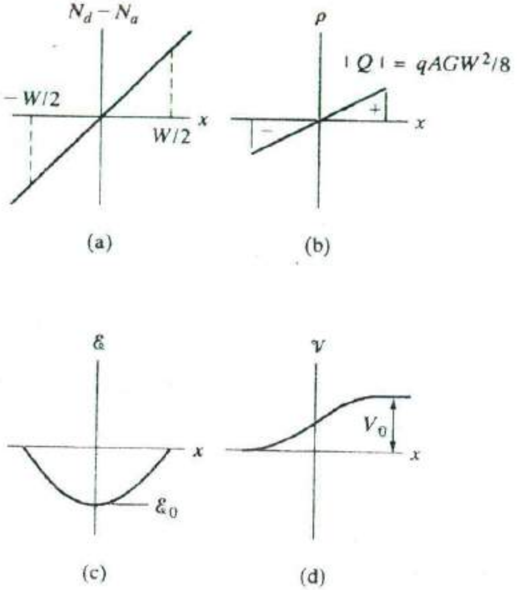


Figure 5–39 Properties of the graded junction transition region: (a) net impurity profile; (b) net charge distribution; (c) electric field; (d) electrostatic potential.

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 4.3V for Al and 4.8V 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 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

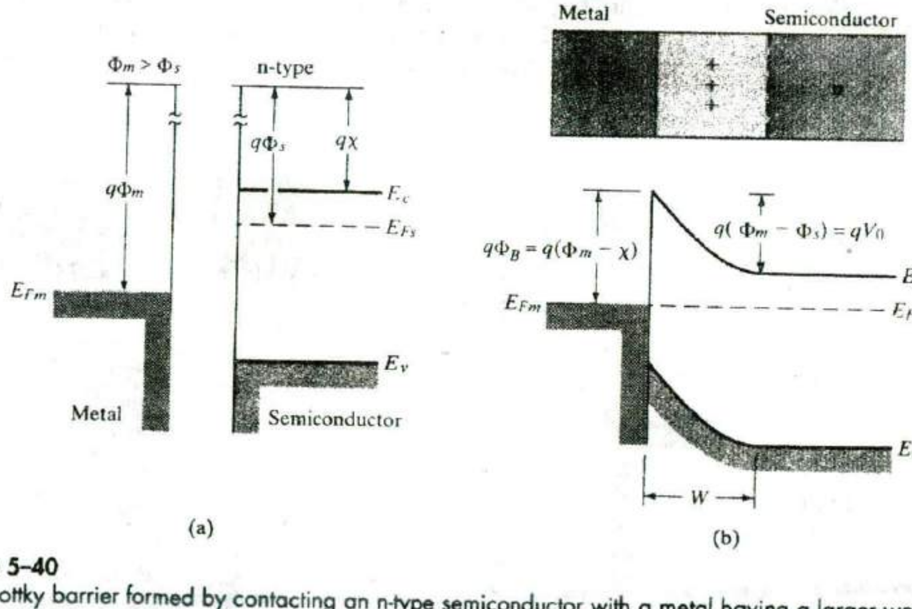


Figure 5–40
A Schottky barrier formed by contacting an n-type semiconductor with a metal having a larger work function: (a) band diagrams for the metal and the semiconductor before joining; (b) equilibrium band diagram for the junction.

[17] While the properties of the Schottky barrier depletion region are similar to the p^+-n, it is clear that the analogy does not include forward-bias hole injection, which is dominant for the p^+-n but not for the contact of Fig. 5–40.

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 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

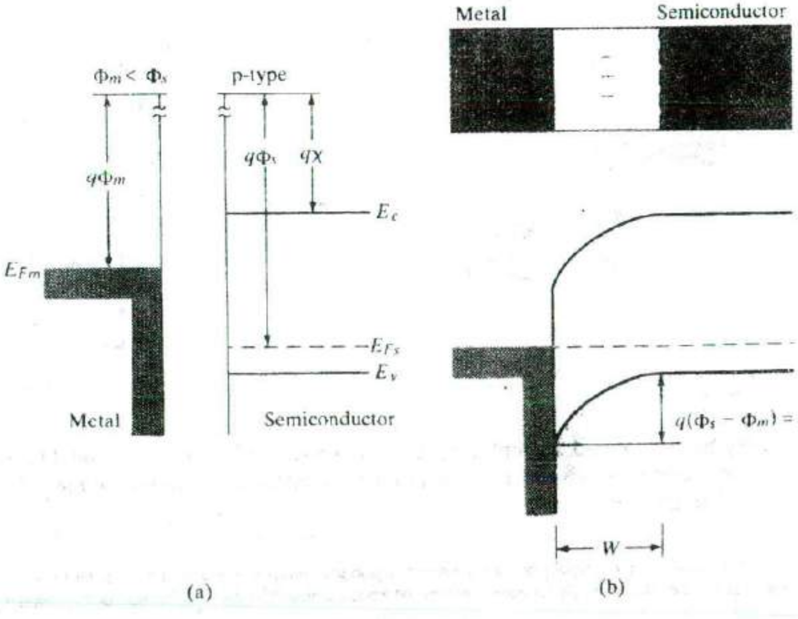


Figure 5–41 Schottky barrier between a p-type semiconductor and a metal having a smaller work function: (a) band diagrams before joining; (b) band diagram for the junction at equilibrium.

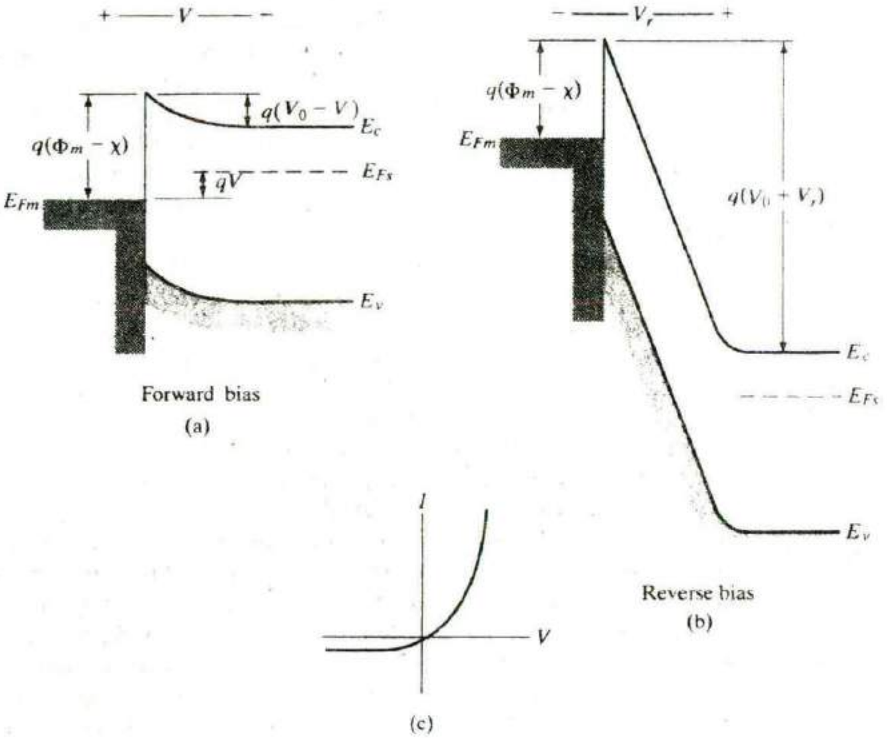


Figure 5–42
Effects of forward and reverse bias on the junction of Fig. 5–40: (a) forward bias; (b) reverse bias; (c) typical current-voltage characteristic.

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 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 *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 dis-

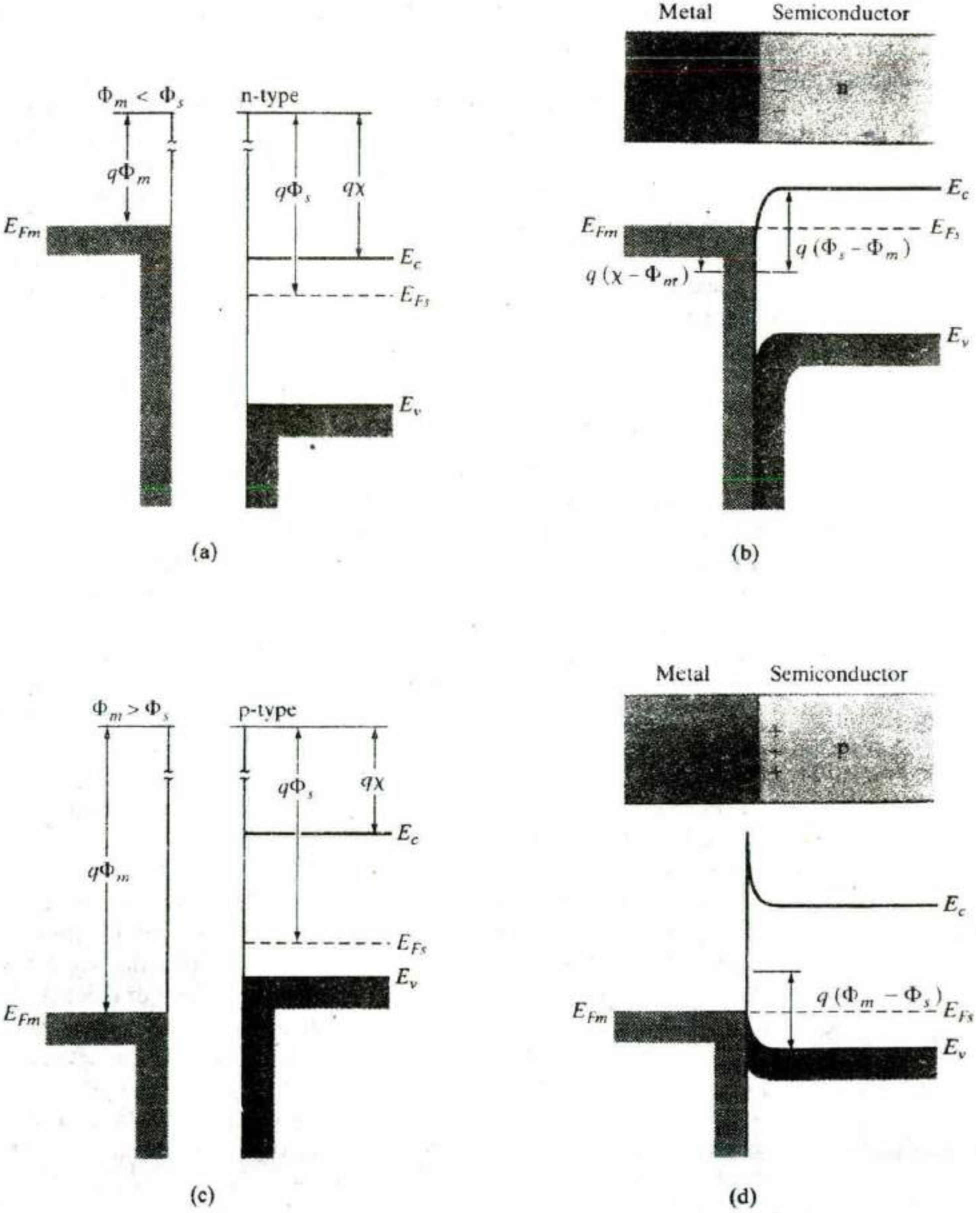


Figure 5–43
Ohmic metal–semiconductor contacts: (a) $\Phi_m < \Phi_s$ for an n-type semiconductor, and (b) the equilibrium band diagram for the junction; (c) $\Phi_m > \Phi_s$ for a p-type semiconductor, and (d) the junction at equilibrium.

cussed 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.

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 $0.7 \sim 0.9$ 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 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

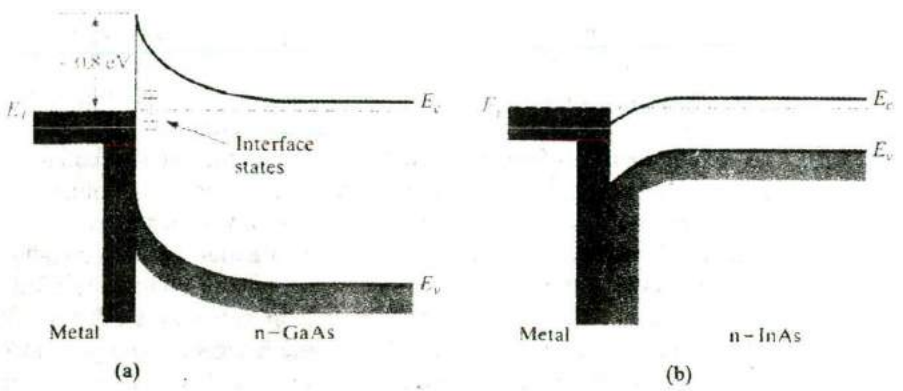


Figure 5–44
Fermi level pinning by interface states in compound semiconductors: (a) E_F is pinned near $E_C - 0.8$ eV in n-type GaAs, regardless of the choice of metal; (b) E_F is pinned above E_C in n-type InAs, providing an excellent ohmic contact.

the case of Pt, heat treatment results in a platinum silicide layer, which provides a reliable Schottky barrier with $\Phi_B \simeq 0.85$ V on n-type Si.

A full treatment of Schottky barrier diodes results in a forward current equation of the form

$$I = ABT^2 e^{-q\Phi_B/kT} e^{qV/\mathbf{n}kT} \qquad (5\text{–}81)$$

where B is a constant containing parameters of the junction properties and $\mathbf{n}$ 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 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 bandgap and the electron affinity which depend on the semiconductor material but not on the doping, and the workfunction which depends on the semiconductor as well as the doping. The electron affinity and workfunction 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 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.

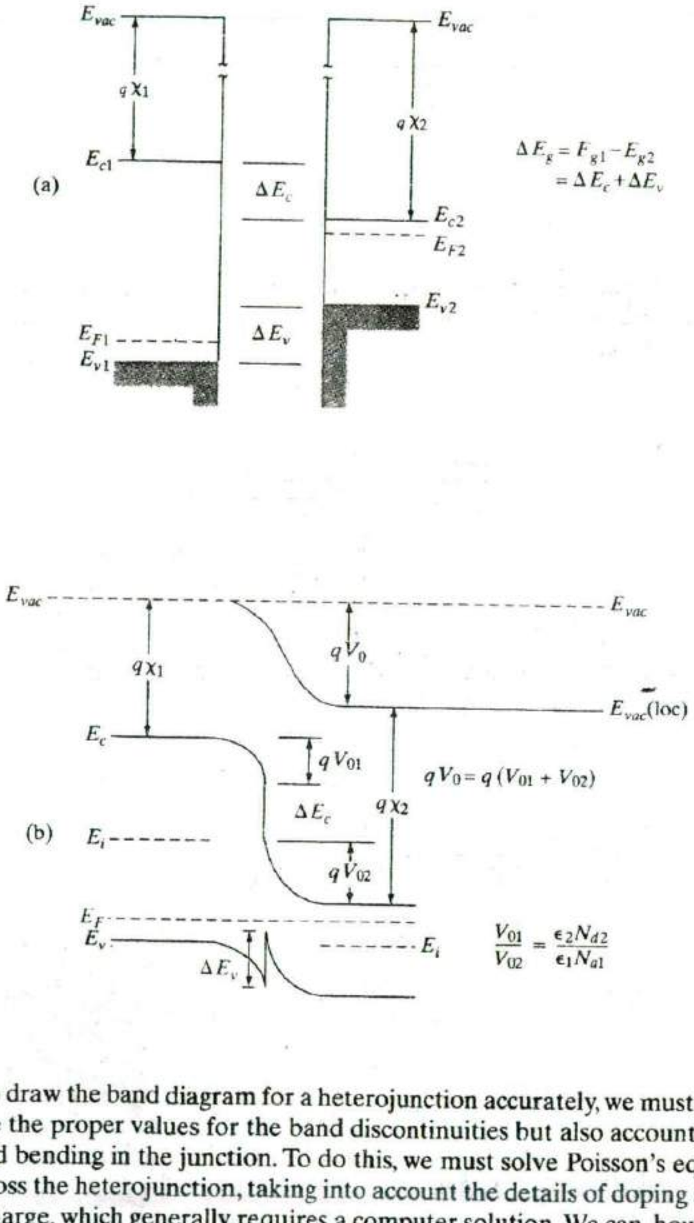
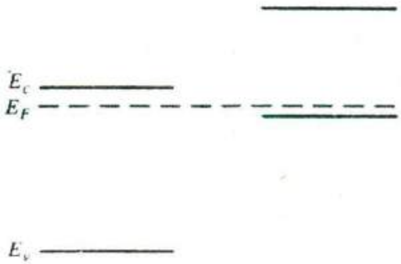


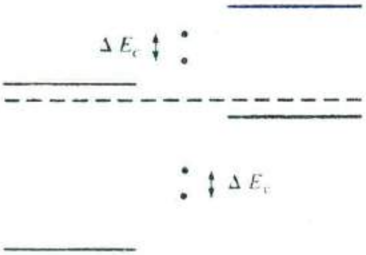
Figure 5–45 An ideal heterojunction between a p-type, wide band gap semiconductor an n-type narrower band gap semiconductor: (a) band diagrams before joining; (b) band discontinuities and band bending at equilibrium.

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:

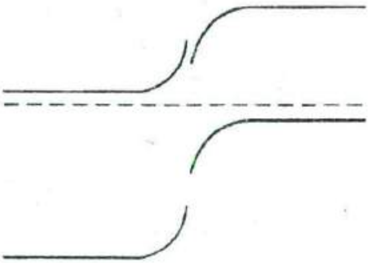
1. Align the Fermi level with the two semiconductor bands separated. Leave space for the transition region.



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.



3. Connect the conduction band and valence band regions, keeping the band gap constant in each material.

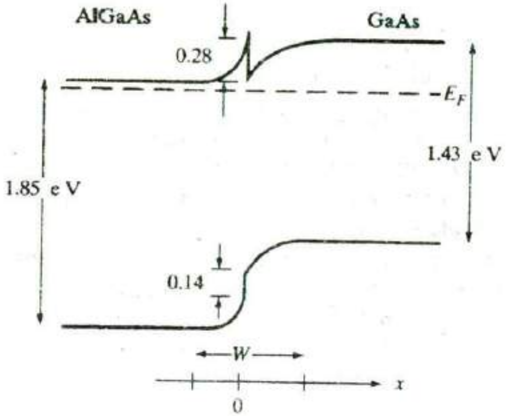


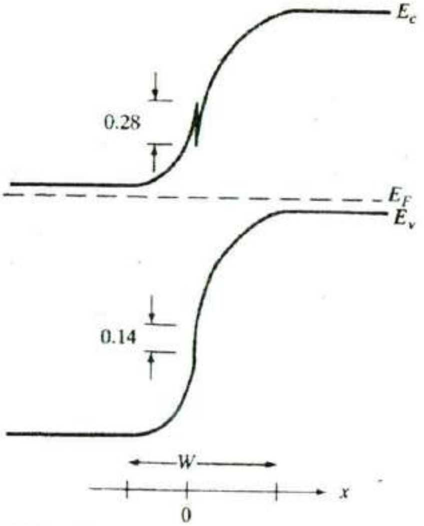
Steps 2 and 3 of this procedure are where the exact band bending is important and must be obtained by solving Poisson's equation. In step 2 we must use the band offset values ΔE_c and ΔE_v for the specific pair of semiconductors in the heterojunction.

EXAMPLE 5–6 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 capital N or P to designate the wide band gap material.

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 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.

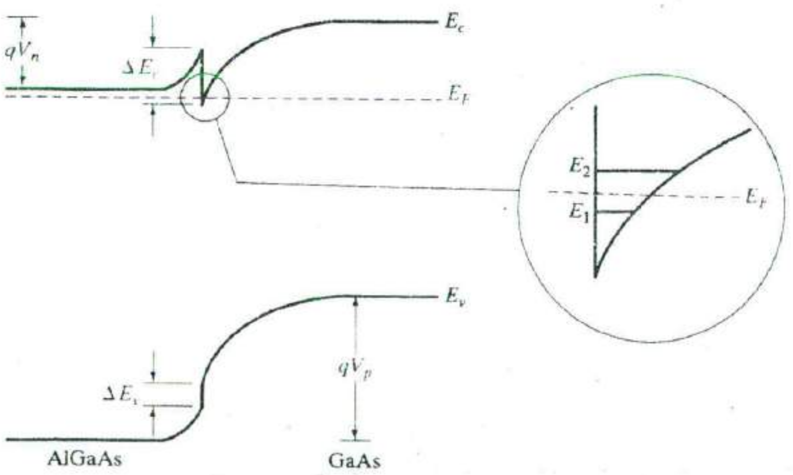


Figure 5–46 A heterojunction between N^+-AlGaAs and lightly doped GaAs, illustrating the potential well for electrons formed in the GaAs conduction band. If this well is sufficiently thin, discrete states (such as E_1 and E_2) are formed, as discussed in Section 2.4.3.

PROBLEMS

5.1 Design an oxide mask to block P diffusion in Si at 1000°C for 30 minutes using a design criterion that the mask thickness should be eight times the diffusion length. If we grow this oxide using a wet oxidation process at 1100°C, how long must we do the oxidation? Calculate the total number of Si atoms that are consumed from the wafer in the process, for a 200 mm diameter wafer.

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:

$$N(x, t) = \frac{N_s}{\sqrt{\pi D t}} e^{-(x/2\sqrt{Dt})^2}$$

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 $N_d = 5 \times 10^{16}\text{cm}^{-3}$) 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.

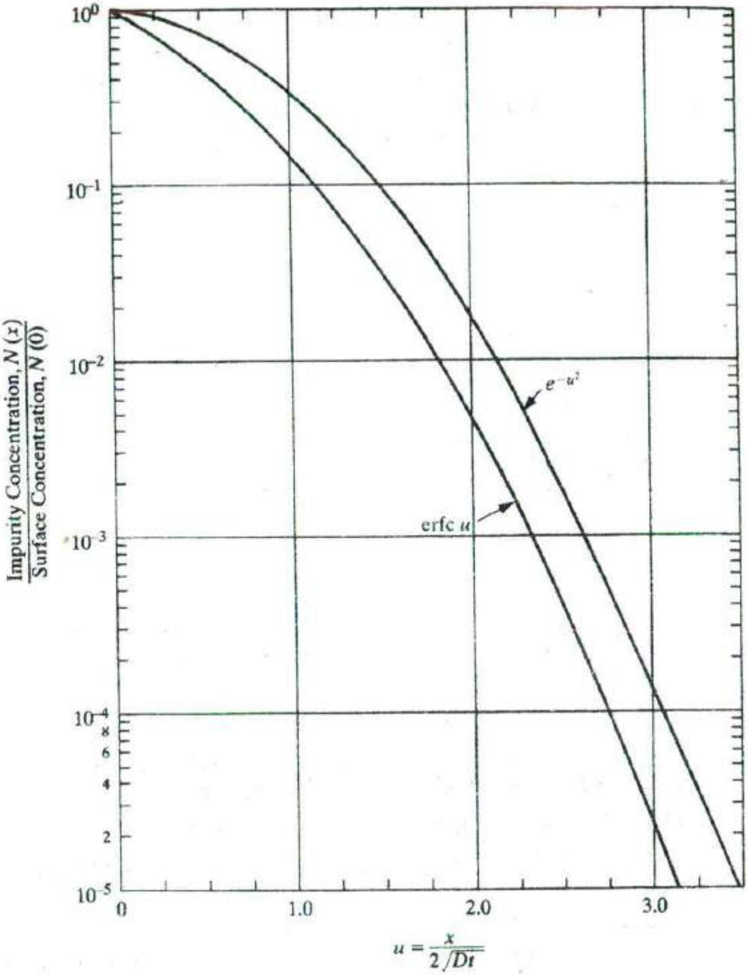
(b) Plot $N_a(x)$ after the diffusion, assuming that B is deposited in a thin layer on the surface prior to diffusion ($N_s = 5 \times 10^{13}$ cm^{-2}), and no additional B atoms are available during the diffusion. Locate the junction for this case.

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 A 900 nm oxide is grown on (100) Si in wet oxygen at 1100°C. 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.

Figure P5–2



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 We want to implant 5×10^{14} cm^{-2} B into Si at an average depth of 0.5 μm. We have an implanter which has a *maximum* acceleration voltage of 150 kV. How can we achieve this profile if we have singly and doubly charged B in the machine? Suppose the doubly ionized beam current is 0.1 mA, how long will the implant take if the scan area is 100 cm^2? By doing clever ion implanter source design. Dr. Boron Maximus has increased the beam current by a factor of 1000. From a dose uniformity point of view is this good or bad?

5.6 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 equations in Prob. 5.2.

5.7 We are interested in patterning the structure shown in Fig. P5-7. Design the mask aligner optics in terms of numerical aperture of the lens and the wavelength of the source.

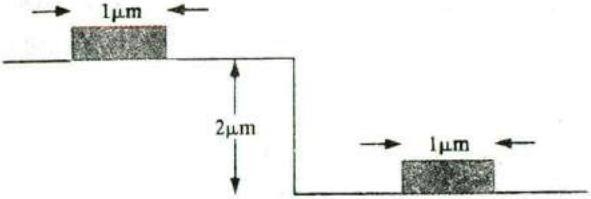


Figure P5–7

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 $N_a = 10^{18}$ 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).

5.10 The junction described in Prob. 5.9 has a circular cross section with 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.

5.11 The electron injection efficiency of a junction is I_n/I at $x_p = 0$.

(a) Assuming 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?

5.12 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.

(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 An abrupt Si p-n junction has the following properties at 300 K:

p side	n side	$A = 10^{-4}$ cm^2
$N_a = 10^{17}$ cm^{-3}	$N_d = 10^{15}$	
$\tau_n = 0.1$ μs	$\tau_p = 10$ μs	
$\mu_p = 200$ cm^2/V-s	$\mu_n = 1300$	
$\mu_n = 700$	$\mu_p = 450$	

Draw an equilibrium band diagram for this junction, including numerical values for the Fermi level position relative to the intrinsic level on each side. Find the contact potential from the diagram and check your answer with the analytical expression for V_0.

5.17 A long p$^+$-n diode is forward biased with current I flowing. 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 the voltage is always $\gg kT/q$, relate the final junction voltage (at $t = \infty$) to the initial voltage (before $t = 0$).

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-s, $\mu_p = 450$ cm^2/V-s, $\tau_n = \tau_p = 2.5$ μs. 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 An n$^+$-p junction with a long p-region has the following properies: $N_a = 10^{16}$ cm^{-3}; $D_p = 13$ cm^2/s; $\mu_n = 1000$ cm^2/V-s; $\tau_n = 2$μs; $n_i = 10^{10}$ cm^{-3}. If we apply 0.7 V forward bias to the junction at 300 K, what is the electric field in the p-region far from the junction?

5.21 For the diode in Problem 5.16, draw the band diagram qualitatively under forward and reverse bias showing the quasi-Fermi levels.

5.22 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

5.23 The junction of problem (5.16) is forward biased by 0.5 V. What is the forward current? What is the current at a reverse bias of –0.5 V?

5.24 In the junction of problem (5.16), what is the total depletion capacitance at –4 V?

5.25 A p^+-n Si diode ($V_0 = 0.956$ V) has a donor doping of 10^{17} cm^{-3} and an n-region width = 1μm. Does it break down by avalanche or punchthrough?

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

$$L_D = \left[\frac{\epsilon_s kT}{q^2 N_d}\right]^{1/2} \quad \text{on the n side.}$$

Calculate the Debye length on the n side for Si junctions having $N_a = 10^{18}$ cm^{-3} on the p-side and $N_d = 10^{14}, 10^{16}$, and 10^{18} 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 ($N_a = N_d = 10^{17}$ 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 We wish to design a p$^+$-n diode such that the avalanche breakdown and punchthrough both occur at 15 V. Assume the relative dielectric constant of the semiconductor is 10, V_0 is 0.5 V, and the breakdown field is 1 MV/cm. Determine the width and doping of the n-region.

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 long p$^+$-n diode is forward biased with current I flowing. The current is suddenly doubled at $t = 0$.

Assume that the stored charge in the n region can be represented by an exponential at each instant, for simplicity. Write the expression for the instantaneous current as a sum of recombination current and current due to changes in the stored charge. Using proper boundary conditions, solve this equation for the instantaneous hole distribution and find the expression for the instantaneous junction voltage.

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 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 to ensure avalanche breakdown rather than punchthrough?

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

$$\delta p(x_n) = \frac{\Delta p_n[e^{(l-x_n)/L_p} - e^{(x_n-l)/L_p}]}{e^{l/L_p} - e^{-l/L_p}}$$

(b) Show that the current in the diode is

$$I = \left(\frac{qAD_p p_n}{L_p}\,\text{ctnh}\,\frac{l}{L_p}\right)(e^{qV/kT} - 1)$$

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

$$I(\text{ohmic contact}) = \left(\frac{qAD_p p_n}{L_p}\,\text{csch}\,\frac{l}{L_p}\right)(e^{qV/kT} - 1)$$

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 $\mathscr{E}_0 = -qGW^{(\mathbf{m}+1)}/\epsilon(\mathbf{m} + 1)$.

(b) Find the expression for $\mathscr{E}(x)$, and use the result to obtain $V_0 - V = qGW^{(\mathbf{m}+2)}/\epsilon(\mathbf{m} + 2)$.

(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

$$C_j = A\left[\frac{qG\epsilon^{(m+1)}}{(\mathbf{m} + 2)(V_0 - V)}\right]^{1/(\mathbf{m}+2)}$$

5.38 Assume a linearly graded junction as in Fig. 5–39, with a doping distribution described by Eq. (5–77). The doping is symmetrical, so that $x_{p0} = x_{n0} = W/2$.

(a) Integrate Eq. (5–78) to show that

$$\mathscr{E}(x) = \frac{q}{2\epsilon}G\left[x^2 - \left(\frac{W}{2}\right)^2\right]$$

(b) Show that the width of the depletion region is

$$W = \left[\frac{12\epsilon(V_0 - V)}{qG}\right]^{1/3}$$

(c) Show that the junction capacitance is

$$C_j = A\left[\frac{qG\epsilon^2}{12(V_0 - V)}\right]^{1/3}$$

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) 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 vs. N_d.

(b) Plot the maximum electric field $\mathscr{E}_0$ vs. N_d for the junctions described in (a).

5.41 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.

5.42 What is the conductivity of a piece of Ge ($n_i = 2.5 \times 10^{13}$ cm^{-3}) doped with 5×10^{13} cm^{-3} donors and 2.5×10^{13} cm^{-3} acceptors? ($D_n = 100$ cm^2/s, $D_p = 50$ cm^2/s). If the electron affinity of Ge = 4.0 eV, and we put down a metal electrode with work function = 4.5 eV, what is the work function difference? Do you expect this to be a Schottky barrier or an ohmic contact?

READING LIST

Arienzo, M., S. S. Iyer, B. S. Meyerson, G. L. Patton, and J. M. C. Stork. "Si-Ge Alloys: Growth, Properties, and Applications." *Solid Surface Science* 48/49 (May 1991): 377–86.

Chang, L. L., and L. Esaki. "Semiconductor Quantum Heterostructures." *Physics Today* 45 (October 1992): 36–43.

Ghandhi, S. K. *VLSI Fabrication Principles*, 2nd ed. New York: Wiley, 1994.

Hummel, R. E. *Electronic Properties of Materials*, 2nd ed. Berlin: Springer-Verlag, 1993.

Jaeger, R. C. *Modular Series on Solid State Devices: Vol V. Introduction to Microelectronic Fabrication*. Reading, MA: Addison-Wesley, 1988.

Karunasiri, R. P. G., and K. L. Wang. "Quantum Devices Using SiGe/Si Heterostructures." *Journal of Vacuum Science and Technology* B 9 (July–August 1992): 2064–71.

Li, S. S. *Semiconductor Physical Electronics*. New York: Plenum Press, 1993.

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.

Neudeck, G. W. *Modular Series on Solid State Devices: Vol II. The PN Junction Diode.* Reading, MA: Addison-Wesley, 1983.

Shockley, W. "The Theory of P-N Junctions in Semiconductors and P-N Junction Transistors." *Bell Syst. Tech. J.* 28 (1949), 435.

Ryssell, H., and I. Ruge. *Ion-Implantation.* Chichester: Wiley, 1986.

Tove, P. A. "Formation and Characterization of Metal–Semiconductor Junctions." *Vacuum* 36 (October 1986): 659–67.

Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

Wolf, S., and R. N. Tauber. *Silicon Processing for the VLSI Era.* Sunset Beach, CA: Lattice Press, 1986.

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

Chapter 6

Field-Effect Transistors

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 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

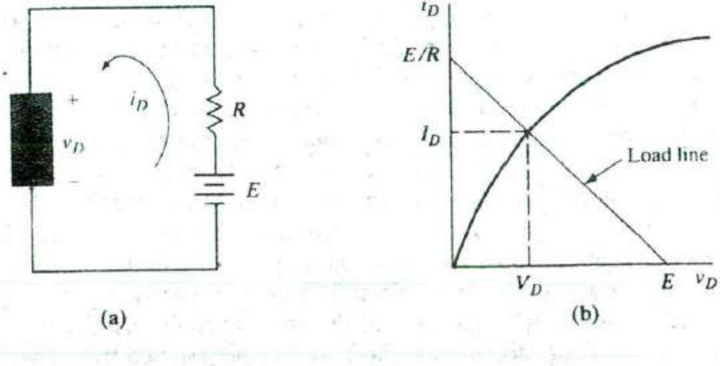


Figure 6–1
A two-terminal nonlinear device: (a) biasing circuit; (b) I–V characteristic and load line.

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]

$$E = i_D R + v_D \tag{6-1}$$

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 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*.

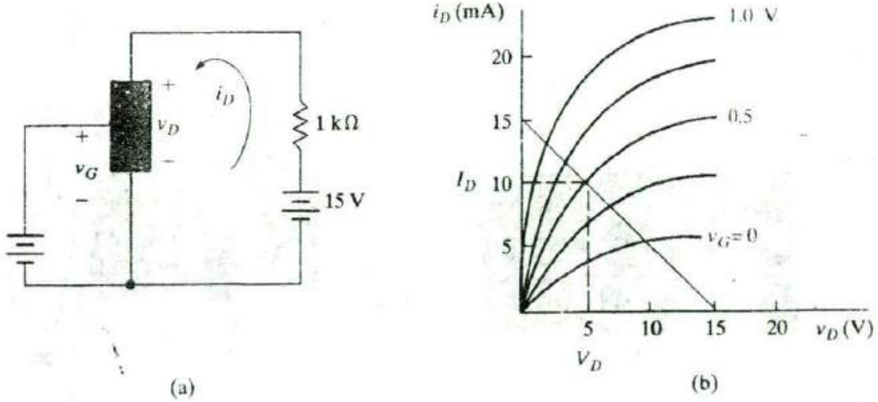


Figure 6–2
A three-terminal nonlinear device that can be controlled by the voltage at the third terminal v_G: (a) biasing circuit; (b) I–V characteristic and load line. If $V_G = 0.5$ V, the d-c values of I_D and V_D are as shown by the dashed lines.

[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.

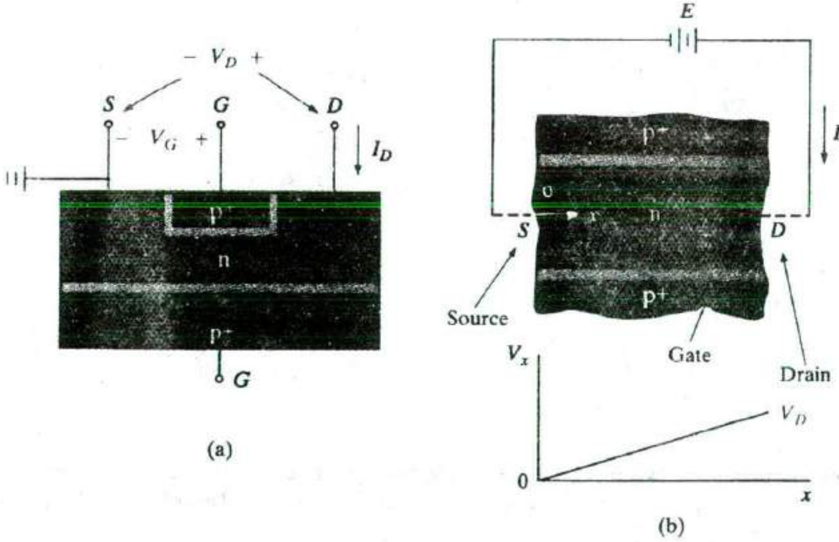
6.1.2 Amplification and Switching

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 equai 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



(a)

(b)

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.

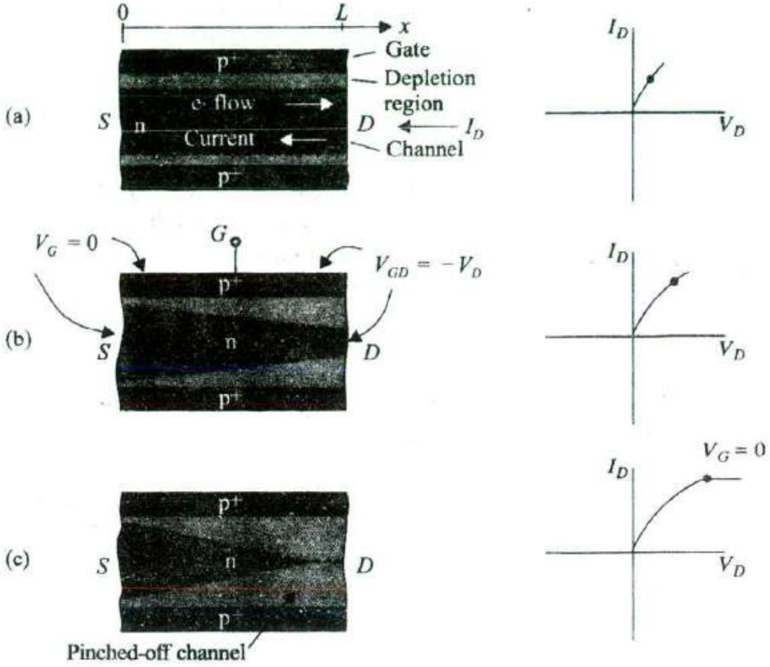
and the channel causes the depletion regions to intrude into the n material, and therefore the effective width of the channel can be restricted. Since the resistivity of the channel region is fixed by its doping, the channel resistance varies with changes in the effective cross-sectional area. By analogy, 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 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 ($V_{GD} = - V_D$) 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.

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.



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 calculate the current at the critical pinch-off voltage and assume there is no further increase in current as V_D is increased.

[2] *Saturation* is used by device engineers in more different contexts than any other word. We have discussed velocity saturation, reverse saturation current of a junction, and now the saturation of FET characteristics. In Chapter 7, we will discuss saturation of a BJT. The student has probably also reached saturation by now in trying to absorb these various meanings.

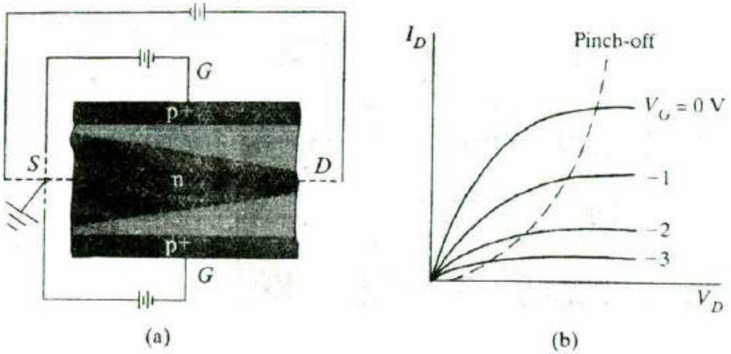


Figure 6–5 Effects of a negative gate bias: (a) increase of depletion region widths with V_G negative; (b) family of current–voltage curves for the channels as V_G is varied.

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):

$$W(x = L) = \left[\frac{2\epsilon(-V_{GD})}{qN_d}\right]^{1/2} \quad (V_{GD}\text{ negative}) \qquad (6\text{–}2)$$

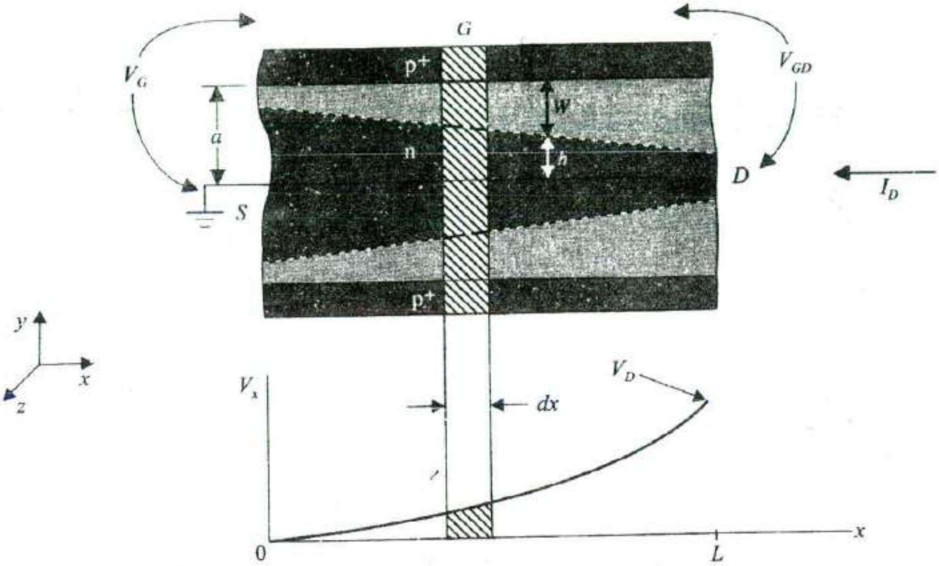


Figure 6–6
Simplified diagram of the channel with definitions of dimensions and differential volume for calculations.

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. 6.2.

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 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

$$\boxed{I_D = G_0 V_P\left[\frac{V_D}{V_P} + \frac{2}{3}\left(-\frac{V_G}{V_P}\right)^{3/2} - \frac{2}{3}\left(\frac{V_D - V_G}{V_P}\right)^{3/2}\right]} \tag{6–9}$$

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

$$\begin{aligned} I_D(\text{sat.}) &= G_0 V_P\left[\frac{V_D}{V_P} + \frac{2}{3}\left(-\frac{V_G}{V_P}\right)^{3/2} - \frac{2}{3}\right] \\ &= G_0 V_P\left[\frac{V_G}{V_P} + \frac{2}{3}\left(+\frac{V_G}{V_P}\right)^{3/2} + \frac{1}{3}\right] \end{aligned} \tag{6–10}$$

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

$$g_m(\text{sat.}) = \frac{\partial I_D(sat.)}{\partial V_G} = G_0\left[1 - \left(-\frac{V_G}{V_P}\right)^{1/2}\right] \tag{6–11}$$

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:

$$I_D(\text{sat.}) \simeq I_{DSS}\left(1 + \frac{V_G}{V_P}\right)^2, \quad (V_G\ negative) \tag{6–12}$$

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 Sec. 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. 6–4(c) suggests. In short-channel devices this effect can cause I_D to increase beyond pinch-off, since L appears in the denominator 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 (~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 $L \lesssim 0.25$ 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.

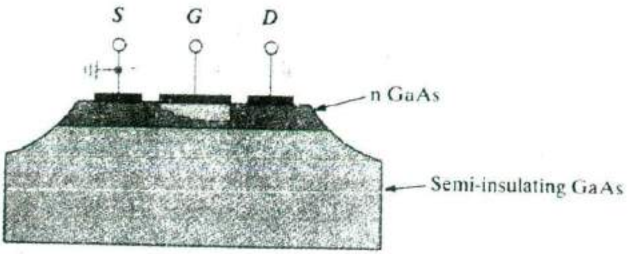


Figure 6–7 GaAs MESFET formed on an n-type GaAs layer grown epitaxially on a semi-insulating substrate. Common metals for the Schottky gate in GaAs are Al or alloys of Ti, W, and Au. The ohmic source and drain contacts may be an alloy of Au and Ge. In this example the device is isolated from others on 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.

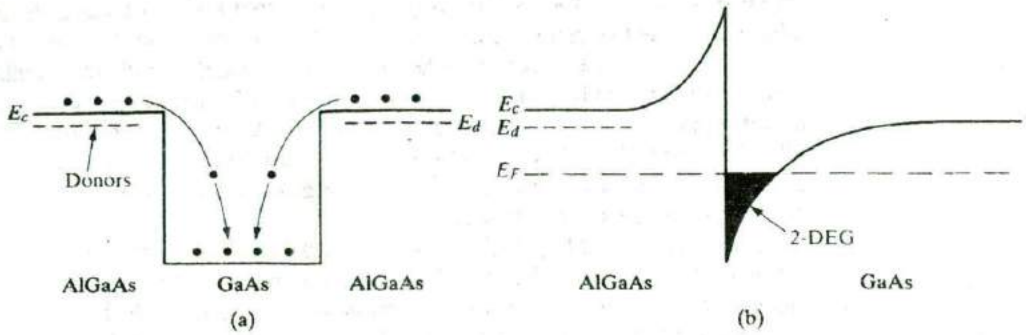


Figure 6–8
(a) Simplified view of modulation doping, showing only the conduction band. Electrons in the donor-doped AlGaAs fall into the GaAs potential well and become trapped. As a result, the undoped GaAs becomes n-type, without the scattering by ionized donors which is typical of bulk n-type material. (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 ($\sim 10^{12}\ \text{cm}^{-2}$) 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 $< 1\ \mu m$). 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

$$v_d = \frac{\mu\mathscr{E}}{1 + \mu\mathscr{E}/v_s} \qquad (6\text{–}13)$$

where μ is the low-field mobility. These two approximations are shown in Fig. 6–9 (a). If we assume that the electrons passing through the channel drift with a constant saturation velocity v_s, the current takes a simple form

$$I_D = qnv_sA = qN_dv_sZh \qquad (6\text{–}14)$$

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. 6–9(b), 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. 6–5(b) for the long-channel constant-mobility case.

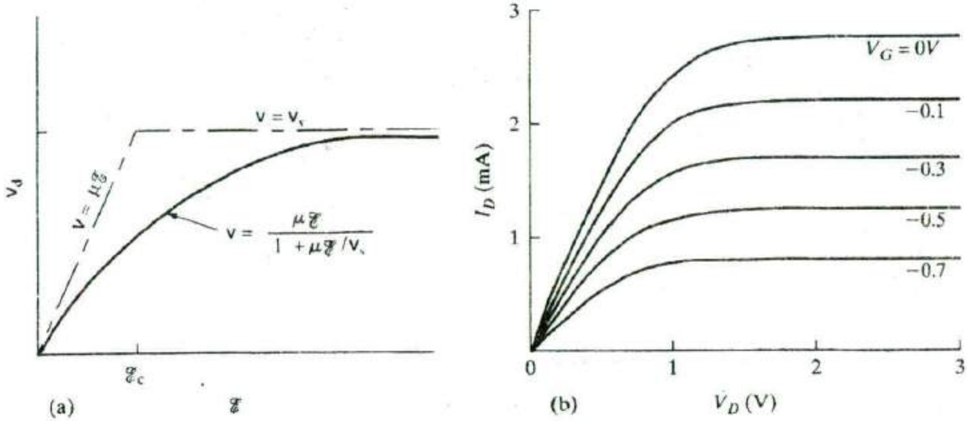


Figure 6–9
Effects of electron velocity saturation at high electric fields: (a) approximations to the saturation of drift velocity with increasing field; (b) drain current–voltage characteristics for the saturated velocity case, showing almost equally spaced curves with increasing gate voltage.

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.

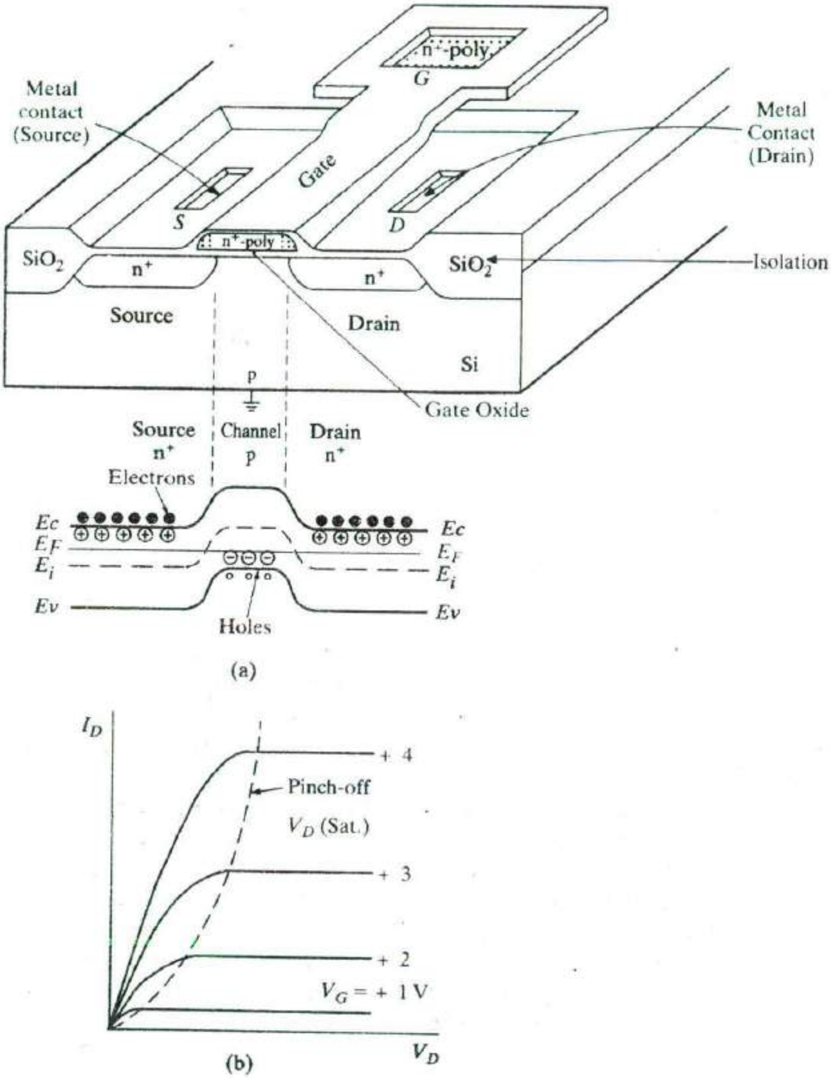


Figure 6–10 An enhancement-type n-channel MOSFET: (a) isometric view of device and equilibrium band diagram along channel; (b) drain current–voltage output characteristics as a function of gate voltage.

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 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, 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

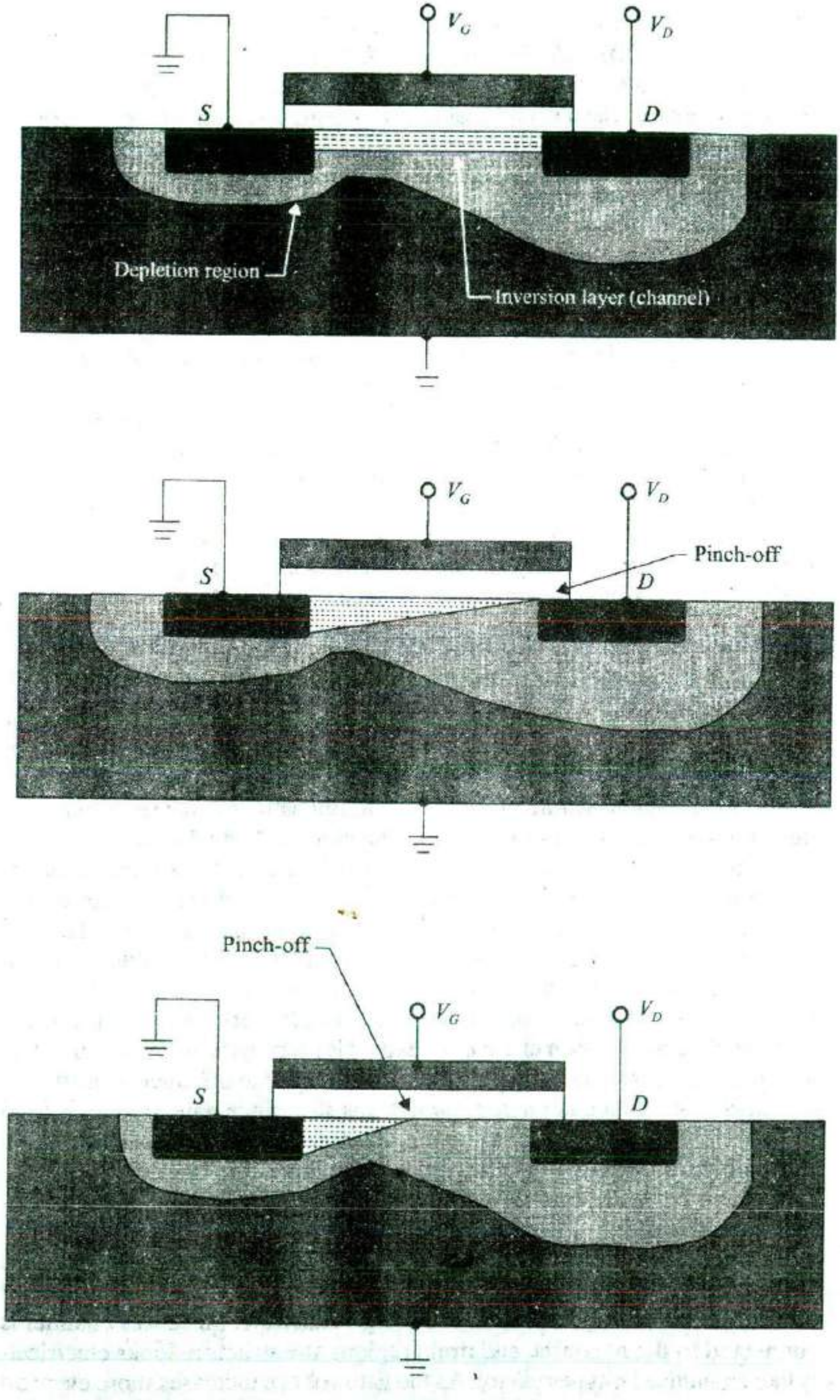


Figure 6–11 n-channel MOSFET cross-sections under different operating conditions: (a) linear region for $V_G > V_T$ and $V_D < (V_G - V_T)$; (b) onset of saturation at pinch-off, $V_G > V_T$ and $V_D = (V_G - V_T)$; (c) strong saturation, $V_G > V_T$ and $V_D > (V_G - V_T)$.

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_{DSAT}) 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 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 thermial 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 $q\phi_F$, 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.

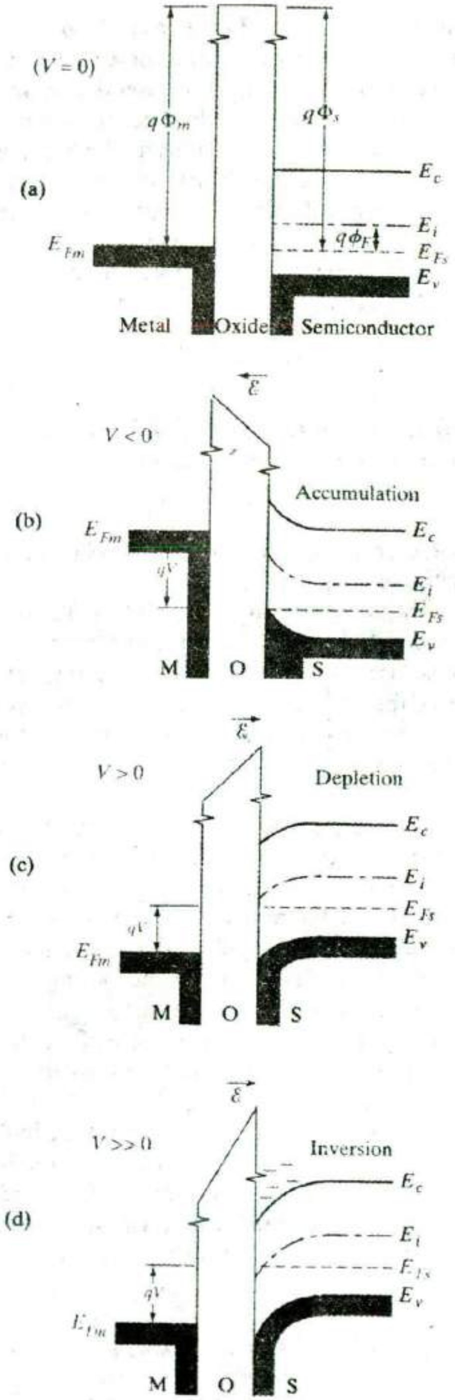


Figure 6–12 Band diagram for the ideal MOS structure at: (a) equilibrium; (b) negative voltage causes hole accumulation in the p-type semiconductor; (c) positive voltage depletes holes from the semiconductor surface: (d) a larger positive voltage causes inversion—a "n-type" layer at the semiconductor surface.

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 volt-

[5] Recall that an electrostatic potential diagram is drawn for positive test charges, in contrast with an electron energy diagram which is drawn for negative charges.

age can 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

$$\phi_s(\text{inv.}) = 2\phi_F = 2\frac{kT}{q}\ln\frac{N_a}{n_i} \tag{6–15}$$

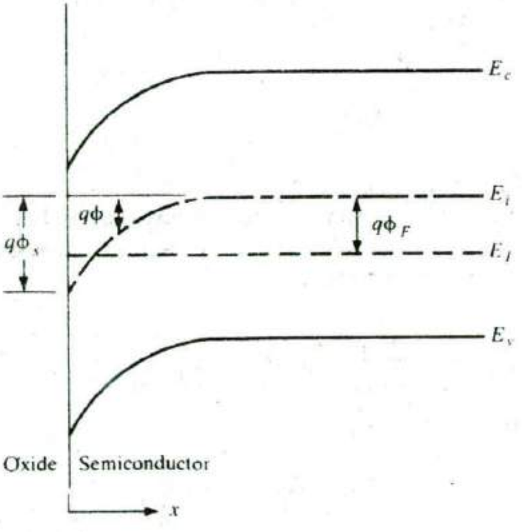


Figure 6–13 Bending of the semiconductor bands at the onset of strong inversion: the surface potential ϕ_s is twice the value of ϕ_F in the neutral p material.

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:

$$p_0 = n_i e^{q\phi_F/kT} = N_a^- \tag{6–18a}$$

$$p = p_0 e^{-q\phi/kT} \tag{6–18b}$$

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

$$\mathscr{E}^2 = \left(\frac{2kT\,p_0}{\epsilon_s}\right)\left[\left(e^{-\frac{q\phi}{kT}} + \frac{q\phi}{kT} - 1\right) + \frac{n_0}{p_0}\left(e^{\frac{q\phi}{kT}} - \frac{q\phi}{kT} - 1\right)\right] \tag{6–23}$$

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' 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 (flatband conditions), the net space charge is zero. This is because the fixed dopant charges are cancelled by the mobile carrier charges at flatband. 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 bandbending 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 bandbending 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

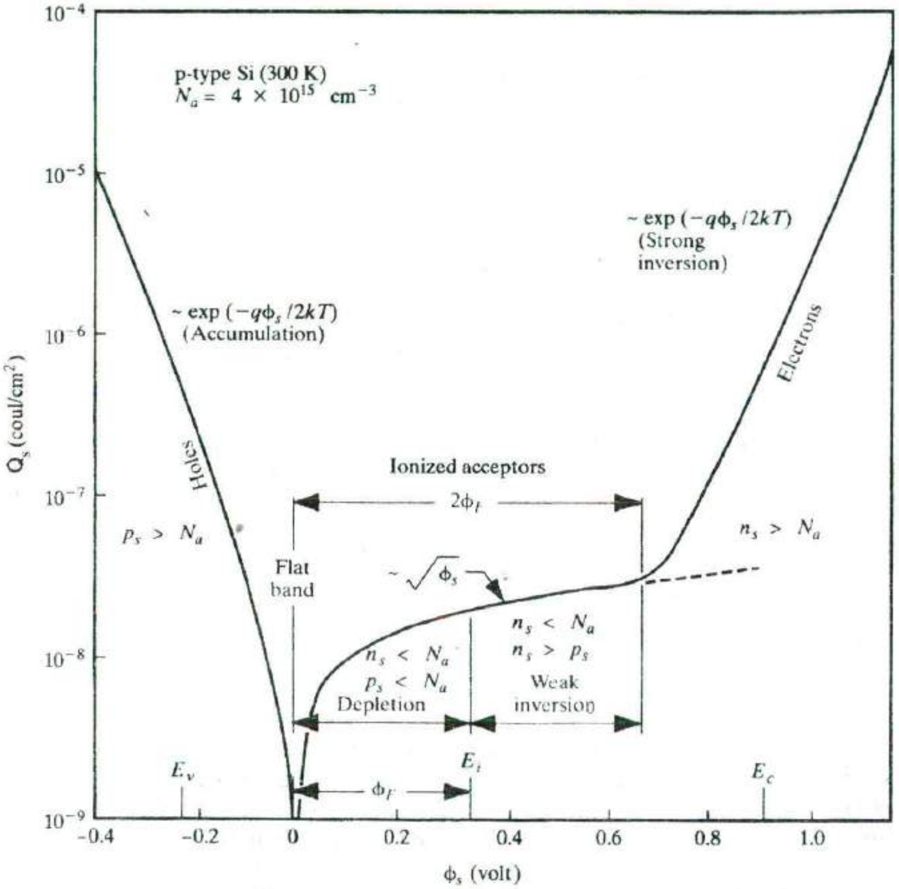


Figure 6–14
Variation of space-charge density in the semiconductor as a function of the surface potential ϕ_s for p-type silicon with $N_a = 4 \times 10^{15}$ cm^{-3} at room temperature. p_s and n_s are the hole and electron concentrations at the surface, ϕ_F is the potential difference between the Fermi level and the intrinsic level of the bulk. (Garrett and Brattain, Phys. Rev., 99, 376 (1995).)

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

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 2-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:

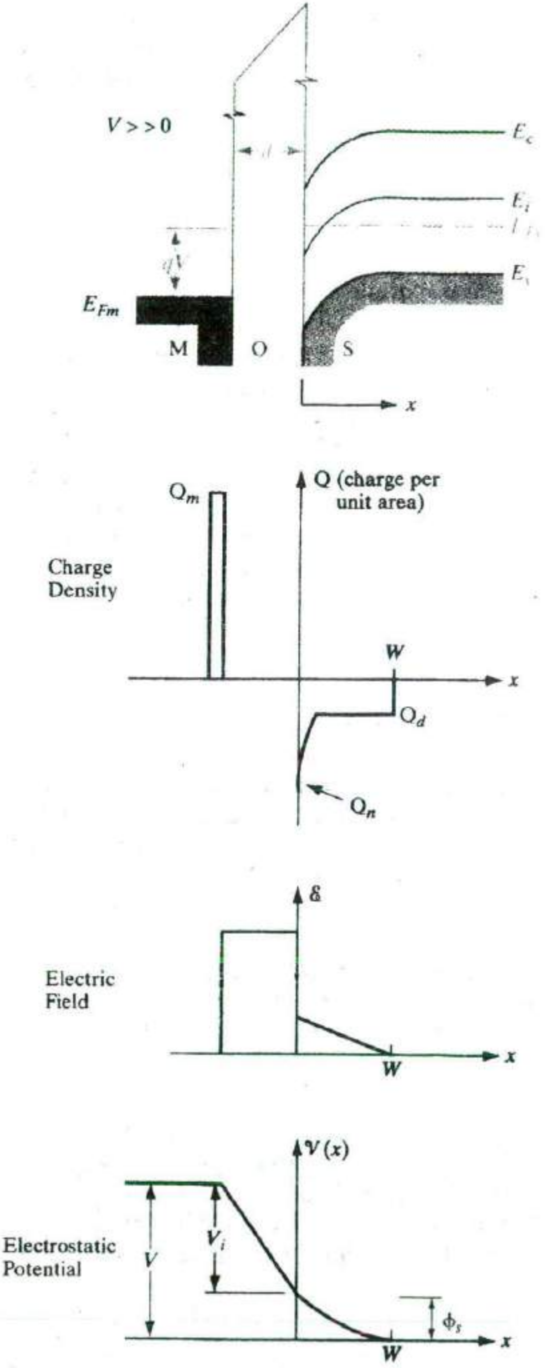
$$V_i = \frac{-Q_s d}{\epsilon_i} = \frac{-Q_s}{C_i} \tag{6–29}$$

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 n-channel, giving a positive V_i.

Using the depletion approximation, we can solve for W as a function of ϕ_s (Prob. 6.8). 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 carrying A throughout the discussion.

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.



$$W = \left[\frac{2\epsilon_s\phi_s}{qN_a}\right]^{1/2} \qquad (6\text{–}30)$$

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

$$W_m = \left[\frac{2\epsilon_s\phi_s(\text{inv.})}{qN_a}\right]^{1/2} = 2\left[\frac{\epsilon_s kT\ \ln(N_a/n_i)}{q^2N_a}\right]^{1/2} \qquad (6\text{–}31)$$

using Eq. (6–15). We know the quantities in this expression, so W_m can be calculated.

EXAMPLE 6–1

Find the maximum width of the depletion region for an ideal MOS capacitor on p-type Si with $N_a = 10^{16}\ \text{cm}^{-3}$.

SOLUTION

The relative dielectric constant of Si is 11.8 from Appendix III. We get ϕ_F from Eq. (6–15):

$$\phi_F = \frac{kT}{q}\ln\frac{N_a}{n_i} = 0.0259\ln\frac{10^{16}}{1.5\times10^{10}} = 0.347\ \text{V}$$

Thus

$$W_m = 2\sqrt{\frac{\epsilon_s\phi_F}{qN_a}} = 2\left[\frac{(11.8)(8.85\times10^{-14})(0.347)}{(1.6\times10^{-19})(10^{16})}\right]^{1/2}$$
$$= 3.01\times10^{-5}\ \text{cm} = 0.301\ \mu\text{m}$$

The charge per unit area in the depletion region Q_d at strong inversion is[7]

$$Q_d = -qN_aW_m = -2(\epsilon_s qN_a\phi_F)^{1/2} \qquad (6\text{–}32)$$

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

$$V_T = -\frac{Q_d}{C_i} + 2\phi_F \quad \textit{(ideal case)} \qquad (6\text{–}33)$$

[7] In the p-channel (n-type substrate) case, for which ϕ_F is negative, we use $Q_d = +qN_dW_m = 2(\epsilon_s qN_d|\phi_F|)^{1/2}$.

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 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

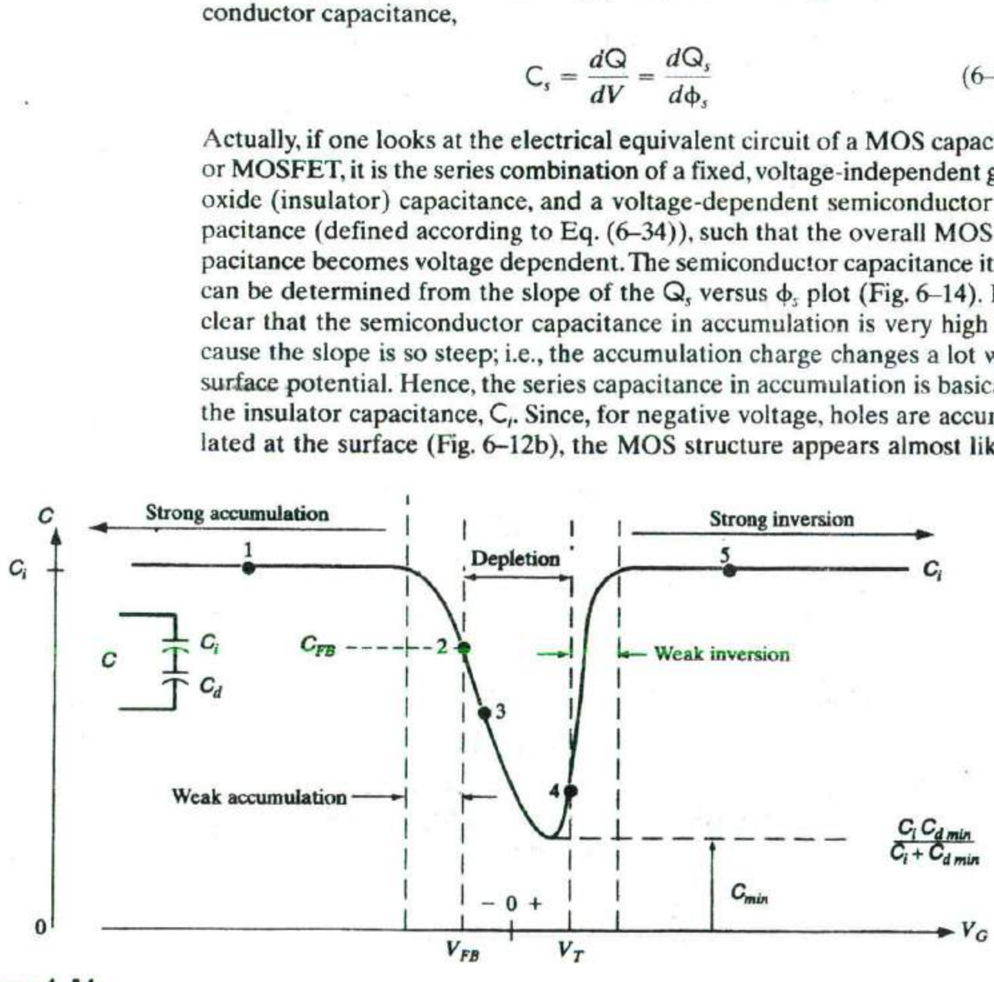


Figure 6–16
Capacitance–voltage relation for an n-channel (p-substrate) MOS capacitor. The dashed curve for $V > V_T$ is observed at high measurement frequencies. The flat band voltage V_{FB} will be discussed in Section 6.4.3. When the semiconductor is in depletion, the semiconductor capacitance C_s is denoted as C_d.

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:

$$C_d = \frac{\epsilon_s}{W} \tag{6–35}$$

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 (point 3) until finally 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 $\sim\sqrt{\phi_s}$, 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).

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.

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 (~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 bulk.

EXAMPLE 6–2 Using the conditions of Example 6–1 and a 100-Å-thick SiO_2 layer, we can calculate major points on the C–V curve of Fig. 6–16. The relative dielectric constant of SiO_2 is 3.9.

$$C_i = \frac{\epsilon_i}{d} = \frac{(3.9)(8.85 \times 10^{-14})}{10^{-6}} = 3.45 \times 10^{-7}\text{F/cm}^2$$

$$Q_d = -qN_aW_m = -(1.6 \times 10^{-19})(10^{16})(0.301 \times 10^{-4})$$

$$= -4.82 \times 10^{-8}\text{C/cm}^2$$

$$V_T = -\frac{Q_d}{C_i} + 2\phi_F = \frac{4.82 \times 10^{-8}}{34.5 \times 10^{-8}} + 2(0.347) = 0.834\text{ V}$$

At maximum depletion

$$C_d = \frac{\epsilon_s}{W_m} = \frac{(11.8)(8.85 \times 10^{-14})}{0.301 \times 10^{-4}} = 3.47 \times 10^{-8}\text{ F/cm}^2$$

$$C_{\min} = \frac{C_iC_d}{C_i + C_d} = \frac{34.5 \times 3.47}{34.5 + 3.47}10^{-8} \cong 3.15 \times 10^{-8}\text{ F/cm}^2$$

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).

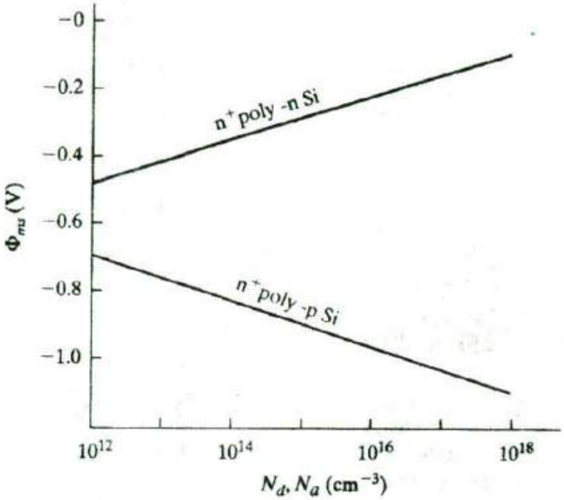
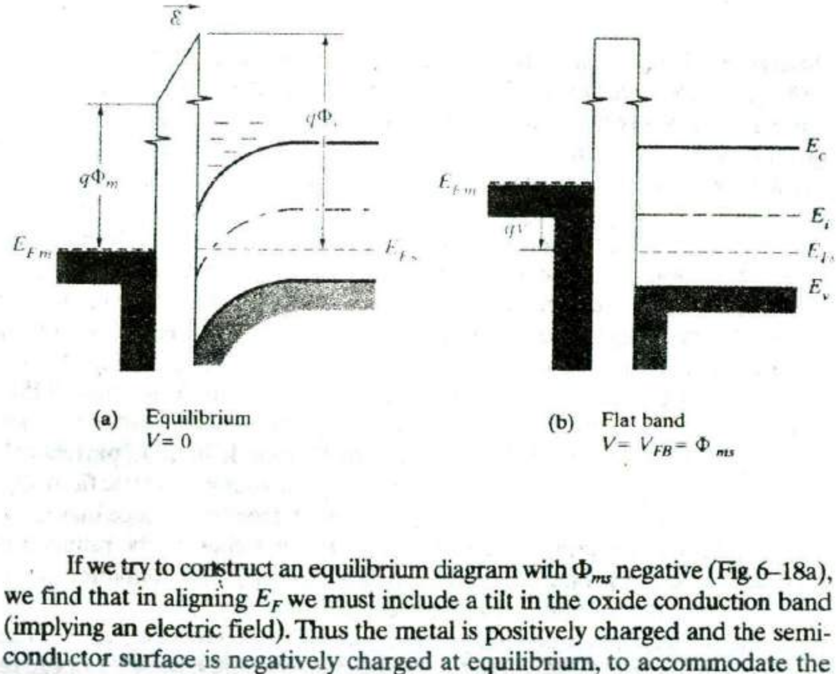


Figure 6–17 Variation of the metal–semiconductor work function potential difference Φ_{ms} with substrate doping concentration, for n^+ poly-Si.



(a) Equilibrium
$V = 0$

(b) Flat band
$V = V_{FB} = \Phi_{ms}$

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.

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 (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 ($V_{FB} = \Phi_{ms}$).

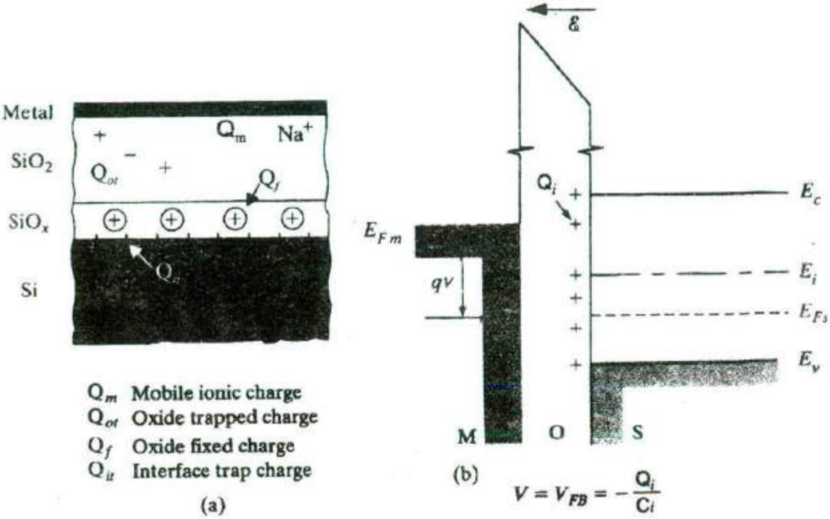


Figure 6–19 Effects of charges in the oxide and at the interface: (a) definitions of charge densities (C/cm^2) due to various sources; (b) representing these charges as an equivalent sheet of positive charge Q_i at the oxide-semiconductor interface. This positive charge induces an equivalent negative charge in the semiconductor, which requires a negative gate voltage to achieve the flat band condition.

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 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. 6.13). 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 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:

$$V_{FB} = \Phi_{ms} - \frac{Q_i}{C_i} \qquad (6\text{–}37)$$

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)

$$\boxed{V_T = \Phi_{ms} - \frac{Q_i}{C_i} - \frac{Q_d}{C_i} + 2\phi_F} \qquad (6\text{–}38)$$

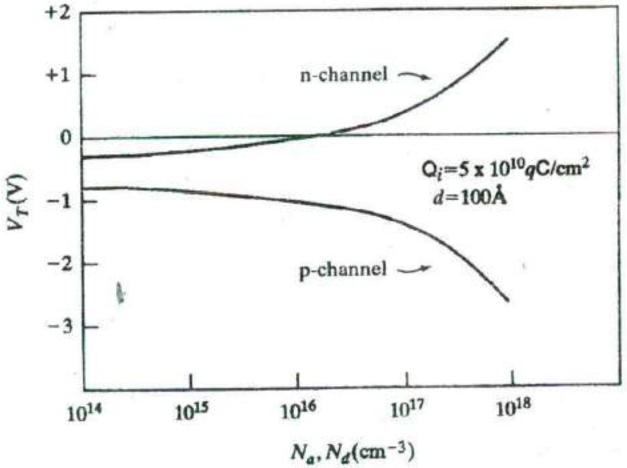
Thus the voltage required to create strong inversion must be large enough to first achieve the flat band condition (Φ_{ms} and Q_i/C_i terms), then accommodate the charge in the depletion region (Q_d/C_i), and finally to induce the inverted region ($2\phi_F$). This equation accounts for the dominant 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,

[8] It is important to remember that n-channel devices are made on p-type substrates, and p-channel devices have n-type substrates.

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.

	$V_T =$	Φ_{ms}	$-\frac{Q_i}{C_i}$	$-\frac{Q_d}{C_i}$	$+ 2\phi_F$
(a)		(−)	(−)	(+) n channel (−) p channel	(+) n channel (−) p channel

(b)



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).

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 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 6–3

We can calculate V_T for the MOS structure described in Examples 6–1 and 6–2, including the effects of flat band voltage. If n$^+$ polysilicon is used for the gate, Fig. 6–17 indicates $\Phi_{ms} = -0.95$ V for $N_a = 10^{16}$ cm^{-3}. Assuming an interface charge of $5 \times 10^{10}\, q(\text{C/cm}^2)$, we obtain

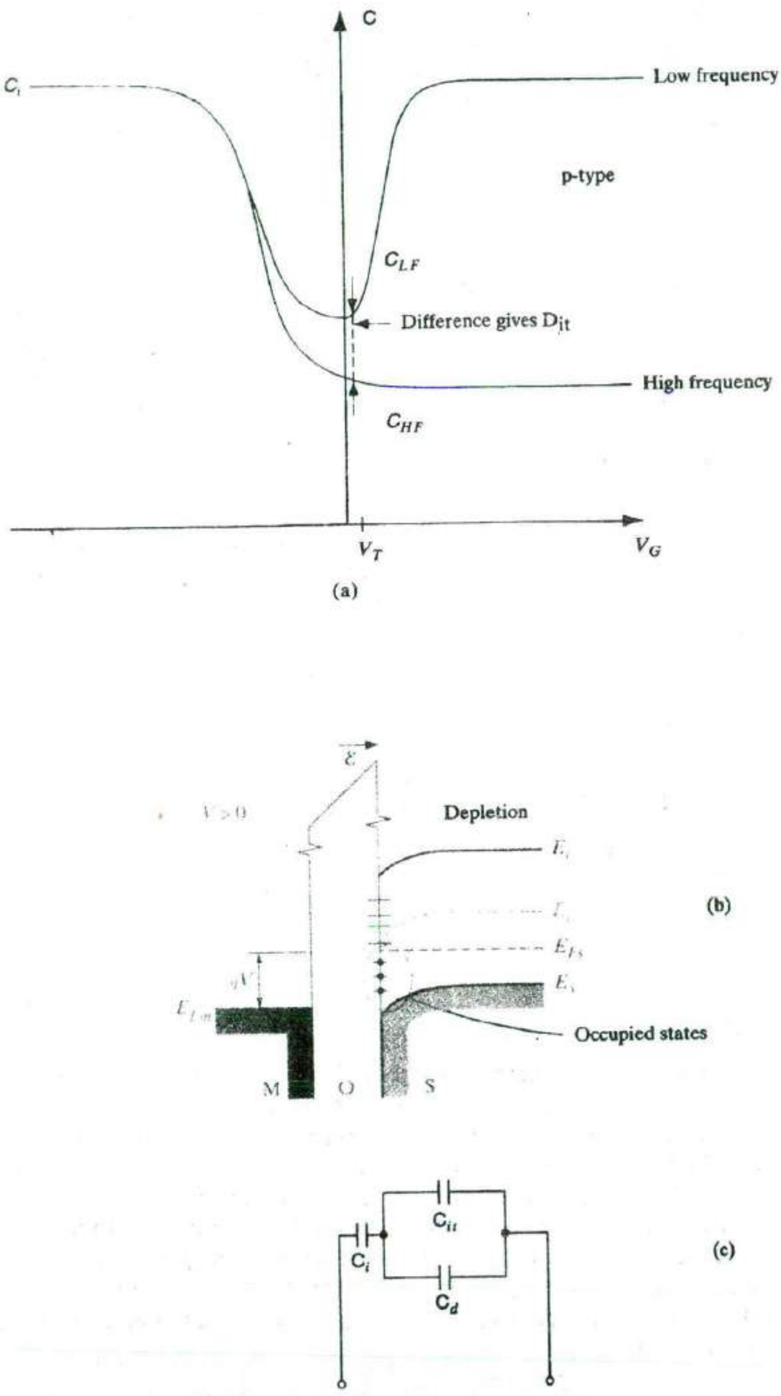
$$\begin{aligned}V_T &= \Phi_{ms} + 2\phi_F - \frac{1}{C_i}(Q_i + Q_d)\\ &= -0.95 + 0.694 - \frac{(5 \times 10^{10} \times 1.6 \times 10^{-19}) - 4.82 \times 10^{-8}}{34.5 \times 10^{-8}}\\ &= -0.14\ V\end{aligned}$$

This value corresponds to the $N_a = 10^{16}$ cm^{-3} point in Fig. 6–20 for the n-channel case.

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 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.

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 bandgap 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}).



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}.

$$N_a = 10^{[30.388 + 1.683 \log C_{dmin} - 0.03177(\log C_{dmin})^2]} \qquad (6\text{–}39)$$

where C_{dmin} is in F/cm^2.

Once the substrate doping is obtained, we can determine the flatband capacitance from it. It can be shown that the semiconductor capacitance at flatband C_{FB} (point 2 in Fig. 6–16) is determined from the Debye length capacitance

$$C_{debye} = \frac{\sqrt{2}\epsilon_s}{L_D} \qquad (6\text{–}40)$$

where the Debye length is dependent on doping as described in Eq. (6–25). The overall MOS flatband 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 states or traps in the bandgap 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 capcitance 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

$$D_{it} = \frac{1}{q}\left(\frac{C_iC_{LF}}{C_i - C_{LF}} - \frac{C_iC_{HF}}{C_i - C_{HF}}\right)\text{cm}^{-2}eV^{-1} \qquad (6\text{-}41)$$

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 flatband 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 flatband 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 bandbending, but induce an equal and opposite charge on the gate electrode. From the resulting *C–V*, the new flatband, V^-_{FB}, is determined. From the difference of the two flatband voltages, we can determine the mobile ion content using

$$Q_m = C_i(V^+_{FB} - V^-_{FB}) \qquad (6\text{–}42)$$

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

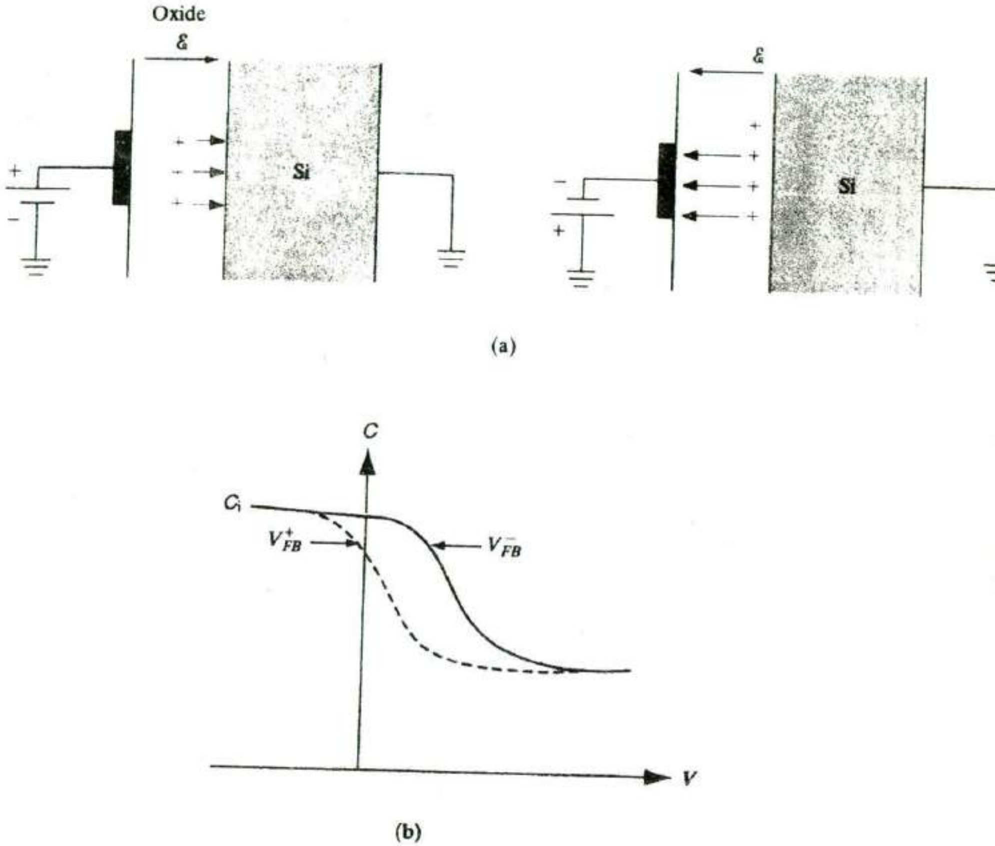
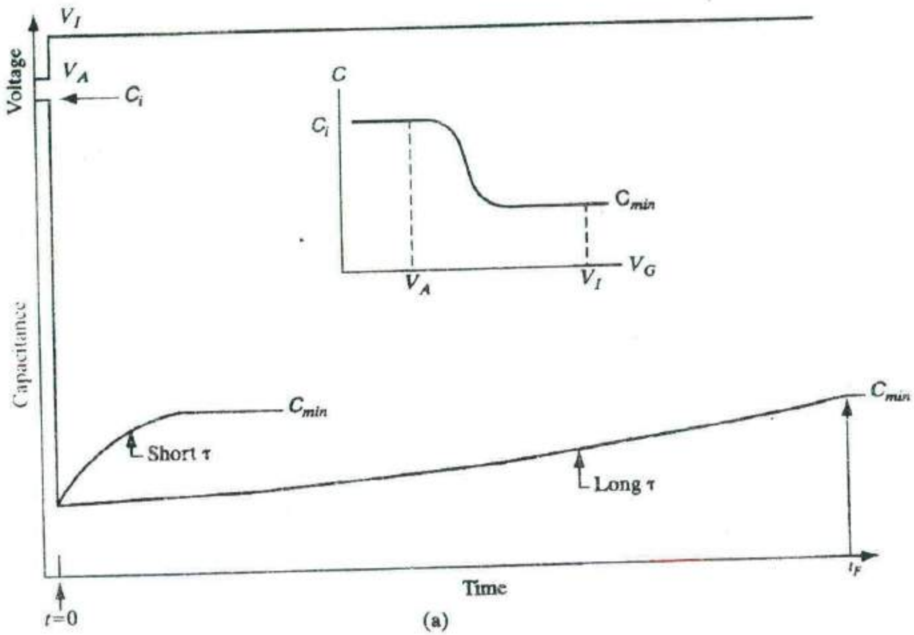


Figure 6–22
Mobile ion determination: (a) Movement of mobile ions due to positive and negative bias-temperatu stress; (b) *C–V* characteristics under positive (dashed line) and negative (solid line) bias-temperature.

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 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}). This capacitance transient, *C–t*, forms the basis of a powerful technique to measure the lifetime, known as the Zerbst technique. It was shown by Zerbst that by plotting the *C–t* data as in Fig. 6–23, the slope is inversely proportional to the lifetime (τ).



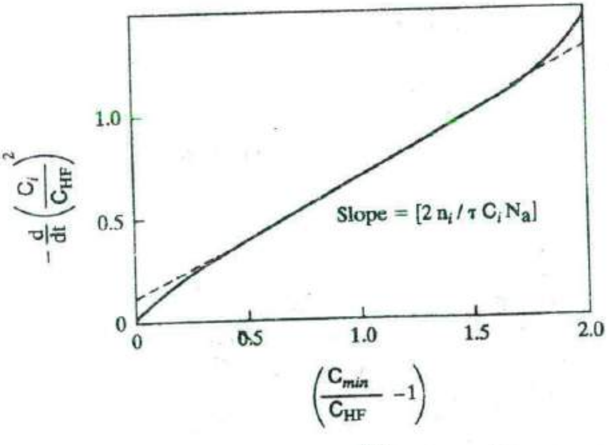


Figure 6–23
Zerbst plots: (a) time-dependent MOS capacitance (C_{HF}) due to the application of a step voltage V_A (which puts the capacitor in accumulation) to V_I (which puts the capacitor in inversion); (b) extraction of minority carrier lifetimes from MOS capacitance-time data.

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* curent 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 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:

$$I_{FN} \propto \mathscr{E}_{ox}^2 \exp\left(\frac{-B}{\mathscr{E}_{ox}}\right) \qquad (6–43)$$

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.

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

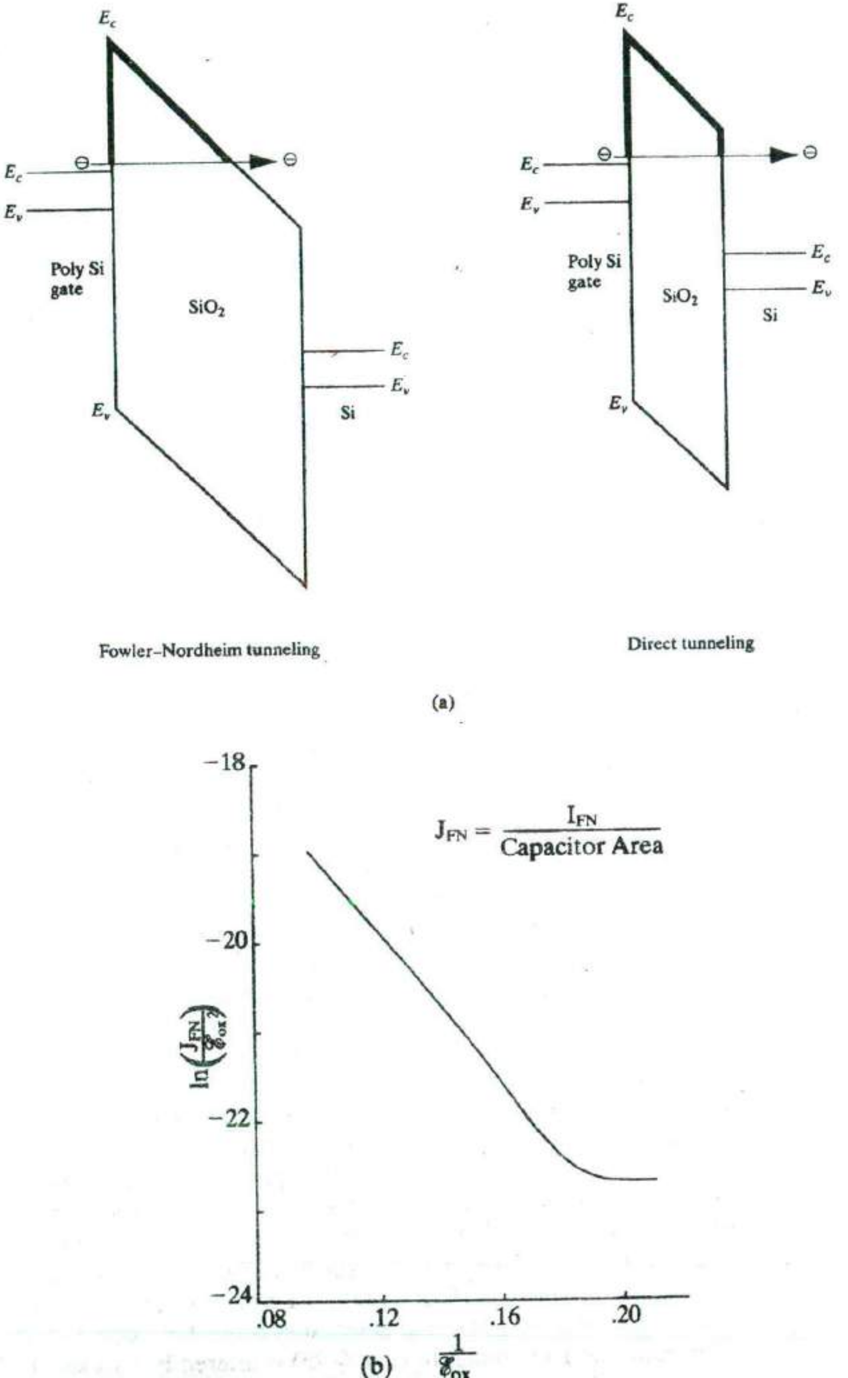

E_c
E_c
E_v
Poly Si gate
SiO_2
E_c
E_v
E_v
Si
Fowler–Nordheim tunneling
E_c
E_c
E_v
Poly Si gate
SiO_2
E_c
E_v
Si
E_v
Direct tunneling
(a)
−18
−20
−22
−24
$\ln\left(\frac{J_{FN}}{\mathscr{E}_{ox}^2}\right)$
$J_{FN} = \frac{I_{FN}}{\text{Capacitor Area}}$
.08
.12
.16
.20
(b)
$\frac{1}{\mathscr{E}_{ox}}$

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.

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 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.

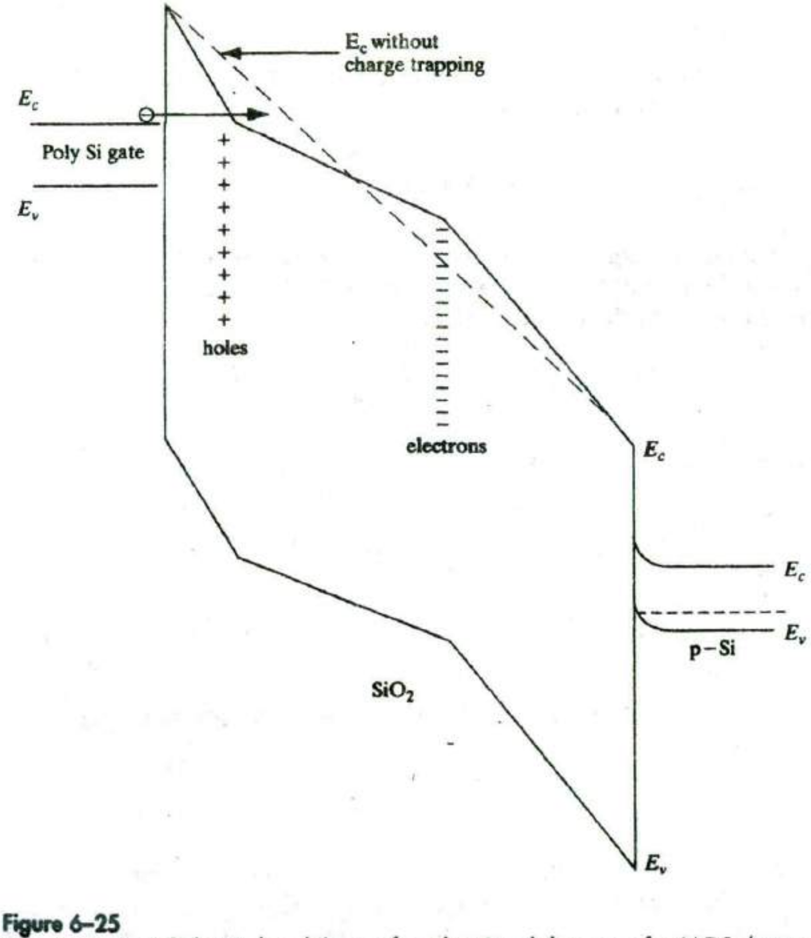


Figure 6–25
Time-dependent dielectric breakdown of oxides: Band diagram of a MOS device showing the band edges in the polysilicon gate, oxide and Si substrate. Trapped holes and electrons in the oxide distort the band edges, and increase the electric field in the oxide near the gate. The tunneling barrier width is seen to be less than if there were no charge trapping (dashed line).

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 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:

$$V_G = V_{FB} - \frac{Q_s}{C_i} + \phi_s \qquad (6\text{–}44)$$

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:

$$Q_n = -C_i\left[V_G - \left(V_{FB} + \phi_s - \frac{Q_d}{C_i}\right)\right] \qquad (6\text{–}45)$$

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:

$$Q_n = -C_i\left[V_G - V_{FB} - 2\phi_F - V_x - \frac{1}{C_i}\sqrt{2q\epsilon_s N_a(2\phi_F + V_x)}\right] \qquad (6\text{–}46)$$

If we neglect the variation of $Q_d(x)$ with bias V_x, Eq. (6–46) can be simplified to

$$Q_n(x) = -C_i(V_G - V_T - V_x) \qquad (6\text{–}47)$$

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 μ_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

$$I_D dx = \bar{\mu}_n Z|Q_n(x)|dV_x \qquad (6\text{–}48)$$

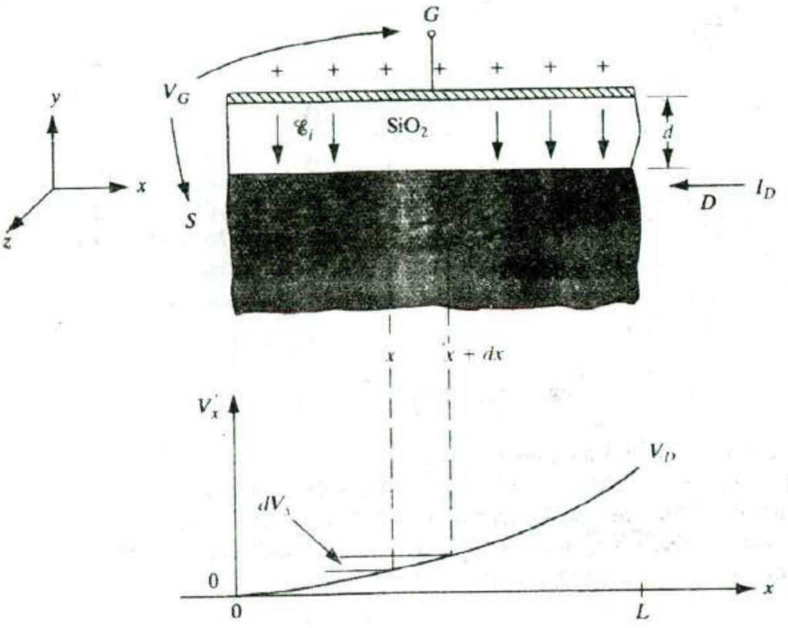


Figure 6–26
Schematic view of the n-channel region of a MOS transistor under bias below pinch-off, and the variation of voltage V_x along the conducting channel.

Integrating from source to drain,

$$\int_0^L I_D dx = \overline{\mu}_n Z C_i \int_0^{V_D} (V_G - V_T - V_x) dV_x$$

$$I_D = \frac{\overline{\mu}_n Z C_i}{L}[(V_G - V_T)V_D - \tfrac{1}{2}V_D^2] \qquad (6\text{–}49)$$

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

$$I_D = \frac{\overline{\mu}_n Z C_i}{L}$$

$$\times\left\{(V_G - V_{FB} - 2\phi_F - \tfrac{1}{2}V_D)V_D - \frac{2}{3}\,\frac{\sqrt{2\epsilon_s qN_a}}{C_i}\left[(V_D + 2\phi_F)^{3/2} - (2\phi_F)^{3/2}\right]\right\} \quad (6\text{–}50)$$

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 $V_D \ll (V_G - V_T)$:

$$g = \frac{\partial I_D}{\partial V_D} \simeq \frac{Z}{L}\overline{\mu}_n C_i (V_G - V_T) \quad (6\text{–}51)$$

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

$$V_D(\text{sat.}) \simeq V_G - V_T \quad (6\text{–}52)$$

The drain current at saturation remains essentially constant for larger values of drain voltage. Substituting Eq. (6–52) into Eq. (6–49), we obtain

$$I_D(\text{sat.}) \simeq \tfrac{1}{2}\overline{\mu}_n C_i \frac{Z}{L}(V_G - V_T)^2 = \frac{Z}{2L}\overline{\mu}_n C_i V_D^2(\text{sat.}) \quad (6\text{–}53)$$

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:

$$g_m(\text{sat.}) = \frac{\partial I_D(\text{sat.})}{\partial V_G} \simeq \frac{Z}{L}\mu_n C_i (V_G - V_T) \quad (6\text{–}54)$$

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

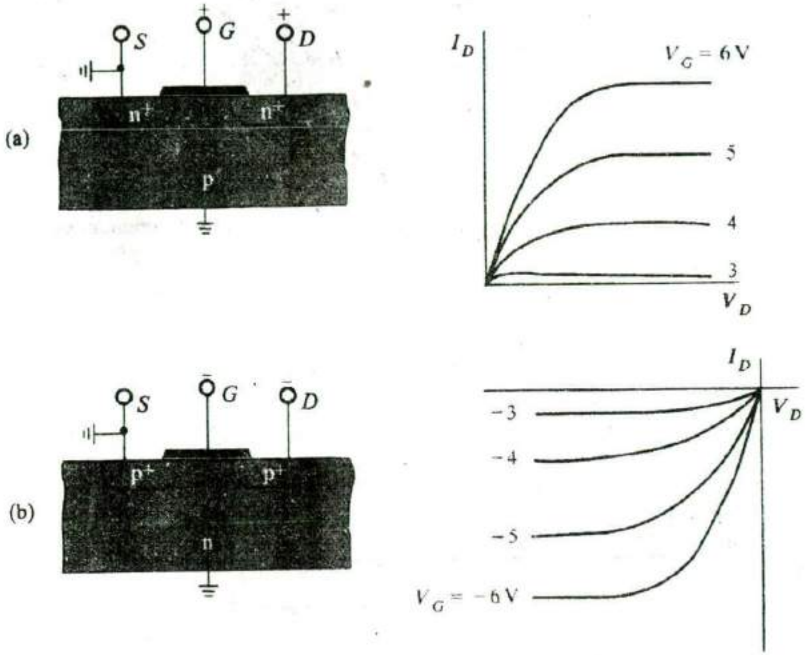


Figure 6–27 Drain current–voltage characteristics for enhancement transistors: (a) for n-channel V_D, V_G, V_T, and I_D are positive; (b) for p-channel all these quantities are negative.

(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. 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 lowing (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.

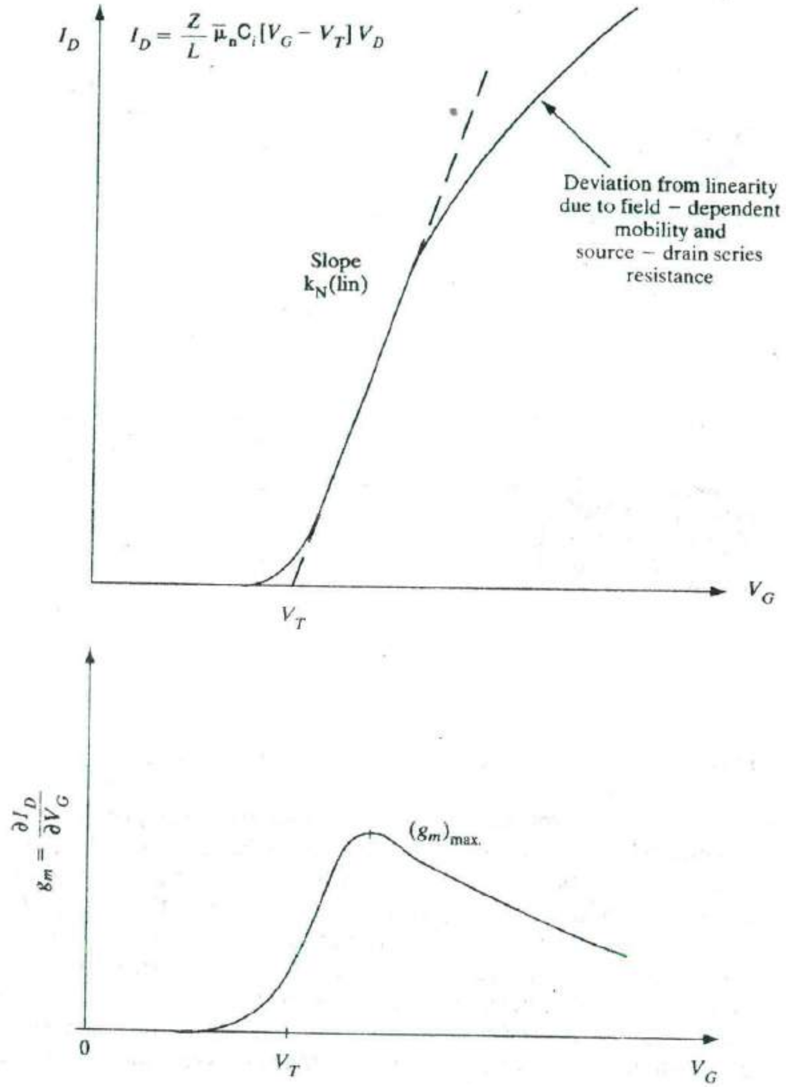


Figure 6–28 Linear region transfer characteristics: (a) plot of drain current versus gate voltage for MOSFETs in the linear region; (b) transconductance as a function of gate bias.

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

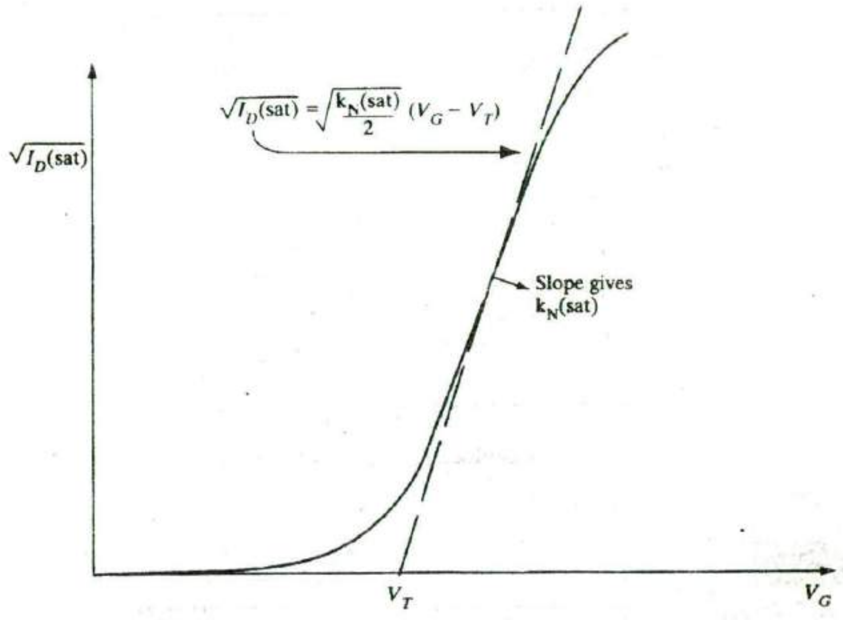


Figure 6–29 Saturation region transfer characteristics: plot of square root of the drain current versus gate voltage for MOSFETs.

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" 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

$$\mathscr{E}_{eff} = \frac{1}{\epsilon_s}\left(Q_d + \frac{1}{3}Q_n\right) \tag{6–55b}$$

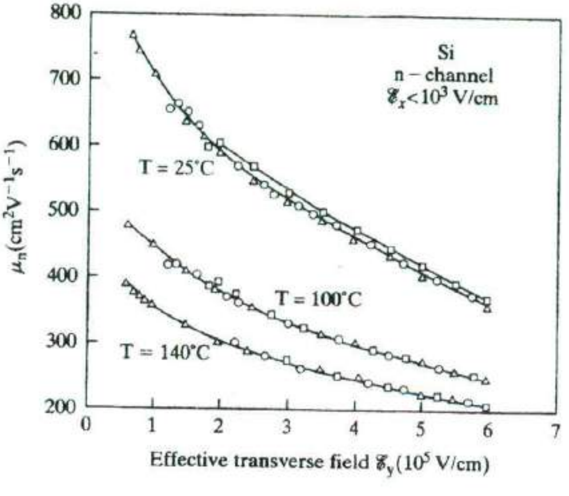


Figure 6–30 Inversion layer electron mobility versus effective transverse field, at various temperatures. The triangles, circles and squares refer to different MOSFETs with different gate oxide thicknesses and channel dopings. (After Sabnis and Clemens, *IEEE IEDM*, 1979).

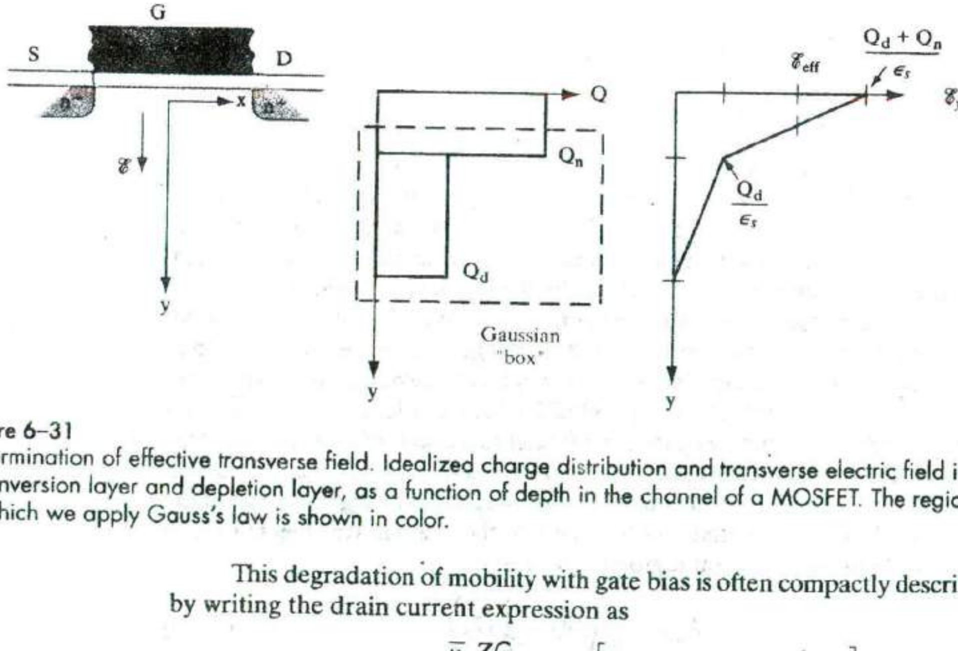


Figure 6–31
Determination of effective transverse field. Idealized charge distribution and transverse electric field in the inversion layer and depletion layer, as a function of depth in the channel of a MOSFET. The region to which we apply Gauss's law is shown in color.

This degradation of mobility with gate bias is often compactly described by writing the drain current expression as

$$I_D = \frac{\overline{\mu}_n Z C_i}{L\{1 + \theta(V_G - V_T)\}}\left[(V_G - V_T)V_D - \frac{1}{2}V_D^2\right] \qquad (6\text{–}56)$$

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 longitudinal electric field. As shown in Fig. 3–24, the carrier drift velocity increases linearly with electric field (ohmic behavior) until the field reaches $\mathscr{E}_{sat}$; in other words, the mobility is constant up to $\mathscr{E}_{sat}$. 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:

$$v = \mu\mathscr{E} \text{ for } \mathscr{E} < \mathscr{E}_{sat} \tag{6–57}$$

$$\text{and } v = v_s \text{ for } \mathscr{E} > \mathscr{E}_{sat} \tag{6–58}$$

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(sat.)), divided by the length of the pinch-off region, ΔL.

$$\mathscr{E}_{\max} = \left(\frac{V_D - V_D(\text{sat.})}{\Delta L}\right) \tag{6–59}$$

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.

$$I_D \approx ZC_i(V_G - V_T)v_s \tag{6–60}$$

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 device. For example, if the transistor is to be used in a circuit driven by a 3-V

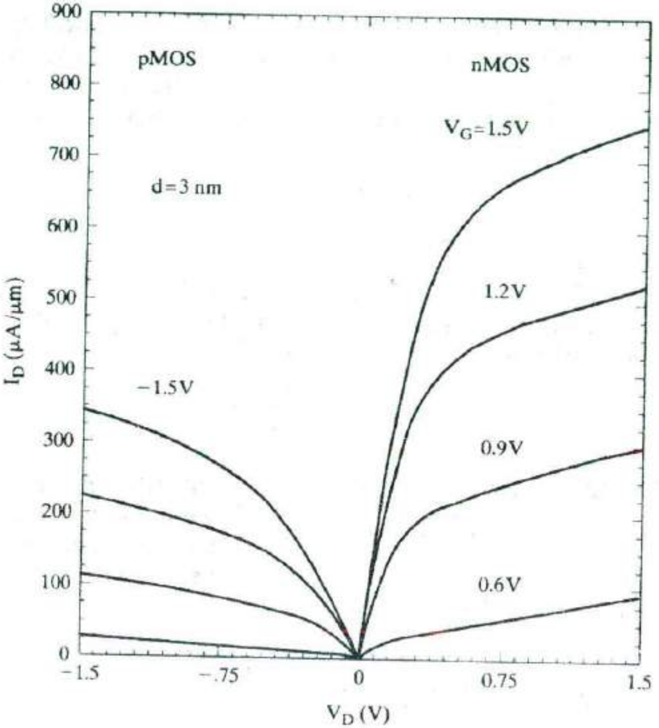


Figure 6–32 Experimental output characteristics of n-channel and p-channel MOSFETs with 0.1 μm channel lengths. The curves exhibit almost equal spacing, indicating a linear dependence of I_D on V_G, rather than a quadratic dependence. We also see that I_D is not constant but increases somewhat with V_D in the saturation region. The p-channel devices have lower currents because hole mobilities are lower than electron mobilities.

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 1960's, 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 mid-gap 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

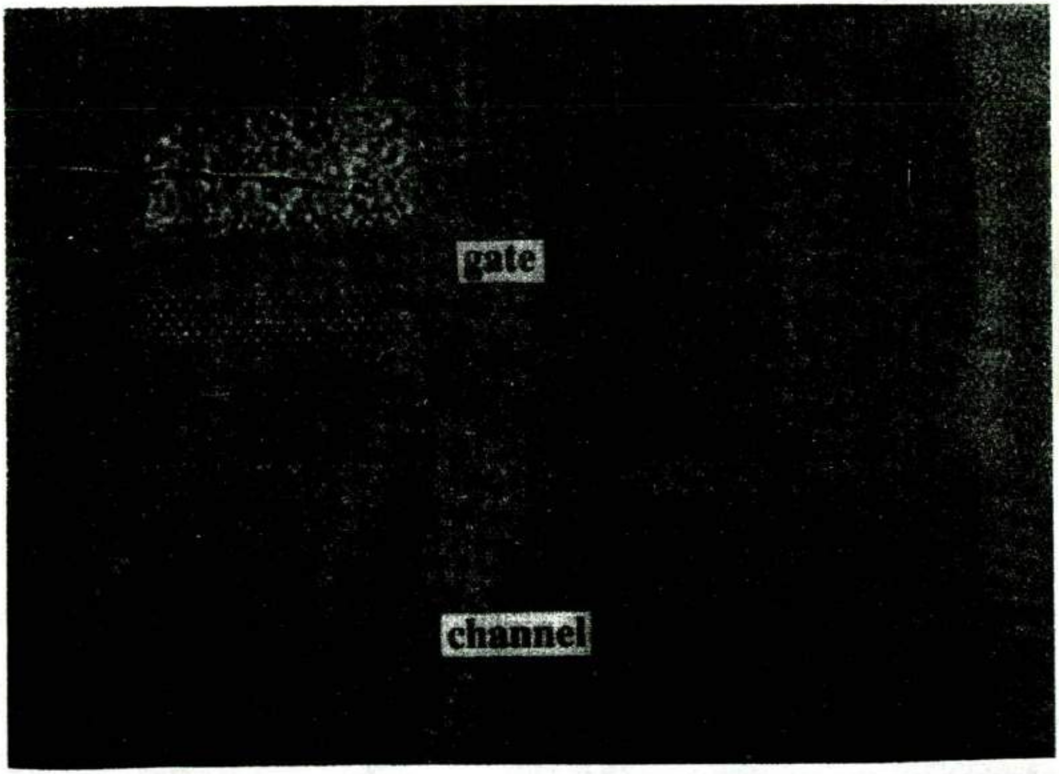


Figure 6–33
Cross section of a MOSFET. This high resolution transmission electron micrograph of a silicon Metal–Oxide Semiconductor Field Effect Transistor shows the silicon channel and metal gate separated by a thin (40Å, 4nm) silicon–dioxide insulator. The inset shows a magnified view of the three regions, in which individual rows of atoms in the crystalline silicon can be distinguished. (Photograph courtesy of AT&T Bell Laboratories.)

this micrograph, is 40Å 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.

EXAMPLE 6–4

Consider an n^+ polysilicon-SiO_2–Si p-channel device with $N_d = 10^{16}\ cm^{-3}$ and $Q_i = 5 \times 10^{10} q\ C/cm^2$. Calculate V_T for a gate oxide thickness of 0.01 μm and repeat for a field oxide thickness of 0.5 μm.

SOLUTION

Values of ϕ_F, Q_i, and Q_d can be obtained from Examples 6–2 and 6–3 if we use appropriate signs as in Fig. 6–20a. The value of C_i for the thin oxide case is the same as in Example 6–2. From Fig. 6–17, $\Phi_{ms} = -0.25$ V.

$$V_T = -0.25 - 0.694 - \frac{8 \times 10^{-9} + 4.82 \times 10^{-8}}{34.5 \times 10^{-8}} = -1.1\ V$$

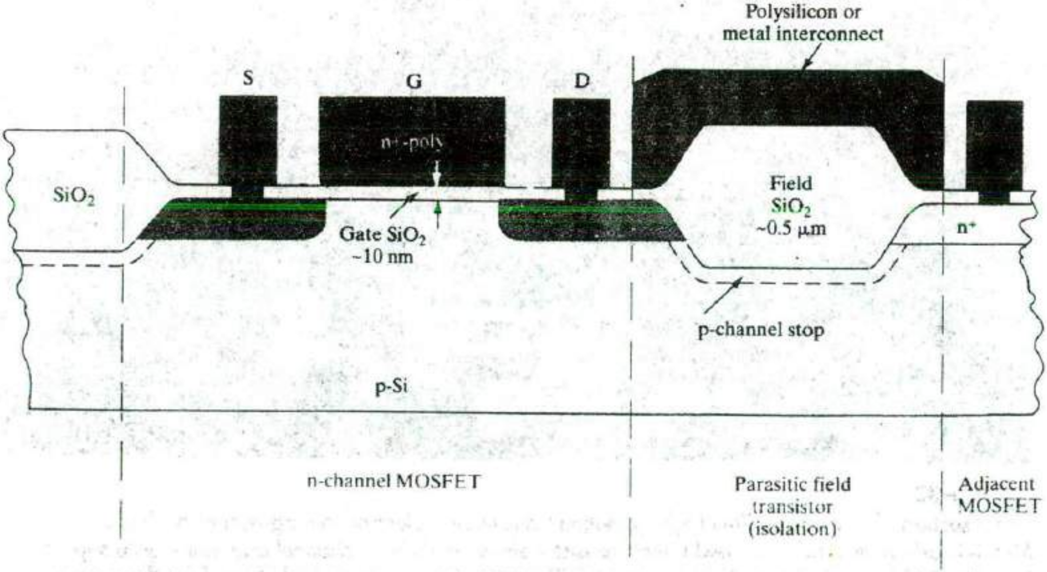


Figure 6–34
Thin oxide in the gate region and thick oxide in the field between transistors for V_T control (not to scale).

This value corresponds to that expected from Fig. 6–20b. In the field region where $d = 0.5\ \mu\text{m}$,

$$V_T = -0.944 - \frac{5.62 \times 10^{-8}}{6.9 \times 10^{-9}} = -9.1\ \text{V}$$

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 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.

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.

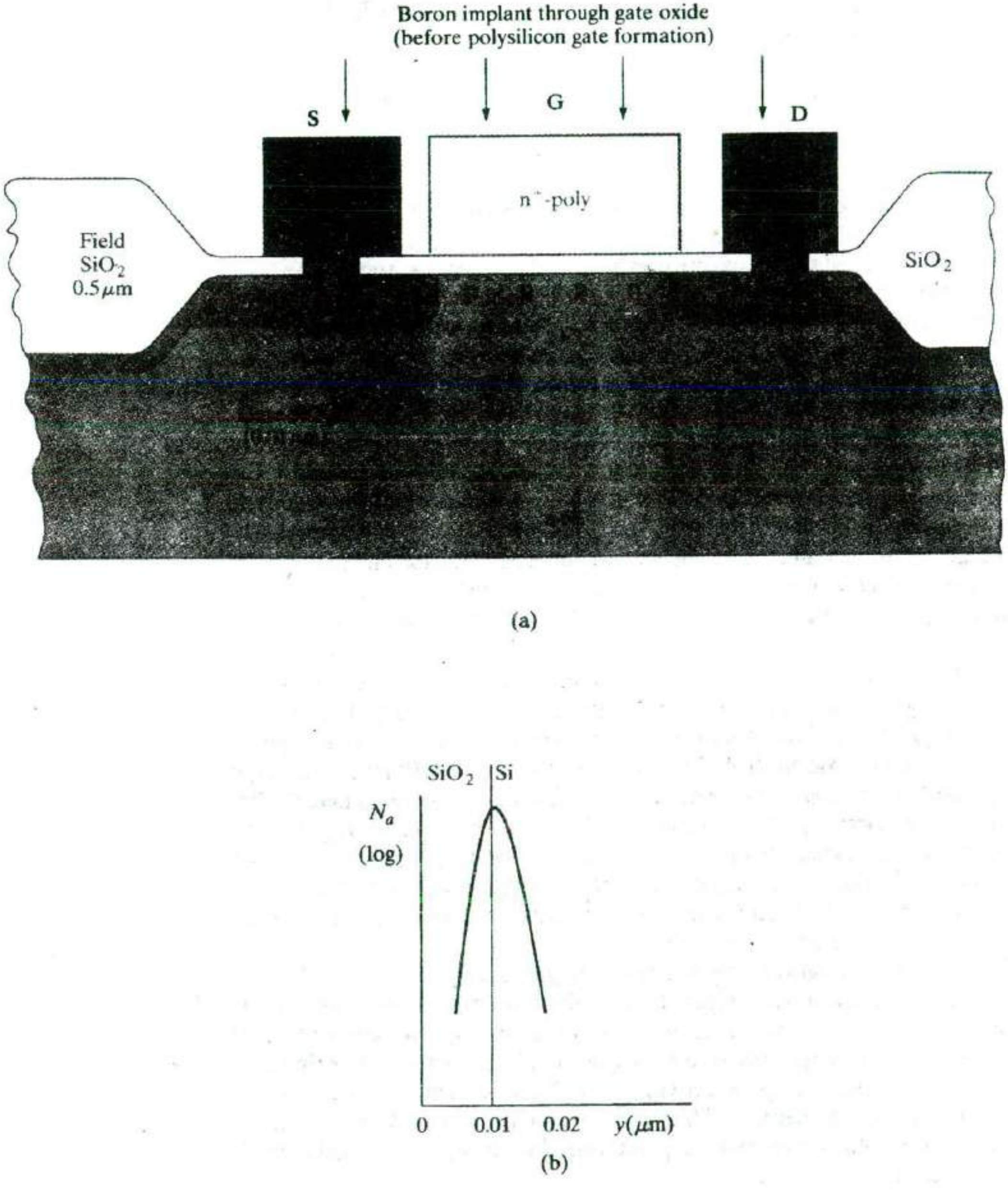


Figure 6–35

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.

EXAMPLE 6–5

For the p-channel transistor of Example 6–4, 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.

SOLUTION

$$-0.5 = -1.1 + \frac{qF_B}{C_i}$$

$$F_B = \frac{3.45 \times 10^{-7}}{1.6 \times 10^{-19}}(0.6) = 1.3 \times 10^{12}\ \text{cm}^{-2}$$

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 above, 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

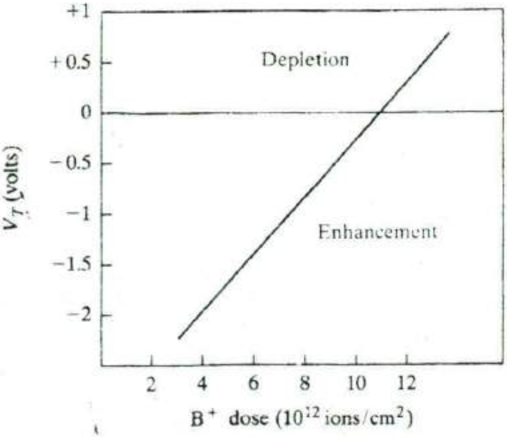


Figure 6–36 Typical variation of V_T for a p-channel device with increased implanted boron dose. The originally enhancement p-channel transistor becomes a depletion-mode device ($V_T > 0$) by sufficient B implantation.

a parasitic channel in the isolation regions) selectively in the isolation region under the field oxide (Fig. 6–34). Generally, a B channel stop implant is used fo n-channel devices. (It must be noted that such an acceptor implant will raise th 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

$$\Delta V_T \simeq \frac{\sqrt{2\epsilon_s q N_a}}{C_i}(-V_B)^{1/2} \quad \text{(n channel)} \tag{6–63}$$

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 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

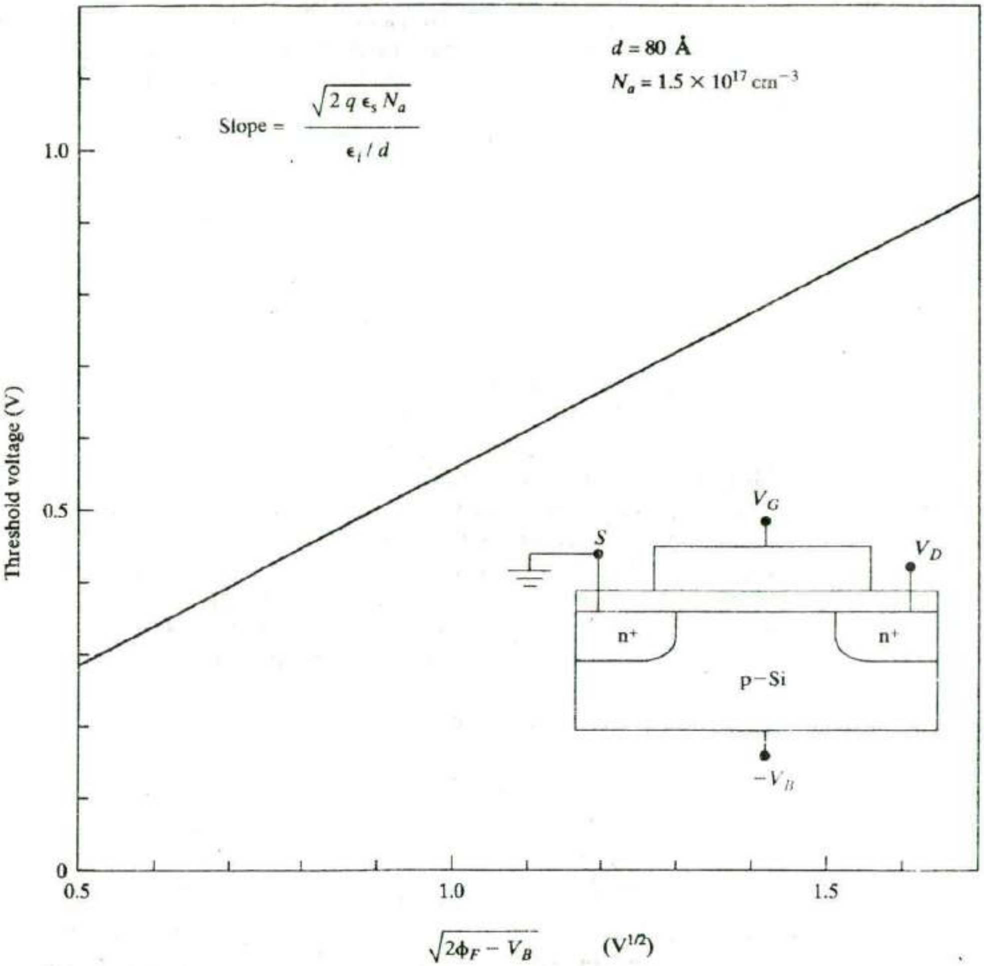


Figure 6–37
Threshold voltage dependence on substrate bias resulting from application of a voltage V_B from the substrate (i.e., bulk) to the source. For n channel, V_B must be zero or negative to avoid forward bias of the source junction. For p channel, V_B must be zero or positive.

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 flatband and threshold (for bandbending 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

$$I_D = \mu(\mathrm{C}_d + \mathrm{C}_{it})\frac{Z}{L}\left(\frac{kT}{q}\right)^2\left(1 - e^{\frac{-qV_D}{kT}}\right)\left(e^{\frac{q(V_G - V_T)}{c_r kT}}\right) \qquad (6\text{–}65)$$

where

$$c_r = \left[1 + \frac{\mathrm{C}_d + \mathrm{C}_{it}}{\mathrm{C}_i}\right]$$

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

$$S = \frac{dV_G}{d(\log I_D)} = \ln 10\,\frac{dV_G}{d(\ln I_D)} = 2.3\,\frac{kT}{q}\left[1 + \frac{\mathrm{C}_d + \mathrm{C}_{it}}{\mathrm{C}_i}\right] \qquad (6\text{–}66)$$

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, $\mathrm{C}_{it} = qD_{it}$. The expression in brackets in Eq. (6–66) is simply the capacitor divider ratio which tells us what 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.

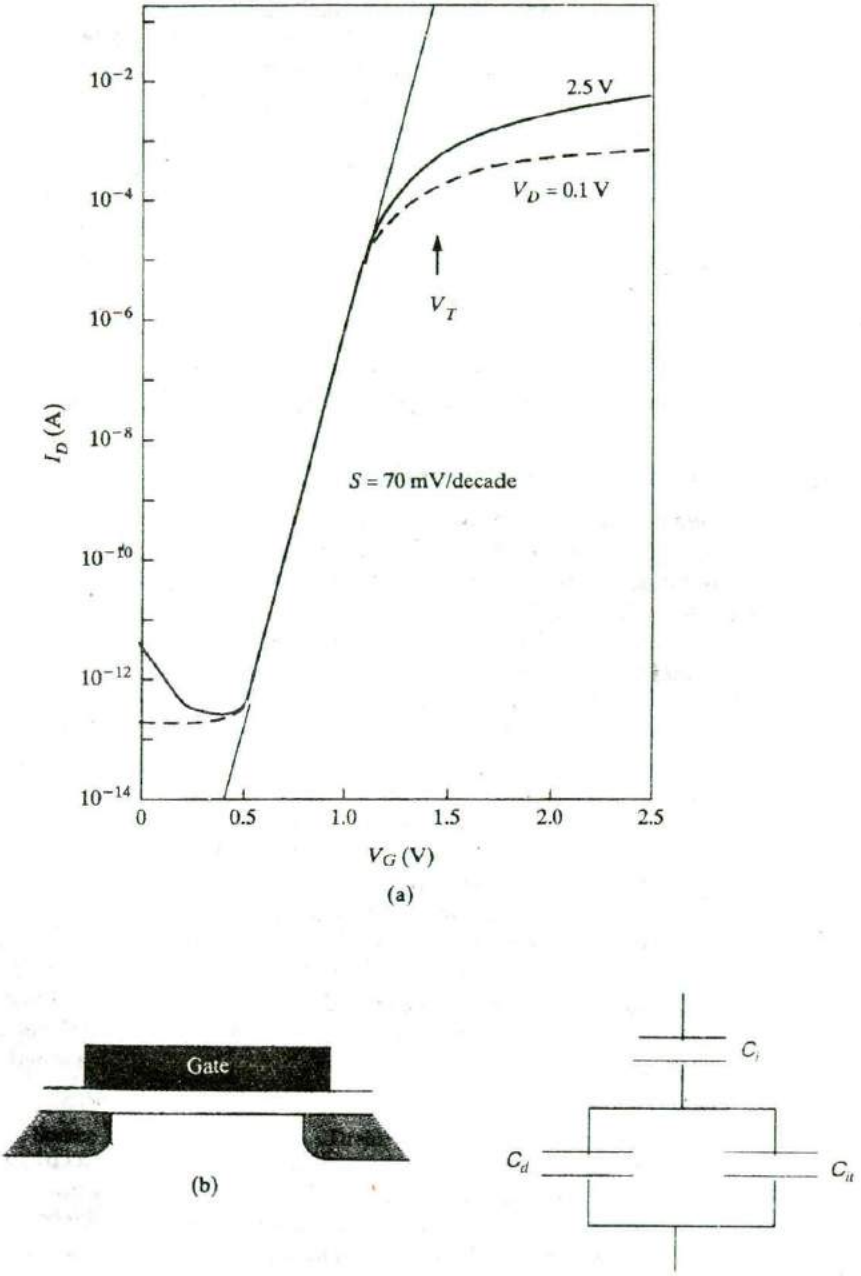


Figure 6–38

Subthreshold conduction in MOSFETs: (a) Semi-log plot of I_D versus V_G; (b) equivalent circuit showing capacitor divider which determines subthreshold slope.

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

$$L_{eff} = L - \Delta L_R \qquad (6\text{–}67)$$

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/

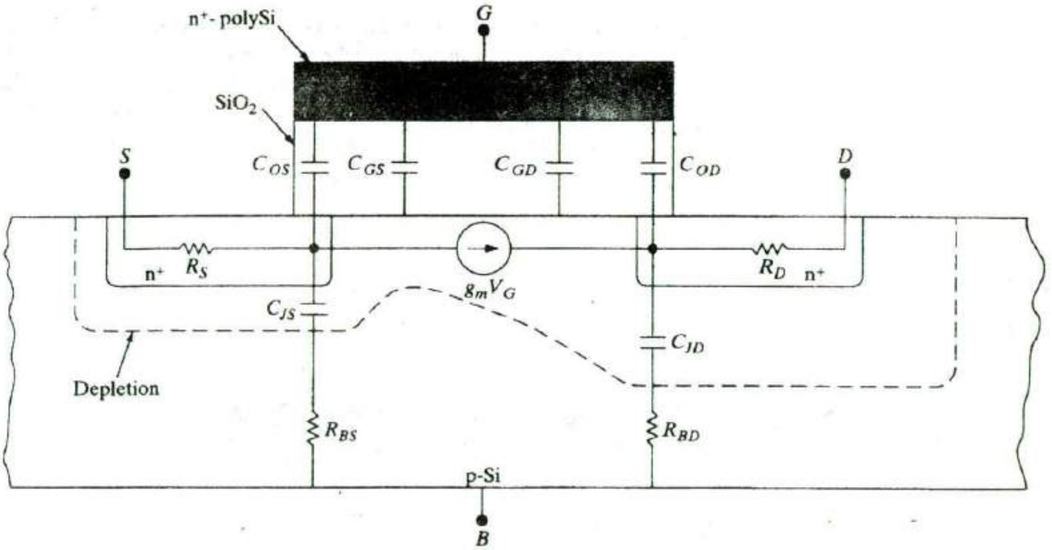


Figure 6–39
Equivalent circuit of a MOSFET, showing the passive capacitive and resistive components. The gate capacitance C_i is the sum of the distributed capacitances from the gate to the source-end of the channel (C_{GS}) and the drain-end (C_{GD}). In addition, we have an overlap capacitance (where 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.

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

$$\frac{V_D}{I_D} = R_{Ch} + R_{SD} = \frac{L - \Delta L_R}{Z - \Delta Z}\,\frac{1}{\bar{\mu}_n C_i(V_G - V_T)} + R_{SD} \qquad (6\text{–}68)$$

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.

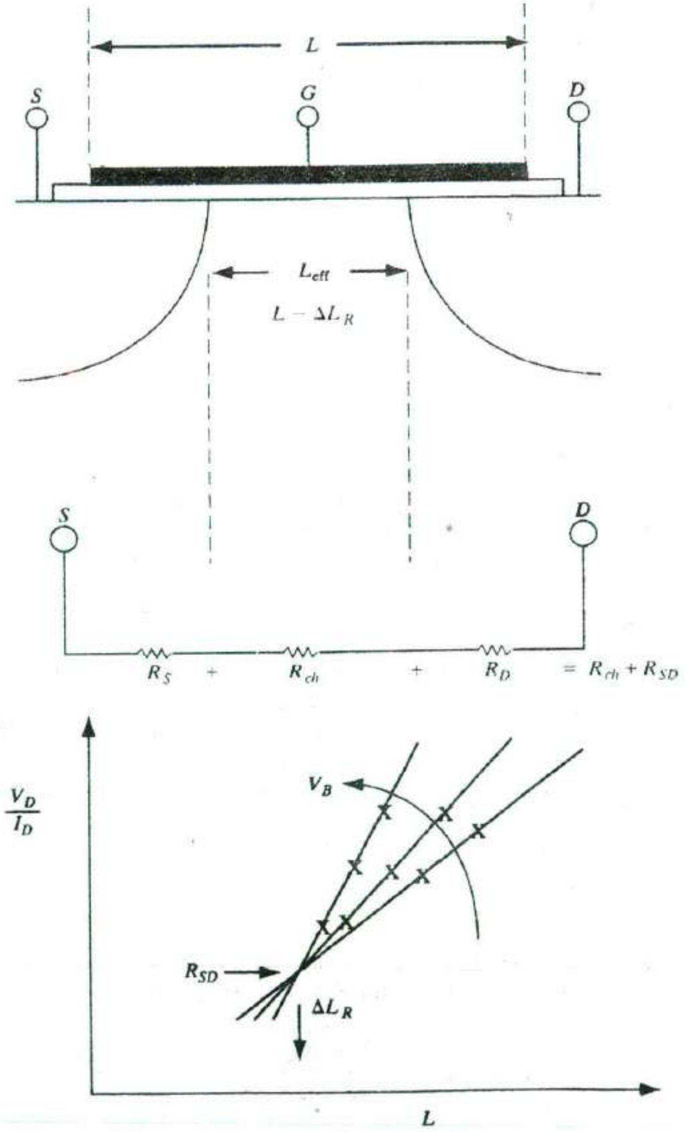


Figure 6–40 Determination of length reduction and source/drain series resistance in a MOSFET. The overall resistance of a MOSFET in the linear region is plotted as a function of channel length, for various substrate biases. The x mark data points for three different physical gate lengths *L*.

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 only has potential energy; as it gains more kinetic energy,

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	$1/K^3$

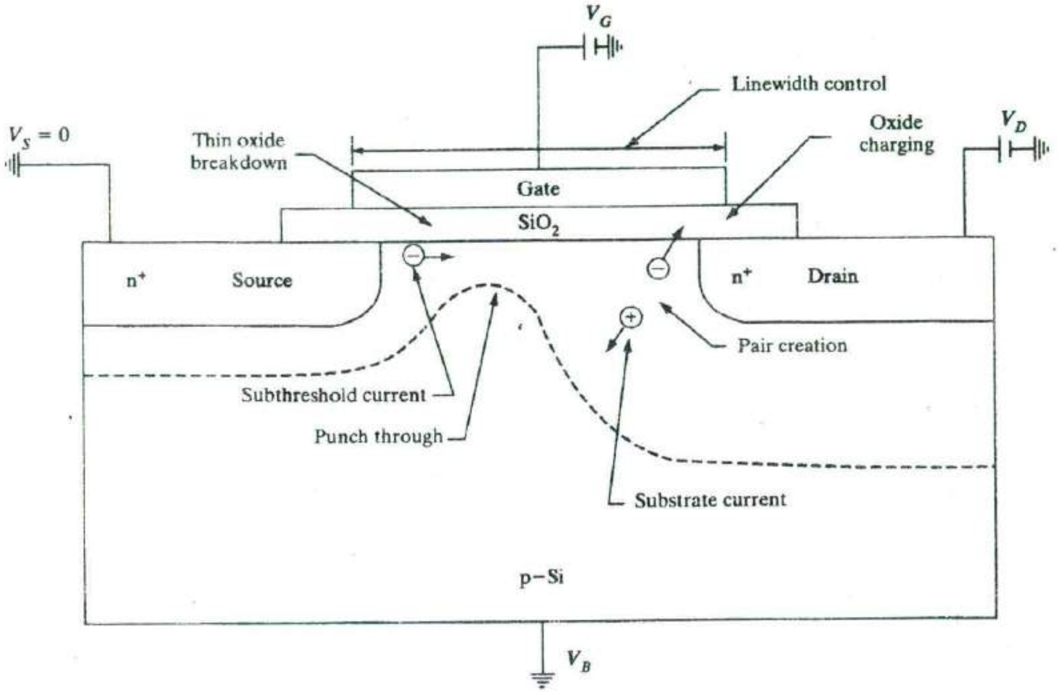


Figure 6 41
Short channel effects in MOSFETs. As MOSFETs are scaled down, potential problems due to short channel effects include hot carrier generation (electron-hole pair creation) in the pinch-off region, punchthrough breakdown between source and drain, and thin gate oxide breakdown.

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 be trapped in the gate oxide as fixed oxide charges. According to Eq. (6–37), this increases the flatband 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).

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

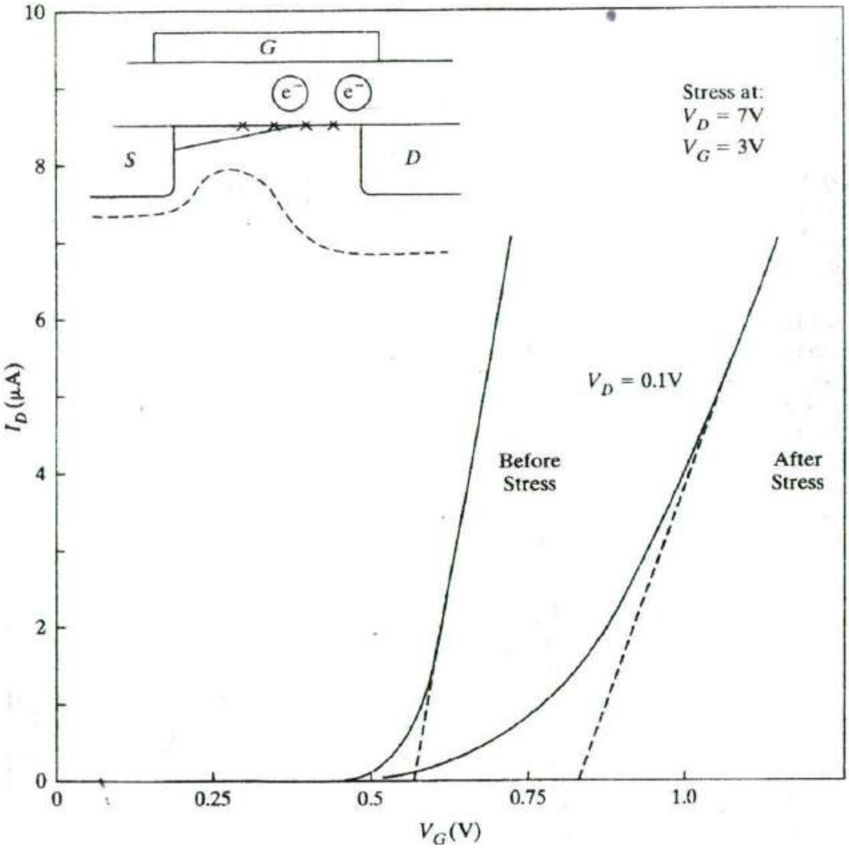


Figure 6–42
Hot carrier degradation in MOSFETs. The linear region transfer characteristics before and after hot carrier stress indicate an increase of V_T and decrease of transconductance (or channel mobility) due to hot electron damage. The damage can be due to hot electron injection into the gate oxide which increases the fixed oxide charge, and increasing fast interface state densities at the oxide-silicon interface (indicated by x).

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

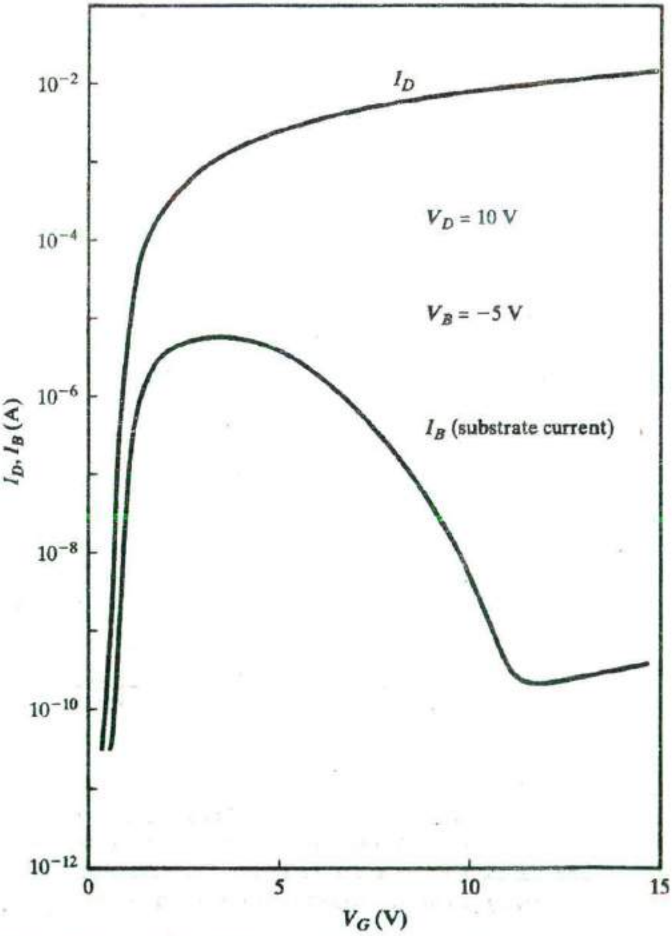


Figure 6–43 Substrate current in a MOSFET. The substrate current in n-channel MOSFETs due to impact-generated holes in the pinch-off region, as a function of gate bias. The substrate current initially increases with V_G because of the corresponding increase of I_D. However, for even higher V_G, the MOSFET 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.)

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 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 punchthrough 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 punchthrough 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

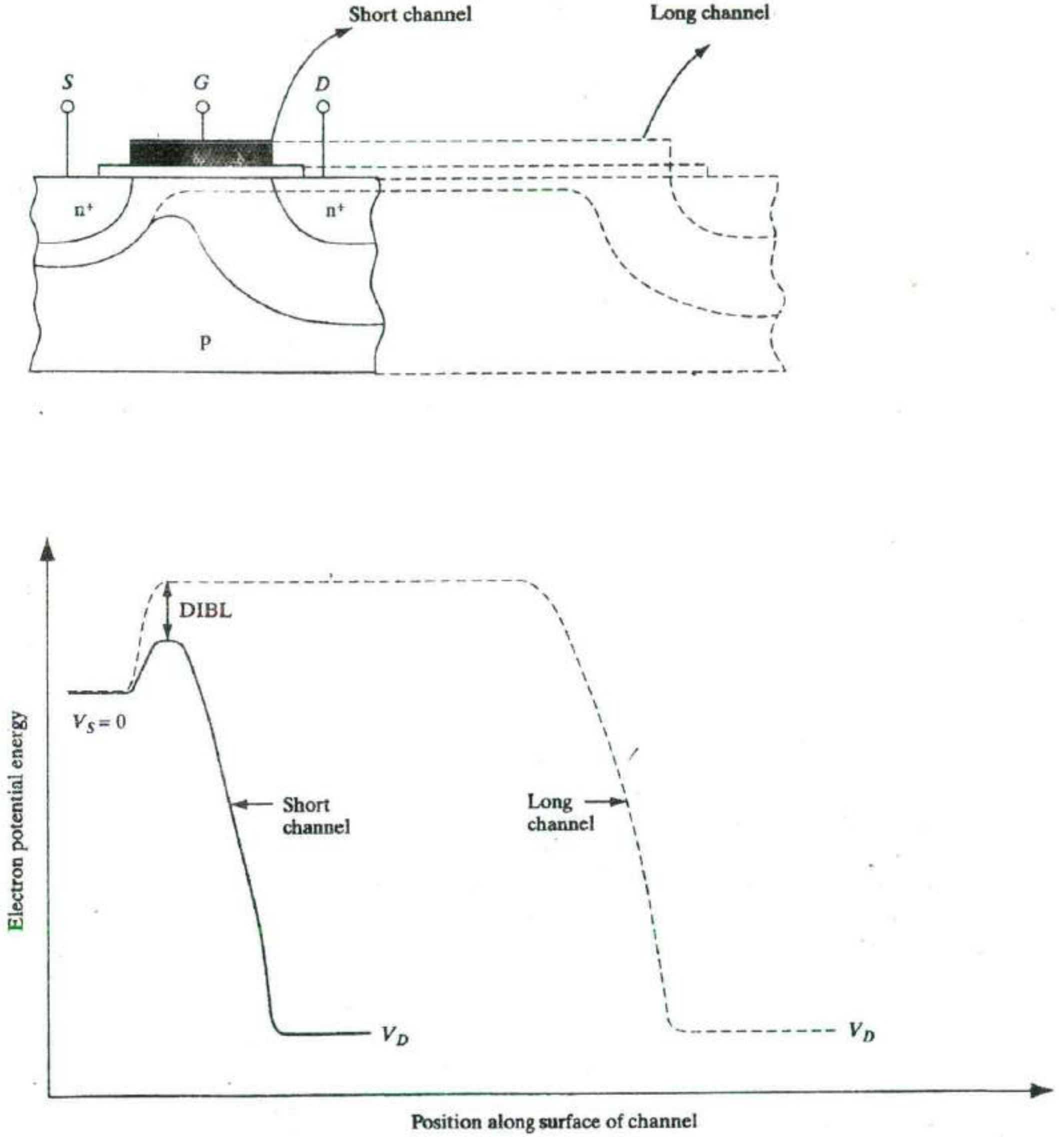


Figure 6–44
Drain-induced barrier lowering in MOSFETs. Cross-sections and potential distribution along the channel for a long channel and short channel MOSFET.

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 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-punchthrough* 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

$$I_D \propto \frac{1}{L - \Delta L} = \frac{1}{L}\left(1 + \frac{\Delta L}{L}\right) \qquad (6\text{–}69)$$

for small pinch-off regions, ΔL. We assume that the fractional change in the channel length is proportional to the drain bias,

$$\frac{\Delta L}{L} = \lambda V_D \qquad (6\text{–}70)$$

where λ is the *channel length modulation parameter*. Hence, in the saturation region, the expression for the drain current becomes

$$I_D = \frac{Z}{2L}\bar{\mu}_n C_i (V_G - V_T)^2(1 + \lambda V_D) \qquad (6\text{–}71)$$

This leads to a slope in the output characteristics, or a lowering of the output impedance (Fig. 6–32).

6.5.11 Short Channel 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.

[9] L. Yau, "A simple theory to predict the threshold voltage of short-channel IGFETs," Solid-State Electronics, 17 (1974): 1059.

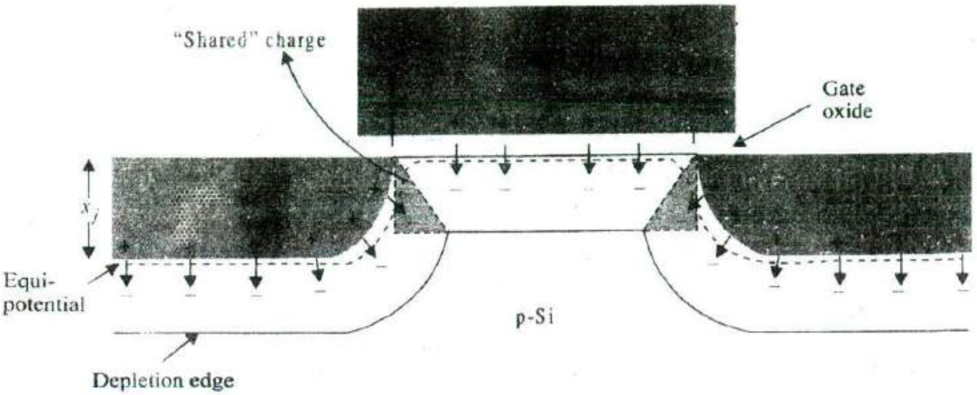


Figure 6–45
Short channel effect in a MOSFET. Cross-sectional view of MOSFET along the length showing depletion charge sharing (colored regions) between the gate, source and drain.

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 electri-

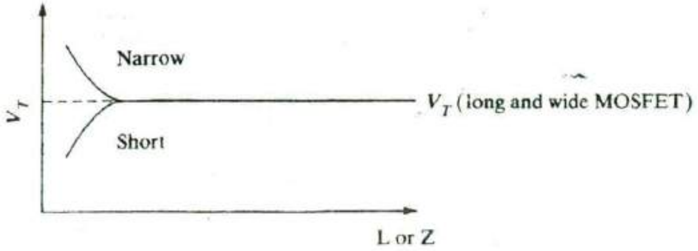


Figure 6–46 Roll-off of V_T with decreasing channel length, and increase of V_T with decreasing width.

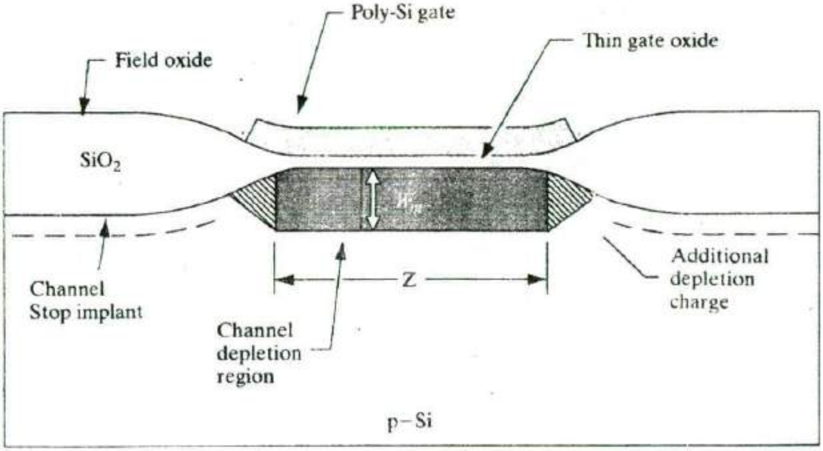


Figure 6–47 Narrow width effect in a MOSFET. Cross-sectional view of MOSFET along the width, showing additional depletion charge (colored regions) underneath the field or the LOCOS isolation regions.

cally 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 1 μm.

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 then bottoms out 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.

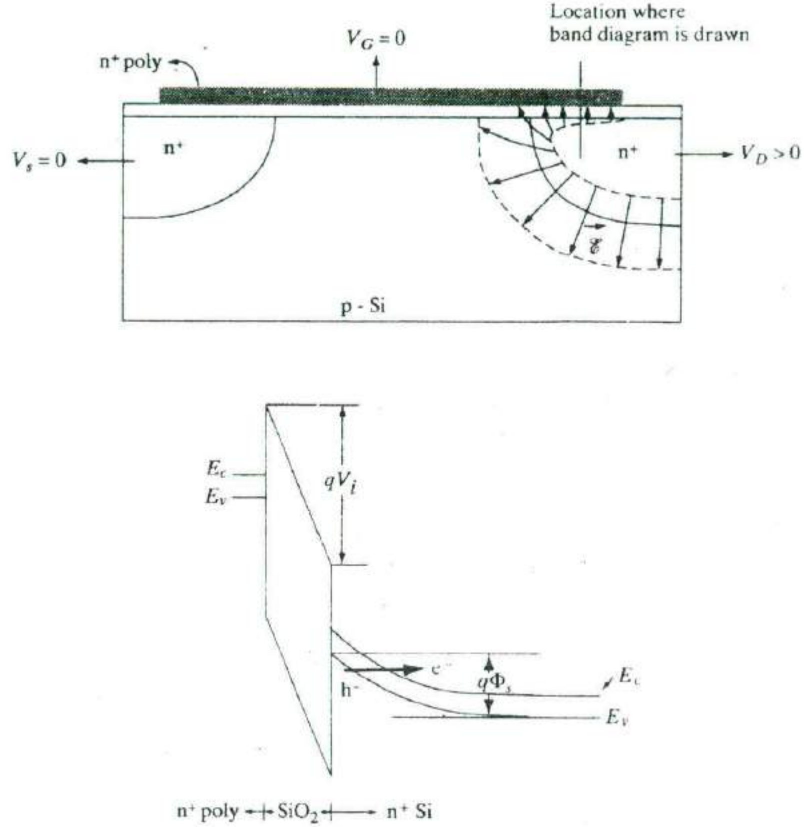


Figure 6–48 Gate-induced drain leakage in MOSFETs. The band diagram for the location shown in color is plotted as a function of depth in the gate–drain overlap region, indicating band-to-band tunneling and creation of electron–hole pairs in the drain region in the Si substrate.

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 bandbending is more than the bandgap 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 bandbending 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.

PROBLEMS

6.1 Modify Eqs. (6–2) through (6–5) to include effects of the contact potential V_0. Define a true pinch-off voltage V_T to distinguish this case from V_P defined in Eq. (6–4).

6.2 Modify Eqs. (6–7) through (6–10) to include V_0. Let V_P be defined as in Eq. (6–4), and call the true pinch-off voltage V_T.

6.3 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?

6.4 If the ratio $Z/L = 10$ for the JFET of Prob. 6.3, 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).

6.5 For the JFET of Prob. 6.4, plot I_D vs. V_D for the same three values of V_G. Terminate each plot at the point of saturation.

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 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 values of $0, -1, -2, -3, -4$, and -5 V.

6.8 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.

6.9 An n$^+$-polysilicon-gate n-channel MOS transistor is made on a p-type Si substrate with $N_a = 5 \times 10^{15}$ cm^{-3}. The SiO$_2$ thickness is 100 Å in the gate region, and the effective interface charge Q_i is $4 \times 10^{10}\ q$C/cm^2.

Find W_m, V_{FB}, and V_T.

6.10 An n$^+$ polysilicon-gate p-channel MOS transistor is made on an n-type Si substrate with $N_d = 5 \times 10^{16}$ cm^{-3}. The SiO$_2$ thickness is 100 Å in the gate region, and the effective interface charge Q_i is $2 \times 10^{11}\ q$ C/cm^2. Sketch the C–V curve for this device and give important numbers for the scale.

6.11 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.

6.12 Calculate the V_T of a Si-MOS transistor for a n$^+$-polysilicon gate with silicon oxide thickness = 50 Å, $N_a = 1 \times 10^{18}$ 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.

6.13 (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.

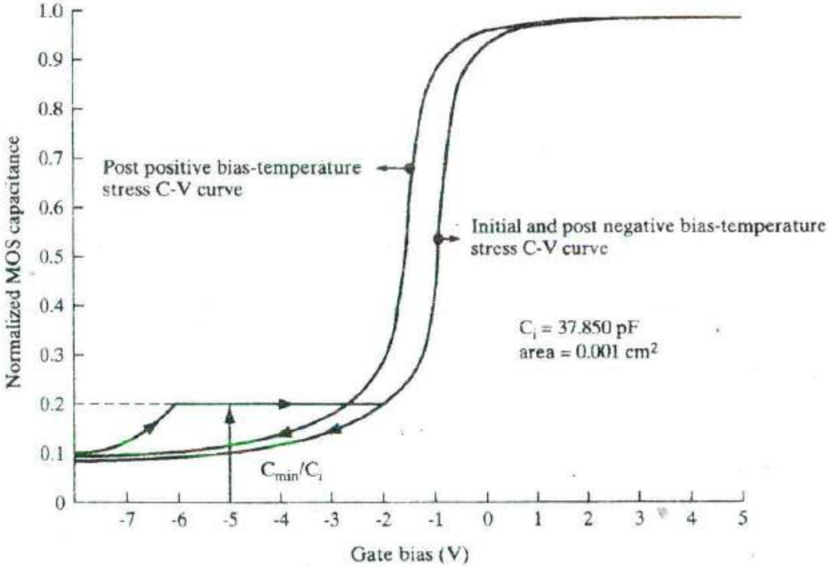
(b) In the case of an arbitrary distribution of charge $\rho(x')$ in the oxide, show that

$$V_{FB} = -\frac{1}{C_i}\int_0^d \frac{x'}{d}\rho(x')dx'$$

6.14 The bias on a Si MOS capacitor is changed from inversion to accumulation mode. If the substrate doping is 10^{16} cm^{-3} donors, what is the change in the surface bandbending at 100°C?

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 −0.35V

Figure P6–15



6.16 For the capacitor in Prob. 6.15, determine the initial flatband voltage.

6.17 For the capacitor in Prob. 6.15, determine the fixed oxide charge, Q_i, and the mobile ion content.

6.18 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.

6.19 Calculate the V_T of a Si n-channel MOSFET for a 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. Repeat for a substrate bias of −2.5V.

6.20 For the MOSFET in Prob. 6.19, and $W = 50$ 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.

6.21 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 singly-charged species and assuming the implant distribution is peaked at the oxide–Si interface and can be regarded as a sheet charge at the 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.

6.22 For the MOSFET characteristics shown in Fig. P6-22, calculate:

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 $V_{FB} = 0$.

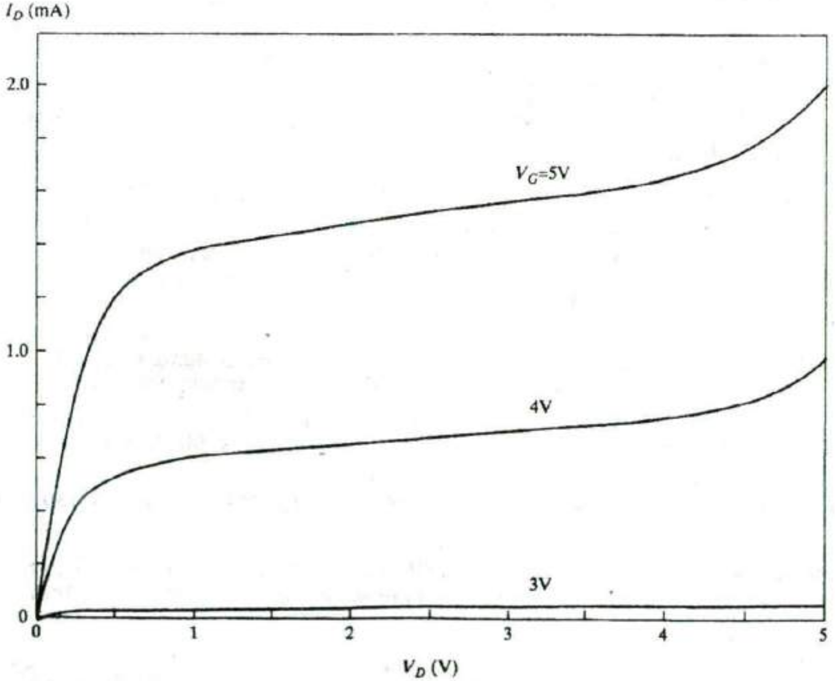


Figure P6–22

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 The flat band voltage is shifted to -2V for an n$^+$-polysilicon-SiO$_2$-Si capacitor with parameters discussed in Example 6–2. Redraw Fig. 6–16 for this case and find the value of interface charge Q_i required to cause this shift in V_{FB}, with Φ_{ms} given by Fig. 6–17.

6.26 Plot I_D vs. V_D with several values of V_G for the thin-oxide p-channel transistor described in Example 6–4. Use the p-channel version of Eq. (6–49), and assume that I_D(sat) remains constant beyond pinch-off. Assume that $\overline{\mu}_p = 200$ 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 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 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?

6.29 Calculate the substrate bias required to achieve enhancement-mode operation with $V_T = +0.5$ V for the n-channel device of Example 6–3. Comment on the practicality of this method of threshold control for thin-oxide transistors.

READING LIST

Dambkes, H. "Gallium Arsenide HEMTs for Low-Noise GHz Communications Engineering." *Microelectronics Journal* 20 (September–October 1989): 1–6.

Drummond, T. J., W. T. Masselink, and H. Morkoc. "Modulation-Doped GaAs/(Al,Ga)As Heterojunction Field-Effect Transistors: MODFETs." *Proceedings of the IEEE* 74 (June 1986): 773–822.

Frensley, W. R. "Gallium Arsenide Transistors." *Scientific American* 257 (August 1987): 80–87.

Inoue, K. "Recent Advances in InP-Based HEMT/HBT Device Technology." *Fourth International Conference on Indium Phosphide and Related Materials* (April 1992): 10–13.

Kahng, D. "A Historical Perspective on the Development of MOS Transistors and Related Devices," *IEEE Trans. Elec. Dev.*, ED-23 (1976): 655.

Morgan, D. V., and R. H. Williams, eds. *Physics and Technology of Heterojunction Devices.* London: P. Peregrinus, 1991.

Morkoc, H. "The HEMT: A Superfast Transistor." *IEEE Spectrum* 21 (February 1984): 28–35.

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.

Nguyen, L. D., L. E. Larson, and U. K. Mishra. "Ultra-High Speed Modulation-Doped Field-Effect Transistors: A Tutorial Review." *Proceedings of the IEEE* 80 (April 1992): 492–518.

Pavlidis, D. "Current Status of Heterojunction Bipolar and High-Electron Mobility Transistor Technologies." *Microelectronic Engineering* 19 (September 1992): 305–12.

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.

Shur, M. *GaAs Devices and Circuits.* New York: Plenum Press, 1987.

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

Smith, R. S. and I. G. Eddison. "Advanced Materials for GaAs Microwave Devices." *Advanced Materials* 4 (December 1992): 786–91.

Sze, S. M. *High-Speed Semiconductor Devices.* New York: Wiley, 1990.

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

Uyemura, J. P. *Fundamentals of MOS Digital Integrated Circuits.* Reading, MA: Addison-Wesley, 1988.

Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

Weisbuch, C., and B. Vinter. *Quantum Semiconductor Structures*. Boston: Academic Press, 1991.

Chapter 7

Bipolar Junction Transistors

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

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

7.1 FUNDAMENTALS OF BJT OPERATION

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

Let us begin the discussion of bipolar transistors by considering the reverse-biased p-n junction diode of Fig. 7–1. According to the theory of Chapter 5, the reverse saturation current through this diode depends on the rate at which minority carriers are generated in the neighborhood of the junction. We found, for example, that the reverse current due to holes being swept from n to p is essentially independent of the size of the junction $\mathscr{E}$ field and hence is independent of the reverse bias. The reason given was that the hole current depends on how often minority holes are generated by EHP creation within a diffusion length of the junction—not upon how fast a particular hole is swept across the depletion layer by the field. As a result, it is

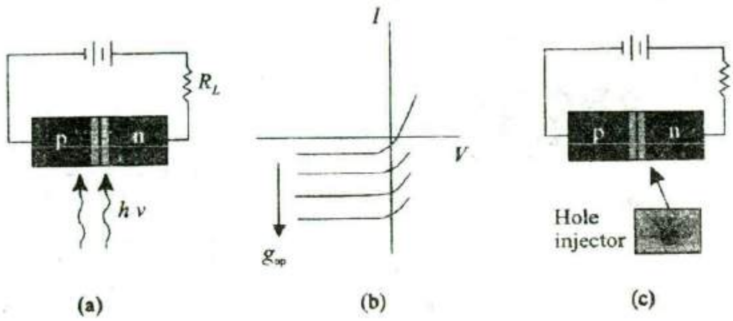


Figure 7–1 External control of the current in a reverse-biased p-n junction: (a) optical generation; (b) junction i–V characteristics as a function of EHP generation; (c) minority carrier injection by a hypothetical device.

possible to increase the reverse current through the diode by increasing the rate of EHP generation (Fig. 7–1b). One convenient method for accomplishing this is optical excitation of EHPs with light ($h\nu > E_g$), as in Section 4.3. With steady photoexcitation the reverse current will still be essentially independent of bias voltage, and if the dark saturation current is negligible, the reverse current is directly proportional to the optical generation rate g_{op}.

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

A convenient hole injection device is a forward-biased p^+-n junction. According to Section 5.3.2, the current in such a junction is due primarily to holes injected from the p^+ region into the n material. If we make the n side of the forward-biased junction the same as the n side of the reverse-biased junction, the p^+-n-p structure of Fig. 7–2 results. With this configuration, injection of holes from the p^+-n junction into the center n region supplies the minority carrier holes to participate in the reverse current through the n-p junction. Of course, it is important that the injected holes do not recombine in the n region before they can diffuse to the depletion layer of the reverse-biased junction. Thus we must make the n region narrow compared with a hole diffusion length.

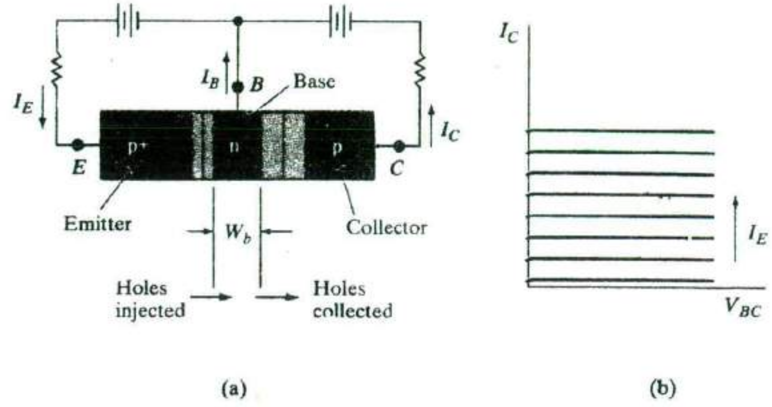


Figure 7–2 A p-n-p transistor: (a) schematic representation of a p-n-p device with a forward-biased emitter junction and a reverse-biased collector junction; (b) I–V characteristics of the reverse-biased n-p junction as a function of emitter current.

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

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

It is clear that current I_E flows into the emitter of a properly biased p-n-p transistor and that I_C flows out at the collector, since the direction of hole flow is from emitter to collector. However, the base current I_B requires a bit more thought. In a good transistor the base current will be very small since I_E is essentially hole current, and the collected hole current I_C is almost equal to I_E. There must be some base current, however, due to requirements

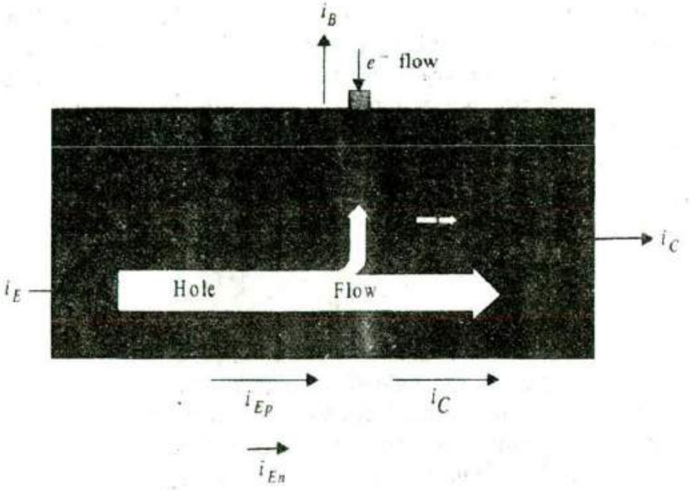


Figure 7–3 Summary of hole and electron flow in a p-n-p transistor with proper biasing: (1) injected holes lost to recombination in the base; (2) holes reaching the reverse-biased collector junction; (3) thermally generated electrons and holes making up the reverse saturation current of the collector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

of electron flow into the n-type base region (Fig. 7–3). We can account for I_B physically by three dominant mechanisms:

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

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

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

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

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

7.2 AMPLIFICATION WITH BJTS

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

$$i_C = Bi_{Ep} \tag{7–1}$$

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

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

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

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

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

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

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

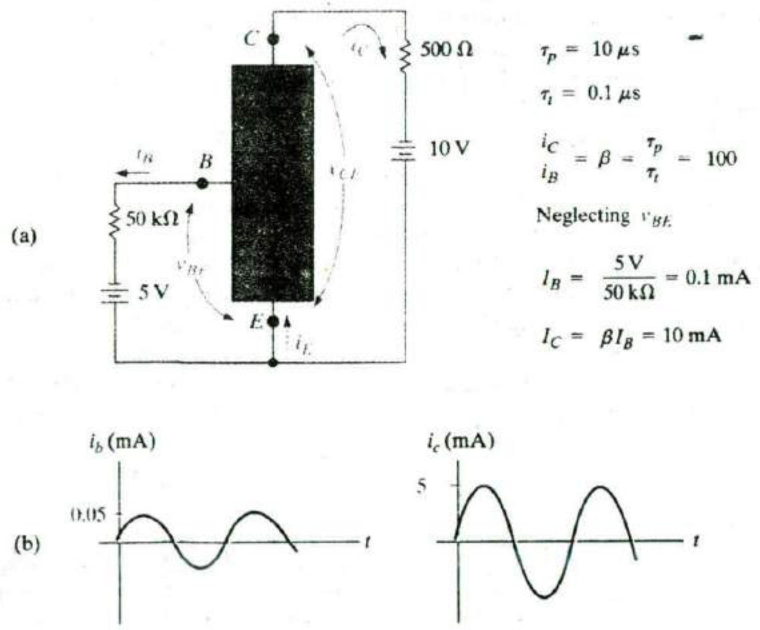
neglecting the collector saturation current. The relation between the collector and base currents is found from Eqs. (7–1) and (7–4):

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

$$\boxed{\frac{i_C}{i_B} = \frac{B\gamma}{1 - B\gamma} = \frac{\alpha}{1 - \alpha} \equiv \beta} \qquad (7\text{–}6)$$

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

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



$$\tau_p = 10\ \mu\text{s}$$
$$\tau_t = 0.1\ \mu\text{s}$$
$$\frac{i_C}{i_B} = \beta = \frac{\tau_p}{\tau_t} = 100$$

Neglecting v_{BE}

$$I_B = \frac{5\text{ V}}{50\text{ k}\Omega} = 0.1\text{ mA}$$
$$I_C = \beta I_B = 10\text{ mA}$$

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

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

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

$$\frac{i_C}{i_B} = \beta = \frac{\tau_p}{\tau_t} \tag{7–7}$$

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

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

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

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

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

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

EXAMPLE 7–1

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

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

SOLUTION

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

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

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

7.3 BJT FABRICATION

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

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

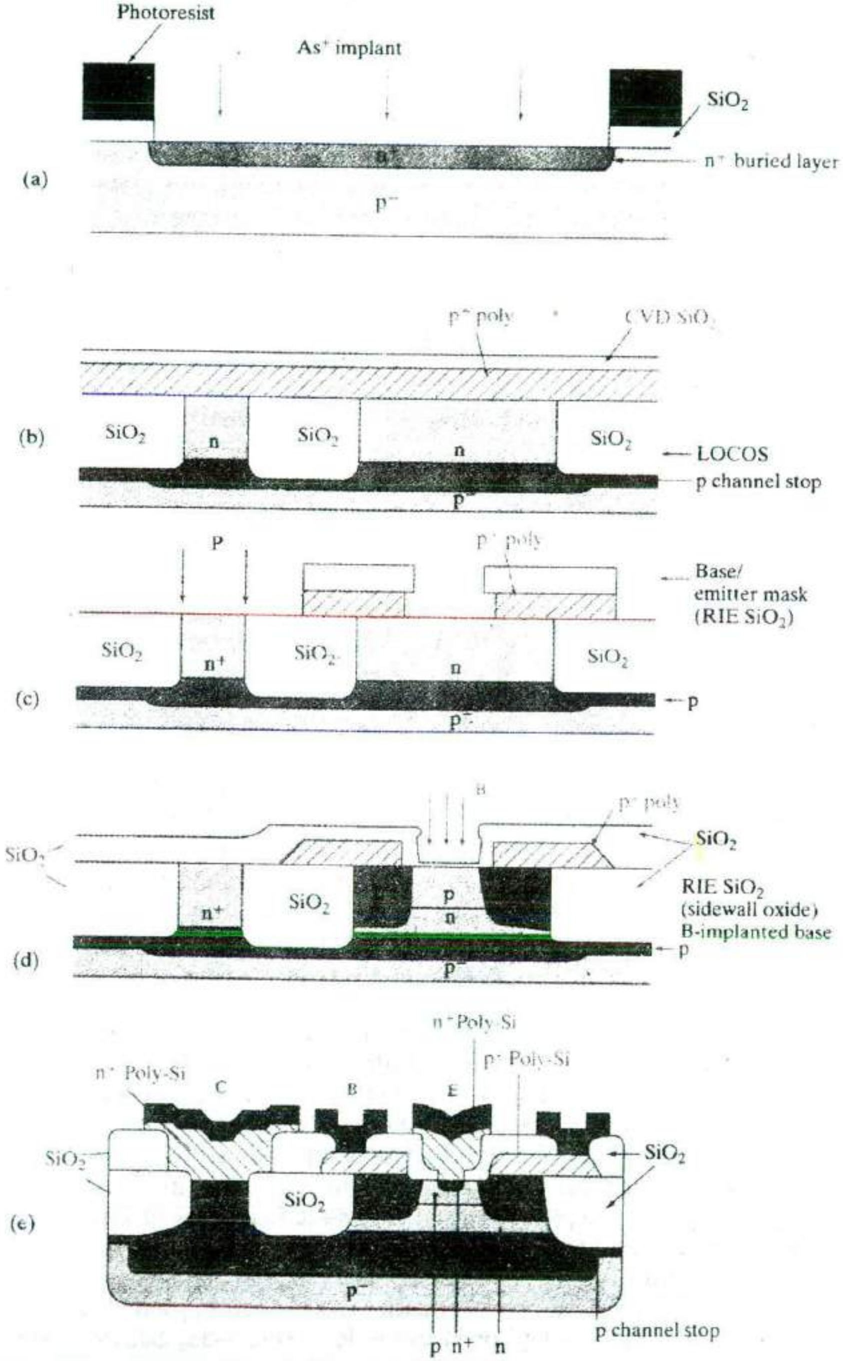


Figure 7–5 Process flow for double polysilicon, self-aligned npn BJT: (a) n$^+$ buried layer formation; (b) n epitaxy followed by LOCOS isolation; (c) base/emitter window definition and (optional) masked "sinker" implant (P) into collector contact region; (d) intrinsic base implant using self-aligned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

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

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

A polysilicon layer is deposited by LPCVD, and doped heavily p^+ with B either during deposition or subsequently by ion implantation. An oxide layer is deposited next by LPCVD. Using photolithography with the base/emitter mask, a window is etched in the polysilicon/oxide stack by RIE (Fig. 7–5c). A heavily doped "extrinsic" p^+ base is formed by diffusion of B from the doped polysilicon layer into the substrate in order to provide a low resistance, high speed base ohmic contact. An oxide layer is then deposited by LPCVD, which has the effect of closing up the base window that was etched previously, and B is implanted into this window (Fig. 7–5d). This base implant forms a more lightly p doped "intrinsic" base through which most of the current flows from the emitter to the collector. The more heavily doped extrinsic base forms a collar around the intrinsic base, and serves to reduce the base series resistance. It is critical that the base be enclosed well within the collector because otherwise it would be shorted to the p^- substrate. Finally, another LPCVD oxide layer is deposited to close up the base window further, and the oxide is etched all the way to the Si substrate by RIE, leaving oxide *spacers* on the sidewalls. Heavily n^+ doped (typically with As) polysilicon is then deposited on the substrate, patterned and etched forming *polysilicon emitter (polyemitter)* and collector contacts, as shown in Fig. 7–5e. (The use of two LPCVD polysilicon layers explains why this process is referred to as the double-polysilicon process.) Arsenic from the polysilicon is diffused into the substrate to form the n$^+$ emitter region nested within the base in a *self-aligned* manner, as well as the n$^+$ collector contact. Self alignment refers to the fact that a separate lithography step is not required to form the n$^+$ emitter region. We cleverly made use of the oxide sidewall spacers to ensure that the n$^+$ emitter region lies within the intrinsic p-type base. This is critical because otherwise the emitter gets shorted to the collector; we also want a gap between the n$^+$ emitter and the p$^+$ extrinsic base, because otherwise the emitter-base junction capacitance becomes too high. In the vertical direction, the difference between the emitter-base junction and the base-collector junction determines the base width. This is made very narrow in high gain, high speed BJTs.

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

7.4 MINORITY CARRIER DISTRIBUTIONS AND TERMINAL CURRENTS

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

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

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

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

7.4.1 Solution of the Diffusion Equation in the Base Region

Since the injected holes are assumed to flow from emitter to collector by diffusion, we can evaluate the currents crossing the two junctions by techniques used in Chapter 5. Neglecting recombination in the two depletion regions, the hole current entering the base at the emitter junction is the current I_E, and the hole current leaving the base at the collector is I_C. If we can solve for the distribution of excess holes in the base region, it is simple to evaluate the gradient of the distribution at the two ends of the base to find the currents. We shall consider the simplified geometry of Fig. 7–6, in which the base width is W_b between the two depletion regions, and the uniform cross-sectional area is A. The excess hole concentration at the edge of the emitter depletion region Δp_E and the corresponding concentration on the collector side of the base Δp_C are found from Eq. (5-29):

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

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

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

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

$$\Delta p_C \simeq -p_n \tag{7–9b}$$

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

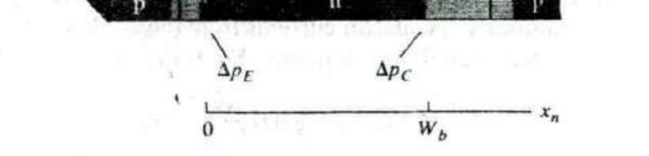


Figure 7–6 Simplified p-n-p transistor geometry used in the calculations.

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

The solution of this equation is

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

where L_p is the diffusion length of holes in the base region. Unlike the simple problem of injection into a long n region, we cannot eliminate one of the constants by assuming the excess holes disappear for large x_n. In fact, since $W_b \ll L_p$ in a properly designed transistor, most of the injected holes reach the collector at W_b. The solution is very similar to that of the narrow base diode problem. In this case the appropriate boundary conditions are

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

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

Solving for the parameters C_1 and C_2 we obtain

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

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

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

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

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

7.4.2 Evaluation of the Terminal Currents

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

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

$$\delta p(x_n) = M_1 \Delta p_E e^{-x_n/L_p} - M_2 \Delta p_E e^{x_n/L_p}$$

where

$$M_1 = \frac{e^{W_b/L_p}}{e^{W_b/L_p} - e^{-W_b/L_p}}$$

$$M_2 = \frac{e^{-W_b/L_p}}{e^{W_b/L_p} - e^{-W_b/L_p}}$$

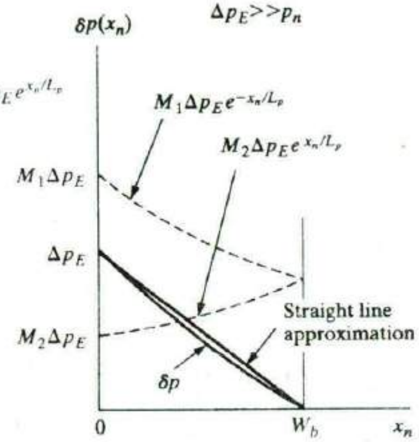


Figure 7–7 Sketch of the terms in Eq. (7–14), illustrating the linearity of the hole distribution in the base region. In this example, $W_b/L_p = \frac{1}{2}$.

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

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

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

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

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

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

$$\boxed{\begin{aligned} I_{Ep} &= qA\frac{D_p}{L_p}\left(\Delta p_E \operatorname{ctnh}\frac{W_b}{L_p} - \Delta p_C \operatorname{csch}\frac{W_b}{L_p}\right) && \text{(7–18a)} \\ I_C &= qA\frac{D_p}{L_p}\left(\Delta p_E \operatorname{csch}\frac{W_b}{L_p} - \Delta p_C \operatorname{ctnh}\frac{W_b}{L_p}\right) && \text{(7–18b)} \end{aligned}}$$

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

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

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

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

EXAMPLE 7–2

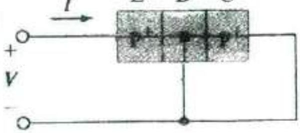
(a) Find the expression for the current I for the transistor connection shown if $\gamma = 1$.

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

SOLUTION

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

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



similarly,

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

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

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

7.4.3 Approximations of the Terminal Currents

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

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

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

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

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

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

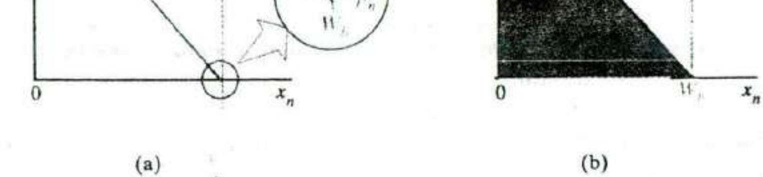


Figure 7–8
Approximate excess hole distributions in the base: (a) forward-biased emitter, reverse-biased collector; (b) triangular distribution for $V_{CB} = 0$ or for negligible p_n.

Table 7–1 Expansions of several pertinent hyperbolic functions.

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

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

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

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

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

$$Q_p \simeq \tfrac{1}{2}qA\,\Delta p_E W_b \qquad (7\text{–}23)$$

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

$$I_B \simeq \frac{Q_p}{\tau_p} = \frac{qAW_b\Delta p_E}{2\tau_p} \qquad (7\text{–}24)$$

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

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

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

7.4.4 Current Transfer Ratio

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

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

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

$$B = \frac{I_C}{I_{Ep}} = \frac{\operatorname{csch} W_b/L_p}{\operatorname{ctnh} W_b/L_p} = \operatorname{sech} \frac{W_b}{L_p} \qquad (7\text{–}26)$$

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

EXAMPLE 7–3

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

SOLUTION

From Eqs. (7–25) and (7–26), we have

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

Using the values given, and taking $L_p^n/L_n^p = \sqrt{\mu_p^n/\mu_n^p}$ for equal lifetimes,

$$\alpha = [\cosh 0.1 + \sqrt{2}(0.1)(0.5) \sinh 0.1]^{-1}$$
$$= [1.005 + 0.0707(0.1))]^{-1} = 0.988$$

We can find β from Eq. (7–6):

$$\beta = \frac{\alpha}{1-\alpha} = 82$$

Thus an incremental change in I_B causes a significant change in I_C.

7.5 GENERALIZED BIASING

The expressions derived in Section 7.4 describe the terminal currents of the transistor. if the device geometry and other factors are consistent with the assumptions. Real transistors may deviate from these approximations, as we shall see in Section 7.7. The collector and emitter junctions may differ in area, saturation current, and other parameters, so that the proper description of the terminal currents may be more complicated than Eqs. (7–18) and (7–19) suggest. For example, if the roles of emitter and collector are reversed, these equations predict that the behavior of the transistor is symmetrical. Real transistors, on the other hand, are generally not symmetrical between emitter and collector. This is a particularly important consideration when the transistor is not biased in the usual way. We have discussed normal biasing (sometimes called the *normal active* mode), in which the emitter junction is forward biased and the collector is reverse biased. In some applications, particularly in switching, this normal biasing rule is violated. In these cases it is important to account for the differences in injection and collection properties of the two junctions. In this section we shall develop a generalized approach which accounts for transistor operation in terms of a coupled-diode model, valid for all combinations of emitter and collector bias. This model involves four measurable parameters that can be related to the geometry and material properties of the device. Using this model in conjunction with the charge control approach, we can describe the physical operation of a transistor in switching circuits and in other applications.

7.5.1 The Coupled-Diode Model

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

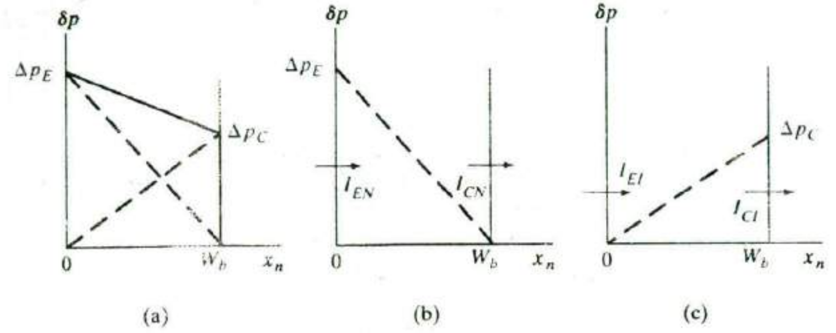


Figure 7–9 Evaluation of a hole distribution in terms of components due to normal and inverted modes: (a) approximate hole distribution in the base with emitter and collector junctions forward biased; (b) component due to injection and collection in the normal mode; (c) component due to the inverted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

injection in the *inverted mode* of injection from collector to emitter.[3] Of course, these inverted components will be negative, since they account for hole flow opposite to our original definitions of I_E and I_C.

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

$$I_{EN} = a\Delta p_E \quad \text{and} \quad I_{CN} = b\Delta p_E \quad \text{with } \Delta p_C = 0 \qquad (7\text{–}27\text{a})$$

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

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

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

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

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

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

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

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

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

the collector junction short circuited. Similarly, the collector current in the inverted mode is

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

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

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

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

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

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

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

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

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

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

$$\alpha_N I_{ES} = \alpha_I I_{CS} \tag{7–33}$$

even for nonsymmetrical transistors.

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

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

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

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

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

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

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

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

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

$$\boxed{\begin{aligned} I_E &= \alpha_I I_C + I_{EO}(e^{qV_{EB}/kT} - 1) \\ I_C &= \alpha_N I_E - I_{CO}(e^{qV_{CB}/kT} - 1) \end{aligned}} \qquad \begin{aligned}(7\text{–}37a)\\(7\text{–}37b)\end{aligned}$$

and the equivalent circuit is shown in Fig. 7–11a. In this form the equations describe both the emitter and collector currents in terms of a simple diode characteristic plus a current generator proportional to the other current. For example, under normal biasing the equivalent circuit reduces to the form shown in Fig. 7–11b. The collector current is α_N times the emitter current plus

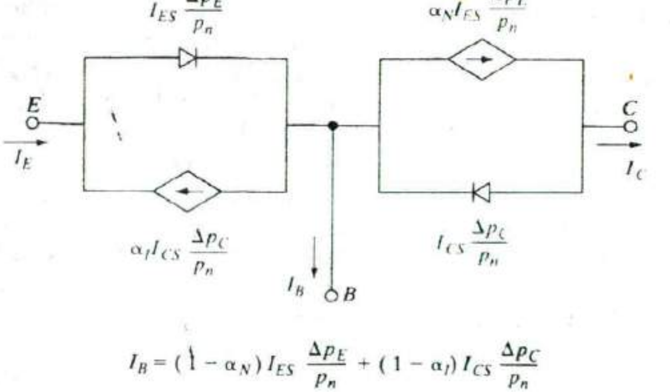

$I_{ES}\frac{\Delta p_E}{p_n}$
$\alpha_N I_{ES}\frac{\Delta p_E}{p_n}$
E
I_E
C
I_C
$\alpha_I I_{CS}\frac{\Delta p_C}{p_n}$
$I_{CS}\frac{\Delta p_C}{p_n}$
I_B
B
$I_B = (1 - \alpha_N) I_{ES}\frac{\Delta p_E}{p_n} + (1 - \alpha_I) I_{CS}\frac{\Delta p_C}{p_n}$

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

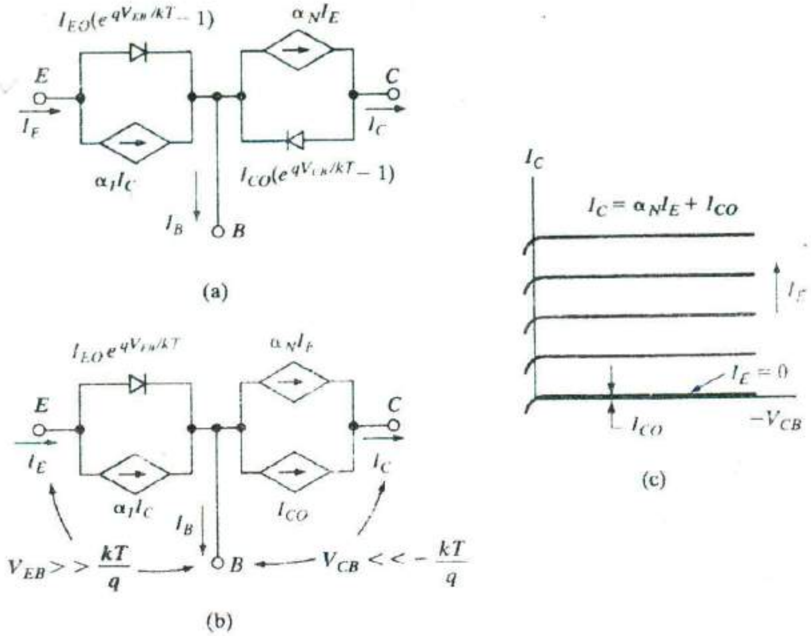


Figure 7–11 Equivalent circuits of the transistor in terms of the terminal currents and the open-circuit saturation currents: (a) synthesis of Eqs. (7–37); (b) equivalent circuit with normal biasing; (c) collector characteristics with normal biasing.

the collector saturation current, as expected. The resulting collector characteristics of the transistor appear as a series of reverse biased diode curves, displaced by increments proportional to the emitter current (Fig. 7–11c).

7.5.2 Charge Control Analysis

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

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

Similarly, the inverted components are

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

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

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

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

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

$$\begin{aligned} \alpha_N &= \frac{\tau_{pN}}{\tau_{tN} + \tau_{pN}} \quad , & \alpha_1 &= \frac{\tau_{pI}}{\tau_{tI} + \tau_{pI}} \\ I_{ES} &= q_N\left(\frac{1}{\tau_{tN}} + \frac{1}{\tau_{pN}}\right) \quad , & I_{CS} &= q_I\left(\frac{1}{\tau_{tI}} + \frac{1}{\tau_{pI}}\right) \\ Q_N &= q_N\frac{\Delta p_E}{p_n} \quad , & Q_I &= q_I\frac{\Delta p_C}{p_n} \end{aligned} \tag{7–40}$$

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

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

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

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

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

The effects of time dependence of stored charge can be included in these equations by the methods introduced in Section 5.5.1. We can include the proper dependencies by adding a rate of change of stored charge to each of the injection currents I_{EN} and I_{CI}:

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

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

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

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

7.6 SWITCHING

In a switching operation a transistor is usually controlled in two conduction states, which can be referred to loosely as the "on" state and the "off" state. Ideally, a switch should appear as a short circuit when turned on and an open circuit when turned off. Furthermore, it is desirable to switch the device from one state to the other with no lost time in between. Transistors do not fit this ideal description of a switch, but they can serve as a useful approximation in practical electronic circuits. The two states of a transistor in switching can be seen in the simple common-emitter example of Fig. 7–12. In this figure the collector current i_C is controlled by the base current i_B over most of the family of characteristic curves. The load line specifies the locus of allowable

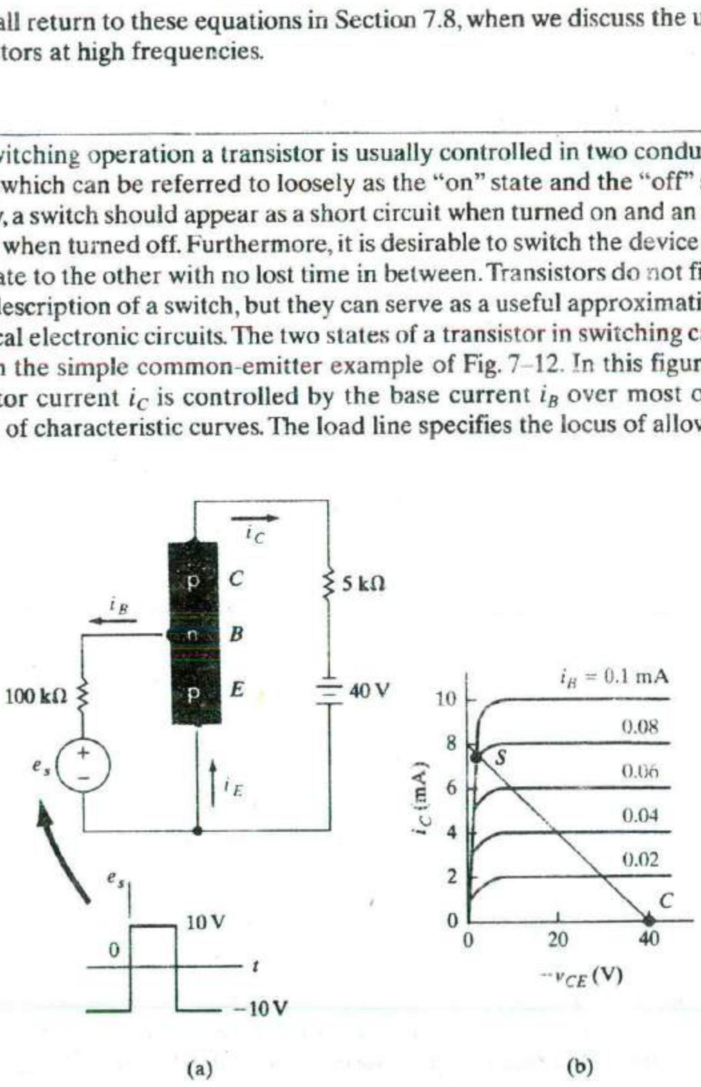


Figure 7–12 Simple switching circuit for a transistor in the common-emitter configuration: (a) biasing circuit; (b) collector characteristics and load line for the circuit, with cutoff and saturation indicated.

$(i_C, -v_{CE})$ points for the circuit, in analogy with Fig. 6-2. If i_B is such that the operating point lies somewhere between the two end points of the load line (Fig. 7–12b), the transistor operates in the normal active mode. That is, the emitter junction is forward biased and the collector is reverse biased, with a reasonable value of i_B flowing out of the base. On the other hand, if the base current is zero or negative, the point C is reached at the bottom end of the load line, and the collector current is negligible. This is the "off" state of the transistor, and the device is said to be operating in the *cutoff* regime. If the base current is positive and sufficiently large, the device is driven to the *saturation* regime, marked S. This is the "on" state of the transistor, in which a large value of i_C flows with only a very small voltage drop v_{CE}. As we shall see below, the beginning of the saturation regime corresponds to the loss of reverse bias across the collector junction. In a typical switching operation the base current swings from positive to negative, thereby driving the device from saturation to cutoff, and vice versa. In this section we shall explore the nature of conduction in the cutoff and saturation regimes; also we shall investigate the factors affecting the speed with which the transistor can be switched between the two states.

7.6.1 Cutoff

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

$$\frac{\Delta p_E}{p_n} \simeq \frac{\Delta p_C}{p_n} \simeq -1 \qquad (7\text{–}44)$$

which implies $p(x_n) = 0$. With a straight-line approximation, the excess hole distribution in the base appears constant at $-p_n$, as shown in Fig. 7–13a. Actually, there will be some slope to the distribution at each edge to account for the reverse saturation current in the junctions, but Fig. 7–13a is approximately correct. The base current i_B can be approximated for a symmetrical transistor on a charge storage basis as $-qAp_nW_b/\tau_p$. In this calculation a negative excess hole concentration corresponds to *generation* in the same way that a positive distribution indicates recombination. This expression is also obtained by applying Eq. (7–44) to Eq. (7–19) with an approximation from Table 7–1. Physically, a small saturation current flows from n to p in each reverse-biased junction, and this current is supplied by the base current i_B (which is negative when flowing into the base of a p-n-p device according to our definitions). A more general evaluation of the currents can be obtained from the Ebers–Moll equations by applying Eq. (7–44) to Eq. (7–34):

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

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

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

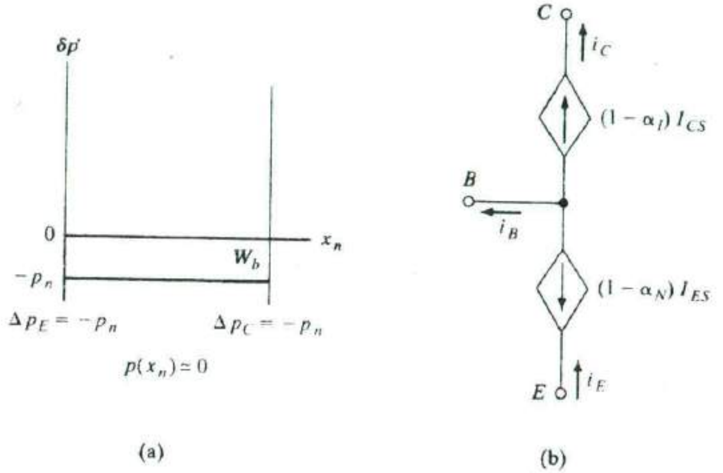


Figure 7–13 The cutoff regime of a p-n-p transistor: (a) excess hole distribution in the base region with emitter and collector junctions reverse biased; (b) equivalent circuit corresponding to Eq. (7–45).

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

7.6.2 Saturation

The saturation regime begins when the reverse bias across the collector junction is reduced to zero, and it continues as the collector becomes forward biased. The excess hole distribution in this case is illustrated in Fig. 7–14. The device is saturated when $\Delta p_C = 0$, and forward bias of the collector junction (Fig. 7–14b) leads to a positive Δp_C, driving the device further into saturation. With the load line fixed by the battery and the 5-kΩ resistor in Fig. 7–12, saturation is reached by increasing the base current i_B. We can see how a large value of i_B leads to saturation by applying the reasoning of charge control to Fig. 7–14. Since a certain amount of stored charge is required to accommodate a given i_B (and vice versa), an increase in i_B calls for an increase in the area under the $\delta p(x_n)$ distribution.

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

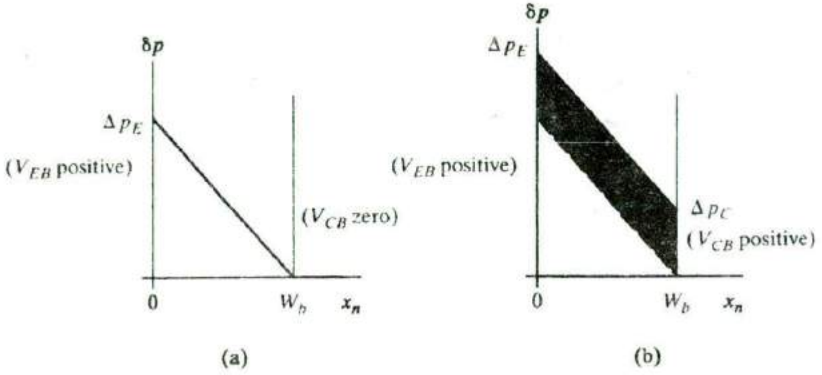


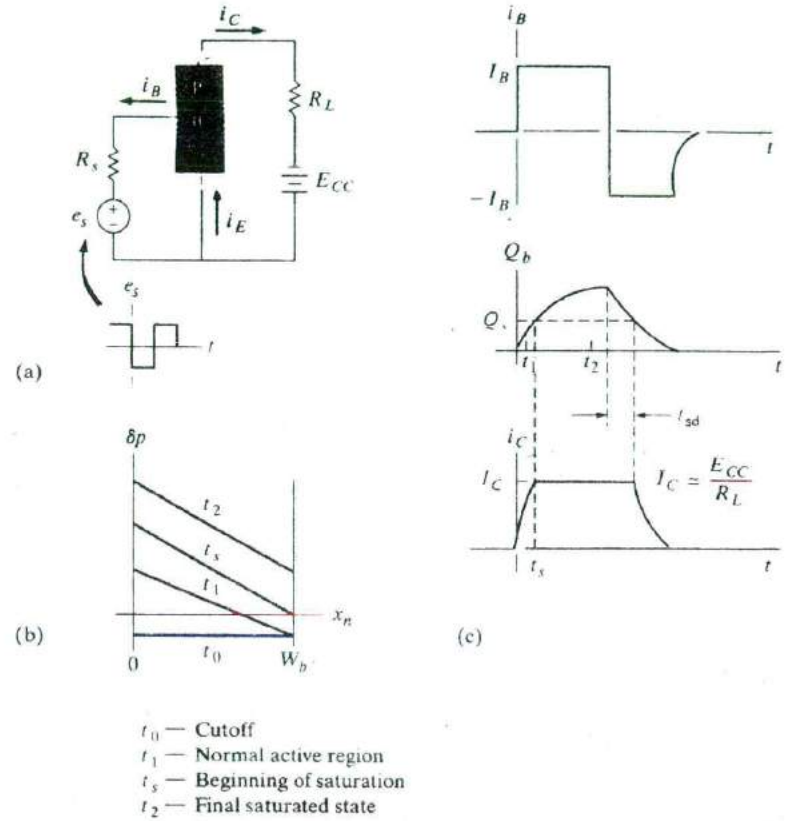
Figure 7–14 Excess hole distribution in the base of a saturated transistor: (a) the beginning of saturation; (b) oversaturation.

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

7.6.3 The Switching Cycle

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

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



below Q_s, i_C drops exponentially with the characteristic fall time. Once the stored charge is withdrawn, the base current cannot be maintained any longer at its large negative value and must decay to the small cutoff value described by Eq. (7–45c).

7.6.4 Specifications for Switching Transistors

We can determine t_s and t_{sd} by solving for the time-dependent base current, $i_B(t)$ given by an expression similar to Eq. (5–47). We must also not neglect the charging time of the emitter junction capacitance in going from cutoff to saturation. Since the emitter junction is reverse biased in cutoff, it is necessary for the emitter space charge layer to be charged to the forward bias condition before collector current can flow. Therefore, we should include a *delay time* t_d as in Fig. 7–16 to account for this effect. Typical values of t_d are given

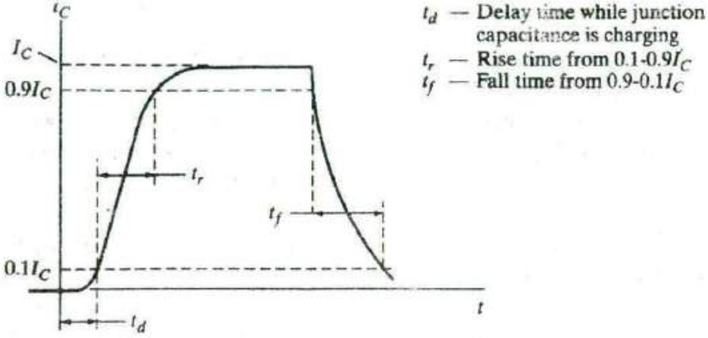


Figure 7–16 Collector current during switching transients, including the delay time required for charging the junction capacitance; definitions of the rise time and fall time.

in the specification information of most switching transistors, along with a *rise time* t_r, defined as the time required for the collector current to rise from 10 to 90 per cent of its final value. A third specification is the *fall time* t_f required for i_C to fall through a similar fraction of its turn-off excursion.

7.7 OTHER IMPORTANT EFFECTS

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

In this section we shall consider the effects of nonuniform doping in the base region of the transistor. In particular, we shall find that graded doping can lead to a drift component of charge transport across the base, adding to the diffusion of carriers from emitter to collector. We shall discuss the effects of large reverse bias on the collector junction, in terms of widening the space charge region about the junction and avalanche multiplication. We shall see that transistor parameters are affected at high current levels by the degree of injection and by heating effects. We shall consider several structural effects that are important in practical devices, such as asymmetry in the areas of the emitter and collector junctions, series resistance between the base contact and the active part of the base region, and nonuniformity of injection at the emitter junction. All these effects are important in understanding the operation of transistors, and proper consideration of their interactions can contribute greatly to the usefulness of practical transistor circuits.

7.7.1 Drift in the Base Region

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

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

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

Therefore, the built-in electric field is

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

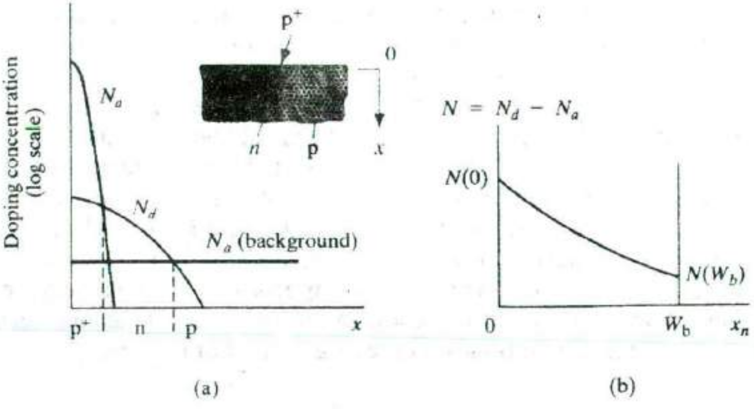


Figure 7–17 Graded doping in the base region of a p-n-p transistor: (a) typical doping profile on a semilog plot; (b) approximate exponential distribution of the net donor concentration in the base region on a linear plot.

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

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

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

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

$$\mathscr{E}(x_n) = \frac{kT}{q}\frac{a}{W_b} \tag{7–49}$$

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

7.7.2 Base Narrowing

In the discussion of transistors thus far, we have assumed that the effective base width W_b is essentially independent of the bias voltages applied to the collector and emitter junctions. This assumption is not always valid; for example, the p$^+$-n-p$^+$ transistor of Fig. 7–18 is affected by the reverse bias applied to the collector. If the base region is lightly doped, the depletion region at the reverse-biased collector junction can extend significantly into the n-type base region. As the collector voltage is increased, the space charge layer takes up more of the metallurgical width of the base L_b, and as a result, the effective base width W_b is decreased. This effect is variously called *base narrowing, base-width modulation*, and the *Early effect* after J.M. Early, who first interpreted it. The effects of base narrowing are apparent in the collector characteristics for the common-emitter configuration (Fig. 7–18b). The decrease in W_b causes β to increase. As a result, the collector current I_C increases with collector voltage rather than staying constant as predicted from the simple treatment. The slope introduced by the Early effect is almost linear with I_C, and the common-emitter characteristics extrapolate to an intersection with the voltage axis at V_A, called the Early voltage.

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

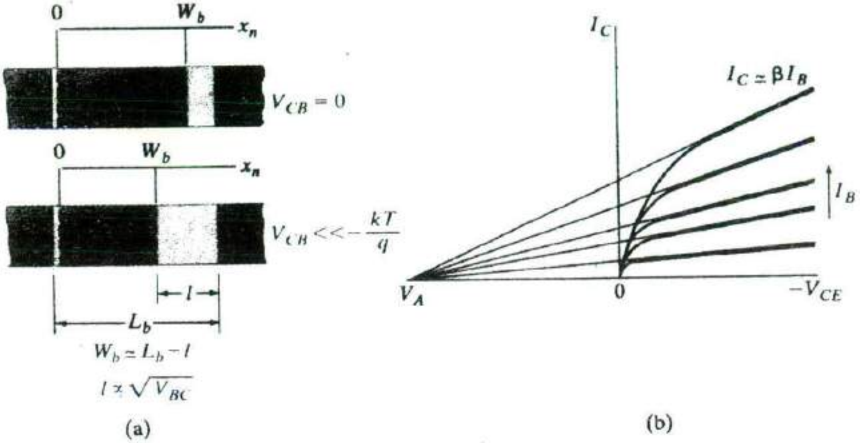


Figure 7–18
The effects of base narrowing on the characteristics of a p⁺-n-p⁺ transistor: (a) decrease in the effective base width as the reverse bias on the collector junction is increased; (b) common-emitter characteristics showing the increase in I_C with increased collector voltage. The black lines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

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

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

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

7.7.3 Avalanche Breakdown

Before punch-through occurs in most transistors, avalanche multiplication at the collector junction becomes important (see Section 5.4.2). As Fig. 7–19 indicates, the collector current increases sharply at a well-defined breakdown voltage BV_{CBO} for the common-base configuration. For the common-emitter case, however, there is a strong influence of carrier multiplication over a fairly broad

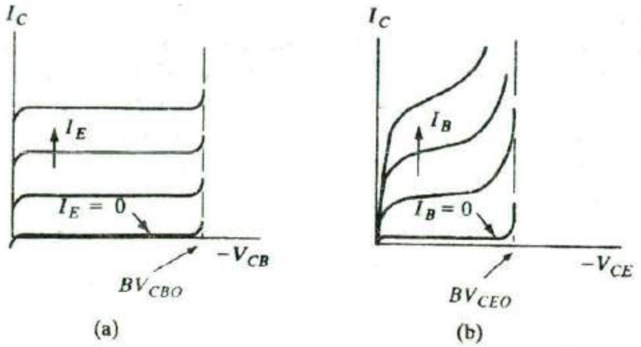


Figure 7–19 Avalanche breakdown in a transistor: (a) common-base configuration; (b) common-emitter configuration.

range of collector voltage. Furthermore, the breakdown voltage in the common-emitter case BV_{CEO} is significantly smaller than BV_{CBO}. We can understand these effects by considering breakdown for the condition $I_E = 0$ in the common-base case and for $I_B = 0$ in the common-emitter case. These conditions are implied by the O in the subscripts of BV_{CEO} and BV_{CBO}. In each case the terminal current I_C is the current entering the collector depletion region multiplied by the factor M. Including multiplication due to impact ionization, Eq.(7–37b) becomes

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

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

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

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

We notice that in this case the collector current increases indefinitely when $M\alpha_N$ approaches unity. By contrast, M must approach infinity in the common-base case before BV_{CBO} is reached. Since α_N is close to unity in most transistors, M need be only slightly larger than unity for Eq. (7–52) to approach breakdown. Avalanche multiplication thus dominates the current in a common-emitter transistor well below the breakdown voltage of the isolated collector junction. The sustaining voltage for avalanching in the common-emitter case BV_{CEO} is therefore smaller than BV_{CBO}.

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

7.7.4 Injection Level; Thermal Effects

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

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

Large values of I_C may be accompanied by significant power dissipation in the transistor and therefore heating of the device. In particular, the product of I_C and the collector voltage V_{BC} is a measure of the power dissipated at the collector junction. This dissipation is due to the fact that carriers swept through the collector junction depletion region are given increased kinetic energy, which in turn is given up to the lattice in scattering collisions. It is very important that the transistor be operated in a range such that $I_C V_{BC}$ does not exceed the maximum power rating of the device. In devices designed for high power capability, the transistor is mounted on an efficient heat sink, so that thermal energy can be transferred away from the junction.

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

7.7.5 Base Resistance and Emitter Crowding

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

If the transistor is designed so that the n-type regions leading from the base to the contacts are large in cross-sectional area, the base resistance r_b may be negligible. On the other hand, the distributed resistance r_b' along the thin base region is almost always important.[5] Since the width of the base between emitter and collector is very narrow, this distributed resistance is usually quite high. Therefore, as base current flows from points within the base region toward each end, a voltage drop occurs along r_b'. In this case the forward bias across the emitter-base junction is not uniform, but instead varies with position according to the voltage drop in the distributed base resistance. In particular, the forward bias of the emitter junction is largest at the corner of the emitter region near the base contact. We can see that this is the case by considering the simplified example of Fig. 7–20e. Neglecting variations in the base current along the path from point A to the contact B, the forward bias of the emitter junction above point A is approximately

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

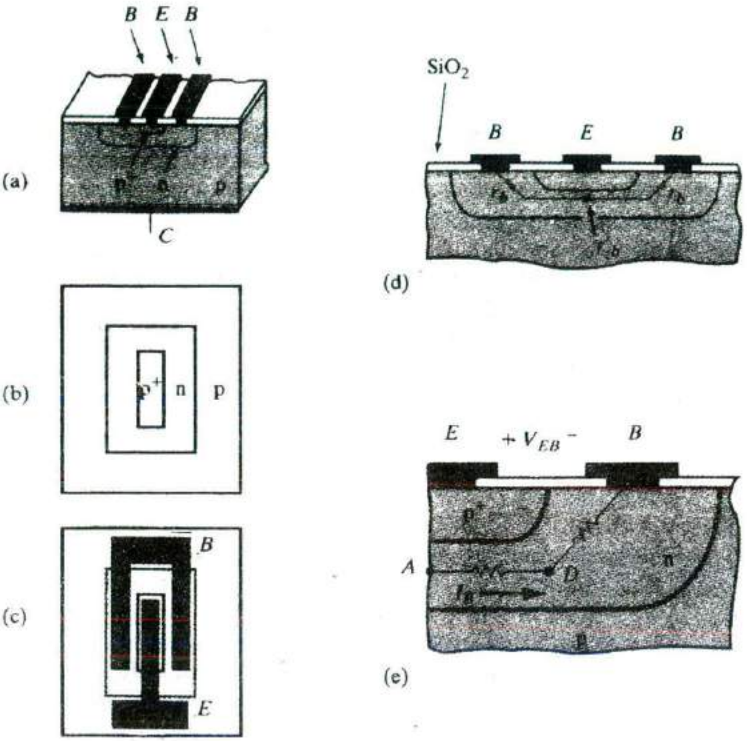


Figure 7–20 Effects of a base resistance: (a) cross section of an implanted transistor; (b) and (c) top view, showing emitter and base areas and metallized contacts; (d) illustration of base resistance; (e) expanded view of distributed resistance in the active part of the base region.

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

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

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

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

Since the forward bias is largest at the edge of the emitter, it follows that the injection of holes is also greatest there. This effect is called *emitter crowding*, and it can strongly affect the behavior of the device. The most important result of emitter crowding is that high-injection effects described in the previous section can become dominant locally at the corners of the emitter before the overall emitter current is very large. In transistors designed to handle appreciable current, this is a problem which must be dealt with by proper structural design. The most effective approach to the problem of emitter crowding is to distribute the emitter current along a relatively large emitter edge, thereby reducing the current density at any one point. Clearly, what

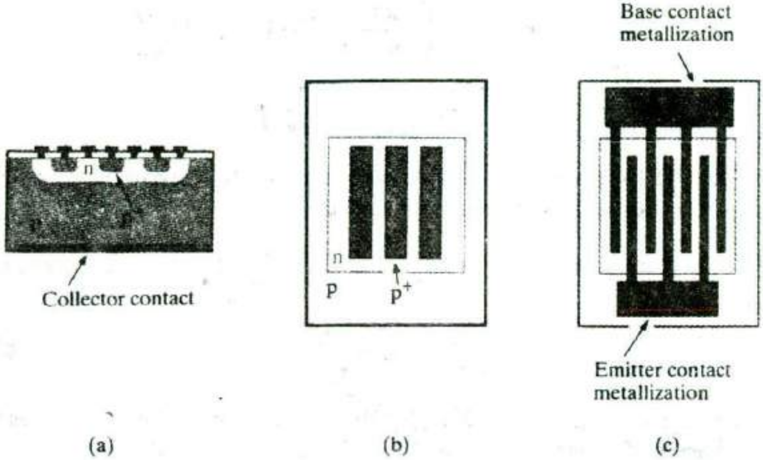


Figure 7–21 An interdigitated geometry to compensate for the effects of emitter crowding in a power transistor: (a) cross section; (b) top view of implanted regions; (c) top view with metallized contacts. The metal interconnections are isolated from the device by an oxide layer except where they contact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

is needed is an emitter region with a large perimeter compared with its area. A likely geometry to accomplish this is a long thin stripe for the emitter, with base contacts on each side (Fig. 7–20b and c). With this geometry the total emitter current I_E is spread out along a rather long edge on each side of the stripe. An even better geometry is several emitter stripes, connected electrically by the metallization and separated by interspersing base contacts (Fig. 7–21). Many such thin emitter and base contact "fingers" can be interlaced to provide for handling large current in a power transistor. This is often called very descriptively an *interdigitated* geometry.

7.7.6 Gummel–Poon Model

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

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

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

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

$$\begin{aligned} I_{Ep} &= qA\mu_p p\left(\frac{-kT}{q}\,\frac{1}{n}\,\frac{dn}{dx_n}\right) - qAD_p\frac{dp}{dx_n} \\ &= \frac{qAD_p}{n}\left(p\frac{dn}{dx_n} + n\frac{dp}{dx_n}\right) \end{aligned} \tag{7–56}$$

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

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

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

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

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

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

$$pn = n_i^2 \tag{7–59a}$$

to the non-equilibrium expression

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hence, for high V_{EB},

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

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

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

$$I_B \propto I_{En} \propto e^{\frac{qV_{EB}}{\mathbf{n}kT}} \tag{7–67a}$$

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

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

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

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

The transistor collector current I_C and the base current I_B are plotted on a semi-log scale as a function of V_{EB} in Fig. 7–22a. This is referred to as a *Gummel* plot. The current gain β is shown as a function of I_C in Fig. 7–22b. We see the dependence of β on I_C in the different bias regions, as described by the Gummel–Poon model. At low injection levels β is degraded by poor emitter injection efficiency [Eq. (7–68)], and at high currents β decreases due to excess majority charge in the base, which degrades γ.

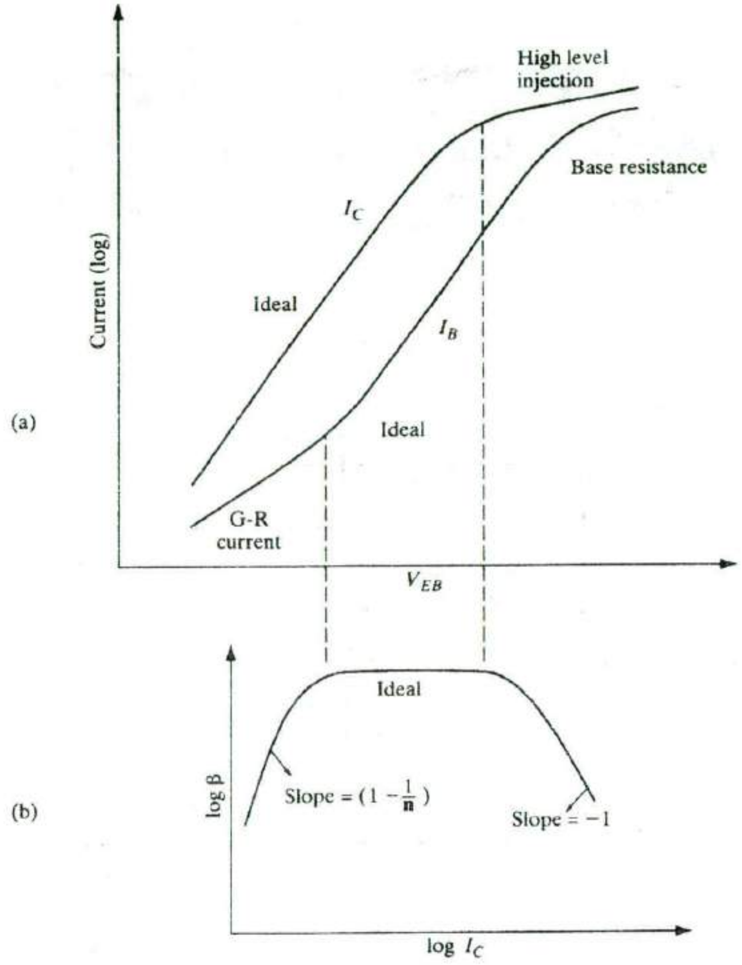


Figure 7–22 Current-voltage characteristics of BJTs: (a) Gummel plot of log of collector and base currents as a function of emitter-base forward bias; (b) d-c common-emitter current gain ($=I_C/I_B$) as a function of I_C. For the intermediate range, we have ideal behavior where both I_C and I_B increase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

7.7.7 Kirk Effect

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

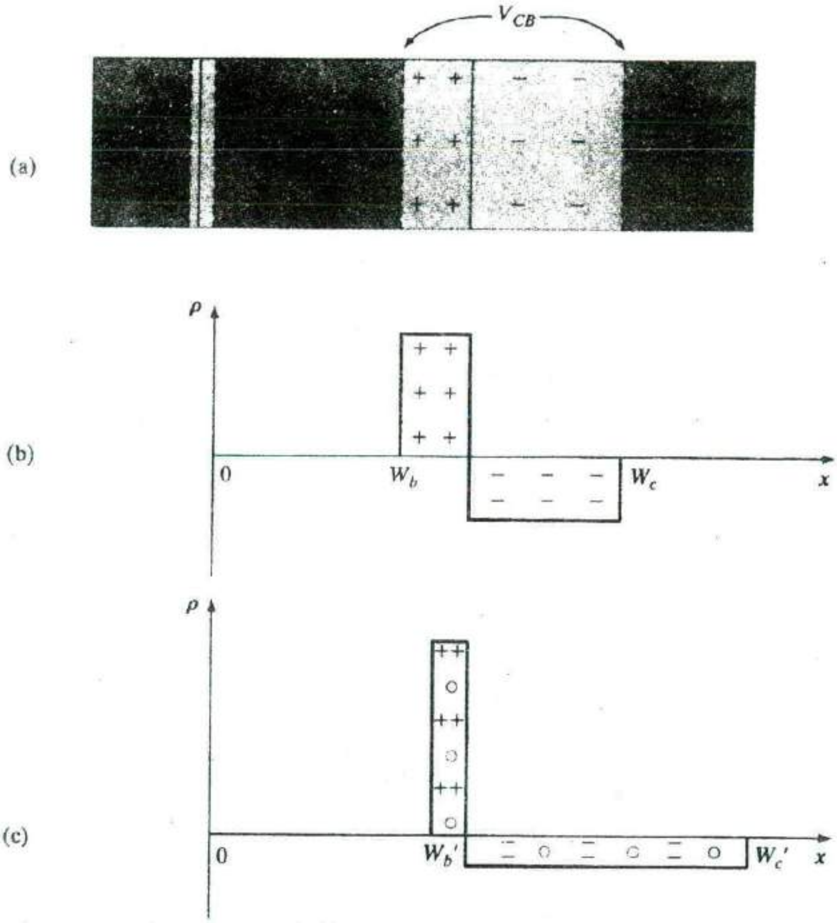


Figure 7–23 Kirk effect: (a) cross-section of p-n-p BJT; (b) space charge distribution in the base-collector reverse biased junction for very low currents; (c) space-charge distribution at the base-collector junction for higher current levels. We see that the injected mobile holes (shown in color) add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

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

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

$$\frac{d\mathscr{E}}{dx} = \frac{1}{\epsilon}\left[q(N_d^+ - N_a^-) + \frac{I_c}{A\mathrm{v}_d}\right] \qquad (7\text{–}69)$$

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

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

$$V_{CB} = -\int_{W_b}^{W_c} \mathscr{E} dx \qquad (7\text{–}70)$$

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

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

7.8 FREQUENCY LIMITATIONS OF TRANSISTORS

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

7.8.1 Capacitance and Charging Times

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

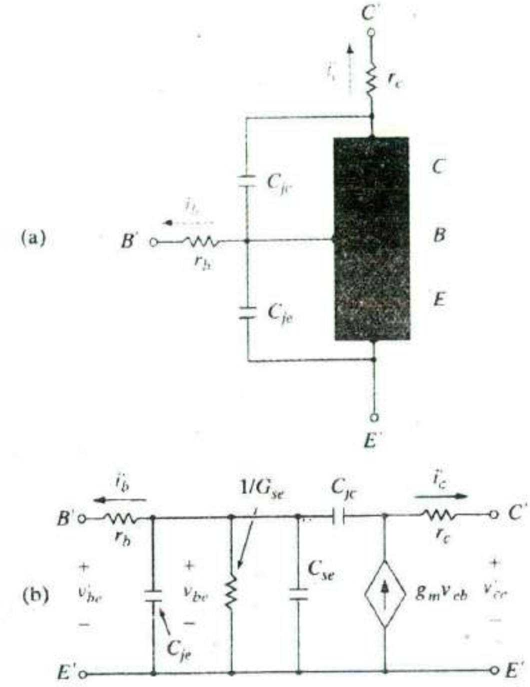


Figure 7–24 Models for a-c operation: (a) inclusion of base and collector resistances and junction capacitances; (b) hybrid-pi model systhesizing Eqs. (7–75) and (7–76).

a series collector resistance.[6] Clearly, the combinations of r_b with C_{je} and r_c with C_{jc} can introduce important time constants into a-c circuit applications of the device.

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

If a small a-c signal is applied to the emitter p-n junction along with a d-c level, we can show (Prob. 7-27) that

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

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

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

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

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

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

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

$$i_B(t) = \frac{Q_N(t)}{\tau_p} + \frac{dQ_N(t)}{dt} \qquad (7\text{–}74a)$$

As discussed in Section 5.5.4, we must be careful about boundary conditions in determining where the stored charges are extracted or "reclaimed" in a diode. For the emitter-base diode in a BJT, we have a "short" diode where Eq. (5–67b) is applicable. The result is that only 2/3 of the stored charge is reclaimed. Hence, we obtain

$$i_B(t) = I_B + \frac{q}{kT}I_Bv_{eb} + \frac{2}{3}\,\frac{q}{kT}I_B\tau_p\frac{dv_{eb}}{dt} \qquad (7\text{–}74b)$$

The a-c component of the base current is

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

where

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

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

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

$$i_c = g_mv_{eb} \quad \text{where } g_m \equiv \frac{q}{kT}\,\beta I_B = \frac{3}{2}\,\frac{C_{se}}{\tau_t} \qquad (7\text{–}76)$$

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

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

7.8.2 Transit Time Effects

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

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

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

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

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

The transit time is

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

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

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

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

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

7.8.3 Webster Effect

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

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

7.8.4 High-Frequency Transistors

The most obvious generality we can make about the fabrication of high-frequency transistors is that the physical size of the device must be kept small. The base width must be narrow to reduce the transit time, and the emitter and collector areas must be small to reduce junction capacitance. Unfortunately, the requirement of small size generally works against the requirements of power rating for the device. Since we usually require a trade-off between frequency and power, the dimensions and other design features of the transistor must be tailored to the specific circuit requirements. On the other hand, many of the fabrication techniques useful for power devices can be adapted to increase the frequency range. For example, the method of interdigitation (Fig. 7–21) provides a means of increasing the useful emitter edge length while keeping the overall emitter area to a minimum. Therefore, some form of interdigitation is generally used in transistors designed for high frequency and reasonable power requirements (Fig. 7–25).

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

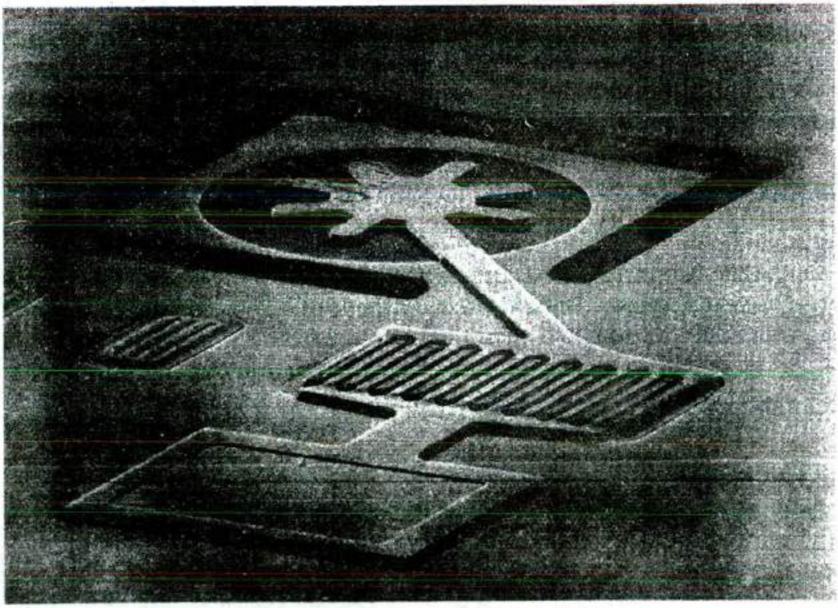


Figure 7–25
A low-noise Si bipolar transistor with f_T = 8 GHz. This device has 9 interdigitated emitter stripes, each 1 μm X 20 μm. (Photograph courtesy of Motorola.)

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

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

7.9 HETEROJUNCTION BIPOLAR TRANSISTORS

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

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

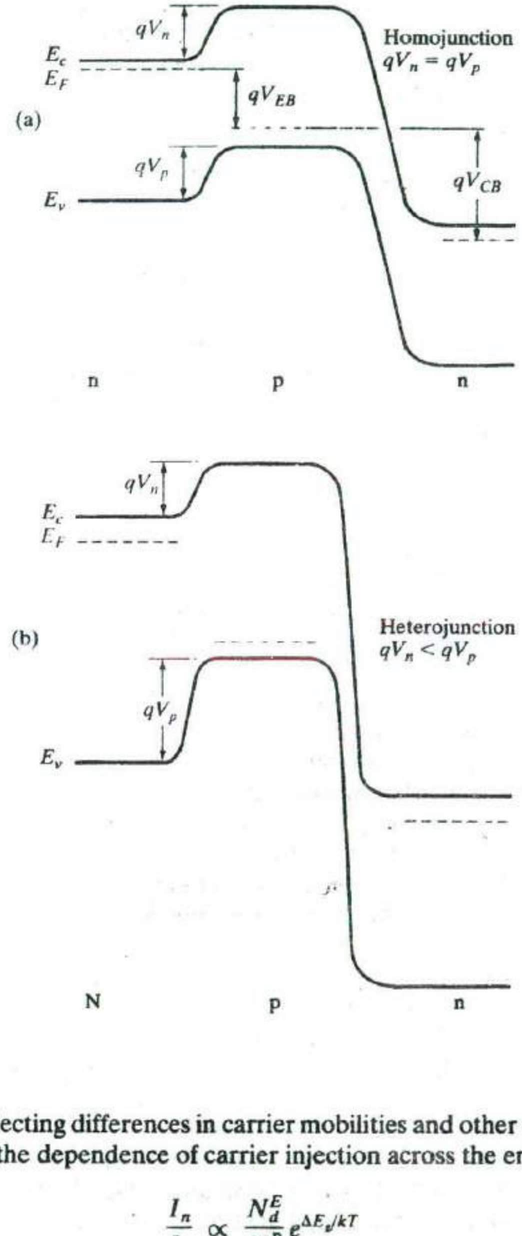


Figure 7–26 Contrast of carrier injection at the emitter of (a) a homojunction BJT and (b) a heterojunction bipolar transistor (HBT). In the forward-biased homojunction emitter, the electron barrier qV_n and the hole barrier qV_p are the same. In the HBT with a wide band gap emitter, the electron barrier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

junction. Neglecting differences in carrier mobilities and other effects, we can approximate the dependence of carrier injection across the emitter as

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

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

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

The heterojunction shown in Fig. 7–26 has a smooth barrier, without the spike and notch commonly observed for heterojunctions (see Fig. 5-46). It is possible to smooth out such discontinuities in the bands by grading the composition of the ternary or quaternary alloy between the two materials (Fig. 7–27). Clearly, grading out the conduction band spike improves the electron injection by reducing the barrier that electrons must overcome. There are some HBT designs, however, that make use of the spike as a "launching ramp" to inject hot electrons into the base.

Materials commonly used in HBTs obviously include the AlGaAs/GaAs system because of its wide range of lattice-matched composition. In addition, the InGaAsP system (including $In_{0.53}Ga_{0.47}As$) grown on InP has become popular in HBT design. InGaAs has much lower surface recombination than GaAs, and the Γ-*L* and Γ-*X* intervalley band separations are much larger than in GaAs (see Fig. 3-6). The lower rate of surface recombination reduces loss of injected carriers at the surface of the device. This is a particularly important effect in small-geometry devices, in which the base perimeter-to-area ratio, and thus the intersection of the base with the surface, is large. The larger intervalley band separation in InGaAs helps ensure that the electrons remain in the low-mass (high mobility) Γ valley during field-enhanced transport through the base region.

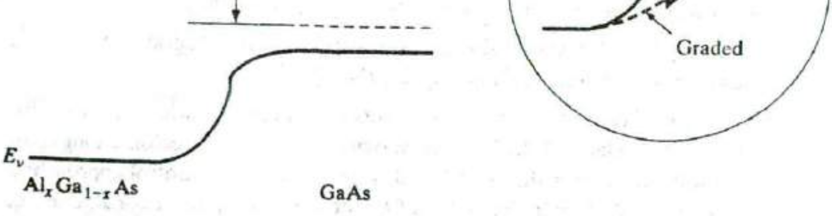


Figure 7–27
Removal of the conduction band spike by grading the alloy composition (x) in the heterojunction. In this example the junction is graded from the composition used in the AlGaAs emitter to $x = 0$ at the GaAs base. This grading typically takes place over a distance of 100 Å or less.

It is also possible to make HBTs using elemental semiconductor heterostructures such as Si/$Si_{1-x}Ge_x$. In this material system, the band gap difference ΔE_g between the Si emitter and the narrower band gap $Si_{1-x}Ge_x$ base occurs primarily in the valence band. As a result, a rather small addition of Ge to the base results in higher electron injection efficiency than is possible in homojunction silicon bipolar transistors.

If the alloy composition in the base of an n-p-n is varied such that E_g decreases slightly from the emitter side to the collector side of the base, a built-in electric field accelerates electrons through the base region. The resulting field-aided base transport is a major advantage of the HBT.

PROBLEMS

7.1 A bipolar junction transistor is fabricated by ion implantation as follows. A 50 keV boron ion implant (dose = 5×10^{14} cm^{-2}) is performed into an n-type silicon substrate with $N_d = 2 \times 10^{15}$ cm^{-3}, followed by a 30 keV phosphorus implant (dose = 1×10^{15} cm^{-2}). A rapid thermal anneal is then performed at 1100°C for 10 s in a N_2 environment. Assume the implants are 100 percent activated by this anneal. The diffusivity of boron at 1100°C can be assumed constant with $D = 10^{-12}$ cm^2/s, and for phosphorus D = 5×10^{-13} cm^2/s. Assume that dopants can outdiffuse from the substrate.

(a) Where are the peaks and what are the widths (at 0.607 of the peak) of the boron and phosphorus implant profiles prior to annealing?

(b) Where are the peaks and what are the widths of the implant profiles after annealing?

(c) What is the emitter junction depth and the collector junction depth (as measured from the silicon surface) after annealing?

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

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

7.4 Derive Eq. (7–19) from the charge control approach by integrating Eq. (7–11) across the base region and applying Eq. (7–13).

7.5 Extend Eq. (7–20a) to include the effects of nonunity emitter injection efficiency ($\gamma < 1$). Derive Eq. (7–25) for γ. Assume the emitter region is long compared with an electron diffusion length. Apply the charge control approach to the linear distribution in Fig. 7–8b to find the base transport factor B. Use Eq. (7–80) for the transit time. Is the result the same as Eq. (7–26) for small W_b/L_p?

7.6 A symmetrical p^+-n-p^+ Si bipolar transistor has the following properties:

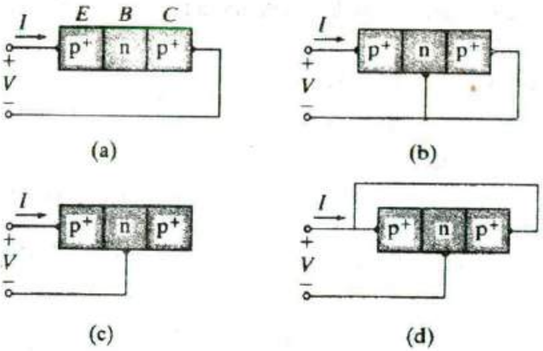
	Emitter	Base
$A = 10^{-4}\ \text{cm}^{-2}$	$N_a = 10^{17}$	$N_d = 10^{15}\ \text{cm}^{-3}$
$W_b = 1\ \mu\text{m}$	$\tau_n = 0.1\ \mu\text{s}$	$\tau_p = 10\ \mu\text{s}$
	$\mu_p = 200$	$\mu_n = 1300\ \text{cm}^2/\text{V-s}$
	$\mu_n = 700$	$\mu_p = 450$

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

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

(c) Calculate the emitter injection efficiency γ and the amplification factor β, assuming the emitter region is long compared to L_n.

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



(a) (b) (c) (d)

Figure P7–7

7.8 For the transistor connection in Fig. P7-7a, (a) show that $V_{EB} = (kT/q)\ \ln 2$; (b) find the expression for I when $V \gg kT/q$ and sketch I vs. V.

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

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

7.10 Suppose that V is negative in Fig. P7-7c.

(a) Find I from the Ebers–Moll equations.

(b) Find the expression for V_{CB}.

(c) Sketch $\delta p(x_n)$ in the base.

7.11 For the transistor connection of Fig. P7-7d, (a) find the expression for $\delta p(x_n)$ in the base region; (b) find the current I.

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

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

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

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

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

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

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

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

7.15 Use Example 5-6 to design an N-p-n heterojunction bipolar transistor with reasonable γ and base resistance.

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

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

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

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

7.17 Calculate and plot the common-base characteristics at 300 K of a symmetrical Si p-n-p BJT with a base area of 10^2 cm^{-2}, a base width of 500 Å, base doping of 10^{14} cm^{-3}, emitter and collector doping of 10^{17} cm^{-3}, and carrier lifetimes $\tau_n = \tau_p = 10^{-6}$ s, for emitter currents of 0, 0.04, 0.08, and 0.12 A.

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

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

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

$\tau_n = \tau_p = 0.1$ μs

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

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

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

$N_C = 10^{16}$ cm^{-3} = collector concentration

W_E = emitter width = 3 μm

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

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

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

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

7.21 For the BJT in Prob. 7.19, calculate α, β, I_E, I_B and I_C for the two values of V_{EB}. What is the base Gummel number in each case?

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

Emitter	Base	Collector
$N_a = 5 \times 10^{18}\ \mathrm{cm}^{-3}$	$N_d = 10^{16}$	$N_a = 10^{15}$
$\tau_n = 1\mu\mathrm{s}$	$\tau_p = 25$	$\tau_n = 2$
$\mu_n = 150\ \mathrm{cm}^2/V\text{-}s$	$\mu_n = 1500$	$\mu_n = 1500$
$\mu_p = 100\ \mathrm{cm}^2/V\text{-}s$	$\mu_p = 400$	$\mu_p = 450$

Base width, $W_b = 0.2\ \mu\mathrm{m}$

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

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

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

7.24 Three n-p-n transistors are identical except that transistor #2 has a base region twice as long as transistor #1, and transistor #3 has a base region doped twice as heavily as transistor #1. All other dopings and lengths are identical for the three transistors. Which transistors have the largest value of each parameter listed below?

Give clear mathematical reasons for each of your answers.

(a) Emitter injection efficiency

(b) Base transport factor

(c) Punch through voltage

(d) Collector junction capacitance with V_{CB} reverse biased at 10V.

(e) Common emitter current gain.

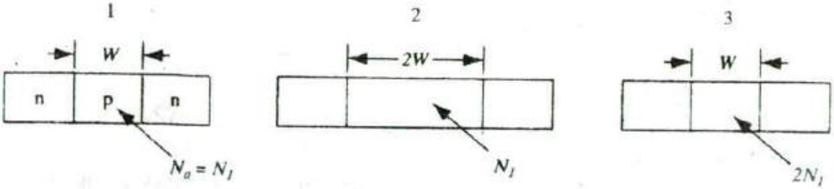


Figure P7–24

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

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

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

7.27 Derive Eq. (7–71) for $\Delta p_E(t)$ assuming that the emitter has an applied voltage

$$v_{EB}(t) = V_{EB} + v_{eb}(t)$$

where $V_{EB} \gg kT/q$. For $v_{eb} \ll kT/q$, the approximation $e^x \simeq 1 + x$ can be employed.

READING LIST

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

Inoue, K. "Recent Advances in InP-Based HEMT/HBT Device Technology." *Fourth International Conference on Indium Phosphide and Related Materials* (April 1992): 10–13.

Jaeger, R.C. *Modular Series on Solid State Devices: Vol V. Introduction to Microelectronic Fabrication.* Reading, MA: Addison-Wesley, 1988.

Levi, A. F. J., R. N. Nottenburg, and Y. K. Chen. "Ultrahigh-Speed Bipolar Transistors." *Physics Today* 43 (February 1990): 58– 64.

Li, S. S. *Semiconductor Physical Electronics.* New York: Plenum Press, 1993.

Morgan, D. V., and R. H. Williams, eds. *Physics and Technology of Heterojunction Devices.* London: P. Peregrinus, 1991.

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.

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

Pavlidis, D. "Current Status of Heterojunction Bipolar and High-Electron Mobility Transistor Technologies." *Microelectronic Engineering* 19 (September 1992): 305– 12.

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

Shur, M. *GaAs Devices and Circuits.* New York: Plenum Press, 1987.

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

Sze, S. M. *High-Speed Semiconductor Devices.* New York: Wiley, 1990.

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

Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

Chapter 8

Optoelectronic Devices

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

8.1 PHOTODIODES

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

8.1.1 Current and Voltage in an Illuminated Junction

In Chapter 5 we identified the current due to drift of minority carriers across a junction as a generation current. In particular, carriers generated within the depletion region W are separated by the junction field, electrons being collected in the n region and holes in the p region. Also, minority carriers generated thermally within a diffusion length of each side of the junction diffuse to the depletion region and are swept to the other side by the electric field. If the junction is uniformly illuminated by photons with $h\nu > E_g$, an added generation rate g_{op} (EHP/cm^3-s) participates in this current (Fig. 8–1). The number of holes created per second within a diffusion length of the transition region on the n side is AL_pg_{op}. Similarly AL_ng_{op} electrons are generated per second within L_n of x_{p0} and AWg_{op} carriers are generated within W. The resulting current due to collection of these optically generated carriers by the junction is

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

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

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

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

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

When the device is short circuited ($V = 0$), the terms from the diode equation cancel in Eq. (8–2), as expected. However, there is a short-circuit current from n to p equal to I_{op}. Thus the I–V characteristics of Fig. 8–1c cross the I-axis at negative values proportional to g_{op}. When there is an open circuit across the device, $I = 0$ and the voltage $V = V_{oc}$ is

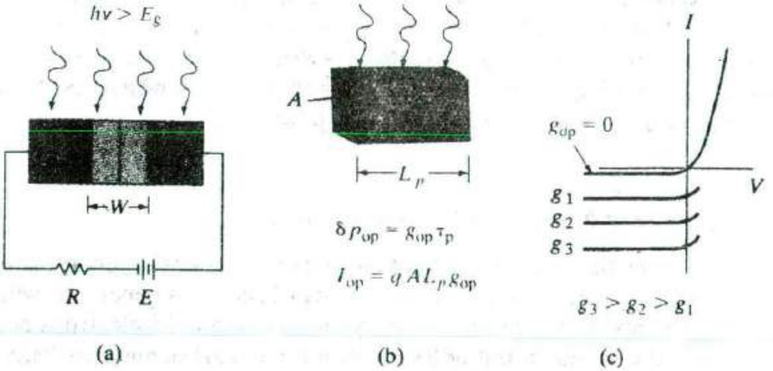


Figure 8–1 Optical generation of carriers in a p-n junction: (a) absorption of light by the device; (b) current I_{op} resulting from EHP generation within a diffusion length of the junction on the n side; (c) I–V characteristics of an illuminated junction.

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

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

$$V_{oc} \simeq \frac{kT}{q}\ln\frac{g_{op}}{g_{th}} \quad \text{for } g_{op} \gg g_{th} \qquad (8\text{–}4)$$

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

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

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

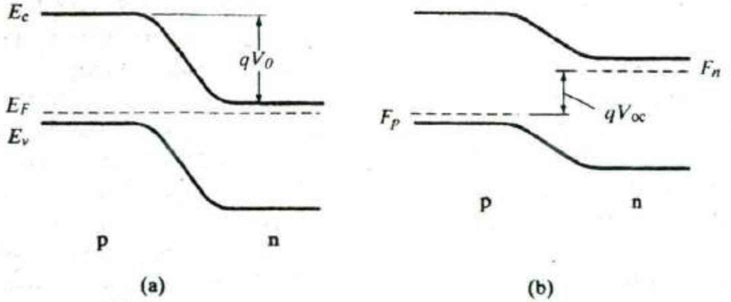


Figure 8–2 Effects of illumination on the open circuit voltage of a junction: (a) junction at equilibrium; (b) appearance of a voltage $V_{oc}$with illumination.

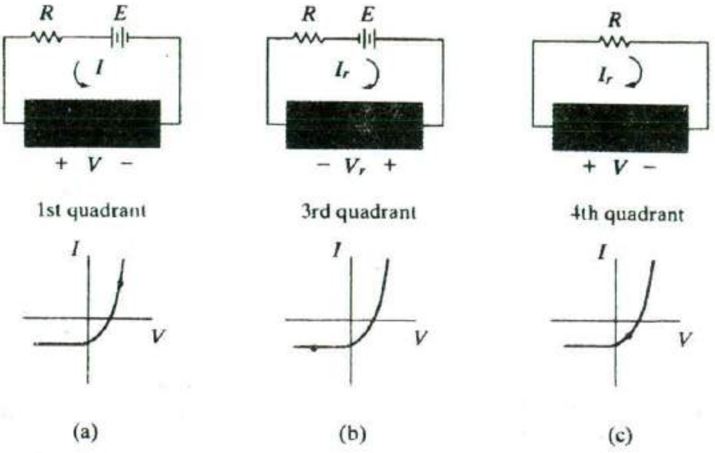


Figure 8–3 Operation of an illuminated junction in the various quadrants of its I–V characteristic; in (a) and (b) power is delivered to the device by the external circuit; in (c) the device delivers power to the load.

8.1.2 Solar Cells

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

To utilize a maximum amount of available optical energy, it is necessary to design a solar cell with a large area junction located near the surface of the device (Fig. 8–5). The planar junction is formed by diffusion or ion implantation, and the surface is coated with appropriate materials to reduce reflection and to decrease surface recombination. Many compromises must be made in solar cell design. In the device shown in Fig. 8–5, for example, the junction depth d must be less than L_p in the n material to allow holes generated near the surface to diffuse to the junction before they recombine; similarly, the thickness of the p region must be such that electrons generated in this region can diffuse to the junction before recombination takes place. This requirement implies a proper match between the electron diffusion length L_n, the thickness of the p region, and the mean optical penetration depth $1/\alpha$ [see Eq. (4–2)]. It is desirable to have a large contact potential V_0 to obtain a large photovoltage, and therefore heavy

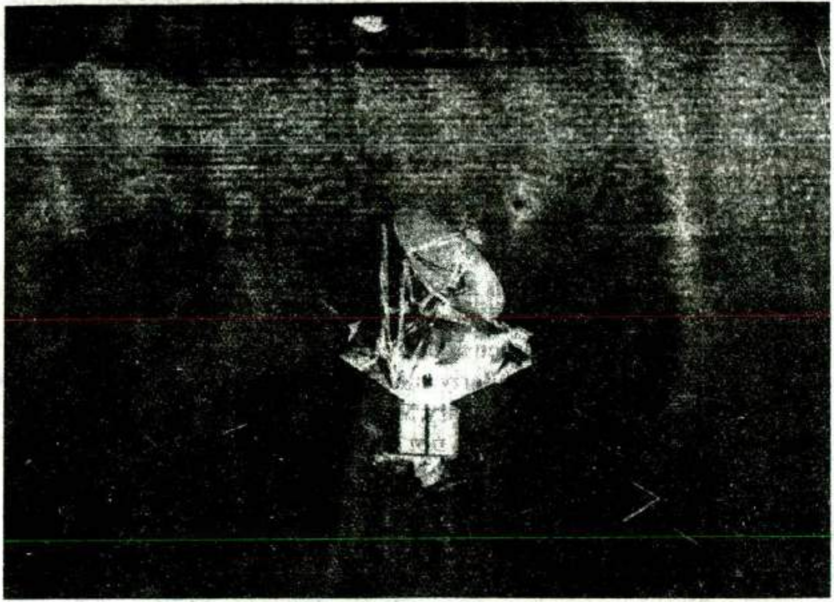


Figure 8–4
Solar cell arrays attatched to the Mars fly-by satellite *Mariner*. (Provided through the courtesy of the National Aeronautics and Space Administration, California Institute of Technology, Jet Propulsion Laboratory.)

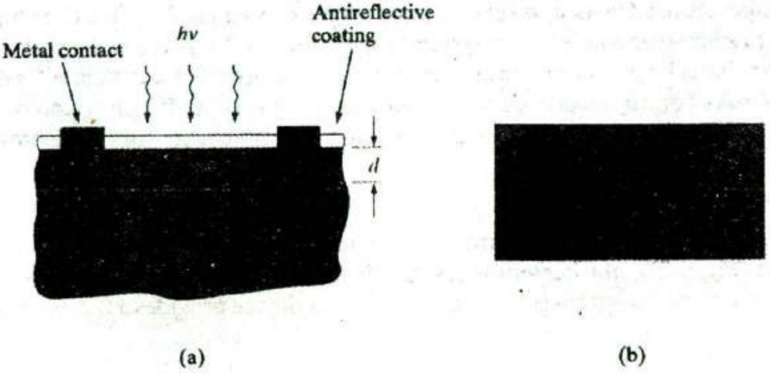


Figure 8–5
Configuration of a solar cell: (a) enlarged view of the planar junction; (b) top view, showing metal contact "fingers."

doping is indicated; on the other hand, long lifetimes are desirable and these are reduced by doping too heavily. It is important that the series resistance of the device be very small so that power is not lost to heat due to ohmic losses in the device itself. A series resistance of only a few ohms can seriously reduce the output power of a solar cell (Prob. 8.4). Since the area is large, the resistance of the p-type body of the device can be made small. However, contacts to the thin n region require special design. If this region is contacted at the edge, current must flow along the thin n region to the contact, resulting in a large series resistance. To prevent this effect, the contact can be distributed over the n surface by providing small contact fingers as in Fig. 8–5b. These narrow contacts serve to reduce the series resistance without interfering appreciably with the incoming light.

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

Applications of solar cells are not restricted to outer space. It is possible to obtain useful power from the sun in terrestrial applications using solar cells, even though the solar intensity is reduced by the atmosphere. About 1 kW/m^2 is available in a particularly sunny location, but not all of this solar power can be converted to electricity. Much of the photon flux is at energies less than the cell band gap, and is not absorbed. High-energy photons are strongly absorbed, and the resulting EHPs may recombine at the surface. A well-made Si cell can have about 10 percent efficiency for solar energy conversion, providing approximately 100 W/m^2 of electrical power under full illumination. This is a modest amount of power per unit solar cell area, considering the effort involved in fabricating a large area of Si cells. One approach to obtaining more power per cell is to focus considerable light onto the cell using mirrors. Although Si cells lose efficiency at the resulting high temperatures, GaAs and related compounds can be used at 100°C or higher. In such solar concentrator systems more effort and expense can be put into the solar cell fabrication, since fewer cells are required. For example, a GaAs–AlGaAs heterojunction cell provides good conversion efficiency and operates at the elevated temperatures common in solar concentrator systems.

8.1.3 Photodetectors

When the photodiode is operated in the third quadrant of its I–V characteristic (Fig. 8–3b), the current is essentially independent of voltage but is proportional to the optical generation rate. Such a device provides a useful means

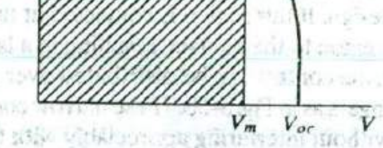


Figure 8–6
I–V characteristics of an illuminated solar cell. The maximum power rectangle is shaded.

of measuring illumination levels or of converting time-varying optical signals into electrical signals.

In most optical detection applications the detector's speed of response is critical. For example, if the photodiode is to respond to a series of light pulses 1 ns apart, the photogenerated minority carriers must diffuse to the junction and be swept across to the other side in a time much less than 1 ns. The carrier diffusion step in this process is time consuming and should be eliminated if possible. Therefore, it is desirable that the width of the depletion region W be large enough so that most of the photons are absorbed within W rather than in the neutral p and n regions. When an EHP is created in the depletion region, the electric field sweeps the electron to the n side and the hole to the p side. Since this carrier drift occurs in a very short time, the response of the photodiode can be quite fast. When the carriers are generated primarily within the depletion layer W, the detector is called a *depletion layer photodiode*. Obviously, it is desirable to dope at least one side of the junction lightly so that W can be made large. The appropriate width for W is chosen as a compromise between sensitivity and speed of response. If W is wide, most of the incident photons will be absorbed in the depletion region. Also, a wide W results in a small junction capacitance [see Eq. (5-62)], thereby reducing the RC time constant of the detector circuit. On the other hand, W must not be so wide that the time required for drift of photogenerated carriers out of the depletion region is excessive.

One convenient method of controlling the width of the depletion region is to build a *p-i-n photodetector* (Fig. 8–7). The "i" region need not be truly intrinsic, as long as the resistivity is high. It can be grown epitaxially on the n-type substrate, and the p region can be obtained by implantation. When this device is reverse biased, the applied voltage appears almost entirely across the i region. If the carrier lifetime within the i region is long compared with the drift time, most of the photogenerated carriers will be collected by the n and p regions.

If low-level optical signals are to be detected, it is often desirable to operate the photodiode in the avalanche region of its characteristic. In this mode each photogenerated carrier results in a significant change in the current because of avalanche multiplication. In the *avalanche photodiode* the

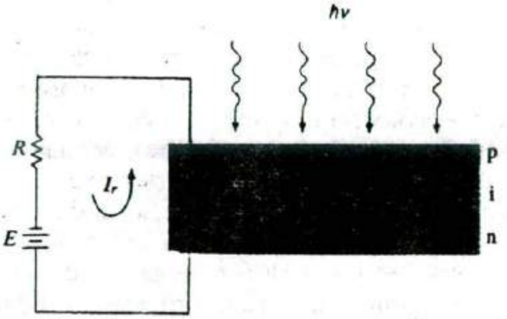


Figure 8–7 Schematic representation of a p-i-n photodiode.

junction must be uniform, and a guard ring is generally used to ensure against edge breakdown. With proper design a Si avalanche photodiode can have high sensitivity to low-level optical signals, and the response time is in the neighborhood of 1 ns. These devices are particularly useful in fiber optic communication systems (Section 8.2.2).

The type of photodiode described here is sensitive to photons with energies near the band gap energy (*intrinsic* detectors). If $h\nu$ is less than E_g, the photons will not be absorbed; on the other hand, if the photons are much more energetic than E_g, they will be absorbed very near the surface, where the recombination rate is high. Therefore, it is necessary to choose a photodiode material with a band gap corresponding to a particular region of the spectrum. Detectors sensitive to longer wavelengths can be designed such that photons can excite electrons into or out of impurity levels (*extrinsic* detectors). However, the sensitivity of such extrinsic detectors is much less than intrinsic detectors, where electron–hole pairs are generated by excitation across the band gap.

By using lattice-matched multilayers of compound semiconductors, the band gap of the absorbing region can be tailored to match the wavelength of light being detected. Wider band gap material can then be used as a window through which the light is transmitted to the absorbing region (Fig. 8–8). For example, we saw in Fig. 1–13 that InGaAs with an In mole fraction of 53 percent can be grown epitaxially on InP with excellent lattice matching. This composition of InGaAs has a band gap of about 0.75 eV, which is sensitive to a useful wavelength for fiber optic systems (1.55 μm), as we shall see in Section 8.2.2. In making a photodiode using InGaAs as the active material, it is possible to bring the light through the wider band gap InP (1.35 eV), thus greatly reducing surface recombination effects. In the case of avalanche photodiodes requiring narrow band gap material, it is often advantageous to absorb the light in the narrow-gap semiconductor (e.g., InGaAs) and transport the resulting carriers to a junction made in wider band gap material (e.g., InP), where the avalanche multiplication takes place (Fig. 8–8b). Such a separation of the absorption and multiplication regions avoids the excessive leakage currents typical of reverse-biased junctions in narrow-gap materials.

8.1.4 Noise and Bandwidth of Photodetectors

In optical communication systems the sensitivity of the photodetector and its response time are of critical importance. Unfortunately, these two properties are generally difficult to optimize without making compromises between them. For example, in a photoconductor the gain depends on the ratio of carrier lifetime to transit time (see Prob. 8–6). On the other hand, the frequency response (and therefore the bandwidth) varies inversely with carrier lifetime. As a result, trade-offs must be made between these two desirable characteristics. It is common to express the *gain-bandwidth product* as a figure of merit for detectors. Designs which increase gain tend to decrease bandwidth and vice versa. Another important property of detectors is the

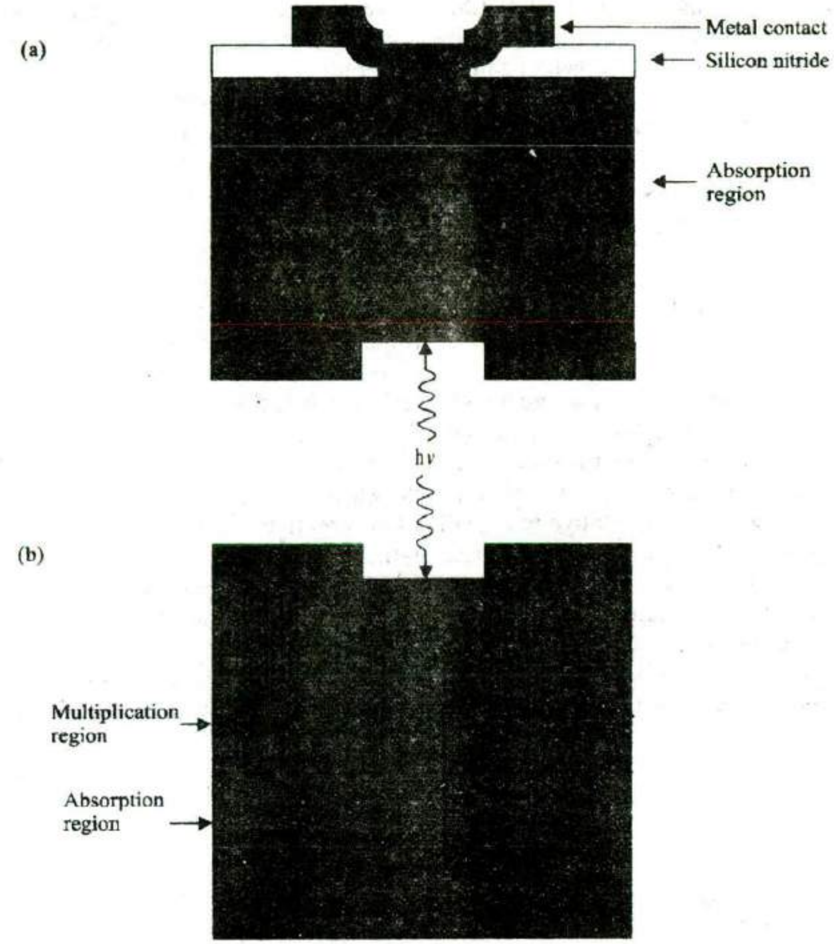


Figure 8–8 Use of multilayer heterojunctions to enhance photodiode operation: (a) a p-i-n photodiode in which light near 1.55 μm is absorbed in a narrow band gap material (InGaAs, $E_g = 0.75$ eV) after passing through a wider-gap material (InP, $E_g = 1.35$ eV); (b) an avalanche photodiode in which light is absorbed in the InGaAs and holes are swept to an InP junction, where the avalanche multiplication takes place. This separation of the absorption and multiplication regions reduces the junction leakage current. In this figure, n^- refers to lightly doped n-type material.

signal-to-noise ratio, which is the amount of usable information compared with the background noise in the detector.

In the case of photoconductors the major source of noise is random fluctuations in the dark current (called *Johnson noise*). The noise current increases with temperature and the conductance of the material in the dark. Therefore, the photoconductor noise at a given temperature can be reduced by increasing the dark resistance. Increased dark resistance also increases the gain of the photoconductor, thereby decreasing the bandwidth.

In a p-i-n diode there is no gain mechanism, since at most one electron–hole pair is collected by the junction for each photon absorbed. Thus the gain is essentially unity, and the gain-bandwidth product is determined by the bandwidth, or frequency response. In a p-i-n the response time is dependent on the width of the depletion region, and the main source of noise is random thermal generation of EHPs within this region (called *shot noise*). The noise in a p-i-n device is considerably lower than that in a photoconductor, which compensates for the lack of gain in the p-i-n.

Avalanche photodiodes have the advantage of providing gain through the avalanche multiplication effect. The disadvantage is increased noise relative to the p-i-n, due to random fluctuations in the avalanche process. This noise is reduced if the impact ionization in the high field region is due to only one type of carrier, since more fluctuations in the ionization process occur when both electrons and holes participate. In Si the ability of electrons to create EHPs in an impact ionization event is much higher than for holes. Therefore, Si avalanche photodiodes can be operated with high gain and relatively low noise. Unfortunately, Si avalanche photodiodes (APDs) cannot be used for most fiber optic transmission because Si is transparent at the wavelengths of low loss and low dispersion ($\lambda = 1.55$ and $1.3\ \mu\text{m}$) for optical fibers. For these longer wavelengths, $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ has become the material of choice. However, the ionization rates of electrons and holes in most compound semiconductors are comparable, which degrades their noise and frequency response relative to Si APDs. One creative way of overcoming this problem is shown in Fig. 8–9. The Si–InGaAs APD is fabricated by using wafer fusion to bond an InGaAs absorption layer to a Si avalanche photodiode. The advantage of this approach is that it utilizes the strengths of both materials systems. In operation, light is absorbed in the narrow bandgap InGaAs layer and the photogenerated electrons are injected into the Si avalanche region, which is better suited for the large fields applied. These

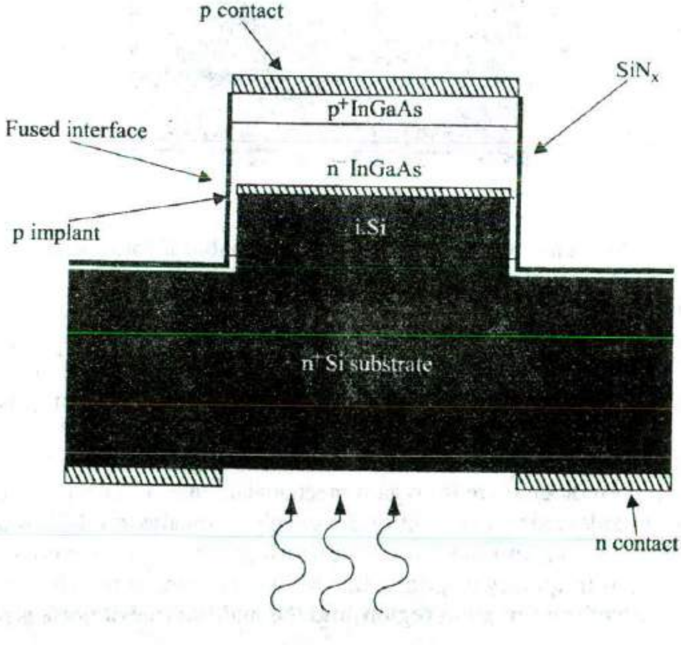


Figure 8–9 Schematic cross section of a silicon hetero-interface photodetector (SHIP). The light passes through the wide-bandgap Si and is absorbed in the InGaAs.

APDs have achieved low dark current, a quantum efficiency of 60 percent at 1.3 μm, and a gain-bandwidth product of 300 GHz.

Another approach that has demonstrated excellent performance utilizes an APD structure in a resonant cavity. The resonant-cavity photodiode (Fig. 8–10) consists of a thin absorbing layer sandwiched between two *distributed Bragg reflector (DBR)* mirrors. The structure is similar to that of a vertical-cavity, surface-emitting laser (to be discussed in Section 8.4.4) except that the active region is an absorber instead of an emitter and the Q of the cavity is typically much lower than that of laser structures. The resonant-cavity structure can provide several performance advantages, one of which is that the tradeoff between responsivity and bandwidth that is inherent to conventional, single-pass p-i-n photodiode structures can be circumvented. For the typical normal-incidence photodetector, a wide bandwidth necessitates a thin absorption layer which, in turn, results in low quantum efficiency. The resonant-cavity structure, on the other hand, effectively decouples the responsivity from the transit-time component of the bandwidth because the optical signal makes multiple passes across the thin absorbing layer inside the microcavity. Consequently, high speed and high quantum efficiency can be achieved simultaneously. In the resonant-cavity APD shown in Fig. 8–10, light is absorbed in the thin InGaAs

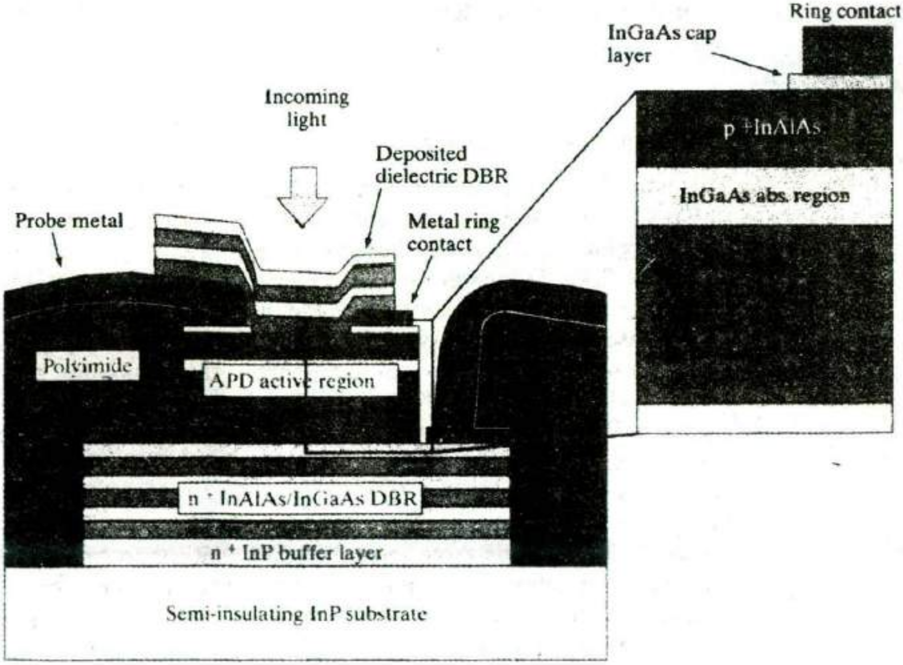


Figure 8–10
Cross section of InAlAs–InGaAs resonant-cavity avalanche photodiode, and a detail of the active region.

layer. The electrons then drift into the InAlAs multiplication region where the field is high enough to provide gain by impact ionization. For very thin multiplication layers there is a size effect that can lead to very low multiplication noise, comparable to that of Si APDs, and high gain-bandwidth products. At low gains where the bandwidth is limited by the transit time and the *RC* time constraint, bandwidths in excess of 30 GHz have been achieved. At higher gains the bandwidth is determined by the gain-bandwidth product which can be in excess of 300 GHz.

8.2 LIGHT-EMITTING DIODES

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

8.2.1 Light-Emitting Materials

The band gaps of various binary compound semiconductors are illustrated in Fig. 4–4 relative to the spectrum. There is a wide variation in band gaps and, therefore, in available photon energies, extending from the ultraviolet (GaN, 3.4 eV) into the infrared (InSb, 0.18 eV). In fact, by utilizing ternary and quaternary compounds the number of available energies can be increased significantly (see Figs. 1–13 and 3–6). A good example of the variation in photon energy obtainable from the compound semiconductors is the ternary alloy gallium arsenide–phosphide, which is illustrated in Fig. 8–11. When the percentage of As is reduced and P is increased in this material, the resulting band gap varies from the direct 1.43-eV gap of GaAs (infrared) to the indirect 2.26-eV gap of GaP (green). The band gap of $GaAs_{1-x}P_x$ varies almost linearly with x until the 0.45 composition is reached, and electron–hole recombination is direct over this range. The most common alloy composition used in LED displays is $x \simeq 0.4$. For this composition the band gap is direct, since the Γ minimum (at $\mathbf{k} = 0$) is the lowest part of the conduction band. This results in efficient radiative recombination, and the emitted photons (~1.9 eV) are in the red portion of the spectrum.

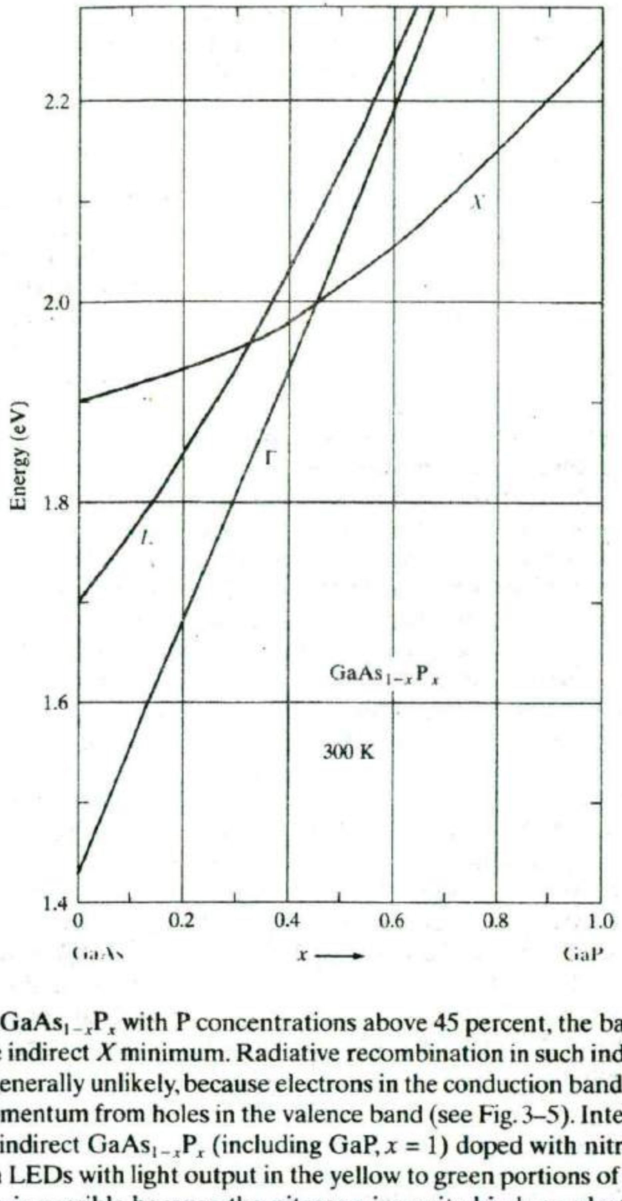


Figure 8–11 Conduction band energies as a function of alloy composition for GaAs$_{1-x}$P$_x$.

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

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

In view of the broad range of applications requiring semiconductor lasers and LEDs with visible and infrared wavelengths, the wide variety of available III–V materials is extremely useful. In addition to the AlGaAs and GaAsP systems shown in Figs. 3–6 and 8–11, the InAlGaP system is useful for yellow and green wavelengths, and GaN is a strong emitter in the blue. Even more wavelengths will be accessible as GaN and related materials become increasingly used in LEDs and lasers. For many years the II–VI semiconductors have been known as efficient light emitters in photoluminescence, but obtaining p-n junctions was extremely difficult. With traditional doping methods, crystal defects tend to compensate the doping impurities such that only n-type (ZnS, ZnSe, CdS, CdSe) or p-type (ZnTe) can be obtained. This frustrating problem prevented the formation of useful p-n junctions until 1990, when the use of nitrogen doping resulted in p-type ZnSe in MBE-grown material. Rapid progress has been made since then, including the use of multilayer heterostructures grown by MBE and OMVPE in the (Zn, Cd)(S, Se) system. Using a nitrogen plasma source, ZnTe can be doped p-type to acceptor concentrations above 10^{19} cm^{-3}. In spite of this research progress, however, II–VI LEDs and lasers lag behind III–V semiconductors in most applications. The availability of blue light from GaN is of particular importance in extending III–V light emission across the entire visible spectrum.

8.2.2 Fiber Optic Communications

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

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

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

$$\mathbf{I}(x) = \mathbf{I}_0 e^{-\alpha x} \tag{8–5}$$

The attenuation is not the same for all wavelengths, however, and it is therefore important to choose a signal wavelength carefully. A plot of $\boldsymbol{\alpha}$ vs. λ for a typical silica glass fiber is shown in Fig. 8–13. It is clear that dips in $\boldsymbol{\alpha}$ near 1.3 and 1.55 μm provide "windows" in the attenuation, which can be exploited to reduce the degradation of signals. The overall decrease in attenuation with increasing wavelength is due to the reduced scattering from small random inhomogeneities which result in fluctuations of the refractive index on a scale comparable to the wavelength. This type of attenuation, called *Rayleigh scattering*, decreases with the fourth power of wavelength. This effect is observed at sunrise and sunset, when attenuation

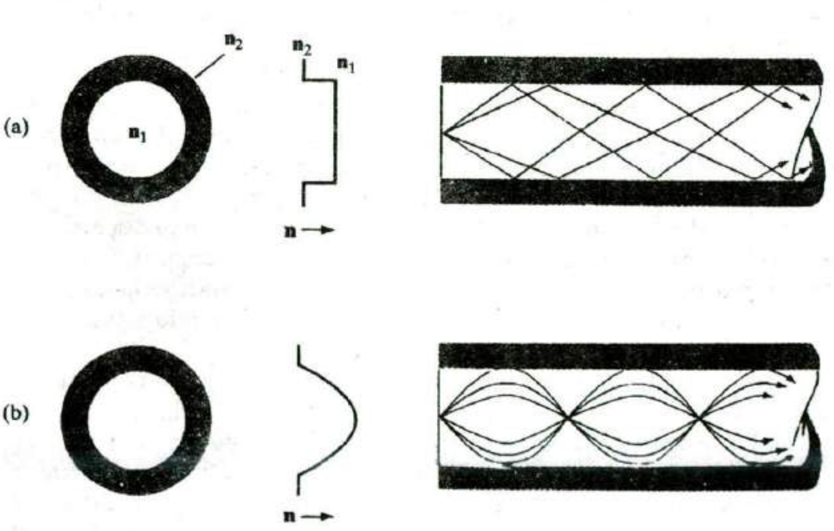


Figure 8–12
Two examples of multimode fibers: (a) *step-index,* having a core with slightly larger refractive index **n**; (b) *graded-index* having in this case a parabolic grading of **n** in the core. The figure illustrates the cross section (left) of the fiber, its index of refraction profile (center),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

[1] The *index of refraction* (or *refractive index*) **n** compares the velocity of light **v** in the material to its velocity c in a vacuum, $\mathbf{n} = c/\mathbf{v}$. Thus if $\mathbf{n}_1 > \mathbf{n}_2$ in Fig. 8–12a, the light velocity is greater in material 2 than in 1. The value of **n** varies somewhat with the wavelength of light.

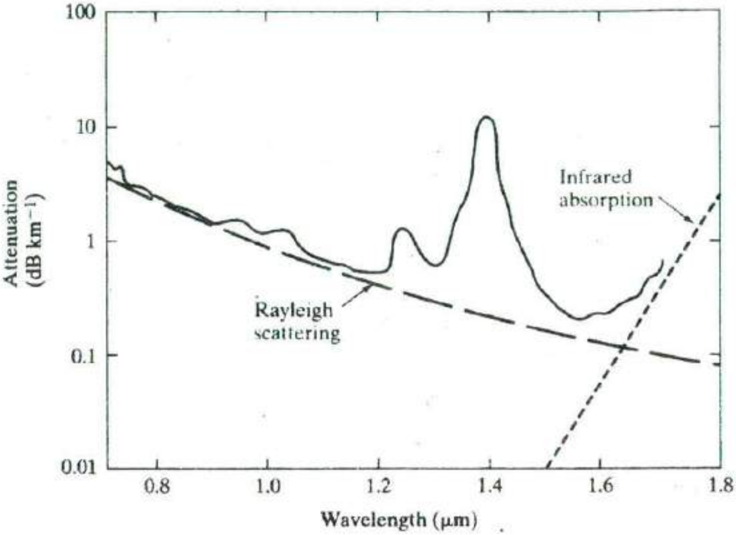


Figure 8–13 Typical plot of attenuation coefficient α vs. wavelength λ for a fused silica optical fiber. Peaks are due primarily to OH$^-$ impurities.

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

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

In early optoelectronic systems for fiber optics, it was most convenient to use the well-established GaAs–AlGaAs system for making lasers and LEDs. These light sources are very efficient, and good detectors can be made using Si p-i-n or avalanche photodiodes. However, these sources operate in the wavelength range near 0.9 μm, where the attenuation is greater than for longer wavelengths. Modern systems, therefore, operate near the 1.3- and 1.55-μm minima in Fig. 8–13. At these wavelengths, sources can be made using InGaAs or InGaAsP grown on InP, and detectors can be made of the same materials (see Fig. 8–8), or using Ge.

8.2.3 Multilayer Heterojunctions for LEDs

The light source in a fiber optic system may be a laser or an LED. In the case of a laser, the light is of essentially a single frequency and allows a very large information bandwidth. Semiconductor lasers suitable for fiber optic communications will be discussed in Section 8.4. An LED designed for a fiber optic system is illustrated in Fig. 8–14. The LED is a multilayer structure of GaAs and AlGaAs. To take advantage of the 1.3- and 1.55-μm windows in Fig. 8–13, similar devices using InGaAs or InGaAsP can be used. The quaternary (four-element) alloy is particularly suitable, in that band gap (and therefore emission wavelength) can be adjusted along with choosing lattice constants for epitaxial growth on convenient substrates. In Fig. 8–14 the fiber is held in an etched well on the back side of the diode by an epoxy resin. This configuration, often called a "Burrus diode" after its developer, is particularly convenient for launching signals from an LED into a fiber, with good mechanical stability.

Although LEDs are less suited to transmission of digital information than are lasers, they are easily modulated by analog signals. The optical power emitted by a properly constructed LED varies linearly with the input current over a wide range. An LED is an *incoherent* light source, in that photons are emitted randomly from the junction in all directions and not in phase with each other. Therefore, transmission of LED-generated signals inherently involves many modes, as in Fig. 8–12. *Multimode* fibers are larger (~25 μm in diameter) than are *single-mode* fibers (~5 μm), which transmit a coherent laser beam.

By forming numerous optical fibers into a bundle, with an appropriate jacket for mechanical strength, an enormous amount of information can be transmitted over long distances.[2] Depending upon the losses in the fibers,

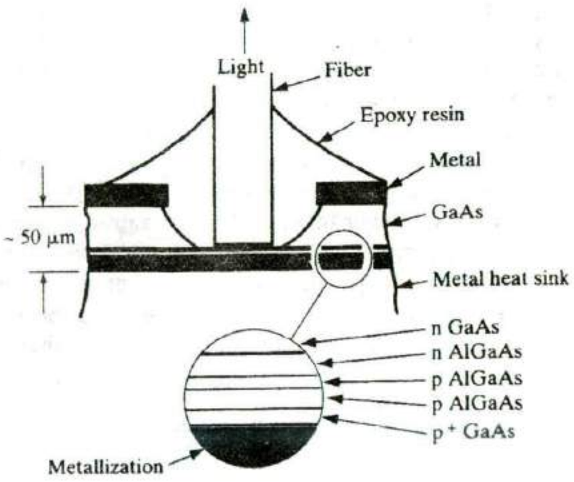


Figure 8–14 Cross section of a GaAs–AlGaAs LED for fiber-optic applications. [After C.A. Burrus and B.I. Miller, *Optics Communications,* vol. 4, p. 307 (1971).]

[2] Transmission rates of many G-bit/s have been achieved. As a convenient calibration of this rate, it is worth noting that the human eye is able to transmit about one G-bit/s to the brain.

repeater stations may be required periodically along the path. Thus many photodetectors and LED or laser sources are required in a fiber optic system. Semiconductor device development, including appropriate binary, ternary, and quaternary compounds for both emitters and detectors, is therefore crucial to the successful implementation of such optical communications systems.

8.3 LASERS

The word LASER is an acronym for *light amplification by stimulated emission of radiation,* which sums up the operation of an important optical and electronic device. The laser is a source of highly directional, monochromatic, coherent light, and as such it has revolutionized some longstanding optical problems and has created some new fields of basic and applied optics. The light from a laser, depending on the type, can be a continuous beam of low or medium power, or it can be a short burst of intense light delivering millions of watts. Light has always been a primary communications link between humans and the environment, but until the invention of the laser, the light sources available for transmitting information and performing experiments were generally neither monochromatic nor coherent, and were of relatively low intensity. Thus the laser is of great interest in optics; but it is equally important in optoelectronics, particularly in fiber optic communications. The last three letters in the word *laser* are intended to imply how the device operates: by the *stimulated emission* of *radiation.* In Chapter 2 we discussed the emission of radiation when excited electrons fall to lower energy states; but generally, these processes occur randomly and can therefore be classed as *spontaneous emission.* This means that the rate at which electrons fall from an upper level of energy E_2 to a lower level E_1 is at every instant proportional to the number of electrons remaining in E_2 (the *population* of E_2). Thus if an initial electron population in E_2 were allowed to decay, we would expect an exponential emptying of the electrons to the lower energy level, with a mean decay time describing how much time an average electron spends in the upper level. An electron in a higher or excited state need not wait for spontaneous emission to occur, however; if conditions are right, it can be *stimulated* to fall to the lower level and emit its photon in a time much shorter than its mean spontaneous decay time. The stimulus is provided by the presence of photons of the proper wavelength. Let us visualize an electron in state E_2 waiting to drop spontaneously to E_1 with the emission of a photon of energy $h\nu_{12} = E_2 - E_1$ (Fig. 8–15). Now we assume that this electron

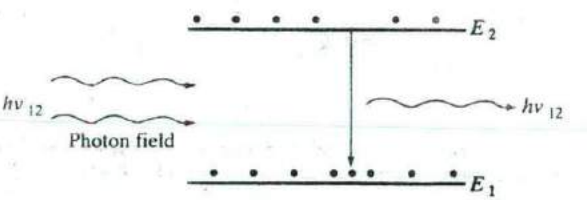


Figure 8–15 Stimulated transition of an electron from an upper state to a lower state, with accompanying photon emission.

in the upper state is immersed in an intense field of photons, each having energy $h\nu_{12} = E_2 - E_1$, and in phase with the other photons. The electron is induced to drop in energy from E_2 to E_1, contributing a photon whose wave is *in phase* with the radiation field. If this process continues and other electrons are stimulated to emit photons in the same fashion, a large radiation field can build up. This radiation will be *monochromatic* since each photon will have an energy of precisely $h\nu_{12} = E_2 - E_1$ and will be *coherent*, because all the photons released will be in phase and reinforcing. This process of stimulated emission can be described quantum mechanically to relate the probability of emission to the intensity of the radiation field. Without quantum mechanics we can make a few observations here about the relative rates at which the absorption and emission processes occur. Let us assume the instantaneous populations of E_1 and E_2 to be n_1 and n_2, respectively. We know from earlier discussions of distributions and the Boltzmann factor that at *thermal equilibrium* the relative population will be

$$\frac{n_2}{n_1} = e^{-(E_2 - E_1)/kT} = e^{-h\nu_{12}/kT} \tag{8–6}$$

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

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

$$\boxed{\begin{array}{lcccc} B_{12}n_1\rho(\nu_{12}) & = & A_{21}n_2 & + & B_{21}n_2\rho(\nu_{12}) \\ \text{Absorption} & = & \text{spontaneous} & + & \text{stimulated} \\ & & \text{emission} & & \text{emission} \end{array}} \tag{8–7}$$

[3] The energy density $\rho(\nu_{12})$ indicates the total energy in the radiation field per unit volume and per unit frequency, due to photons with $h\nu_{12} = E_2 - E_1$.

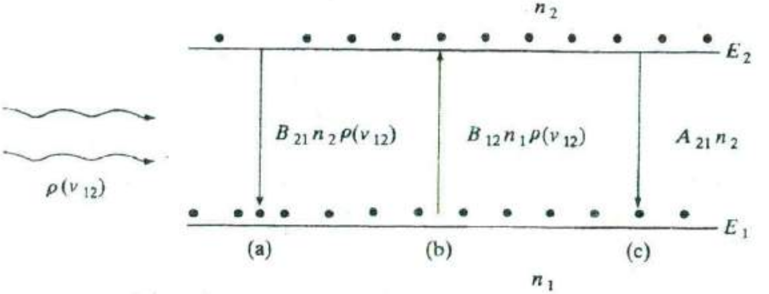


Figure 8–16 Balance of absorption and emission in steady state: (a) stimulated emission; (b) absorption; (c) spontaneous emission.

This relation was described by Einstein, and the coefficients B_{12}, A_{21}, B_{21} are called the *Einstein coefficients.* We notice from Eq. (8–7) that no energy density ρ is required to cause a transition from an upper to a lower state; spontaneous emission occurs without an energy density to drive it. The reverse is not true, however; exciting an electron to a higher state (absorption) requires the application of energy, as we would expect thermodynamically.

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

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

As Eq. (8–8) indicates, the way to enhance the stimulated emission over spontaneous emission is to have a very large photon field energy density $\rho(\nu_{12})$. In the laser, this is encouraged by providing an *optical resonant cavity* in which the photon density can build up to a large value through multiple internal reflections at certain frequencies (ν).

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

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

Thus if stimulated emission is to dominate over absorption of photons from the radiation field, we must have a way of maintaining more electrons in the upper level than in the lower level. This condition is quite unnatural, since Eq. (8–6) indicates that n_2/n_1 is less than unity for any equilibrium case. Because of its unusual nature, the condition $n_2 > n_1$ is called *population inversion.* It is also referred to as a condition of *negative temperature.* This rather startling terminology emphasizes the nonequilibrium nature of population inversion, and refers to the fact that the ratio n_2/n_1 in Eq. (8–6) could be larger than unity only if the temperature were negative. Of course, this manner of speaking does not imply anything about temperature in the usual sense of that word. The fact is that Eq. (8–6) is a thermal equilibrium equation and cannot be applied to the situation of population inversion without invoking the concept of negative temperature.

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

An optical resonant cavity can be obtained using reflecting mirrors to reflect the photons back and forth, allowing the photon energy density to build up. One or both of the end mirrors are constructed to be partially transmitting so that a fraction of the light will "leak out" of the resonant system. This transmitted light is the output of the laser. Of course, in designing such a laser one must choose the amount of transmission to be a small perturbation on the resonant system. The gain in photons per pass between the end plates must be larger than the transmission at the ends, scattering from impurities, absorption, and other losses. The arrangement of parallel plates providing multiple internal reflections is similar to that used in the Fabry–Perot interferometer;[4] thus the reflecting ends of the laser cavity are often referred to as Fabry–Perot faces. As Fig. 8–17 indicates, light of a particular frequency can be reflected back and forth within the resonant cavity in a reinforcing (coherent) manner if an integral number of half-wavelengths fit between the end mirrors. Thus the length of the cavity for stimulated emission must be

$$L = \frac{\mathbf{m}\lambda}{2} \qquad (8\text{–}10)$$

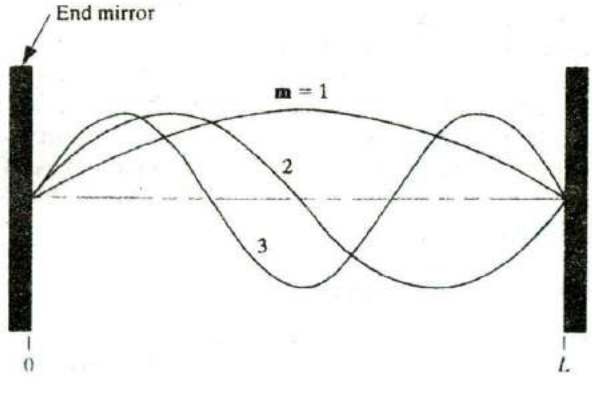


Figure 8–17 Resonant modes within a laser cavity.

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

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

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

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

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

8.4 SEMICONDUCTOR LASERS

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

8.4.1 Population Inversion at a Junction

If a p-n junction is formed between degenerate materials, the bands under forward bias appear as shown in Fig. 8–18. If the bias (and thus the current)

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

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

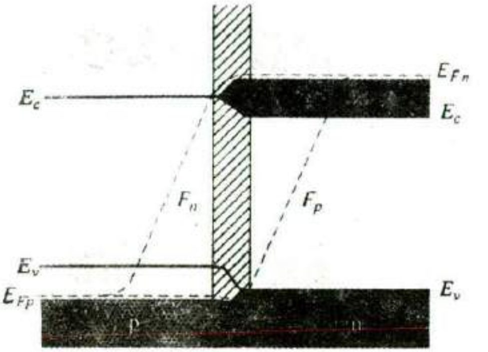


Figure 8–18 Band diagram of a p-n junction laser under forward bias. The cross-hatched region indicates the inversion region at the junction.

is large enough, electrons and holes are injected into and across the transition region in considerable concentrations. As a result, the region about the junction is far from being depleted of carriers. This region contains a large concentration of electrons within the conduction band and a large concentration of holes within the valence band. If these population densities are high enough, a condition of population inversion results, and the region about the junction over which it occurs is called an *inversion region.*[7]

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

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

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

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

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

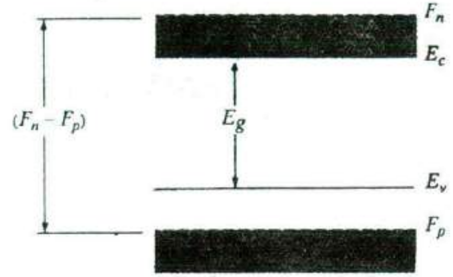


Figure 8–19 Expanded view of the inversion region.

is a measure of the departure from equilibrium at that point. Obviously, this departure is considerable in the inversion region, since F_n and F_p are separated by an energy greater than the band gap (Fig. 8–19).

Unlike the case of the two-level system discussed in Section 8.3, the condition for population inversion in semiconductors must take into account the distribution of energies available for transitions between the bands. The basic definition of population inversion holds—for dominance of stimulated emission between two energy levels separated by energy $h\nu$, the electron population of the upper level must be greater than that of the lower level. The unusual aspect of a semiconductor is that bands of levels are available for such transitions. Population inversion obviously exists for transitions between the bottom of the conduction band E_c and the top of the valence band E_v in Fig. 8–19. In fact, transitions between levels in the conduction band up to F_n and levels in the valence band down to F_p take place under conditions of population inversion. For any given transition energy $h\nu$ in a semiconductor, population inversion exists when

$$(F_n - F_p) > h\nu \tag{8–13a}$$

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

$$(F_n - F_p) > E_g \tag{8–13b}$$

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

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

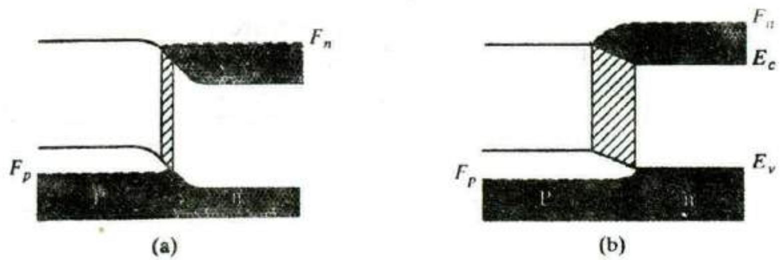


Figure 8–20 Variation of inversion region width with forward bias: $V(a) < V(b)$.

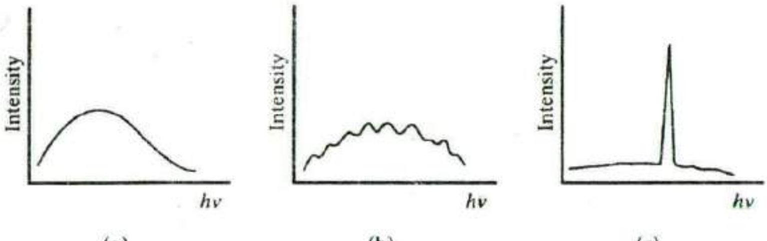


Figure 8–21 Light intensity vs. photon energy $h\nu$ for a junction laser: (a) incoherent emission below threshold; (b) laser modes at threshold; (c) dominant laser mode above threshold. The intensity scales are greatly compressed from (a) to (b) to (c).

8.4.2 Emission Spectra for p-n Junction Lasers

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

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

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

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

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

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

Now reverting to discrete changes in $\mathbf{m}$ and λ_0, we can write

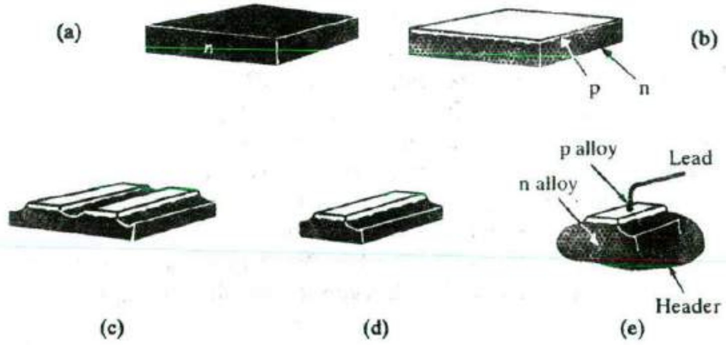
$$-\Delta\lambda_0 = \frac{\lambda_0^2}{2L\mathbf{n}}\left(1 - \frac{\lambda_0}{\mathbf{n}}\frac{d\mathbf{n}}{d\lambda_0}\right)^{-1}\Delta\mathbf{m} \tag{8–16}$$

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

8.4.3 The Basic Semiconductor Laser

To build a p-n junction laser, we need to form a junction in a highly doped, direct semiconductor (GaAs, for example), construct a resonant cavity in the proper geometrical relationship to the junction, and make contact to the junction in a mounting which allows for efficient heat transfer. The first lasers were built as shown in Fig. 8–22. Beginning with a degenerate n-type sample, a p region is formed on one side, for example by diffusing Zn into the n-type GaAs. Since Zn is in column II of the periodic table and is introduced substitutionally on Ga sites, it serves as an acceptor in GaAs; therefore, the heavily doped Zn diffused layer forms a p$^+$ region (Fig. 8–22b). At

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



this point we have a large-area planar p-n junction. Next, grooves are cut or etched along the length of the sample as in Fig. 8–22c, leaving a series of long p regions isolated from each other. These p-n junctions can be cut or broken apart (Fig. 8–22d) and then cleaved into devices of the desired length.

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

8.4.4 Heterojunction Lasers

The device described above was the first type used in the early development of semiconductor lasers. Since the device contains only one junction in a single type of material, it is referred to as a *homojunction* laser. To obtain more efficient lasers, and particularly to build lasers that operate at room temperature, it is necessary to use multiple layers in the laser structure. Such devices, called *heterojunction lasers*, can be made to operate continuously at room temperature to satisfy the requirements of optical communications. An example of a heterojunction laser is shown in Fig. 8–23. In this structure the injected carriers are confined to a narrow region so that population

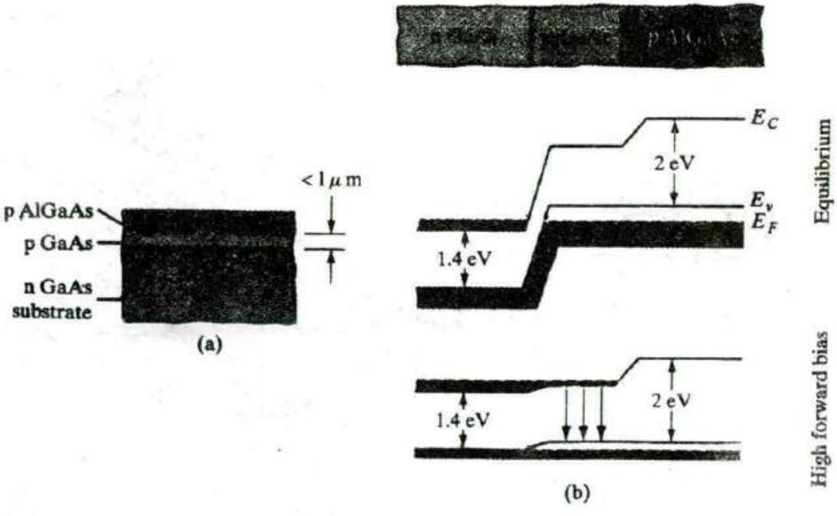


Figure 8–23
Use of a single heterojunction for carrier confinement in laser diodes: (a) AlGaAs heterojunction grown on the thin p-type GaAs layer; (b) band diagrams for the structure of (a), showing confinement of electrons to the thin p region under bias.

inversion can be built up at lower current levels. The result is a lowering of the *threshold current* at which laser action begins. Carrier confinement is obtained in this single-heterojunction laser by the layer of AlGaAs grown epitaxially on the GaAs.

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

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

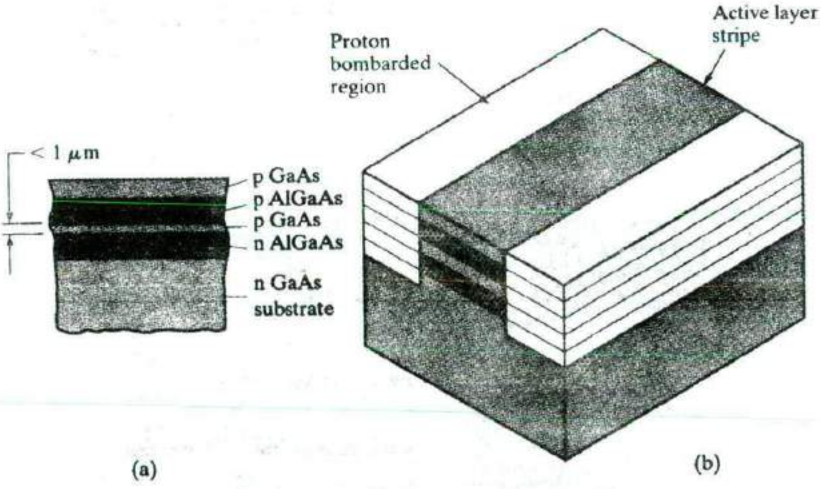


Figure 8–24 A double-heterojunction laser structure: (a) multiple layers used to confine injected carriers and provide waveguiding for the light; (b) a stripe geometry designed to restrict the current injection to a narrow stripe along the lasing direction. One of many methods for obtaining the stripe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

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

Vertical Cavity Surface-Emitting Lasers (VCSELs). There are advantages to laser structures in which light is emitted normal to the surface, including ease of device testing on the wafer before packaging. An interesting approach is the VCSEL, in which the cavity mirrors are replaced by DBRs, which use many partial reflectors spaced to reflect light constructively. DBRs can be grown by MBE or OMVPE. In Fig. 8–26 the bottom DBR mirror of a VCSEL

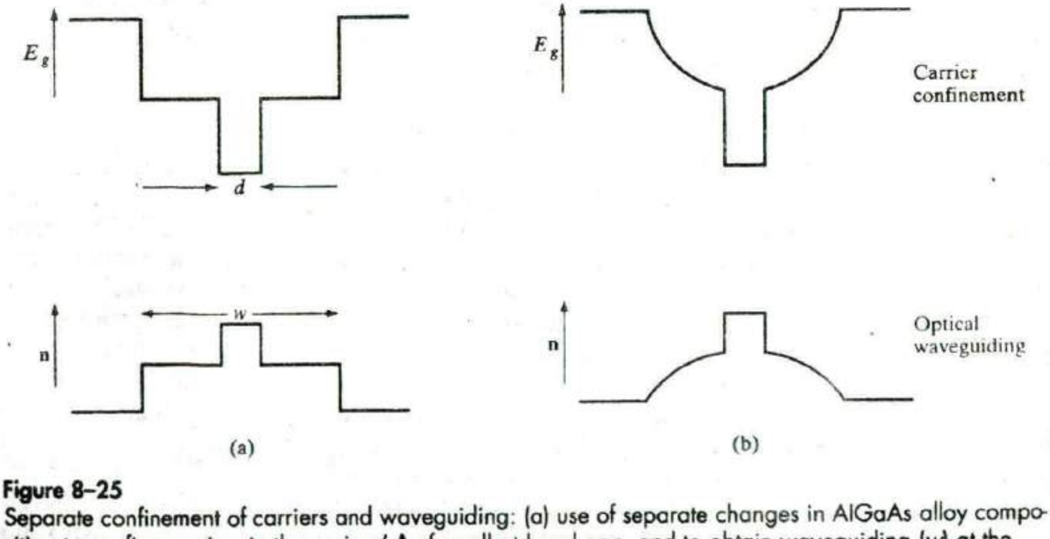


Figure 8–25
Separate confinement of carriers and waveguiding: (a) use of separate changes in AlGaAs alloy composition to confine carriers in the region (d) of smallest band gap, and to obtain waveguiding (w) at the larger step in refractive index; (b) grading the alloy composition, and therefore the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

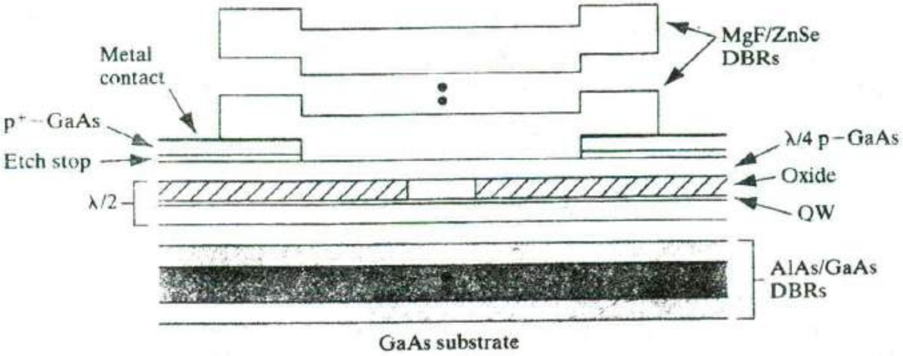


Figure 8–26
Schematic cross section of oxide-confined vertical cavity surface-emitting laser diode. [D.G.Deppe et al., *IEEE J. Selected Topics in Quantum Elec.*, 3(3) (June 1997): 893-904]

is composed of many alternating layers of AlAs and GaAs with thickness one-quarter of a wavelength in each material. The top mirror is composed of deposited dielectric layers (alternating ZnSe and MgF). Current is funneled into the active region from the top contact by using an oxide layer achieved by laterally oxidizing an AlGaAs layer to form an aluminum oxide. The active region of the laser employs InGaAs–GaAs quantum wells, and the GaAs cavity between the two DBRs is one wavelength long. The VCSEL can be made with much shorter cavity length than other structures, and as a result of Eq. (8–16) the laser modes are widely separated in wavelength. Thus single-mode laser operation is more easily achieved with the VCSEL. Lasing can be achieved at very low current ($< 50\ \mu A$) with this device.

8.4.5 Materials for Semiconductor Lasers

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

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

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

Two of the problems which had stymied progress in this field since pioneering work by Pankove in the 1970s was the absence of a suitable substrate having sufficient lattice match with GaN, and the inability to achieve p-type doping in this semiconductor. GaN bulk crystals cannot be grown easily because of the high vapor pressure of the nitrogen-bearing precursor (generally ammonia). This requires growth at high temperature and pressure. This precludes using bulk GaN wafers as substrates for epitaxial growth. However, epitaxial layers can be grown on other substrates with reasonable success, in spite of the lattice mismatch.

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

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

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

Short wavelength emitters such as UV/blue semiconductor lasers are important for storage applications such as digital versatile discs (DVDs), which are higher density versions of compact discs (CDs). The storage density on these discs is inversely proportional to the square of the laser wavelength that is used to read the information. Thus reducing the laser wavelength by a factor of two leads to a four-fold increase of storage density. Such increased storage capacity opens up entirely new applications for DVDs that were not possible previously with conventional CDs, for example, the storage of full-length movies. A recent example of success in this rapidly progressing field is a 417 nm semiconductor laser made with InGaN multi–quantum-well heterostructures.

PROBLEMS

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

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

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

8.4 A major problem with solar cells is internal resistance, generally in the thin region at the surface, which must be only partially contacted, as in Fig. 8–5. Assume that the cell of Prob. 8.3 has a series resistance of 1 Ω, so that the cell voltage is reduced by the *IR* drop. Replot the *I–V* curve for this case and compare with the cell of Prob. 8.3.

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

8.6 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 $\Delta I = qALg_{op}\tau_n/\tau_t$ for this photoconductor, where τ_t is the transit time of electrons drifting down the length of the bar.

8.7 What composition x of $Al_xGa_{1-x}As$ would produce red light emission at 680 nm? What composition of $GaAs_{1-x}P_x$? $In_xGa_{1-x}P$?

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

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

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

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

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

Assume that a long p$^+$-n diode is uniformly illuminated by an optical signal, resuting in g_{op} EHP/cm^3-s.

(a) Show that the excess hole distribution in the n region is

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

(b) Calculate the hole diffusion current $I_p(x_n)$ and evaluate it at $x_n = 0$. Compare the result with Eq. (8–2) evaluated for a p$^+$-n junction.

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

8.11 The maximum power delivered by a solar cell can be found by maximizing the *I–V* product.

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

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

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

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

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

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

8.12 For a solar cell, Eq. (8–2) can be rewritten

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

Given the cell parameters of Prob. 8.11, plot the I–V curve as in Fig. 8–6 and draw the maximum power rectangle. Remember that I is a negative number but is ploted positive as I_r in the figure. I_{th} and I_{sc} are positive magnitudes in the equation.

8.13 Solar cells are severely degraded by unwanted series resistance. For the cell described in Prob. 8–4, include a series resistance R, which reduces the cell voltage by the amount IR. Calculate and plot the fill factor for a series resistance R from 0 to 5 Ω, and comment on the effect of R on cell efficiency.

8.14 Based upon Fig. 1–13, what ternary alloy, composition, and binary substrate can be used for an LED at the 1.55-μm optical fiber window? What type of epitaxial layer/substrate combination would you use for an LED with emission at 1.3 μm?

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

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

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

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

from Planck's radiation law. Given the result of Prob. 8.16, find the value of the ratio A_{21}/B_{12}.

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

READING LIST

Agrawal, G. P., and N. K. Dutta. *Long-Wavelength Semiconductor Lasers.* New York: Van Nostrand Reinhold, 1986.

Baughmann, M. G. D., J. C. Wright, A. B. Ellis, T. Kuech, and G. C. Lisensky. "Diode Lasers." *Journal of Chemical Education* 69 (February 1992): 89–95.

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

Buckley, D. N. "The Light Fantastic: Materials and Processing Technologies for Photonics." *The Electrochemical Society Interface* 1 (Winter 1992): 41+.

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

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

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

Craford, M.G. "LEDs Challenge the Incandescents." *IEEE Circuits and Devices* 8(5) (1992): 24+.

Crow, J.D. "Optical Interconnects Speed Interprocessor Nets." *IEEE Circuits and Devices* 7 (March 1991): 20–5.

Dagenais, M., R. F. Leheny, H. Temkin, and P. Bhattacharya. "Applications and Challenges of OEICs" *Journal of Lightwave Technology* 8 (June 1990): 846–62.

Das, P. *Lasers and Optical Engineering.* New York: Springer-Verlag, 1991.

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

Desurvire, E. "The Golden Age of Optical Fiber Amplifiers." *Physics Today* 47 (January 1994): 20–7.

Dupuis, R. D. "AlGaAs-GaAs Lasers Grown by MOCVD—A Review." *Journal of Crystal Growth* 55 (October 1981): 213–22.

Han, J., L. He, R. L. Gunshor, and A. V. Nurmikko. "Blue/Green Lasers Focus on the Market." *IEEE Circuits and Devices* 10 (March 1994): 18–23.

Hecht, J. "Diode-Laser Performance Rises as Structures Shrink." *Laser Focus World* 28 (May 1992): 127–8+.

Hecht, J. "Laser Action in Fibers Promises a Revolution in Communications." *Laser Focus World* 29 (February 1993): 75–6+.

Hecht, J. "Semiconductor Lasers Shine Out." *Electronics and Wireless World* 97 (April 1992): 302–5.

Hummel, R. E. *Electronic Properties of Materials,* 2nd ed., Berlin: Springer-Verlag, 1993.

Ikegami, T. M. And M. Nakahara. "Optical Fiber Amplifiers." *Proceedings of the SPIE* 1362, pt. 1 (1991): 350–60.

Jahns, J., and S. H. Lee, eds. *Optical Computing Hardware.* Boston: Academic Press, 1993.

Jewell, J. L., and G. R. Olbright. "Surface-Emitting Lasers Emerge from the Laboratory." *Laser Focus World* 28 (May 1992): 217–23.

Jungbluth, E. D. "Crystal Growth Methods Shape Communications Lasers." *Laser Focus World* 29 (February 1993): 61–72.

Leheny, R. F. "Optoelectronic Integration: A Technology for Future Telecommunication Systems." *IEEE Circuits and Devices* 5 (May 1989): 38–41.

Neamen, D. A. *Semiconductor Physics and Devices: Basic Principles.* Homewood, IL: Irwin, 1992.

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

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

Pollack, M. A. "Advances in Materials for Optoelectronic and Photonic Integrated Circuits." *Materials Science & Engineering B* B6 (July 1990): 233–45.

Saleh, B. E. A., and M. C. Teich. *Fundamentals of Photonics.* New York: Wiley, 1991.

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

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

Weisbuch, C., and B. Vinter. *Quantum Semiconductor Structures.* Boston: Academic Press, 1991.

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

Yariv, A. *Optical Electronics,* 3rd ed. New York: Holt, Rinehart, and Winston, 1985.

Zory, P. S., Jr. *Quantum Well Lasers.* Boston: Academic Press, 1993.