CHAPTER 2

Bipolar, MOS, and BiCMOS Integrated-Circuit Technology

2.1 Introduction

For the designer and user of integrated circuits, a knowledge of the details of the fabrication process is important for two reasons. First, IC technology has become pervasive because it provides the economic advantage of the planar process for fabricating complex circuitry at low cost through batch processing. Thus a knowledge of the factors influencing the cost of fabrication of integrated circuits is essential for both the selection of a circuit approach to solve a given design problem by the designer and the selection of a particular circuit for fabrication as a custom integrated circuit by the user. Second, integrated-circuit technology presents a completely different set of cost constraints to the circuit designer from those encountered with discrete components. The optimum choice of a circuit approach to realize a specified circuit function requires an understanding of the degrees of freedom available with the technology and the nature of the devices that are most easily fabricated on the integrated-circuit chip.

At the present time, analog integrated circuits are designed and fabricated in bipolar technology, in MOS technology, and in technologies that combine both types of devices in one process. The necessity of combining complex digital functions on the same integrated circuit with analog functions has resulted in an increased use of digital MOS technologies for analog functions, particularly those functions such as analog-digital conversion required for interfaces between analog signals and digital systems. However, bipolar technology is now used and will continue to be used in a wide range of applications requiring high-current drive capability and the highest levels of precision analog performance.

In this chapter, we first enumerate the basic processes that are fundamental in the fabrication of bipolar and MOS integrated circuits: solid-state diffusion, lithography, epitaxial growth, ion implantation, selective oxidation, and polysilicon deposition. Next, we describe the sequence of steps that are used in the fabrication of bipolar integrated circuits and describe the properties of the passive and active devices that result from the process sequence. Also, we examine several modifications to the basic process. In the next subsection, we consider the sequence of steps in fabricating MOS integrated circuits and describe the types of devices resulting in that technology. This is followed by descriptions of BiCMOS technology, silicon-germanium heterojunction transistors, and interconnect materials under study to replace aluminum wires and silicon-dioxide dielectric. Next, we examine the factors affecting the manufacturing cost of monolithic circuits and, finally, present packaging considerations for integrated circuits.

2.2 Basic Processes in Integrated-Circuit Fabrication

The fabrication of integrated circuits and most modern discrete component transistors is based on a sequence of photomasking, diffusion, ion implantation, oxidation, and epitaxial growth steps applied to a slice of silicon starting material called a wafer.[1,2] Before beginning a description of the basic process steps, we will first review the effects produced on the electrical properties of silicon by the addition of impurity atoms.

2.2.1 Electrical Resistivity of Silicon

The addition of small concentrations of *n*-type or *p*-type impurities to a crystalline silicon sample has the effect of increasing the number of majority carriers (electrons for *n*-type, holes for *p*-type) and decreasing the number of minority carriers. The addition of impurities is called *doping* the sample. For practical concentrations of impurities, the density of majority carriers is approximately equal to the density of the impurity atoms in the crystal. Thus for *n*-type material,

$$n_n \simeq N_D \tag{2.1}$$

where n_n (cm^{-3}) is the equilibrium concentration of electrons and N_D (cm^{-3}) is the concentration of *n*-type donor impurity atoms. For *p*-type material,

$$p_p \simeq N_A \tag{2.2}$$

where p_p (cm^{-3}) is the equilibrium concentration of holes and N_A (cm^{-3}) is the concentration of *p*-type acceptor impurities. Any increase in the equilibrium concentration of one type of carrier in the crystal must result in a decrease in the equilibrium concentration of the other. This occurs because the holes and electrons recombine with each other at a rate that is proportional to the product of the concentration of holes and the concentration of electrons. Thus the number of recombinations per second, R, is given by

$$R = \gamma np \tag{2.3}$$

where γ is a constant, and n and p are electron and hole concentrations, respectively, in the silicon sample. The generation of the hole-electron pairs is a thermal process that depends only on temperature; the rate of generation, G, is not dependent on impurity concentration. In equilibrium, R and G must be equal, so that

$$G = \text{constant} = R = \gamma np \tag{2.4}$$

If no impurities are present, then

$$n = p = n_i(T) \tag{2.5}$$

where n_i (cm^{-3}) is the *intrinsic* concentration of carriers in a pure sample of silicon. Equations 2.4 and 2.5 establish that, for any impurity concentration, $\gamma np = \text{constant} = \gamma n_i^2$, and thus

$$np = n_i^2(T) \tag{2.6}$$

Equation 2.6 shows that as the majority carrier concentration is increased by impurity doping, the minority carrier concentration is decreased by the same factor so that product np is constant in equilibrium. For impurity concentrations of practical interest, the majority carriers outnumber the minority carriers by many orders of magnitude.

The importance of minority- and majority-carrier concentrations in the operation of the transistor was described in Chapter 1. Another important effect of the addition of impurities is

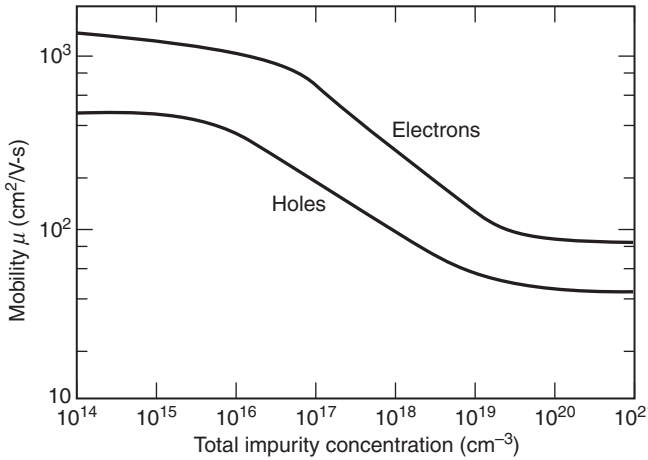


Figure 2.1 Hole and electron mobility as a function of doping in silicon.[3]

an increase in the ohmic conductivity of the material itself. This conductivity is given by

$$\sigma = q(\mu_n n + \mu_p p) \tag{2.7}$$

where μ_n (cm^2/V-s) is the electron mobility, μ_p (cm^2/V-s) is the hole mobility, and σ $(\Omega\text{-cm})^{-1}$ is the electrical conductivity. For an n-type sample, substitution of (2.1) and (2.6) in (2.7) gives

$$\sigma = q\left(\mu_n N_D + \mu_p \frac{n_i^2}{N_D}\right) \simeq q\mu_n N_D \tag{2.8}$$

For a p-type sample, substitution of (2.2) and (2.6) in (2.7) gives

$$\sigma = q\left(\mu_n \frac{n_i^2}{N_A} + \mu_p N_A\right) \simeq q\mu_p N_A \tag{2.9}$$

The mobility μ is different for holes and electrons and is also a function of the impurity concentration in the crystal for high impurity concentrations. Measured values of mobility in silicon as a function of impurity concentration are shown in Fig. 2.1. The resistivity ρ (Ω-cm) is usually specified in preference to the conductivity, and the resistivity of n- and p-type silicon as a function of impurity concentration is shown in Fig. 2.2. The conductivity and resistivity are related by the simple expression $\rho = 1/\sigma$.

2.2.2 Solid-State Diffusion

Solid-state diffusion of impurities in silicon is the movement, usually at high temperature, of impurity atoms from the surface of the silicon sample into the bulk material. During this high-temperature process, the impurity atoms replace silicon atoms in the lattice and are termed *substitutional impurities*. Since the doped silicon behaves electrically as p-type or n-type material depending on the type of impurity present, regions of p-type and n-type material can be formed by solid-state diffusion.

The nature of the diffusion process is illustrated by the conceptual example shown in Figs. 2.3 and 2.4. We assume that the silicon sample initially contains a uniform concentration of n-type impurity of 10^{15} atoms per cubic centimeter. Commonly used n-type impurities in silicon are phosphorus, arsenic, and antimony. We further assume that by some means we deposit atoms of p-type impurity on the top surface of the silicon sample. The most commonly used

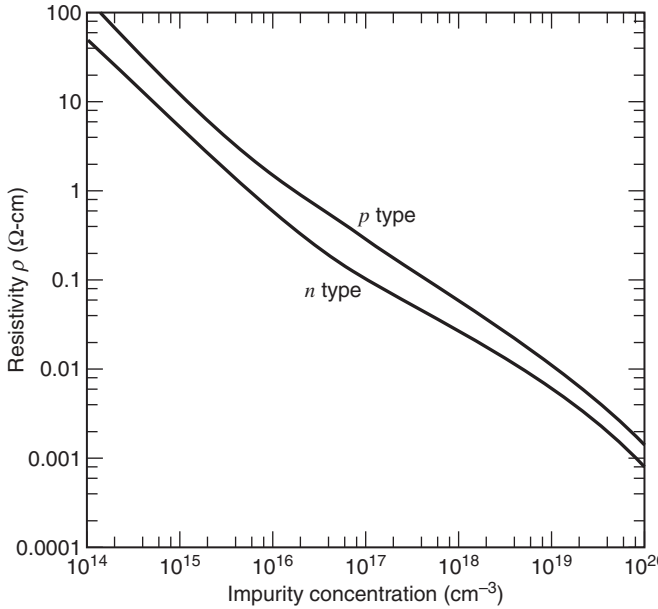


Figure 2.2 Resistivity of *p*- and *n*-type silicon as a function of impurity concentration.[4]

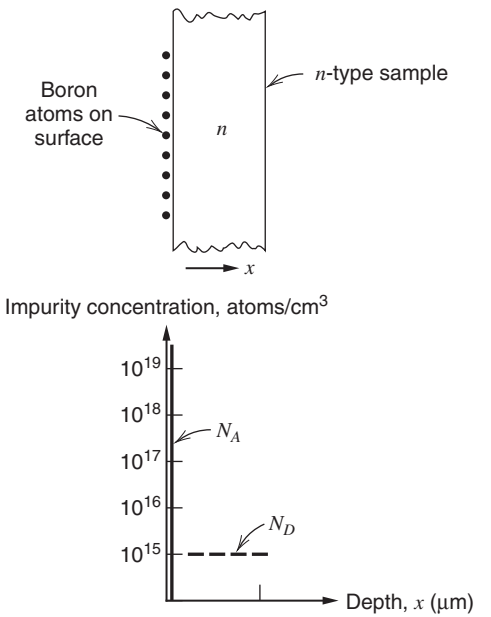


Figure 2.3 An *n*-type silicon sample with boron deposited on the surface.

p-type impurity in silicon device fabrication is boron. The distribution of impurities prior to the diffusion step is illustrated in Fig. 2.3. The initial placement of the impurity atoms on the surface of the silicon is called the *predeposition step* and can be accomplished by a number of different techniques.

If the sample is now subjected to a high temperature of about 1100°C for a time of about one hour, the impurities *diffuse* into the sample, as illustrated in Fig. 2.4. Within the silicon, the regions in which the *p*-type impurities outnumber the original *n*-type impurities display *p*-type electrical behavior, whereas the regions in which the *n*-type impurities are more numerous

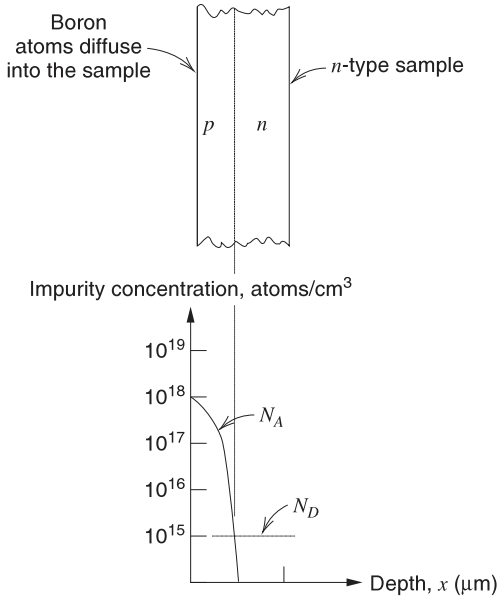


Figure 2.4 Distribution of impurities after diffusion.

display *n*-type electrical behavior. The diffusion process has allowed the formation of a *pn* junction within the continuous crystal of silicon material. The depth of this junction from the surface varies from 0.1 μm to 20 μm for silicon integrated-circuit diffusions (where 1 μm = 1 micrometer = 10^{-6} m).

2.2.3 Electrical Properties of Diffused Layers

The result of the diffusion process is often a thin layer near the surface of the silicon sample that has been converted from one impurity type to another. Silicon devices and integrated circuits are constructed primarily from these layers. From an electrical standpoint, if the *pn* junction formed by this diffusion is reverse biased, then the layer is electrically isolated from the underlying material by the reverse-biased junction, and the electrical properties of the layer itself can be measured. The electrical parameter most often used to characterize such layers is the *sheet resistance*. To define this quantity, consider the resistance of a uniformly doped sample of length L, width W, thickness T, and *n*-type doping concentration N_D, as shown in Fig. 2.5. The resistance is

$$R = \frac{\rho L}{WT} = \frac{1}{\sigma}\frac{L}{WT}$$

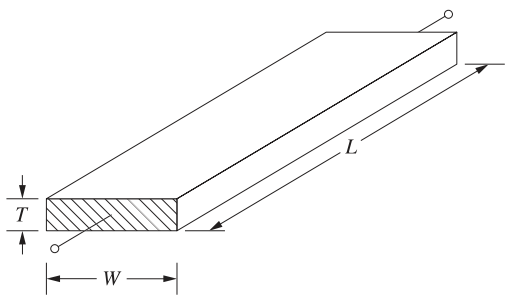


Figure 2.5 Rectangular sample for calculation of sheet resistance.

Substitution of the expression for conductivity σ from (2.8) gives

$$R = \left(\frac{1}{q\mu_n N_D}\right)\frac{L}{WT} = \frac{L}{W}\left(\frac{1}{q\mu_n N_D T}\right) = \frac{L}{W}R_\square \tag{2.10}$$

Quantity $R_\square$ is the *sheet resistance* of the layer and has units of Ohms. Since the sheet resistance is the resistance of any *square* sheet of material with thickness T, its units are often given as *Ohms per square* ($\Omega/\square$) rather than simply Ohms. The sheet resistance can be written in terms of the resistivity of the material, using (2.8), as

$$R_\square = \frac{1}{q\mu_n N_D T} = \frac{\rho}{T} \tag{2.11}$$

The diffused layer illustrated in Fig. 2.6 is similar to this case except that the impurity concentration is not uniform. However, we can consider the layer to be made up of a parallel combination of many thin conducting sheets. The conducting sheet of thickness dx at depth x has a conductance

$$dG = q\left(\frac{W}{L}\right)\mu_n N_D(x)\,dx \tag{2.12}$$

To find the total conductance, we sum all the contributions.

$$G = \int_0^{x_j} q\frac{W}{L}\mu_n N_D(x)\,dx = \frac{W}{L}\int_0^{x_j} q\mu_n N_D(x)\,dx \tag{2.13}$$

Inverting (2.13), we obtain

$$R = \frac{L}{W}\left[\frac{1}{\displaystyle\int_0^{x_j} q\mu_n N_D(x)\,dx}\right] \tag{2.14}$$

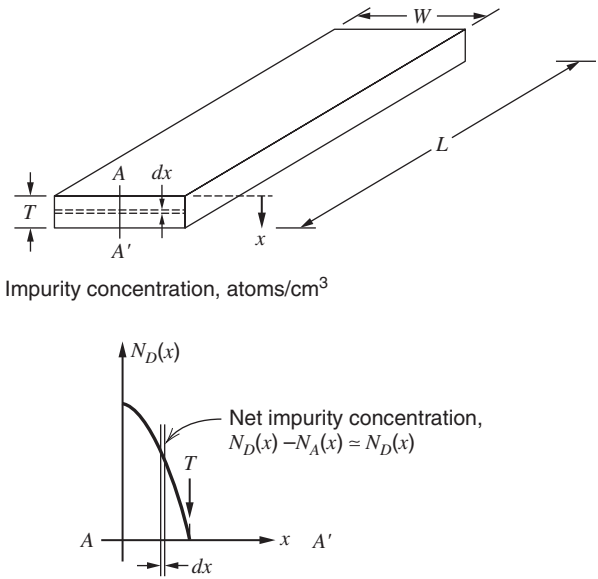


Figure 2.6 Calculation of the resistance of a diffused layer.

Comparison of (2.10) and (2.14) gives

$$R_{\square} = \left[\int_0^{x_j} q\mu_n N_D(x)\,dx\right]^{-1} \simeq \left[q\overline{\mu}_n \int_0^{x_j} N_D(x)\,dx\right]^{-1} \tag{2.15}$$

where $\overline{\mu}_n$ is the average mobility. Thus (2.10) can be used for diffused layers if the appropriate value of $R_{\square}$ is used. Equation 2.15 shows that the sheet resistance of the diffused layer depends on the total number of impurity atoms in the layer per unit area. The depth x_j in (2.13), (2.14), and (2.15) is actually the distance from the surface to the edge of the junction depletion layer, since the donor atoms within the depletion layer do not contribute to conduction. Sheet resistance is a useful parameter for the electrical characterization of diffusion processes and is a key parameter in the design of integrated resistors. The sheet resistance of a diffused layer is easily measured in the laboratory; the actual evaluation of (2.15) is seldom necessary.

■ **EXAMPLE**

Calculate the resistance of a layer with length 50 μm and width 5 μm in material of sheet resistance 200 Ω/□.

From (2.10)

$$R = \frac{50}{5} \times 200\ \Omega = 2\ \text{k}\Omega$$

Note that this region constitutes 10 squares in series, and R is thus 10 times the sheet resistance. ■

In order to use these diffusion process steps to fabricate useful devices, the diffusion must be restricted to a small region on the surface of the sample rather than the entire planar surface. This restriction is accomplished with photolithography.

2.2.4 Photolithography

When a sample of crystalline silicon is placed in an oxidizing environment, a layer of silicon dioxide will form at the surface. This layer acts as a barrier to the diffusion of impurities, so that impurities separated from the surface of the silicon by a layer of oxide do not diffuse into the silicon during high-temperature processing. A *pn* junction can thus be formed in a selected location on the sample by first covering the sample with a layer of oxide (called an *oxidation step*), removing the oxide in the selected region, and then performing a predeposition and diffusion step. The selective removal of the oxide in the desired areas is accomplished with photolithography. This process is illustrated by the conceptual example of Fig. 2.7. Again we assume the starting material is a sample of *n*-type silicon. We first perform an oxidation step in which a layer of silicon dioxide (SiO_2) is thermally grown on the top surface, usually of thickness of 0.2 μm to 1 μm. The wafer following this step is shown in Fig. 2.7*a*. Then the sample is coated with a thin layer of photosensitive material called photoresist. When this material is exposed to a particular wavelength of light, it undergoes a chemical change and, in the case of positive photoresist, becomes soluble in certain chemicals in which the unexposed photoresist is insoluble. The sample at this stage is illustrated in Fig. 2.7*b*. To define the desired diffusion areas on the silicon sample, a photomask is placed over the surface of the sample; this photomask is opaque except for clear areas where the diffusion is to take place. Light of the appropriate wavelength is directed at the sample, as shown in Fig. 2.7*c*, and falls on the photoresist only in the clear areas of the mask. These areas of the resist are then chemically

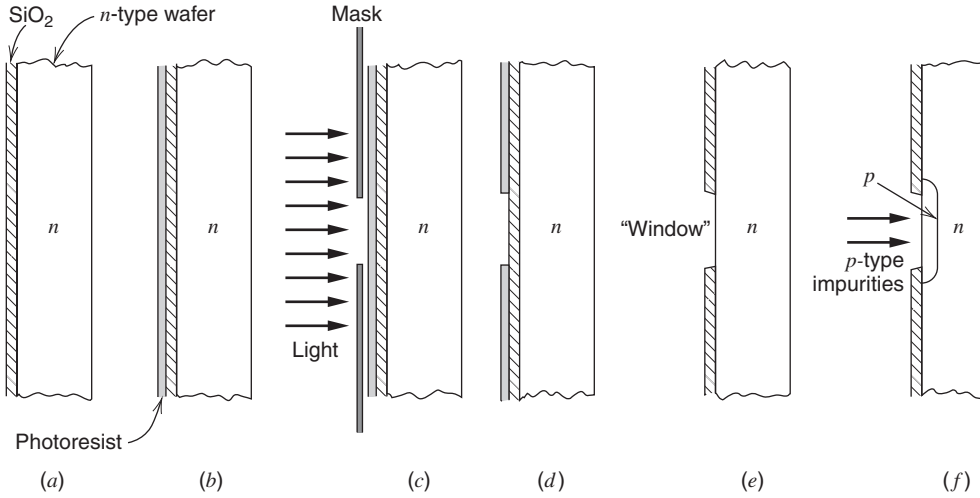


Figure 2.7 Conceptual example of the use of photolithography to form a *pn* junction diode. (*a*) Grow SiO_2. (*b*) Apply photoresist. (*c*) Expose through mask. (*d*) Develop photoresist. (*e*) Etch SiO_2 and remove photoresist. (*f*) Predeposit and diffuse impurities.

dissolved in the development step, as shown in Fig. 2.7*d*. The unexposed areas of the photoresist are impervious to the developer.

Since the objective is the formation of a region clear of SiO_2, the next step is the etching of the oxide. This step can be accomplished by dipping the sample in an etching solution, such as hydrofluoric acid, or by exposing it to an electrically produced plasma in a plasma etcher. In either case, the result is that in the regions where the photoresist has been removed, the oxide is etched away, leaving the bare silicon surface.

The remaining photoresist is next removed by a chemical stripping operation, leaving the sample with holes, or *windows,* in the oxide at the desired locations, as shown in Fig. 2.7*e*. The sample now undergoes a predeposition and diffusion step, resulting in the formation of *p*-type regions where the oxide had been removed, as shown in Fig. 2.7*f*. In some instances, the impurity to be locally added to the silicon surface is deposited by using ion implantation (see Section 2.2.6). This method of insertion can often take place through the silicon dioxide so that the oxide-etch step is unnecessary.

The minimum dimension of the diffused region that can be routinely formed with this technique in device production has decreased with time, and at present is approximately 0.1 μm $\times$ 0.1 μm. The number of such regions that can be fabricated simultaneously can be calculated by noting that the silicon sample used in the production of integrated circuits is a round slice, typically 4 inches to 12 inches in diameter and 250 μm thick. Thus the number of electrically independent *pn* junctions of dimension 0.1 μm $\times$ 0.1 μm spaced 0.1 μm apart that can be formed on one such wafer is on the order of 10^{12}. In actual integrated circuits, a number of masking and diffusion steps are used to form more complex structures such as transistors, but the key points are that photolithography is capable of defining a large number of devices on the surface of the sample and that all of these devices are batch fabricated at the same time. Thus the cost of the photomasking and diffusion steps applied to the wafer during the process is divided among the devices or circuits on the wafer. This ability to fabricate hundreds or thousands of devices at once is the key to the economic advantage of IC technology.

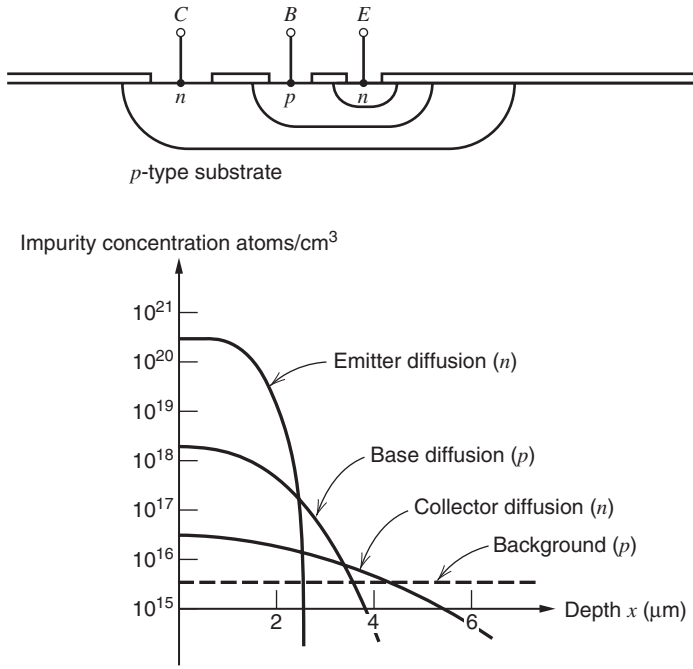


Figure 2.8 Triple-diffused transistor and resulting impurity profile.

2.2.5 Epitaxial Growth

Early planar transistors and the first integrated circuits used only photomasking and diffusion steps in the fabrication process. However, all-diffused integrated circuits had severe limitations compared with discrete component circuits. In a triple-diffused bipolar transistor, as illustrated in Fig. 2.8, the collector region is formed by an *n*-type diffusion into the *p*-type wafer. The drawbacks of this structure are that the series collector resistance is high and the collector-to-emitter breakdown voltage is low. The former occurs because the impurity concentration in the portion of the collector diffusion below the collector-base junction is low, giving the region high resistivity. The latter occurs because the concentration of impurities near the surface of the collector is relatively high, resulting in a low breakdown voltage between the collector and base diffusions at the surface, as described in Chapter 1. To overcome these drawbacks, the impurity concentration should be low at the collector-base junction for high breakdown voltage but high below the junction for low collector resistance. Such a concentration profile cannot be realized with diffusions alone, and the epitaxial growth technique was adopted as a result.

Epitaxial (epi) growth consists of formation of a layer of single-crystal silicon on the surface of the silicon sample so that the crystal structure of the silicon is continuous across the interface. The impurity concentration in the epi layer can be controlled independently and can be greater or smaller than in the substrate material. In addition, the epi layer is often of opposite impurity type from the substrate on which it is grown. The thickness of epi layers used in integrated-circuit fabrication varies from 1 μm to 20 μm, and the growth of the layer is accomplished by placing the wafer in an ambient atmosphere containing silicon tetrachloride ($SiCl_4$) or silane (SiH_4) at an elevated temperature. A chemical reaction takes place in which elemental silicon is deposited on the surface of the wafer, and the resulting surface layer of

silicon is crystalline in structure with few defects if the conditions are carefully controlled. Such a layer is suitable as starting material for the fabrication of bipolar transistors. Epitaxy is also utilized in some CMOS and most BiCMOS technologies.

2.2.6 Ion Implantation

Ion implantation is a technique for directly inserting impurity atoms into a silicon wafer.[5,6] The wafer is placed in an evacuated chamber, and ions of the desired impurity species are directed at the sample at high velocity. These ions penetrate the surface of the silicon wafer to an average depth of from less than 0.1 μm to about 0.6 μm, depending on the velocity with which they strike the sample. The wafer is then held at a moderate temperature for a period of time (for example, 800°C for 10 minutes) in order to allow the ions to become mobile and fit into the crystal lattice. This is called an *anneal step* and is essential to allow repair of any crystal damage caused by the implantation. The principal advantages of ion implantation over conventional diffusion are (1) that small amounts of impurities can be reproducibly deposited, and (2) that the amount of impurity deposited per unit area can be precisely controlled. In addition, the deposition can be made with a high level of uniformity across the wafer. Another useful property of ion-implanted layers is that the peak of the impurity concentration profile can be made to occur below the surface of the silicon, unlike with diffused layers. This allows the fabrication of implanted bipolar structures with properties that are significantly better than those of diffused devices. This technique is also widely applied in MOS technology where small, well-controlled amounts of impurity are required at the silicon surface for adjustment of device thresholds, as described in Section 1.5.1.

2.2.7 Local Oxidation

In both MOS and bipolar technologies, the need often arises to fabricate regions of the silicon surface that are covered with relatively thin silicon dioxide, adjacent to areas covered by relatively thick oxide. Typically, the former regions constitute the active-device areas, whereas the latter constitute the regions that electrically isolate the devices from each other. A second requirement is that the transition from thick to thin regions must be accomplished without introducing a large vertical step in the surface geometry of the silicon, so that the metallization and other patterns that are later deposited can lie on a relatively planar surface. Local oxidation is used to achieve this result. The local oxidation process begins with a sample that already has a thin oxide grown on it, as shown in Fig. 2.9*a*. First a layer of silicon nitride (SiN) is deposited on the sample and subsequently removed with a masking step from all areas where thick oxide is to be grown, as shown in Fig. 2.9*b*. Silicon nitride acts as a barrier to oxygen atoms that might otherwise reach the Si-SiO_2 interface and cause further oxidation. Thus when a subsequent long, high-temperature oxidation step is carried out, a thick oxide is grown in the regions where there is no nitride, but no oxidation takes place under the nitride. The resulting geometry after nitride removal is shown in Fig. 2.9*c*. Note that the top surface of the silicon dioxide has a smooth transition from thick to thin areas and that the height of this transition is less than the oxide thickness difference because the oxidation in the thick oxide regions consumes some of the underlying silicon.

2.2.8 Polysilicon Deposition

Many process technologies utilize layers of polycrystalline silicon that are deposited during fabrication. After deposition of the polycrystalline silicon layer on the wafer, the desired features are defined by using a masking step and can serve as gate electrodes for silicon-gate MOS

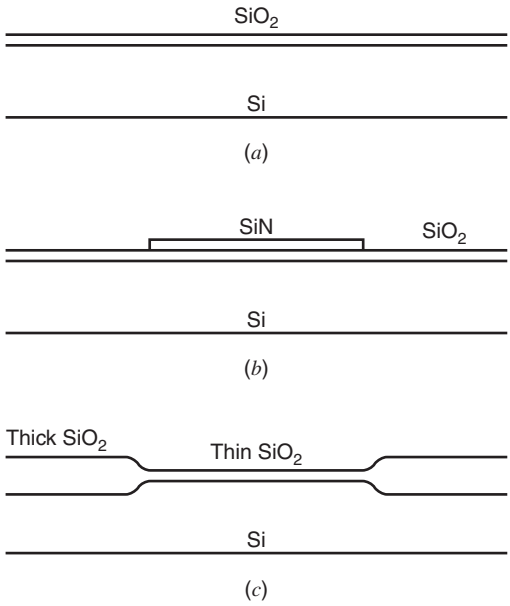


Figure 2.9 Local oxidation process. (*a*) Silicon sample prior to deposition of nitride. (*b*) After nitride deposition and definition. (*c*) After oxidation and nitride removal.

transistors, emitters of bipolar transistors, plates of capacitors, resistors, fuses, and interconnect layers. The sheet resistance of such layers can be controlled by the impurity added, much like bulk silicon, in a range from about 20 Ω/□ up to very high values. The process that is used to deposit the layer is much like that used for epitaxy. However, since the deposition is usually over a layer of silicon dioxide, the layer does not form as a single-crystal extension of the underlying silicon but forms as a granular (or polysilicon) film. Some MOS technologies contain as many as three separate polysilicon layers, separated from one another by layers of SiO_2.

2.3 High-Voltage Bipolar Integrated-Circuit Fabrication

Integrated-circuit fabrication techniques have changed dramatically since the invention of the basic planar process. This change has been driven by developments in photolithography, processing techniques, and also the trend to reduce power-supply voltages in many systems. Developments in photolithography have reduced the minimum feature size attainable from tens of microns to the submicron level. The precise control allowed by ion implantation has resulted in this technique becoming the dominant means of predepositing impurity atoms. Finally, many circuits now operate from 3 V or 5 V power supplies instead of from the ± 15 V supplies used earlier to achieve high dynamic range in stand-alone integrated circuits, such as operational amplifiers. Reducing the operating voltages allows closer spacing between devices in an IC. It also allows shallower structures with higher frequency capability. These effects stem from the fact that the thickness of junction depletion layers is reduced by reducing operating voltages, as described in Chapter 1. Thus the highest-frequency IC processes are designed to operate from 5-V supplies or less and are generally not usable at higher supply voltages. In fact, a fundamental trade-off exists between the frequency capability of a process and its breakdown voltage.

In this section, we examine first the sequence of steps used in the fabrication of high-voltage bipolar integrated circuits using junction isolation. This was the original IC process

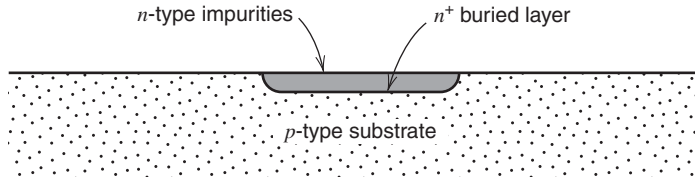


Figure 2.10 Buried-layer diffusion.

and is useful as a vehicle to illustrate the basic methods of IC fabrication. It is still used in various forms to fabricate high-voltage circuits.

The fabrication of a junction-isolated bipolar integrated circuit involves a sequence of from six to eight masking and diffusion steps. The starting material is a wafer of *p*-type silicon, usually 250 μm thick and with an impurity concentration of approximately 10^{16} atoms/cm^3. We will consider the sequence of diffusion steps required to form an *npn* integrated-circuit transistor. The first mask and diffusion step, illustrated in Fig. 2.10, forms a low-resistance *n*-type layer that will eventually become a low-resistance path for the collector current of the transistor. This step is called the *buried-layer diffusion*, and the layer itself is called the *buried layer*. The sheet resistance of the layer is in the range of 20 to 50 Ω/□, and the impurity used is usually arsenic or antimony because these impurities diffuse slowly and thus do not greatly redistribute during subsequent processing.

After the buried-layer step, the wafer is stripped of all oxide and an epi layer is grown, as shown in Fig. 2.11. The thickness of the layer and its *n*-type impurity concentration determine the collector-base breakdown voltage of the transistors in the circuit since this material forms the collector region of the transistor. For example, if the circuit is to operate at a power-supply voltage of 36 V, the devices generally are required to have BV_{CEO} breakdown voltages above this value. As described in Chapter 1, this implies that the plane breakdown voltage in the collector-base junction must be several times this value because of the effects of collector avalanche multiplication. For $BV_{CEO} = 36$ V, a collector-base plane breakdown voltage of approximately 90 V is required, which implies an impurity concentration in the collector of approximately 10^{15} atoms/cm^3 and a resistivity of 5 Ω-cm. The thickness of the epitaxial layer then must be large enough to accommodate the depletion layer associated with the collector-base junction. At 36 V, the results of Chapter 1 can be used to show that the depletion-layer thickness is approximately 6 μm. Since the buried layer diffuses outward approximately 8 μm during subsequent processing, and the base diffusion will be approximately 3 μm deep, a total epitaxial layer thickness of 17 μm is required for a 36-V circuit. For circuits with lower operating voltages, thinner and more heavily doped epitaxial layers are used to reduce the transistor collector series resistance, as will be shown later.

Following the epitaxial growth, an oxide layer is grown on the top surface of the epitaxial layer. A mask step and boron (*p*-type) predeposition and diffusion are performed, resulting in the structure shown in Fig. 2.12. The function of this diffusion is to isolate the collectors of the transistors from each other with reverse-biased *pn* junctions, and it is termed the *isolation*

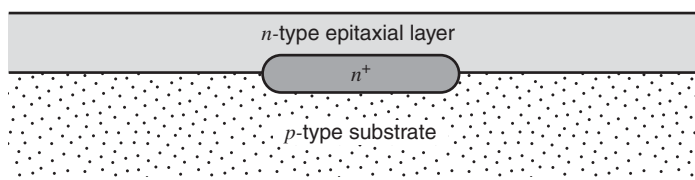


Figure 2.11 Bipolar integrated-circuit wafer following epitaxial growth.

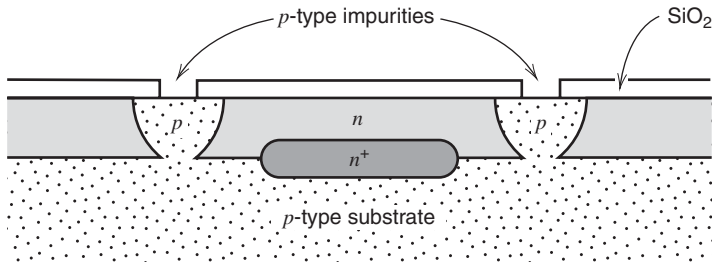


Figure 2.12 Structure following isolation diffusion.

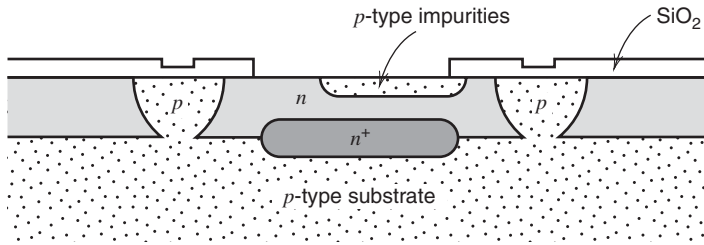


Figure 2.13 Structure following base diffusion.

diffusion. Because of the depth to which the diffusion must penetrate, this diffusion requires several hours in a diffusion furnace at temperatures of about 1200°C. The isolated diffused layer has a sheet resistance from 20 Ω/□ to 40 Ω/□.

The next steps are the base mask, base predeposition, and base diffusion, as shown in Fig. 2.13. The latter is usually a boron diffusion, and the resulting layer has a sheet resistance of from 100 Ω/□ to 300 Ω/□ , and a depth of 1 μm to 3 μm at the end of the process. This diffusion forms not only the bases of the transistors, but also many of the resistors in the circuit, so that control of the sheet resistance is important.

Following the base diffusion, the emitters of the transistors are formed by a mask step, *n*-type predeposition, and diffusion, as shown in Fig. 2.14. The sheet resistance is between 2 Ω/□ and 10 Ω/□, and the depth is 0.5 μm to 2.5 μm after the diffusion. This diffusion step is also used to form a low-resistance region, which serves as the contact to the collector region. This is necessary because ohmic contact is difficult to accomplish between aluminum metallization and the high-resistivity epitaxial material directly. The next masking step, the contact mask, is used to open holes in the oxide over the emitter, the base, and the collector of the transistors so that electrical contact can be made to them. Contact windows are also opened

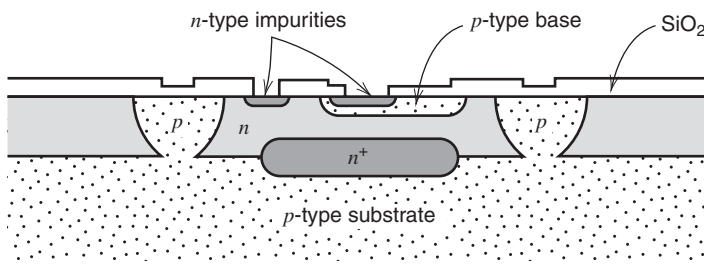


Figure 2.14 Structure following emitter diffusion.

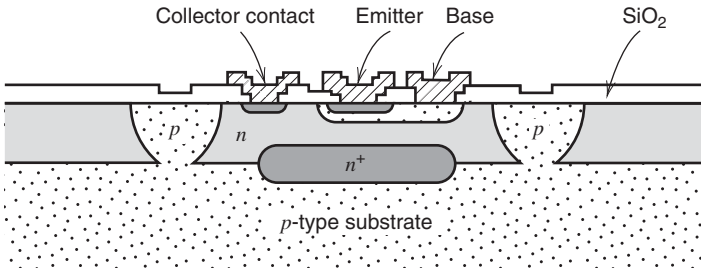


Figure 2.15 Final structure following contact mask and metallization.

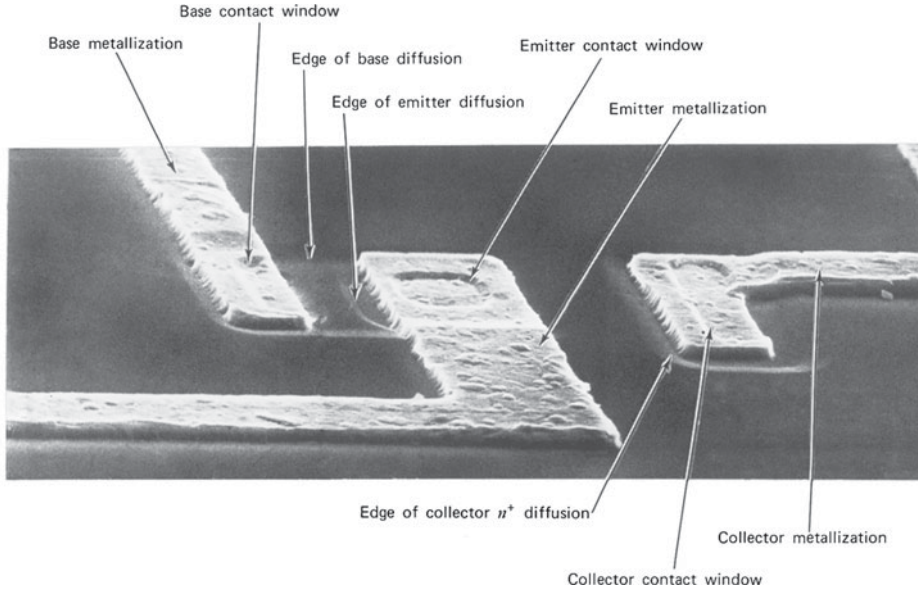


Figure 2.16 Scanning electron microscope photograph of *npn* transistor structure.

for the passive components on the chip. The entire wafer is then coated with a thin (about 1 μm) layer of aluminum that will interconnect the circuit elements. The actual interconnect pattern is defined by the last mask step, in which the aluminum is etched away in the areas where the photoresist is removed in the develop step. The final structure is shown in Fig. 2.15. A microscope photograph of an actual structure of the same type is shown in Fig. 2.16. The terraced effect on the surface of the device results from the fact that additional oxide is grown during each diffusion cycle, so that the oxide is thickest over the epitaxial region, where no oxide has been removed, is less thick over the base and isolation regions, which are both opened at the base mask step, and is thinnest over the emitter diffusion. A typical diffusion profile for a high-voltage, deep-diffused analog integrated circuit is shown in Fig. 2.17.

This sequence allows simultaneous fabrication of a large number (often thousands) of complex circuits on a single wafer. The wafer is then placed in an automatic tester, which checks the electrical characteristics of each circuit on the wafer and puts an ink dot on circuits that fail to meet specifications. The wafer is then broken up, by sawing or scribing and breaking, into individual circuits. The resulting silicon chips are called *dice*, and the singular is *die*. Each good die is then mounted in a package, ready for final testing.

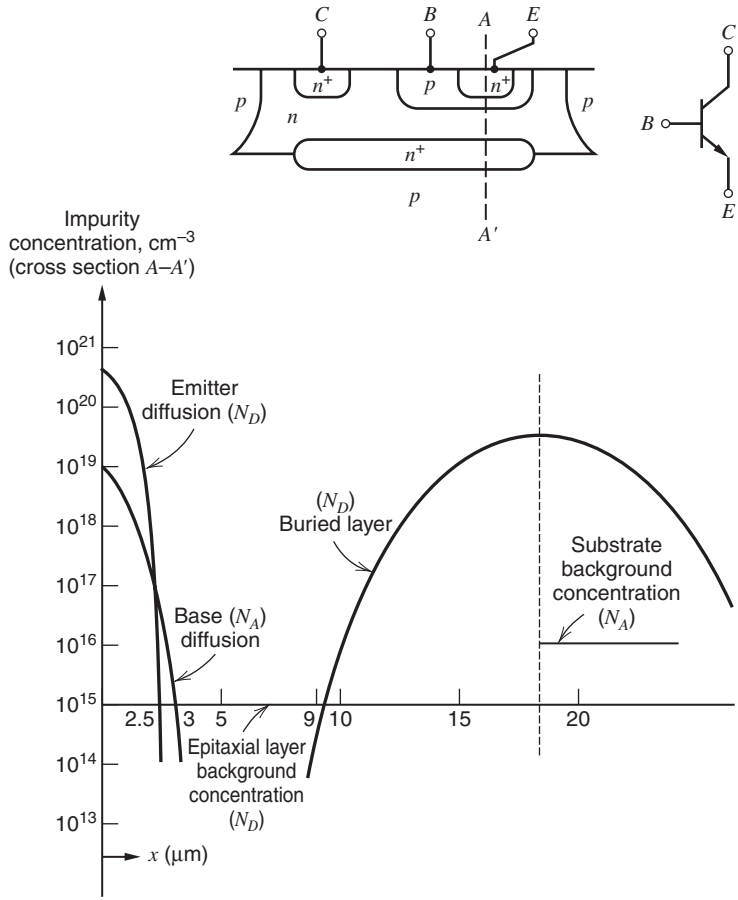


Figure 2.17 Typical impurity concentration for a monolithic *npn* transistor in a high-voltage, deep-diffused process.

2.4 Advanced Bipolar Integrated-Circuit Fabrication

A large fraction of bipolar analog integrated circuits currently manufactured uses the basic technology described in the previous section, or variations thereof. The fabrication sequence is relatively simple and low in cost. However, many of the circuit applications of commercial importance have demanded steadily increasing frequency response capability, which translates directly to a need for transistors of higher frequency-response capability in the technology. The higher speed requirement dictates a device structure with thinner base width to reduce base transit time and smaller dimensions overall to reduce parasitic capacitances. The smaller device dimensions require that the width of the junction depletion layers within the structure be reduced in proportion, which in turn requires the use of lower circuit operating voltages and higher impurity concentrations in the device structure. To meet this need, a class of bipolar fabrication technologies has evolved that, compared to the high-voltage process sequence described in the last section, use much thinner and more heavily doped epitaxial layers, selectively oxidized regions for isolation instead of diffused junctions, and a polysilicon layer as the source of dopant for the emitter. Because of the growing importance of this class of bipolar process, the sequence for such a process is described in this section.

The starting point for the process is similar to that for the conventional process, with a mask and implant step resulting in the formation of a heavily-doped n^+ buried layer in a p-type

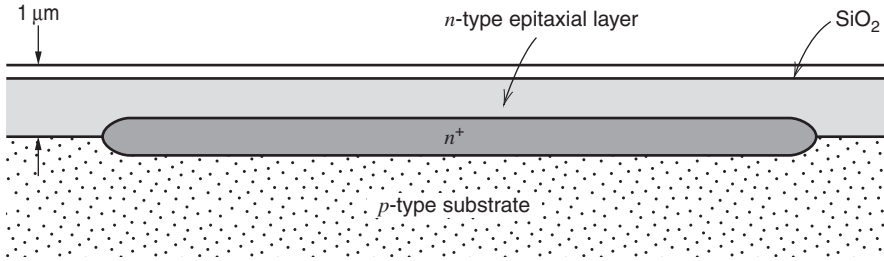


Figure 2.18 Device cross section following initial buried-layer mask, implant, and epitaxial-layer growth.

substrate. Following this step, a thin n-type epitaxial layer is grown, about 1 μm in thickness and about 0.5 Ω-cm in resistivity. The result after these steps is shown in cross section in Fig. 2.18.

Next, a selective oxidation step is carried out to form the regions that will isolate the transistor from its neighbors and also isolate the collector-contact region from the rest of the transistor. The oxidation step is as described in Section 2.2.7, except that prior to the actual growth of the thick SiO_2 layer, an etching step is performed to remove silicon material from the regions where oxide will be grown. If this is not done, the thick oxide growth results in elevated *humps* in the regions where the oxide is grown. The steps around these humps cause difficulty in coverage by subsequent layers of metal and polysilicon that will be deposited. The removal of some silicon material before oxide growth results in a nearly planar surface after the oxide is grown and removes the step coverage problem in subsequent processing. The resulting structure following this step is shown in Fig. 2.19. Note that the SiO_2 regions extend all the way down to the p-type substrate, electrically isolating the n-type epi regions from one another. These regions are often referred to as *moats*. Because growth of oxide layers thicker than a micron or so requires impractically long times, this method of isolation is practical only for very thin transistor structures.

Next, two mask and implant steps are performed. A heavy n^+ implant is made in the collector-contact region and diffused down to the buried layer, resulting in a low-resistance path to the collector. A second mask is performed to define the base region, and a thin-base p-type implant is performed. The resulting structure is shown in Fig. 2.20.

A major challenge in fabricating this type of device is the formation of very thin base and emitter structures, and then providing low-resistance ohmic contact to these regions. This is most often achieved using polysilicon as a doping source. An n^+ doped layer of polysilicon is deposited and masked to leave polysilicon only in the region directly over the emitter. During subsequent high-temperature processing steps, the dopant (usually arsenic) diffuses out of the polysilicon and into the crystalline silicon, forming a very thin, heavily doped emitter region. Following the poly deposition, a heavy p-type implant is performed, which results in a more

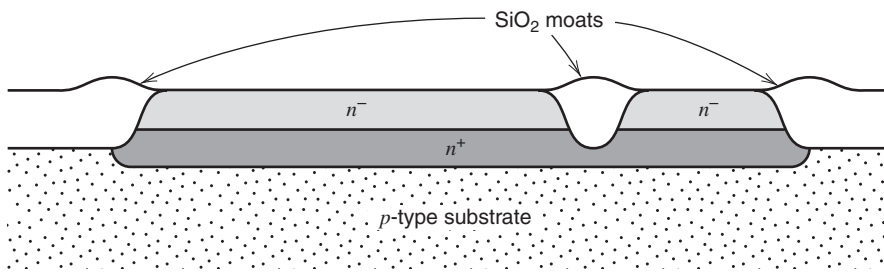


Figure 2.19 Device cross section following selective etch and oxidation to form thick-oxide moats.

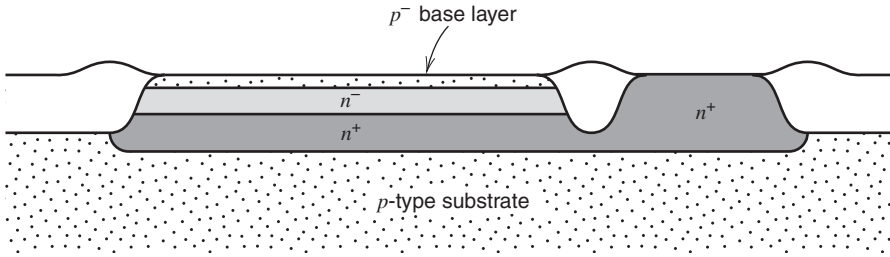


Figure 2.20 Device cross section following mask, implant, and diffusion of collector n^+ region, and mask and implant of base p-type region.

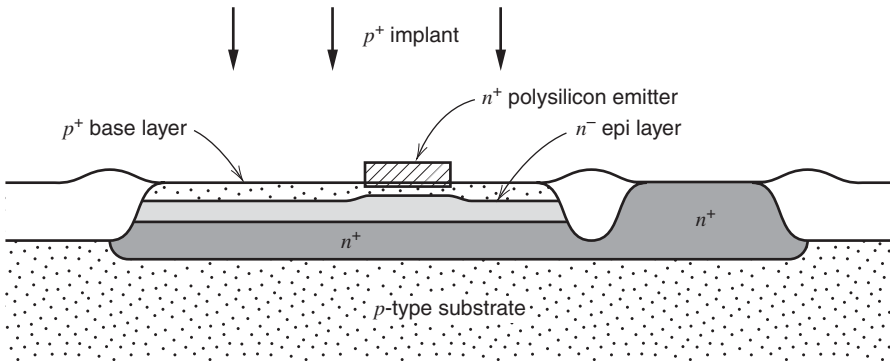


Figure 2.21 Device cross section following poly deposition and mask, base p-type implant, and thermal diffusion cycle.

heavily doped p-type layer at all points in the base region except directly under the polysilicon, where the polysilicon itself acts as a mask to prevent the boron atoms from reaching this part of the base region. The structure that results following this step is shown in Fig. 2.21.

This method of forming low-resistance regions to contact the base is called a *self-aligned structure* because the alignment of the base region with the emitter happens automatically and does not depend on mask alignment. Similar processing is used in MOS technology, described later in this chapter.

The final device structure after metallization is shown in Fig. 2.22. Since the moats are made of SiO_2, the metallization contact windows can overlap into them, a fact that dramatically reduces the minimum achievable dimensions of the base and collector regions. All exposed

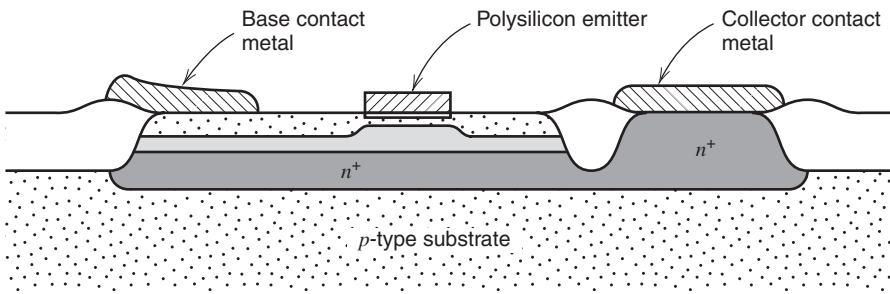


Figure 2.22 Final device cross section. Note that collector and base contact windows can overlap moat regions. Emitter contact for the structure shown here would be made on an extension of the polysilicon emitter out of the device active area, allowing the minimum possible emitter size.

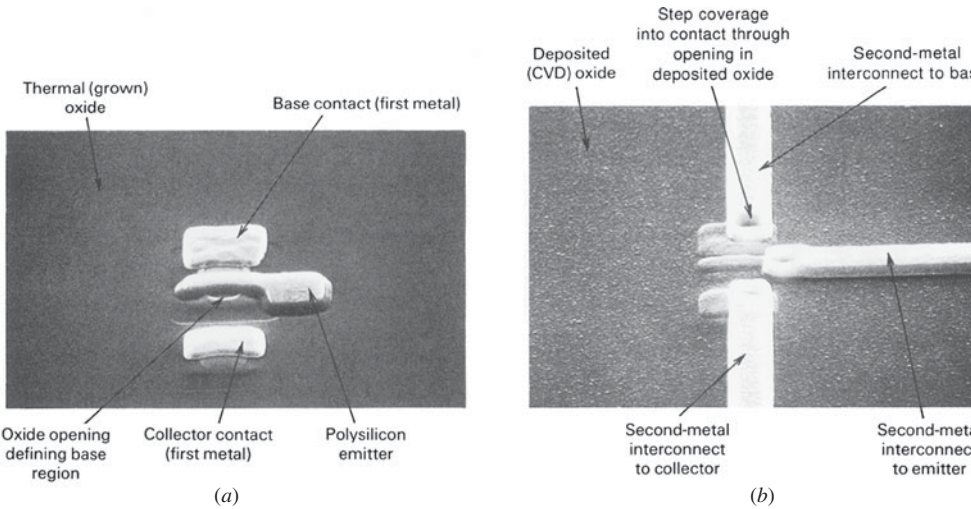


Figure 2.23 Scanning-electron-microscope photographs of a bipolar transistor in an advanced, polysilicon-emitter, oxide-isolated process. (*a*) After polysilicon emitter definition and first-metal contact to the base and collector. The polysilicon emitter is 1 μm wide. (*b*) After oxide deposition, contact etch, and second-metal interconnect. [QUBic process photograph courtesy of Signetics.]

silicon and polysilicon is covered with a highly conductive silicide (a compound of silicon and a refractory metal such as tungsten) to reduce series and contact resistance. For minimum-dimension transistors, the contact to the emitter is made by extending the polysilicon to a region outside the device active area and forming a metal contact to the polysilicon there. A photograph of such a device is shown in Fig. 2.23, and a typical impurity profile is shown in Fig. 2.24. The use of the remote emitter contact with polysilicon connection does add some series emitter resistance, so for larger device geometries or cases in which emitter resistance is critical, a larger emitter is used and the contact is placed directly on top of the polysilicon emitter itself. Production IC processes[7,8] based on technologies similar to the one just described yield bipolar transistors having f_T values well in excess of 10 GHz, compared to a typical value of 500 MHz for deep-diffused, high-voltage processes.

2.5 Active Devices in Bipolar Analog Integrated Circuits

The high-voltage IC fabrication process described previously is an outgrowth of the one used to make *npn* double-diffused discrete bipolar transistors, and as a result the process inherently produces double-diffused *npn* transistors of relatively high performance. The advanced technology process improves further on all aspects of device performance except for breakdown voltage. In addition to *npn* transistors, *pnp* transistors are also required in many analog circuits, and an important development in the evolution of analog IC technologies was the invention of device structures that allowed the standard technology to produce *pnp* transistors as well. In this section, we will explore the structure and properties of *npn*, lateral *pnp*, and substrate *pnp* transistors. We will draw examples primarily from the high-voltage technology. The available structures in the more advanced technology are similar, except that their frequency response is correspondingly higher. We will include representative device parameters from these newer technologies as well.

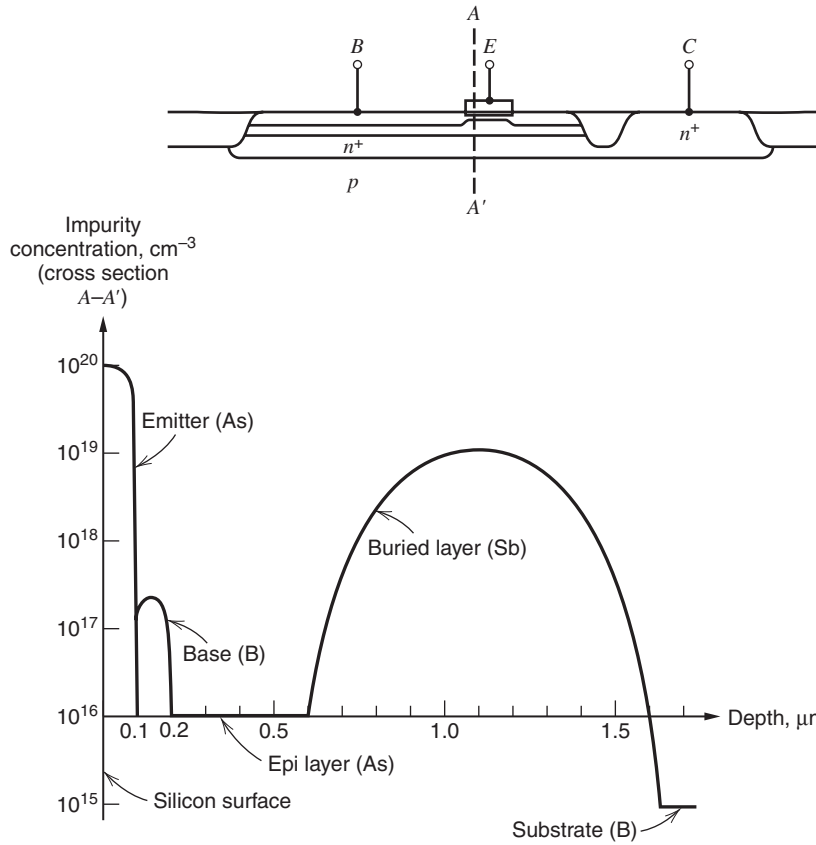


Figure 2.24 Typical impurity profile in a shallow oxide-isolated bipolar transistor.

2.5.1 Integrated-Circuit *npn* Transistors

The structure of a high-voltage, integrated-circuit *npn* transistor was described in the last section and is shown in plan view and cross section in Fig. 2.25. In the forward-active region of operation, the only electrically active portion of the structure that provides current gain is that portion of the base immediately under the emitter diffusion. The rest of the structure provides a top contact to the three transistor terminals and electrical isolation of the device from the rest of the devices on the same die. From an electrical standpoint, the principal effect of these regions is to contribute parasitic resistances and capacitances that must be included in the small-signal model for the complete device to provide an accurate representation of high-frequency behavior.

An important distinction between integrated-circuit design and discrete-component circuit design is that the IC designer has the capability to utilize a device geometry that is specifically optimized for the particular set of conditions found in the circuit. Thus the circuit-design problem involves a certain amount of device design as well. For example, the need often exists for a transistor with a high current-carrying capability to be used in the output stage of an amplifier. Such a device can be made by using a larger device geometry than the standard one, and the transistor then effectively consists of many standard devices connected in parallel. The larger geometry, however, will display larger base-emitter, collector-base, and collector-substrate capacitance than the standard device, and this must be taken into

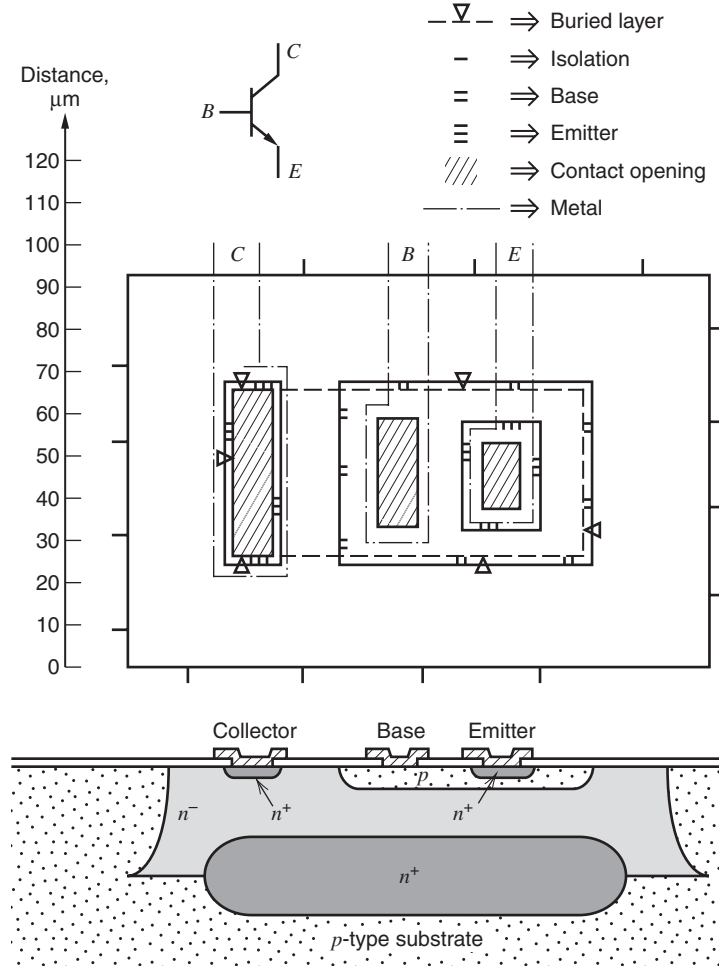


Figure 2.25 Integrated-circuit *npn* transistor. The mask layers are coded as shown.

account in analyzing the frequency response of the circuit. The circuit designer then must be able to determine the effect of changes in device geometry on device characteristics and to estimate the important device parameters when the device structure and doping levels are known. To illustrate this procedure, we will calculate the model parameters of the *npn* device shown in Fig. 2.25. This structure is typical of the devices used in circuits with a 5-Ω-cm, 17-μm epitaxial layer. The emitter diffusion is 20 μm × 25 μm, the base diffusion is 45 μm × 60 μm, and the base-isolation spacing is 25 μm. The overall device dimensions are 140 μm × 95 μm. Device geometries intended for lower epi resistivity and thickness can be much smaller; the base-isolation spacing is dictated by the side diffusion of the isolation region plus the depletion layers associated with the base-collector and collector-isolation junctions.

Saturation Current I_S. In Chapter 1, the saturation current of a graded-base transistor was shown to be

$$I_S = \frac{qA\,\overline{D}_n n_i^2}{Q_B} \tag{2.16}$$

where A is the emitter-base junction area, Q_B is the total number of impurity atoms per unit area in the base, n_i is the intrinsic carrier concentration, and $\overline{D}_n$ is the effective diffusion constant for electrons in the base region of the transistor. From Fig. 2.17, the quantity Q_B can be identified as the area under the concentration curve in the base region. This could be determined graphically but is most easily determined experimentally from measurements of the base-emitter voltage at a constant collector current. Substitution of (2.16) in (1.35) gives

$$\frac{Q_B}{\overline{D}_n} = A\frac{qn_i^2}{I_C}\exp\frac{V_{BE}}{V_T} \tag{2.17}$$

and Q_B can be determined from this equation.

■ **EXAMPLE**

A base-emitter voltage of 550 mV is measured at a collector current of 10 μA on a test transistor with a 100 μm × 100 μm emitter area. Estimate Q_B if $T = 300°$K. From Chapter 1, we have $n_i = 1.5 \times 10^{10}\ \text{cm}^{-3}$. Substitution in (2.17) gives

$$\frac{Q_B}{\overline{D}_n} = (100 \times 10^{-4})^2\ \frac{1.6 \times 10^{-19} \times 2.25 \times 10^{20}}{10^{-5}}\ \exp(550/26)$$
$$= 5.54 \times 10^{11}\text{cm}^{-4}\text{s}$$

At the doping levels encountered in the base, an approximate value of $\overline{D}_n$, the electron diffusivity, is

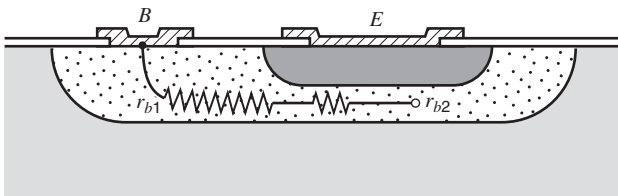
$$\overline{D}_n = 13\,\text{cm}^2\text{s}^{-1}$$

Thus for this example,

$$Q_B = 5.54 \times 10^{11} \times 13\ \text{cm}^{-2} = 7.2 \times 10^{12}\ \text{atoms/cm}^2$$

Note that Q_B depends on the diffusion profiles and will be different for different types of processes. Generally speaking, fabrication processes intended for lower voltage operation use thinner base regions and display lower values of Q_B. Within one nominally fixed process, Q_B can vary by a factor of two or three to one because of diffusion process variations. The principal significance of the numerical value for Q_B is that it allows the calculation of the saturation current I_S for any device structure once the emitter-base junction area is known. ■

Series Base Resistance r_b. Because the base contact is physically removed from the active base region, a significant series ohmic resistance is observed between the contact and the active base. This resistance can have a significant effect on the high-frequency gain and on the noise performance of the device. As illustrated in Fig. 2.26*a*, this resistance consists of two parts. The first is the resistance r_{b1} of the path between the base contact and the edge of the emitter



(*a*)

Figure 2.26 (*a*) Base resistance components for the *npn* transistor.

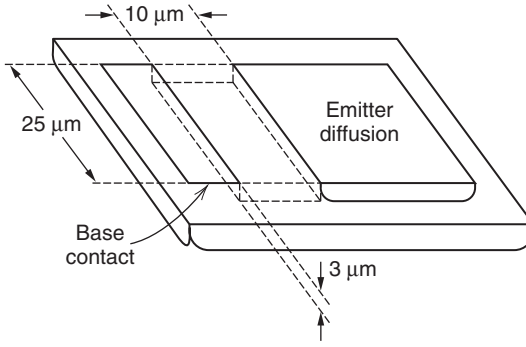


Figure 2.26 (*b*) Calculation of r_{b1}. The r_{b1} component of base resistance can be estimated by calculating the resistance of the rectangular block above.

diffusion. The second part r_{b2} is that resistance between the edge of the emitter and the site within the base region at which the current is actually flowing. The former component can be estimated by neglecting fringing and by assuming that this component of the resistance is that of a rectangle of material as shown in Fig. 2.26*b*. For a base sheet resistance of 100 Ω/□ and typical dimensions as shown in Fig. 2.26*b*, this would give a resistance of

$$r_{b1} = \frac{10\ \mu\text{m}}{25\ \mu\text{m}} 100\ \Omega = 40\ \Omega$$

The calculation of r_{b2} is complicated by several factors. First, the current flow in this region is not well modeled by a single resistor because the base resistance is distributed throughout the base region and two-dimensional effects are important. Second, at even moderate current levels, the effect of current crowding[9] in the base causes most of the carrier injection from the emitter into the base to occur near the periphery of the emitter diffusion. At higher current levels, essentially all of the injection takes place at the periphery and the effective value of r_b approaches r_{b1}. In this situation, the portion of the base directly beneath the emitter is not involved in transistor action. A typically observed variation of r_b with collector current for the *npn* geometry of Fig. 2.25 is shown in Fig. 2.27. In transistors designed for low-noise and/or high-frequency applications where low r_b is important, an effort is often made to maximize the periphery of the emitter that is adjacent to the base contact. At the same time, the emitter-base

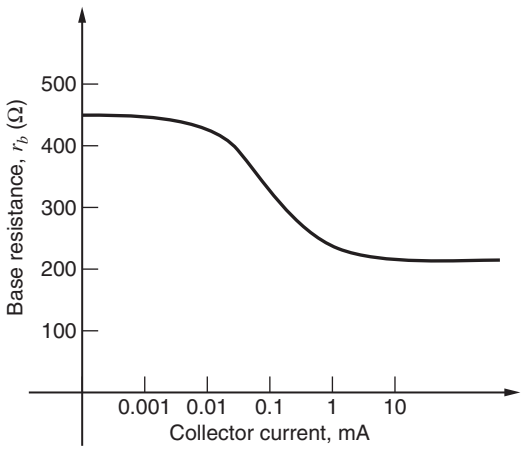
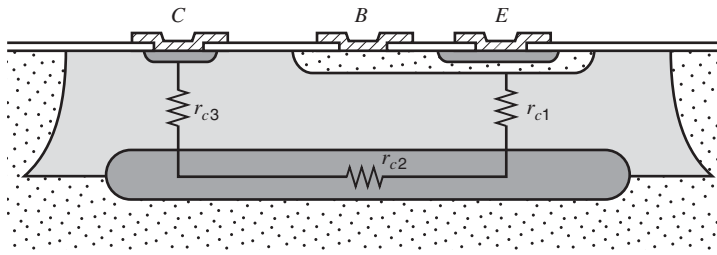


Figure 2.27 Typical variation of effective small-signal base resistance with collector current for integrated-circuit *npn* transistor.



(*a*)

Figure 2.28 (*a*) Components of collector resistance r_c.

junction and collector-base junction areas must be kept small to minimize capacitance. In the case of high-frequency transistors, this usually dictates the use of an emitter geometry that consists of many narrow stripes with base contacts between them. The ease with which the designer can use such device geometries is an example of the flexibility allowed by monolithic IC construction.

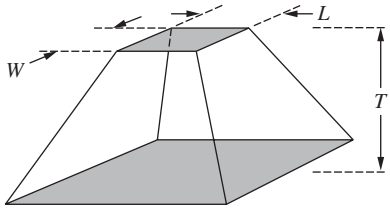
Series Collector Resistance r_c. The series collector resistance is important both in high-frequency circuits and in low-frequency applications where low collector-emitter saturation voltage is required. Because of the complex three-dimensional shape of the collector region itself, only an approximate value for the collector resistance can be obtained by hand analysis. From Fig. 2.28, we see that the resistance consists of three parts: that from the collector-base junction under the emitter down to the buried layer, r_{c1}; that of the buried layer from the region under the emitter over to the region under the collector contact, r_{c2}; and finally, that portion from the buried layer up to the collector contact, r_{c3}. The small-signal series collector resistance in the forward-active region can be estimated by adding the resistance of these three paths.

■ **EXAMPLE**

Estimate the collector resistance of the transistor of Fig. 2.25, assuming the doping profile of Fig. 2.17. We first calculate the r_{c1} component. The thickness of the lightly doped epi layer between the collector-base junction and the buried layer is 6 μm. Assuming that the collector-base junction is at zero bias, the results of Chapter 1 can be used to show that the depletion layer is about 1 μm thick. Thus the undepleted epi material under the base is 5 μm thick.

The effective cross-sectional area of the resistance r_{c1} is larger at the buried layer than at the collector-base junction. The emitter dimensions are 20 μm × 25 μm, while the buried layer dimensions are 41 μm × 85 μm on the mask. Since the buried layer side-diffuses a distance roughly equal to the distance that it out-diffuses, about 8 μm must be added on each edge, giving an effective size of 57 μm × 101 μm. An exact calculation of the ohmic resistance of this three-dimensional region would require a solution of Laplace's equation in the region, with a rather complex set of boundary conditions. Consequently, we will carry out an approximate analysis by modeling the region as a rectangular parallelepiped, as shown in Fig. 2.28*b*. Under the assumptions that the top and bottom surfaces of the region are equipotential surfaces, and that the current flow in the region takes place only in the vertical direction, the resistance of the structure can be shown to be

$$R = \frac{\rho T}{WL} \frac{\ln\left(\frac{a}{b}\right)}{(a-b)} \tag{2.18}$$



(*b*)

Figure 2.28 (*b*) Model for calculation of collector resistance.

where

T = thickness of the region
ρ = resistivity of the material
W, L = width, length of the top rectangle
a = ratio of the width of the bottom rectangle to the width of the top rectangle
b = ratio of the length of the bottom rectangle to the length of the top rectangle

Direct application of this expression to the case at hand would give an unrealistically low value of resistance, because the assumption of one-dimensional flow is seriously violated when the dimensions of the lower rectangle are much larger than those of the top rectangle. Equation 2.18 gives realistic results when the sides of the region form an angle of about 60° or less with the vertical. When the angle of the sides is increased beyond this point, the resistance does not decrease very much because of the long path for current flow between the top electrode and the remote regions of the bottom electrode. Thus the limits of the bottom electrode should be determined either by the edges of the buried layer or by the edges of the emitter plus a distance equal to about twice the vertical thickness T of the region, whichever is smaller. For the case of r_{c1},

$$T = 5\ \mu\text{m} = 5 \times 10^{-4}\ \text{cm}$$
$$\rho = 5\ \Omega\text{-cm}$$

We assume that the effective emitter dimensions are those defined by the mask plus approximately 2 μm of side diffusion on each edge. Thus

$$W = 20\ \mu\text{m} + 4\ \mu\text{m} = 24 \times 10^{-4}\ \text{cm}$$
$$L = 25\ \mu\text{m} + 4\ \mu\text{m} = 29 \times 10^{-4}\ \text{cm}$$

For this case, the buried-layer edges are further away from the emitter edge than twice the thickness T on all four sides when side diffusion is taken into account. Thus the *effective* buried-layer dimensions that we use in (2.18) are

$$W_{BL} = W + 4T = 24\ \mu\text{m} + 20\ \mu\text{m} = 44\ \mu\text{m}$$
$$L_{BL} = L + 4T = 29\ \mu\text{m} + 20\ \mu\text{m} = 49\ \mu\text{m}$$

and

$$a = \frac{44\ \mu\text{m}}{24\ \mu\text{m}} = 1.83$$

$$b = \frac{49\ \mu\text{m}}{29\ \mu\text{m}} = 1.69$$

Thus from (2.18),

$$r_{c1} = \frac{(5)(5 \times 10^{-4})}{(24 \times 10^{-4})(29 \times 10^{-4})}(0.57)\ \Omega = 204\ \Omega$$

We will now calculate r_{c2}, assuming a buried-layer sheet resistance of 20 Ω/□. The distance from the center of the emitter to the center of the collector-contact diffusion is 62 μm, and the width of the buried layer is 41 μm. The r_{c2} component is thus, approximately,

$$r_{c2} = (20\ \Omega/\square)\left(\frac{L}{W}\right) = 20\ \Omega/\square\left(\frac{62\ \mu\text{m}}{41\ \mu\text{m}}\right) = 30\ \Omega$$

Here the buried-layer side diffusion was not taken into account because the ohmic resistance of the buried layer is determined entirely by the number of impurity atoms actually diffused [see (2.15)] into the silicon, which is determined by the mask dimensions and the sheet resistance of the buried layer.

For the calculation of r_{c3}, the dimensions of the collector-contact n^+ diffusion are 18 μm × 49 μm, including side diffusion. The distance from the buried layer to the bottom of the n^+ diffusion is seen in Fig. 2.17 to be 6.5 μm, and thus $T = 6.5$ μm in this case. On the three sides of the collector n^+ diffusion that do not face the base region, the out-diffused buried layer extends only 4 μm outside the n^+ diffusion, and thus the effective dimension of the buried layer is determined by the actual buried-layer edge on these sides. On the side facing the base region, the effective edge of the buried layer is a distance $2T$, or 13 μm, away from the edge of the n^+ diffusion. The effective buried-layer dimensions for the calculation of r_{c3} are thus 35 μm × 57 μm. Using (2.18),

$$r_{c3} = \frac{(5)(6.5 \times 10^{-4})}{(18 \times 10^{-4})(49 \times 10^{-4})}0.66 = 243\ \Omega$$

The total collector resistance is thus

$$r_c = r_{c1} + r_{c2} + r_{c3} = 531\ \Omega$$

The value actually observed in such devices is somewhat lower than this for three reasons. First, we have approximated the flow as one-dimensional, and it is actually three-dimensional. Second, for larger collector-base voltages, the collector-base depletion layer extends further into the epi, decreasing r_{c1}. Third, the value of r_c that is important is often that for a saturated device. In saturation, holes are injected into the epi region under the emitter by the forward-biased, collector-base function, and they modulate the conductivity of the region even at moderate current levels.[10] Thus the collector resistance one measures when the device is in saturation is closer to $(r_{c2} + r_{c3})$, or about 250 to 300 Ω. Thus r_c is smaller in saturation than in the forward-active region. ■

Collector-Base Capacitance. The collector-base capacitance is simply the capacitance of the collector-base junction including both the flat bottom portion of the junction and the sidewalls. This junction is formed by the diffusion of boron into an n-type epitaxial material that we will assume has a resistivity of 5 Ω-cm, corresponding to an impurity concentration of 10^{15} atoms/cm^3. The uniformly doped epi layer is much more lightly doped than the ρ-diffused region, and as a result, this junction is well approximated by a step junction in which the depletion layer lies almost entirely in the epitaxial material. Under this assumption, the results of Chapter 1 regarding step junctions can be applied, and for convenience this relationship has been plotted in nomograph form in Fig. 2.29. This nomograph is a graphical representation of the relation

$$\frac{C_j}{A} = \sqrt{\frac{q\epsilon N_B}{2(\psi_0 + V_R)}} \tag{2.19}$$

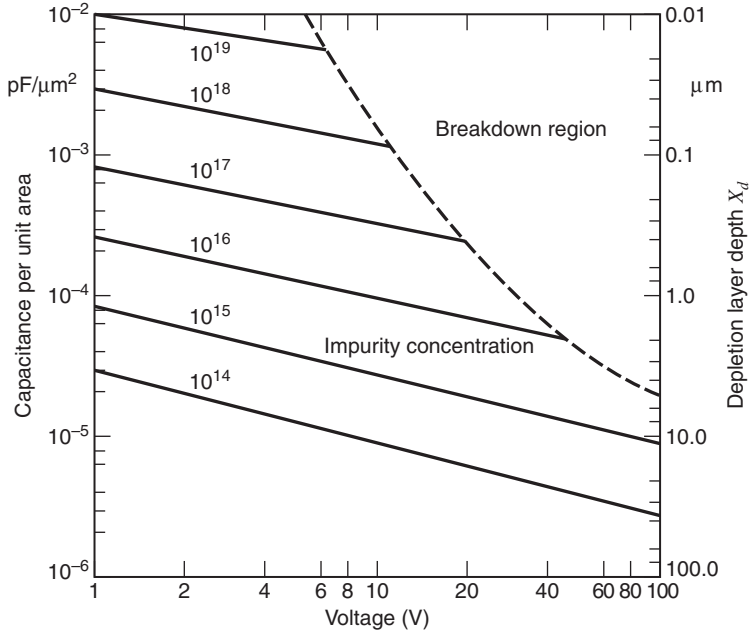


Figure 2.29 Capacitance and depletion-layer width of an abrupt *pn* junction as a function of applied voltage and doping concentration on the lightly doped side of the junction.[11]

where N_B is the doping density in the epi material and V_R is the reverse bias on the junction. The nomograph of Fig. 2.29 can also be used to determine the junction depletion-region width as a function of applied voltage, since this width is inversely proportional to the capacitance. The width in microns is given on the axis on the right side of the figure.

Note that the horizontal axis in Fig. 2.29 is the *total* junction potential, which is the applied potential plus the built-in voltage ψ_0. In order to use the curve, then, the built-in potential must be calculated. While this would be an involved calculation for a diffused junction, the built-in potential is actually only weakly dependent on the details of the diffusion profile and can be assumed to be about 0.55 V for the collector-base junction, 0.52 V for the collector-substrate junction, and about 0.7 V for the emitter-base junction.

■ **EXAMPLE**

Calculate the collector-base capacitance of the device of Fig. 2.25. The zero-bias capacitance per unit area of the collector-base junction can be found from Fig. 2.29 to be approximately 10^{-4} pF/μm^2. The total area of the collector-base junction is the sum of the area of the bottom of the base diffusion plus the base sidewall area. From Fig. 2.25, the bottom area is

$$A_{\text{bottom}} = 60\ \mu\text{m} \times 45\ \mu\text{m} = 2700\ \mu\text{m}^2$$

The edges of the base region can be seen from Fig. 2.17 to have the shape similar to one-quarter of a cylinder. We will assume that the region is cylindrical in shape, which yields a sidewall area of

$$A_{\text{sidewall}} = P \times d \times \frac{\pi}{2}$$

where

P = base region periphery
d = base diffusion depth

Thus we have

$$A_{\text{sidewall}} = 3\ \mu\text{m} \times (60\ \mu\text{m} + 60\ \mu\text{m} + 45\ \mu\text{m} + 45\ \mu\text{m}) \times \frac{\pi}{2} = 989\ \mu\text{m}^2$$

and the total capacitance is

$$C_{\mu 0} = (A_{\text{bottom}} + A_{\text{sidewall}})(10^{-4}\ \text{pF}/\mu\text{m}^2) = 0.36\ \text{pF}$$

■

Collector-Substrate Capacitance. The collector-substrate capacitance consists of three portions: that of the junction between the buried layer and the substrate, that of the sidewall of the isolation diffusion, and that between the epitaxial material and the substrate. Since the substrate has an impurity concentration of about 10^{16} cm^{-3}, it is more heavily doped than the epi material, and we can analyze both the sidewall and epi-substrate capacitance under the assumption that the junction is a one-sided step junction with the epi material as the lightly doped side. Under this assumption, the capacitance per unit area in these regions is the same as in the collector-base junction.

■ **EXAMPLE**

Calculate the collector-substrate capacitance of the standard device of Fig. 2.25. The area of the collector-substrate sidewall is

$$A_{\text{sidewall}} = (17\ \mu\text{m})(140\ \mu\text{m} + 140\ \mu\text{m} + 95\ \mu\text{m} + 95\ \mu\text{m})\left(\frac{\pi}{2}\right) = 12,550\ \mu\text{m}^2$$

We will assume that the actual buried layer covers the area defined by the mask, indicated on Fig. 2.25 as an area of 41 μm × 85 μm, plus 8 μm of side-diffusion on each edge. This gives a total area of 57 μm × 101 μm. The area of the junction between the epi material and the substrate is the total area of the isolated region, minus that of the buried layer.

$$\begin{aligned} A_{\text{epi-substrate}} &= (140\ \mu\text{m} \times 95\ \mu\text{m}) - (57\ \mu\text{m} \times 101\ \mu\text{m}) \\ &= 7543\ \mu\text{m}^2 \end{aligned}$$

The capacitances of the sidewall and epi-substrate junctions are, using a capacitance per unit area of 10^{-4}pF/μm^2

$$\begin{aligned} C_{cs0}\ (\text{sidewall}) &= (12,550\ \mu\text{m}^2)(10^{-4}\text{pF}/\mu\text{m}^2) = 1.26\ \text{pF} \\ C_{cs0}(\text{epi-substrate}) &= (7543\ \mu\text{m}^2)(10^{-4}\ \text{pF}/\mu\text{m}^2) = 0.754\ \text{pF} \end{aligned}$$

For the junction between the buried layer and the substrate, the lightly doped side of the junction is the substrate. Assuming a substrate doping level of 10^{16} atoms/cm^3, and a built-in voltage of 0.52 V, we can calculate the zero-bias capacitance per unit area as 3.3×10^{-4} pF/μm^2. The area of the buried layer is

$$A_{BL} = 57\ \mu\text{m} \times 101\ \mu\text{m} = 5757\ \mu\text{m}^2$$

and the zero-bias capacitance from the buried layer to the substrate is thus

$$C_{cs0}(BL) = (5757\ \mu\text{m}^2)(3.3 \times 10^{-4}\ \text{pF}/\mu\text{m}^2) = 1.89\ \text{pF}$$

The total zero-bias, collector-substrate capacitance is thus

$$C_{cs0} = 1.26\ \text{pF} + 0.754\ \text{pF} + 1.89\ \text{pF} = 3.90\ \text{pF}$$

■

Emitter-Base Capacitance. The emitter-base junction of the transistor has a doping profile that is not well approximated by a step junction because the impurity concentration on both sides of the junction varies with distance in a rather complicated way. Furthermore, the sidewall capacitance per unit area is not constant but varies with distance from the surface because the base impurity concentration varies with distance. A precise evaluation of this capacitance can be carried out numerically, but a first-order estimate of the capacitance can be obtained by calculating the capacitance of an abrupt junction with an impurity concentration on the lightly doped side that is equal to the concentration in the base at the edge of the junction. The sidewall contribution is neglected.

■ **EXAMPLE**

Calculate the zero-bias, emitter-base junction capacitance of the standard device of Fig. 2.25.

We first estimate the impurity concentration at the emitter edge of the base region. From Fig. 2.17, it can be seen that this concentration is approximately 10^{17} atoms/cm^3. From the nomograph of Fig. 2.29, this abrupt junction would have a zero-bias capacitance per unit area of 10^{-3} pF/μm^2. Since the area of the bottom portion of the emitter-base junction is 25 μm $\times$ 20 μm, the capacitance of the bottom portion is

$$C_{\text{bottom}} = (500\ \mu\text{m}^2)(10^{-3}\ \text{pF}/\mu\text{m}^2) = 0.5\ \text{pF}$$

Again assuming a cylindrical cross section, the sidewall area is given by

$$A_{\text{sidewall}} = 2\,(25\ \mu\text{m} + 20\ \mu\text{m})\left(\frac{\pi}{2}\right)(2.5\ \mu\text{m}) = 353\ \mu\text{m}^2$$

Assuming that the capacitance per unit area of the sidewall is approximately the same as the bottom,

$$C_{\text{sidewall}} = (353\ \mu\text{m}^2)(10^{-3}\ \text{pF}/\mu\text{m}^2) = 0.35\ \text{pF}$$

The total emitter-base capacitance is

$$C_{je0} = 0.85\ \text{pF}$$

■

Current Gain. As described in Chapter 1, the current gain of the transistor depends on minority-carrier lifetime in the base, which affects the base transport factor, and on the diffusion length in the emitter, which affects the emitter efficiency. In analog IC processing, the base minority-carrier lifetime is sufficiently long that the base transport factor is not a limiting factor in the forward current gain in *npn* transistors. Because the emitter region is heavily doped with phosphorus, the minority-carrier lifetime is degraded in this region, and current gain is limited primarily by emitter efficiency.[12] Because the doping level, and hence lifetime, vary with distance in the emitter, the calculation of emitter efficiency for the *npn* transistor is difficult, and measured parameters must be used. The room-temperature current gain typically lies between 200 and 1000 for these devices. The current gain falls with decreasing temperature, usually to a value of from 0.5 to 0.75 times the room temperature value at -55°C.

Summary of High-Voltage *npn* Device Parameters. A typical set of device parameters for the device of Fig. 2.25 is shown in Fig. 2.30. This transistor geometry is typical of that used for circuits that must operate at power supply voltages up to 40 V. For lower operating voltages,

	Parameter	Typical Value, 5-Ω-cm,17-μm epi 44-V Device	Typical Value, 1-Ω-cm,10-μm epi 20-V Device
	β_F	200	200
	B_R	2	2
	V_A	130 V	90 V
	η	2×10^{-4}	2.8×10^{-4}
	I_S	5×10^{-15}A	1.5×10^{-15} A
	I_{CO}	10^{-10}A	10^{-10} A
	BV_{CEO}	50 V	25 V
	BV_{CBO}	90 V	50 V
	BV_{EBO}	7 V	7 V
	τ_F	0.35 ns	0.25 ns
	τ_R	400 ns	200 ns
	β_0	200	150
	r_b	200 Ω	200 Ω
	r_c (saturation)	200 Ω	75 Ω
	r_{ex}	2 Ω	2 Ω
Base-emitter junction	C_{je0}	1 pF	1.3 pF
	ψ_{0e}	0.7 V	0.7 V
	n_e	0.33	0.33
Base-collector junction	$C_{\mu 0}$	0.3 pF	0.6 pF
	ψ_{0c}	0.55 V	0.6 V
	n_c	0.5	0.5
Collector-substrate junction	C_{cs0}	3 pF	3 pF
	ψ_{0s}	0.52 V	0.58 V
	n_s	0.5	0.5

Figure 2.30 Typical parameters for high-voltage integrated *npn* transistors with 500 μm^2 emitter area. The thick epi device is typical of those used in circuits operating at up to 44 V power-supply voltage, while the thinner device can operate up to about 20 V. While the geometry of the thin epi device is smaller, the collector-base capacitance is larger because of the heavier epi doping. The emitter-base capacitance is higher because the base is shallower, and the doping level in the base at the emitter-base junction is higher.

thinner epitaxial layers can be used, and smaller device geometries can be used as a result. Also shown in Fig. 2.30 are typical parameters for a device made with 1-Ω-cm epi material, which is 10 μm thick. Such a device is physically smaller and has a collector-emitter breakdown voltage of about 25 V.

Advanced-Technology Oxide-Isolated *npn* Bipolar Transistors. The structure of an advanced oxide-isolated, poly-emitter *npn* bipolar transistor is shown in plan view and cross section in Fig. 2.31. Typical parameters for such a device are listed in Fig. 2.32. Note the enormous reduction in device size, transit time, and parasitic capacitance compared to the high-voltage, deep-diffused process. These very small devices achieve optimum performance characteristics at relatively low bias currents. The value of β for such a device typically peaks at a collector current of about 50 μA. For these advanced-technology transistors, the use of ion implantation allows precise control of very shallow emitter (0.1 μm) and base (0.2 μm)

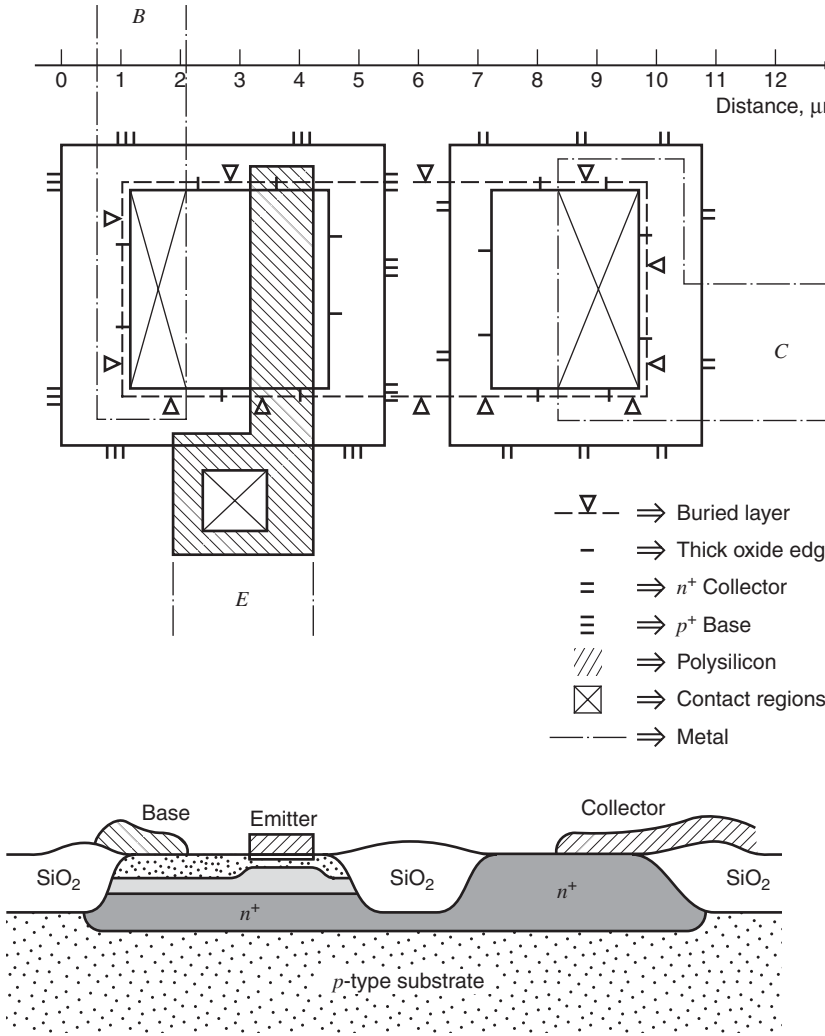


Figure 2.31 Plan view and cross section of a typical advanced-technology bipolar transistor. Note the much smaller dimensions compared with the high-voltage device.

regions. The resulting base width is of the order of 0.1 μm, and (1.99) predicts a base transit time about 25 times smaller than the deep-diffused device of Fig. 2.17. This is observed in practice, and the ion-implanted transistor has a peak f_T of about 13 GHz.

2.5.2 Integrated-Circuit *pnp* Transistors

As mentioned previously, the integrated-circuit bipolar fabrication process is an outgrowth of that used to build double-diffused epitaxial *npn* transistors, and the technology inherently produces *npn* transistors of high performance. However, *pnp* transistors of comparable performance are not easily produced in the same process, and the earliest analog integrated circuits used no *pnp* transistors. The lack of a complementary device for use in biasing, level shifting, and as load devices in amplifier stages proved to be a severe limitation on the performance attainable in analog circuits, leading to the development of several *pnp* transistor structures

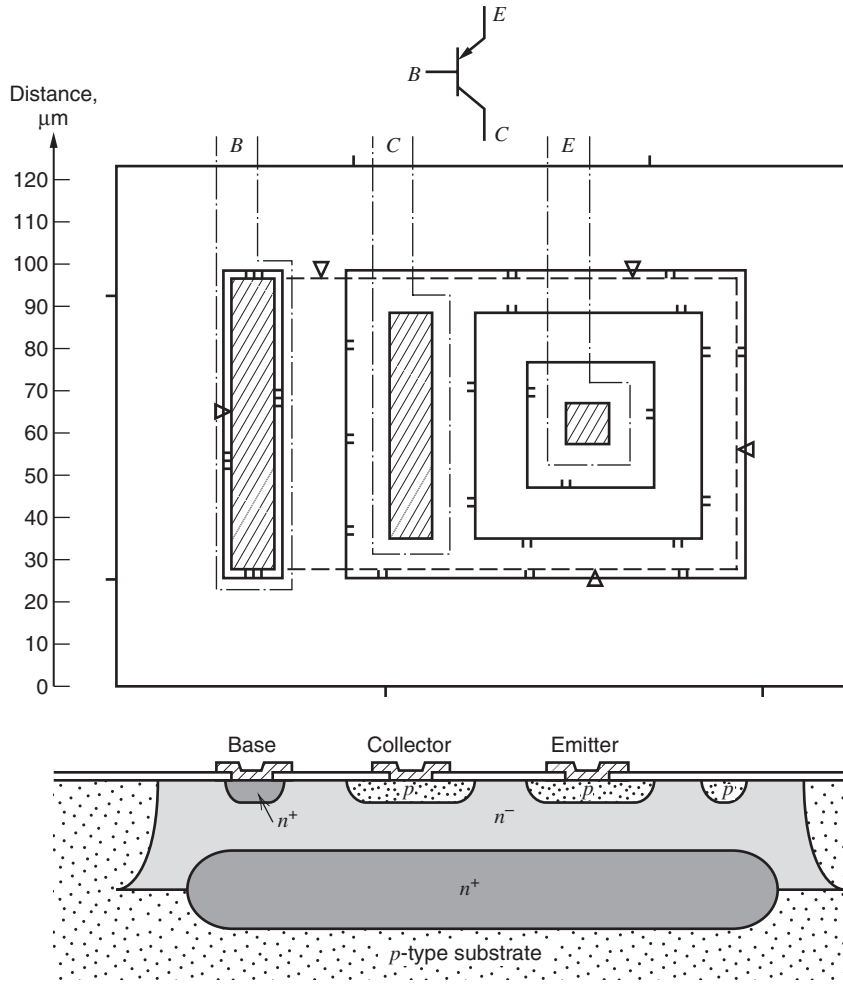
Parameter	Vertical *npn* Transistor with 2 μm^2 Emitter Area	Lateral *pnp* Transistor with 2 μm^2 Emitter Area
β_F	120	50
β_R	2	3
V_A	35 V	30 V
I_S	6×10^{-18}A	6×10^{-18}A
I_{CO}	1 pA	1 pA
BV_{CEO}	8 V	14 V
BV_{CBO}	18 V	18 V
BV_{EBO}	6 V	18 V
τ_F	10 ps	650 ps
τ_R	5 ns	5 ns
r_b	400 Ω	200 Ω
r_c	100 Ω	20 Ω
r_{ex}	40 Ω	10 Ω
C_{je0}	5 fF	14 fF
ψ_{0e}	0.8 V	0.7 V
n_e	0.4	0.5
$C_{\mu 0}$	5 fF	15 fF
ψ_{0c}	0.6 V	0.6 V
n_c	0.33	0.33
C_{cs0} (C_{bs0})	20 fF	40 fF
ψ_{0s}	0.6 V	0.6 V
n_s	0.33	0.4

Figure 2.32 Typical device parameters for bipolar transistors in a low-voltage, oxide-isolated, ion-implanted process.

that are compatible with the standard IC fabrication process. Because these devices utilize the lightly doped *n*-type epitaxial material as the base of the transistor, they are generally inferior to the *npn* devices in frequency response and high-current behavior, but are useful nonetheless. In this section, we will describe the lateral *pnp* and substrate *pnp* structures.

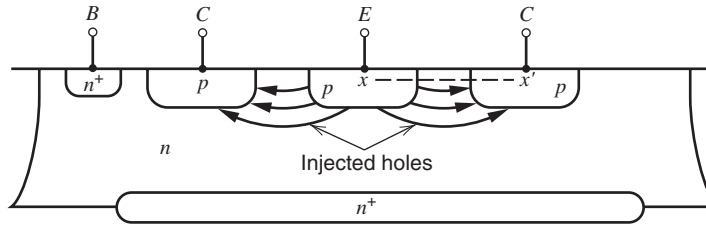
Lateral *pnp* Transistors. A typical lateral *pnp* transistor structure fabricated in a high-voltage process is illustrated in Fig. 2.33*a*.[13] The emitter and collector are formed with the same diffusion that forms the base of the *npn* transistors. The collector is a *p*-type ring around the emitter, and the base contact is made in the *n*-type epi material *outside* the collector ring. The flow of minority carriers across the base is illustrated in Fig. 2.33*b*. Holes are injected from the emitter, flow parallel to the surface across the *n*-type base region, and ideally are collected by the *p*-type collector before reaching the base contact. Thus the transistor action is *lateral* rather than *vertical* as in the case for *npn* transistors. The principal drawback of the structure is the fact that the base region is more lightly doped than the collector. As a result, the collector-base depletion layer extends almost entirely into the base. The base region must then be made wide enough so that the depletion layer does not reach the emitter when the maximum collector-emitter voltage is applied. In a typical analog IC process, the width of this depletion layer is 6 μm to 8 μm when the collector-emitter voltage is in the 40-V range. Thus the minimum base width for such a device is about 8 μm, and the minimum base transit time can be estimated from (1.99) as

$$\tau_F = \frac{W_B^2}{2D_p} \tag{2.20}$$



(*a*)

Figure 2.33 (*a*) Lateral *pnp* structure fabricated in a high-voltage process.



(*b*)

Figure 2.33 (*b*) Minority-carrier flow in the lateral *pnp* transistor.

Use of $W_B = 8\ \mu\text{m}$ and $D_p = 10\ \text{cm}^2/\text{s}$ (for holes) in (2.20) gives

$$\tau_F = 32\ \text{ns}$$

This corresponds to a peak f_T of 5 MHz, which is a factor of 100 lower than a typical *npn* transistor in the same process.

The current gain of lateral *pnp* transistors tends to be low for several reasons. First, minority carriers (holes) in the base are injected downward from the emitter as well as laterally, and some of them are collected by the substrate, which acts as the collector of a parasitic vertical *pnp* transistor. The buried layer sets up a retarding field that tends to inhibit this process, but it still produces a measurable degradation of β_F. Second, the emitter of the *pnp* is not as heavily doped as is the case for the *npn* devices, and thus the emitter injection efficiency given by (1.51b) is not optimized for the *pnp* devices. Finally, the wide base of the lateral *pnp* results in both a low emitter injection efficiency and also a low base transport factor as given by (1.51a).

Another drawback resulting from the use of a lightly doped base region is that the current gain of the device falls very rapidly with increasing collector current due to high-level injection. The minority-carrier distribution in the base of a lateral *pnp* transistor in the forward-active region is shown in Fig. 2.34. The collector current per unit of cross-sectional area can be obtained from (1.32) as

$$J_p = qD_p \frac{p_n(0)}{W_B} \tag{2.21}$$

Inverting this relationship, we can calculate the minority-carrier density at the emitter edge of the base as

$$p_n(0) = \frac{J_p W_B}{qD_p} \tag{2.22}$$

As long as this concentration is much less than the majority-carrier density in the base, low-level injection conditions exist and the base minority-carrier lifetime remains constant. However, when the minority-carrier density becomes comparable with the majority-carrier density, the majority-carrier density must increase to maintain charge neutrality in the base. This causes a decrease in β_F for two reasons. First, there is a decrease in the effective lifetime of minority carriers in the base since there is an increased number of majority carriers with which recombination can occur. Thus the base transport factor given by (1.51*a*) decreases. Second, the increase in the majority-carrier density represents an effective increase in base doping density. This causes a decrease in emitter injection efficiency given by (1.51*b*). Both these mechanisms are also present in *npn* transistors, but occur at much higher current levels due to the higher doping density in the base of the *npn* transistor.

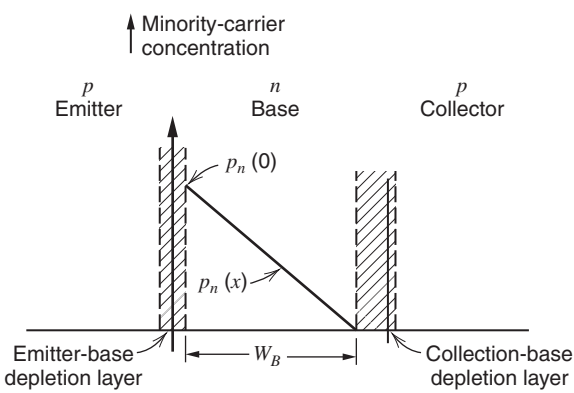


Figure 2.34 Minority-carrier distribution in the base of a lateral *pnp* transistor in the forward-active region. This distribution is that observed through section *x*-*x*′ in Fig. 2.33*b*.

The collector current at which these effects become significant can be calculated for a lateral *pnp* transistor by equating the minority-carrier concentration given by (2.22) to the equilibrium majority-carrier concentration. Thus

$$\frac{J_p W_B}{qD_p} = n_n \simeq N_D \tag{2.23}$$

where (2.1) has been substituted for n_n, and N_D is the donor density in the *pnp* base (*npn* collector). From (2.23), we can calculate the collector current for the onset of high-level injection in a *pnp* transistor as

$$I_C = \frac{qAN_D D_p}{W_B} \tag{2.24}$$

where A is the effective area of the emitter-base junction. Note that this current depends directly on the base doping density in the transistor, and since this is quite low in a lateral *pnp* transistor, the current density at which this fall-off begins is quite low.

Lateral *pnp* transistors are also widely used in shallow oxide-isolated bipolar IC technologies. The device structure used is essentially identical to that of Fig. 2.33, except that the device area is orders of magnitude smaller and the junction isolation is replaced by oxide isolation. Typical parameters for such a device are listed in Fig. 2.32. As in the case of *npn* transistors, we see dramatic reductions in device transit time and parasitic capacitance compared to the high-voltage, thick-epi process. The value of β for such a device typically peaks at a collector current of about 50 nA.

■ **EXAMPLE**

Calculate the collector current at which the current gain begins to fall for the *pnp* structure of Fig. 2.33*a*. The effective cross-sectional area A of the emitter is the sidewall area of the emitter, which is the p-type diffusion depth multiplied by the periphery of the emitter multiplied by $\pi/2$.

$$A = (3\ \mu\text{m})(30\ \mu\text{m} + 30\ \mu\text{m} + 30\ \mu\text{m} + 30\ \mu\text{m})\left(\frac{\pi}{2}\right) = 565\ \mu\text{m}^2 = 5.6 \times 10^{-6}\ \text{cm}^2$$

The majority-carrier density is 10^{15} atoms/cm^3 for an epi-layer resistivity of 5 Ω-cm. In addition, we can assume $W_B = 8\ \mu$m and $D_p = 10$ cm^2/s. Substitution of this data in (2.24) gives

$$I_C = 5.6 \times 10^{-6} \times 1.6 \times 10^{-19} \times 10^{15} \times 10 \frac{1}{8 \times 10^{-4}} \text{A} = 11.2\ \mu\text{A}$$

■

The typical lateral *pnp* structure of Fig. 2.33*a* shows a low-current beta of approximately 30 to 50, which begins to decrease at a collector current of a few tens of microamperes, and has fallen to less than 10 at a collector current of 1 mA. A typical set of parameters for a structure of this type is shown in Fig. 2.35. Note that in the lateral *pnp* transistor, the substrate junction capacitance appears between the *base* and the substrate.

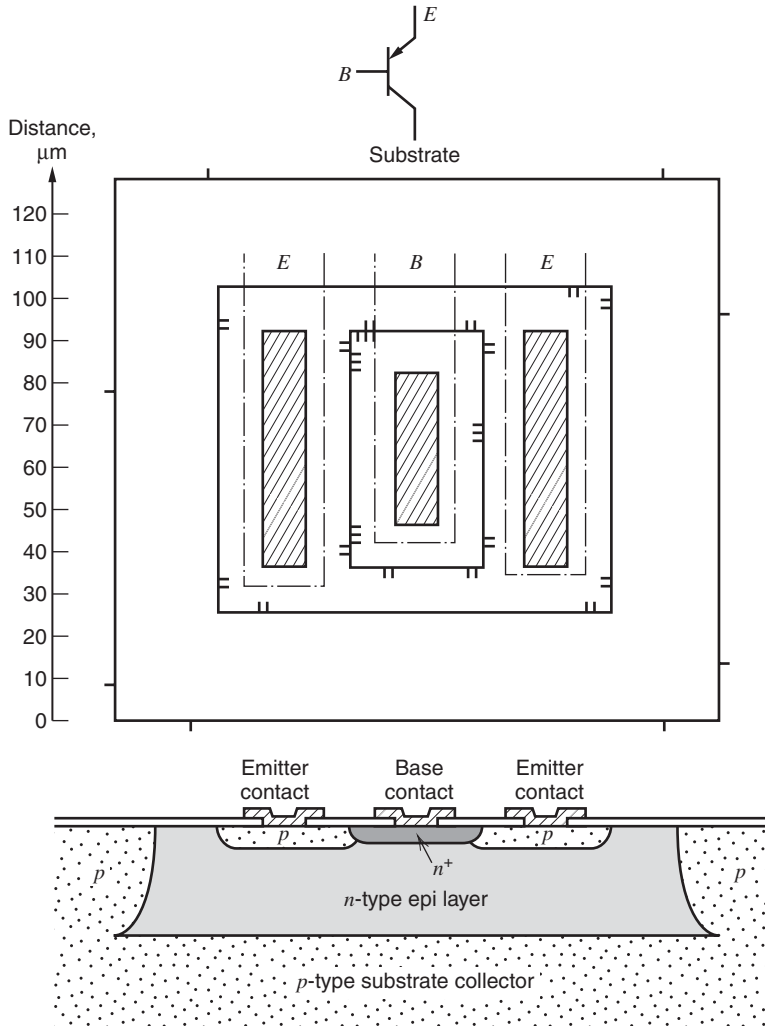
Substrate *pnp* Transistors. One reason for the poor high-current performance of the lateral *pnp* is the relatively small effective cross-sectional area of the emitter, which results from the lateral nature of the injection. A common application for a *pnp* transistor is in a Class-B output stage where the device is called on to operate at collector currents in the 10-mA range. A lateral *pnp* designed to do this would require a large amount of die area. In this application, a different structure is usually used in which the substrate itself is used as the collector instead of a diffused p-type region. Such a substrate *pnp* transistor in a high-voltage, thick-epi process is

	Parameter	Typical Value, 5-Ω-cm,17-μm epi 44-V Device	Typical Value, 1-Ω-cm, 10-μm epi 20-V Device
	β_F	50	20
	β_R	4	2
	V_A	50 V	50 V
	η	5×10^{-4}	5×10^{-4}
	I_S	2×10^{-15}A	2×10^{-15}A
	I_{CO}	10^{-10}A	5×10^{-9}A
	BV_{CEO}	60 V	30 V
	BV_{CBO}	90 V	50 V
	BV_{EBO}	90 V	50 V
	τ_F	30 ns	20 ns
	τ_R	3000 ns	2000 ns
	β_0	50	20
	r_b	300 Ω	150 Ω
	r_c	100 Ω	75 Ω
	r_{ex}	10 Ω	10 Ω
Base-emitter junction	C_{je0}	0.3 pF	0.6 pF
	ψ_{0e}	0.55 V	0.6 V
	n_e	0.5	0.5
Base-collector junction	$C_{\mu 0}$	1 pF	2 pF
	ψ_{0c}	0.55 V	0.6 V
	n_c	0.5	0.5
Base-substrate junction	C_{bs0}	3 pF	3.5 pF
	ψ_{0s}	0.52 V	0.58 V
	n_s	0.5	0.5

Figure 2.35 Typical parameters for lateral *pnp* transistors with 900 μm^2 emitter area in a high-voltage, thick-epi process.

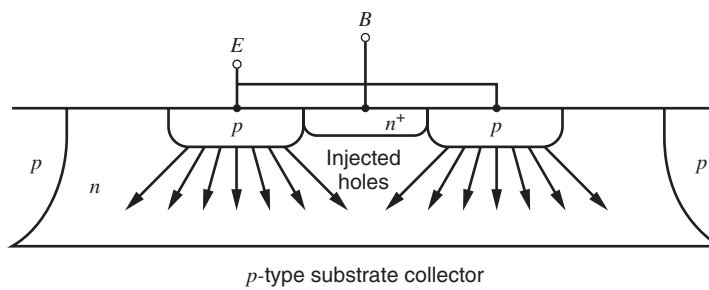
shown in Fig. 2.36*a*. The *p*-type emitter diffusion for this particular substrate *pnp* geometry is rectangular with a rectangular hole in the middle. In this hole an n^+ region is formed with the *npn* emitter diffusion to provide a contact for the *n*-type base. Because of the lightly doped base material, the series base resistance can become quite large if the base contact is far removed from the active base region. In this particular structure, the n^+ base contact diffusion is actually allowed to come in contact with the *p*-type emitter diffusion, in order to get the low-resistance base contact diffusion as close as possible to the active base. The only drawback of this, in a substrate *pnp* structure, is that the emitter-base breakdown voltage is reduced to approximately 7 V. If larger emitter-base breakdown is required, then the *p*-emitter diffusion must be separated from the n^+ base contact diffusion by a distance of about 10 μm to 15 μm. Many variations exist on the substrate *pnp* geometry shown in Fig. 2.36*a*. They can also be realized in thin-epi, oxide-isolated processes.

The minority-carrier flow in the forward-active region is illustrated in Fig. 2.36*b*. The principal advantage of this device is that the current flow is vertical and the effective cross-sectional area of the emitter is much larger than in the case of the lateral *pnp* for the same overall device size. The device is restricted to use in emitter-follower configurations, however, since the collector is electrically identical with the substrate that must be tied to the most negative circuit potential. Other than the better current-handling capability, the properties of substrate *pnp* transistors are similar to those for lateral *pnp* transistors since the base width is similar



(*a*)

Figure 2.36 (*a*) Substrate *pnp* structure in a high-voltage, thick-epi process.



(*b*)

Figure 2.36 (*b*) Minority-carrier flow in the substrate *pnp* transistor.

in both cases. An important consideration in the design of substrate *pnp* structures is that the collector current flows in the *p*-substrate region, which usually has relatively high resistivity. Thus, unless care is taken to provide an adequate low-resistance path for the collector current, a high series collector resistance can result. This resistance can degrade device performance in two ways. First, large collector currents in the *pnp* can cause enough voltage drop in the substrate region itself that other substrate-epitaxial layer junctions within the circuit can become forward biased. This usually has a catastrophic effect on circuit performance. Second, the effects of the collector-base junction capacitance on the *pnp* are multiplied by the Miller effect resulting from the large series collector resistance, as described further in Chapter 7. To minimize these effects, the collector contact is usually made by contacting the isolation diffusion immediately adjacent to the substrate *pnp* itself with metallization. For high-current devices, this isolation diffusion contact is made to surround the device to as great an extent as possible.

The properties of a typical substrate *pnp* transistor in a high-voltage, thick-epi process are summarized in Fig. 2.37. The dependence of current gain on collector current for a typical *npn*, lateral *pnp*, and substrate *pnp* transistor in a high-voltage, thick-epi process are shown in Fig. 2.38. The low-current reduction in β, which is apparent for all three devices, is due to recombination in the base-emitter depletion region, described in Section 1.3.5.

	Parameter	Typical Value, 5-Ω-cm, 17-μm epi 44-V Device 5100 μm² Emitter Area	Typical Value, 1-Ω-cm, 10-μm epi 20-V Device 5100 μm² Emitter Area
	β_F	50	30
	β_R	4	2
	V_A	50 V	30 V
	η	5×10^{-4}	9×10^{-4}
	I_S	10^{-14} A	10^{-14} A
	I_{CO}	2×10^{-10}A	2×10^{-10} A
	BV_{CEO}	60 V	30 V
	BV_{CBO}	90 V	50 V
	BV_{EBO}	7 V or 90 V	7 V or 50 V
	τ_F	20 ns	14 ns
	τ_R	2000 ns	1000 ns
	β_0	50	30
	r_b	150 Ω	50 Ω
	r_c	50 Ω	50 Ω
	r_{ex}	2 Ω	2 Ω
Base-emitter junction	C_{je0}	0.5 pF	1 pF
	ψ_{0e}	0.55 V	0.58 V
	n_e	0.5	0.5
Base-collector junction	$C_{\mu 0}$	2 pF	3 pF
	ψ_{0c}	0.52 V	0.58 V
	n_c	0.5	0.5

Figure 2.37 Typical device parameters for a substrate *pnp* with 5100 μm² emitter area in a high-voltage, thick-epi process.

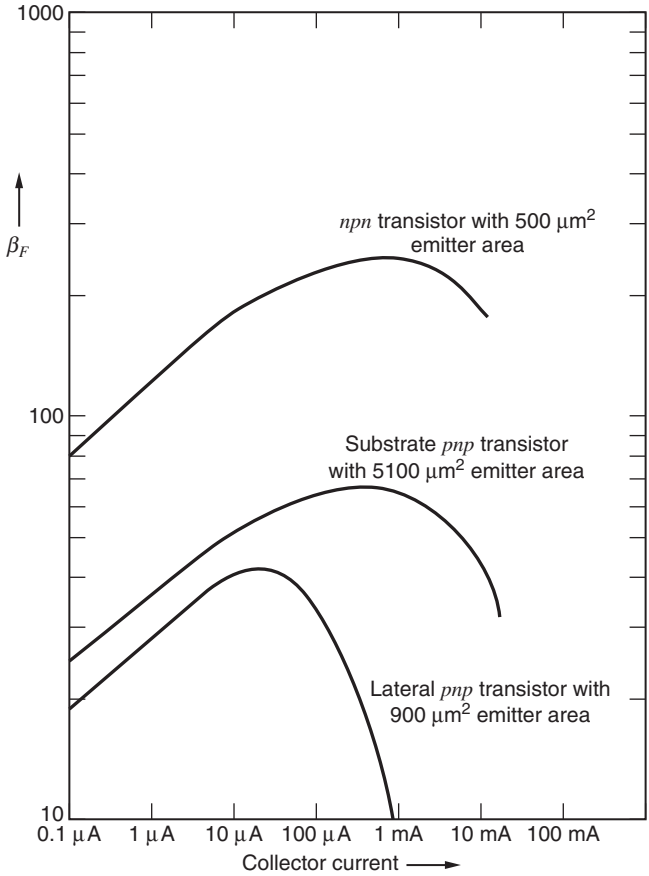


Figure 2.38 Current gain as a function of collector current for typical lateral *pnp*, substrate *pnp*, and *npn* transistor geometries in a high-voltage, thick-epi process.

2.6 Passive Components in Bipolar Integrated Circuits

In this section, we describe the structures available to the integrated-circuit designer for realization of resistance and capacitance. Resistor structures include base-diffused, emitter-diffused, ion-implanted, pinch, epitaxial, and pinched epitaxial resistors. Other resistor technologies, such as thin-film resistors, are considered in Section 2.7.3. Capacitance structures include MOS and junction capacitors. Inductors with values larger than a few nanohenries have not proven to be feasible in monolithic technology. However, such small inductors are useful in very high frequency integrated circuits.[14,15,16]

2.6.1 Diffused Resistors

In an earlier section of this chapter, the sheet resistance of a diffused layer was calculated. Integrated-circuit resistors are generally fabricated using one of the diffused or ion-implanted layers formed during the fabrication process, or in some cases a combination of two layers. The layers available for use as resistors include the base, the emitter, the epitaxial layer, the buried layer, the active-base region layer of a transistor, and the epitaxial layer pinched between the base diffusion and the *p*-type substrate. The choice of layer generally depends on the value, tolerance, and temperature coefficient of the resistor required.

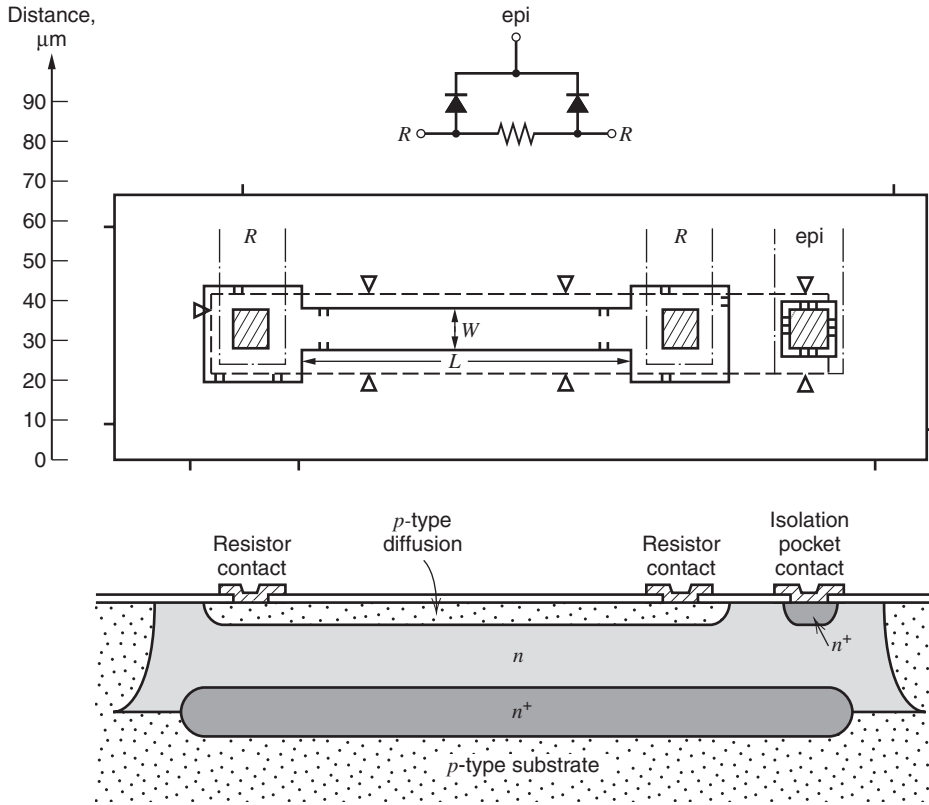


Figure 2.39 Base-diffused resistor structure.

Base and Emitter Diffused Resistors. The structure of a typical base-diffused resistor in a high-voltage process is shown in Fig. 2.39. The resistor is formed from the *p*-type base diffusion for the *npn* transistors and is situated in a separate isolation region. The epitaxial region into which the resistor structure is diffused must be biased in such a way that the *pn* junction between the resistor and the epi layer is always reverse biased. For this reason, a contact is made to the *n*-type epi region as shown in Fig. 2.39, and it is connected either to that end of the resistor that is most positive or to a potential that is more positive than either end of the resistor. The junction between these two regions contributes a parasitic capacitance between the resistor and the epi layer, and this capacitance is distributed along the length of the resistor. For most applications, this parasitic capacitance can be adequately modeled by separating it into two lumped portions and placing one lump at each end of the resistor as illustrated in Fig. 2.40.

The resistance of the structure shown in Fig. 2.39 is given by (2.10) as

$$R = \frac{L}{W} R_{\square}$$

where L is the resistor length and W is the width. The base sheet resistance $R_{\square}$ lies in the range 100 to 200 Ω/□, and thus resistances in the range 50 Ω to 50 kΩ are practical using the base diffusion. The resistance contributed by the *clubheads* at each end of the resistor can be significant, particularly for small values of L/W. The clubheads are required to allow space for ohmic contact to be made at the ends of the resistor.

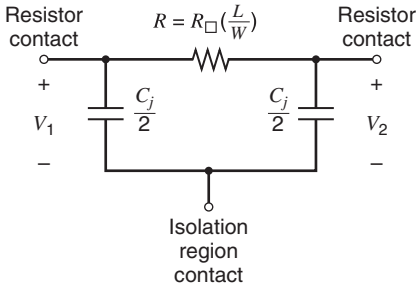


Figure 2.40 Lumped model for the base-diffused resistor.

Since minimization of die area is an important objective, the width of the resistor is kept as small as possible, the minimum practical width being limited to about 1 μm by photolithographic considerations. Both the tolerance on the resistor value and the precision with which two identical resistors can be matched can be improved by the use of wider geometries. However, for a given base sheet resistance and a given resistor value, the area occupied by the resistor increases as the *square* of its width. This can be seen from (2.10) since the ratio L/W is constant.

In shallow ion-implanted processes, the ion-implanted base can be used in the same way to form a resistor.

■ **EXAMPLE**

Calculate the resistance and parasitic capacitance of the base-diffused resistor structure shown in Fig. 2.39 for a base sheet resistance of 100 Ω/□, and an epi resistivity of 2.5 Ω-cm. Neglect end effects. The resistance is simply

$$R = 100\ \Omega/\square \left(\frac{100\ \mu\text{m}}{10\ \mu\text{m}}\right) = 1\ \text{k}\Omega$$

The capacitance is the total area of the resistor multiplied by the capacitance per unit area. The area of the resistor body is

$$A_1 = (10\ \mu\text{m})(100\ \mu\text{m}) = 1000\ \mu\text{m}^2$$

The area of the clubheads is

$$A_2 = 2\ (30\ \mu\text{m} \times 30\ \mu\text{m}) = 1800\ \mu\text{m}^2$$

The total zero-bias capacitance is, from Fig. 2.29,

$$C_{j0} = (10^{-4}\ \text{pF}/\mu\text{m}^2)(2800\ \mu\text{m}^2) = 0.28\ \text{pF}$$

As a first-order approximation, this capacitance can be divided into two parts, one placed at each end. Note that this capacitance will vary depending on the voltage at the clubhead with respect to the epitaxial pocket. ■

Emitter-diffused resistors are fabricated using geometries similar to the base resistor, but the emitter diffusion is used to form the actual resistor. Since the sheet resistance of this diffusion is in the 2 to 10 Ω/□ range, these resistors can be used to advantage where very low resistance values are required. In fact, they are widely used simply to provide a crossunder beneath an aluminum metallization interconnection. The parasitic capacitance can be

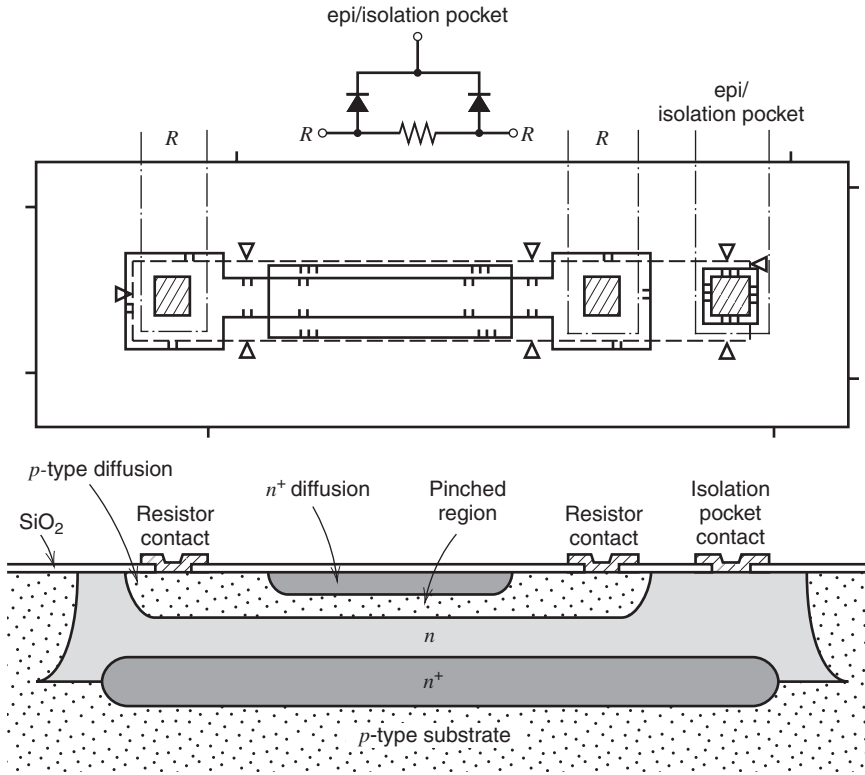


Figure 2.41 Pinch resistor structure.

calculated in a way similar to that for the base diffusion. However, these resistors have different temperature dependence from base-diffused resistors and the two types do not track with temperature.

Base Pinch Resistors. A third layer available for use as a resistor is the layer that forms the active base region in the *npn* transistor. This layer is *pinched* between the n^+ emitter and the *n*-type collector regions, giving rise to the term *pinch resistor*. The layer can be electrically isolated by reverse biasing the emitter-base and collector-base junctions, which is usually accomplished by connecting the *n*-type regions to the most positive end of the resistor. The structure of a typical pinch resistor is shown in Fig. 2.41; the n^+ diffusion overlaps the *p*-diffusion so that the n^+ region is electrically connected to the *n*-type epi region. The sheet resistance is in the 5 kΩ/□ to 15 kΩ/□ range. As a result, this resistor allows the fabrication of large values of resistance. Unfortunately, the sheet resistance undergoes the same process-related variations as does the Q_B of the transistor, which is approximately ± 50 percent. Also, because the material making up the resistor itself is relatively lightly doped, the resistance displays a relatively large variation with temperature. Another significant drawback is that the maximum voltage that can be applied across the resistor is limited to around 6 V because of the breakdown voltage between the emitter-diffused top layer and the base diffusion. Nonetheless, this type of resistor has found wide application where the large tolerance and low breakdown voltage are not significant drawbacks.

2.6.2 Epitaxial and Epitaxial Pinch Resistors

The limitation of the pinch resistor to low operating voltages disallows its use in circuits where a small bias current is to be derived directly from a power-supply voltage of more than about 7 V using a large-value resistor. The epitaxial layer itself has a sheet resistance much larger than the base diffusion, and the epi layer is often used as a resistor for this application. For example, the sheet resistance of a 17-μm thick, 5-Ω-cm epi layer can be calculated from (2.11) as

$$R_{\square} = \frac{\rho_{\text{epi}}}{T} = \frac{5\ \Omega\text{-cm}}{(17\ \mu\text{m}) \times (10^{-4}\ \text{cm}/\mu\text{m})} = 2.9\ \text{k}\Omega/\square \tag{2.25}$$

Large values of resistance can be realized in a small area using structures of the type shown in Fig. 2.42. Again, because of the light doping in the resistor body, these resistors display a rather large temperature coefficient. A still larger sheet resistance can be obtained by putting a *p*-type base diffusion over the top of an epitaxial resistor, as shown in Fig. 2.42. The depth of the *p*-type base and the thickness of the depletion region between the *p*-type base and the

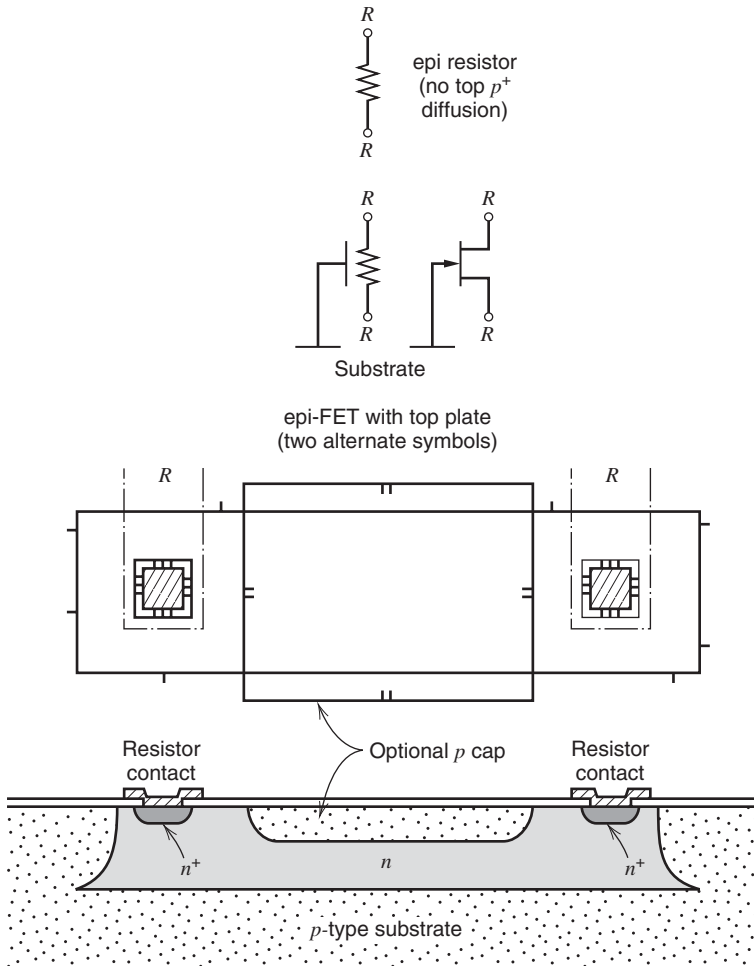


Figure 2.42 Epitaxial resistor structure. The *p*-cap diffusion is optional and forms an epitaxial pinch resistor.

Resistor Type	Sheet ρ Ω/□	Absolute Tolerance (%)	Matching Tolerance (%)	Temperature Coefficient
Base diffused	100 to 200	± 20	± 2(5 μm wide) ± 0.2(50 μm wide)	(+1500 to +2000) ppm/°C
Emitter diffused	2 to 10	± 20	± 2	+600 ppm/°C
Ion implanted	100 to 1000	± 3	± 1(5 μm wide) ± 0.1(50 μm wide)	Controllable to ± 100ppm/°C
Base pinch	2k to 10k	± 50	± 10	+2500 ppm/°C
Epitaxial	2k to 5k	± 30	± 5	+3000 ppm/°C
Epitaxial pinch	4k to 10k	± 50	± 7	+3000 ppm/°C
Thin film	0.1k to 2k	± 5 to ± 20	± 0.2 to ± 2	(± 10 to ± 200) ppm/°C

Figure 2.43 Summary of resistor properties for different types of IC resistors.

n-type epi together reduce the thickness of the resistor, increasing its sheet resistance. Such a structure actually behaves as a junction FET, in which the *p*-type gate is tied to the substrate.[17]

The properties of the various diffused and pinch-resistor structures are summarized in Fig. 2.43.

2.6.3 Integrated-Circuit Capacitors

Early analog integrated circuits were designed on the assumption that capacitors of usable value were impractical to integrate on the chip because they would take too much area, and external capacitors were used where required. Monolithic capacitors of value larger than a few tens of picofarads are still expensive in terms of die area. As a result, design approaches have evolved for monolithic circuits that allow small values of capacitance to be used to perform functions that previously required large capacitance values. The compensation of operational amplifiers is perhaps the best example of this result, and monolithic capacitors are now widely used in all types of analog integrated circuits. These capacitors fall into two categories. First, *pn* junctions under reverse bias inherently display depletion capacitance, and in certain circumstances this capacitance can be effectively utilized. The drawbacks of junction capacitance are that the junction must always be kept reverse biased, that the capacitance varies with reverse voltage, and that the breakdown voltage is only about 7 V for the emitter-base junction. For the collector-base junction, the breakdown voltage is higher, but the capacitance per unit area is quite low.

By far the most commonly used monolithic capacitor in bipolar technology is the MOS capacitor structure shown in Fig. 2.44. In the fabrication sequence, an additional mask step is inserted to define a region over an emitter diffusion on which a thin layer of silicon dioxide is grown. Aluminum metallization is then placed over this thin oxide, producing a capacitor between the aluminum and the emitter diffusion, which has a capacitance of 0.3 fF/μm^2 to 0.5 fF/μm^2 and a breakdown voltage of 60 V to 100 V. This capacitor is extremely linear and has a low temperature coefficient. A sizable parasitic capacitance C_{ISO} is present between the *n*-type bottom plate and the substrate because of the depletion capacitance of the epi-substrate junction, but this parasitic is unimportant in many applications.

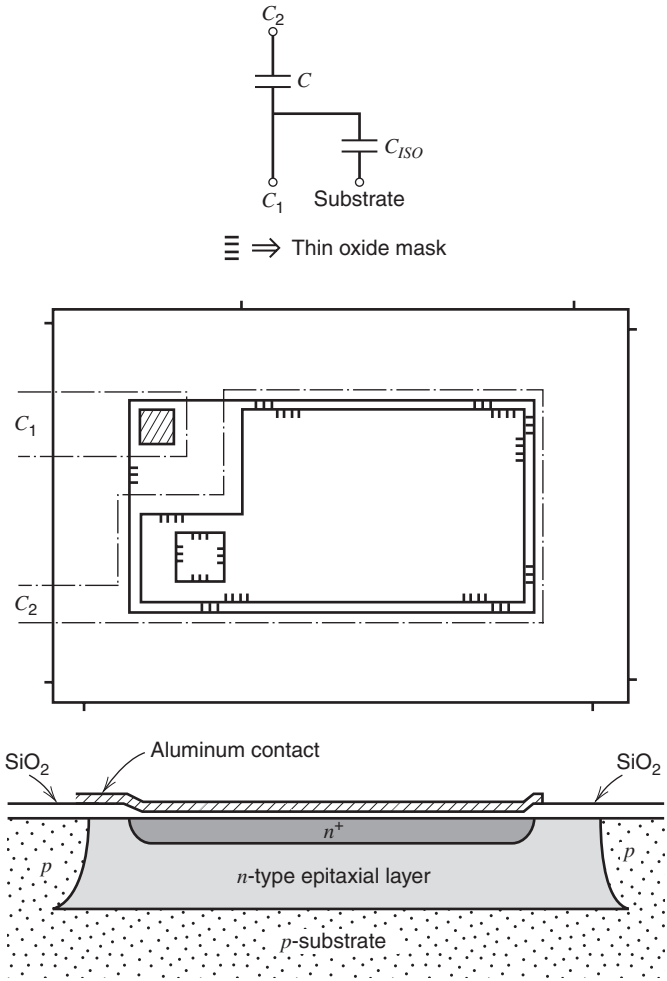
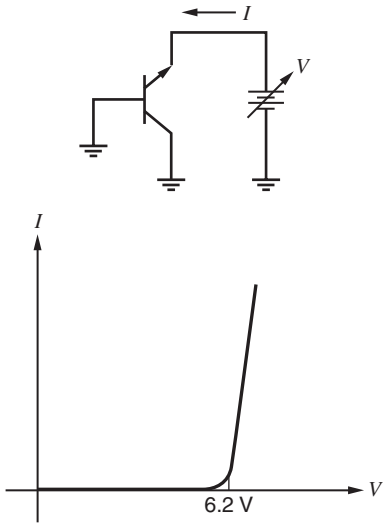


Figure 2.44 MOS capacitor structure.

2.6.4 Zener Diodes

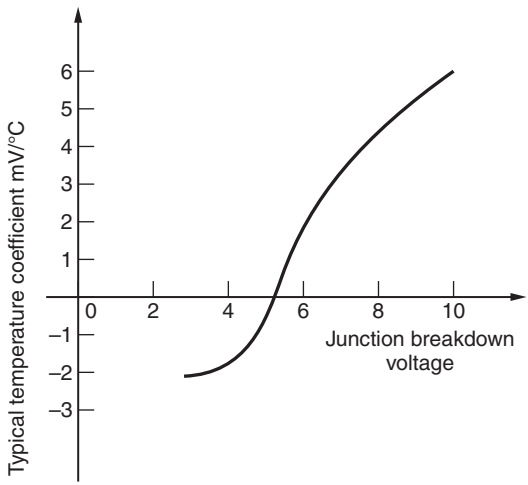
As described in Chapter 1, the emitter-base junction of the *npn* transistor structure displays a reverse breakdown voltage of between 6 V and 8 V, depending on processing details. When the total supply voltage is more than this value, the reverse-biased, emitter-base junction is useful as a voltage reference for the stabilization of bias reference circuits, and for such functions as level shifting. The reverse bias *I-V* characteristic of a typical emitter-base junction is illustrated in Fig. 2.45*a*.

An important aspect of the behavior of this device is the temperature sensitivity of the breakdown voltage. The actual breakdown mechanism is dominated by quantum mechanical tunneling through the depletion layer when the breakdown voltage is below about 6 V; it is dominated by avalanche multiplication in the depletion layer at the larger breakdown voltages. Because these two mechanisms have opposite temperature coefficients of breakdown voltage, the actually observed breakdown voltage has a temperature coefficient that varies with the value of breakdown voltage itself, as shown in Fig. 2.45*b*.



(*a*)

Figure 2.45 (*a*) Current-voltage characteristic of a typical emitter-base Zener diode.



(*b*)

Figure 2.45 (*b*) Temperature coefficient of junction breakdown voltage as a function of breakdown voltage.

2.6.5 Junction Diodes

Junction diodes can be formed by various connections of the *npn* and *pnp* transistor structures, as illustrated in Fig. 2.46. When the diode is forward biased in the diode connections *a, b*, and *d* of Fig. 2.46, the collector-base junction becomes forward biased as well. When this occurs, the collector-base junction injects holes into the epi region that can be collected by the reverse-biased, epi-isolation junction or by other devices in the same isolation region. A similar phenomenon occurs when a transistor enters saturation. As a result, substrate currents can flow that can cause voltage drops in the high-resistivity substrate material, and other epi-isolation junctions within the circuit can become inadvertently forward biased. Thus the diode connections of Fig. 2.46*c* are usually preferable since they keep the base-collector junction at zero bias. These connections have the additional advantage of resulting in the smallest amount of minority charge storage within the diode under forward-bias conditions.

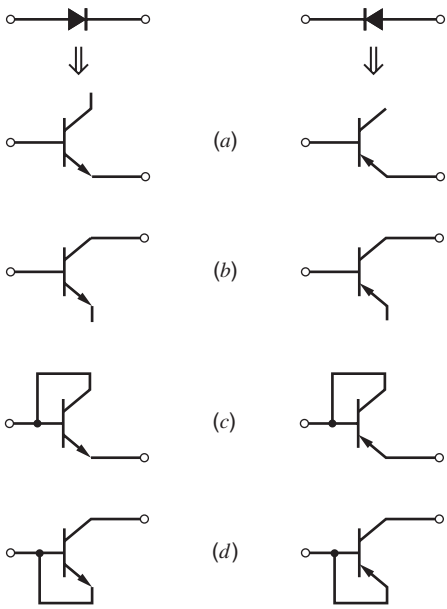


Figure 2.46 Diode connections for *npn* and *pnp* transistors.

2.7 Modifications to the Basic Bipolar Process

The basic high-voltage bipolar IC fabrication process described previously can be modified by the addition of extra processing steps to produce special devices or characteristics.

2.7.1 Dielectric Isolation

We first consider a special isolation technique—*dielectric isolation*—that has been used in digital and analog integrated circuits that must operate at very high speed and/or must operate in the presence of large amounts of radiation. The objective of the isolation technique is to electrically isolate the collectors of the devices from each other with a layer of silicon dioxide rather than with a *pn* junction. This layer has much lower capacitance per unit area than a *pn* junction, and as a result, the collector-substrate capacitance of the transistors is greatly reduced. Also, the reverse photocurrent that occurs with junction-isolated devices under intense radiation is eliminated.

The fabrication sequence used for dielectric isolation is illustrated in Figs. 2.47*a–d*. The starting material is a wafer of *n*-type material of resistivity appropriate for the collector region of the transistor. The first step is to etch grooves in the back side of the starting wafer, which will become the isolation regions in the finished circuit. These grooves are about 20 μm deep for typical analog circuit processing. This step, called *moat etch*, can be accomplished with a variety of techniques, including a preferential etch that allows precise definition of the depth of the moats. Next, an oxide is grown on the surface and a thick layer of polycrystalline silicon is deposited on the surface. This layer will be the mechanical support for the finished wafer and thus must be on the order of 200 μm thick. Next, the starting wafer is etched or ground from the top side until it is entirely removed except for the material left in the isolated islands between the moats, as illustrated in Fig. 2.47*c*. After the growth of an oxide, the wafer is ready for the rest of the standard process sequence. Note that the isolation of each device is accomplished by means of an oxide layer.

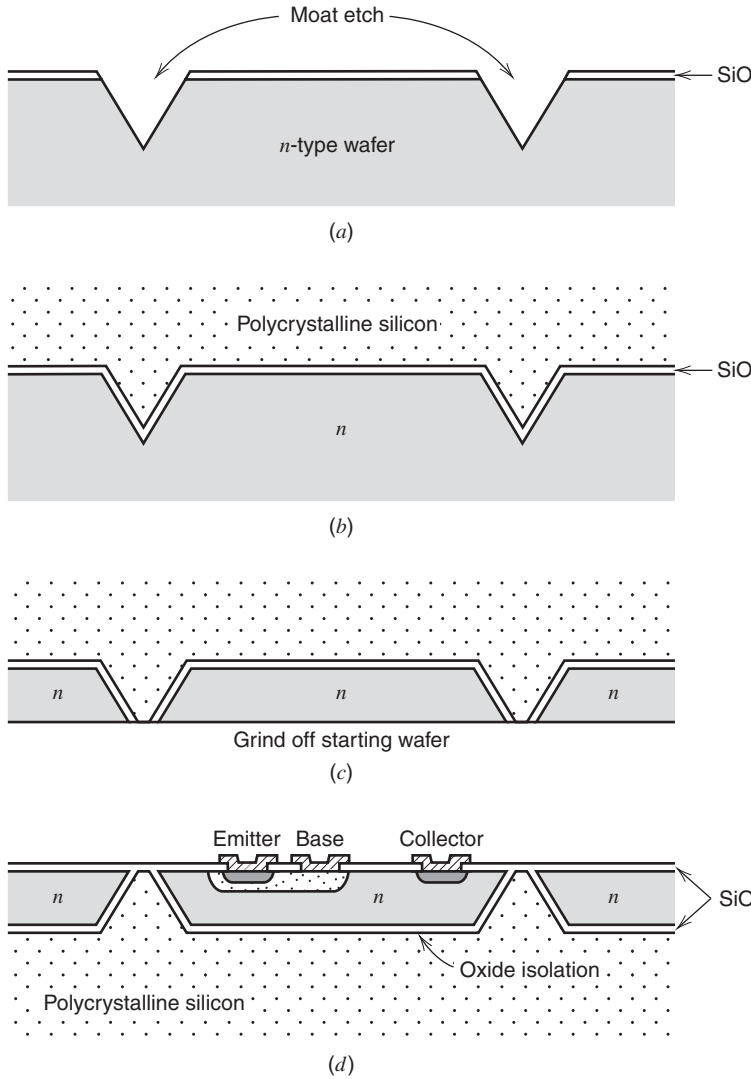


Figure 2.47 Fabrication steps in dielectric isolation. (*a*) Moat etch on bottom of starting wafer. (*b*) Deposit polycrystalline silicon support layer. (*c*) Grind off starting wafer and polish. (*d*) Carry out standard process, starting with base mask.

2.7.2 Compatible Processing for High-Performance Active Devices

Many specialized circuit applications require a particular type of active device other than the *npn* and *pnp* transistors that result from the standard process schedule. These include high-beta (*superbeta*) *npn* transistors for low-input-current amplifiers, MOSFETs for analog switching and low-input-current amplifiers, and high-speed *pnp* transistors for fast analog circuits. The fabrication of these devices generally requires the addition of one or more mask steps to the basic fabrication process. We now describe these special structures.

Superbeta Transistors. One approach to decreasing the input bias current in amplifiers is to increase the current gain of the input stage transistors.[18] Since a decrease in the base width

of a transistor improves both the base transport factor and the emitter efficiency (see Section 1.3.1), the current gain increases as the base width is made smaller. Thus the current gain of the devices in the circuit can be increased by simply increasing the emitter diffusion time and narrowing the base width in the resulting devices. However, any increase in the current gain also causes a reduction in the breakdown voltage BV_{CEO} of the transistors. Section 1.3.4 shows that

$$BV_{CEO} = \frac{BV_{CBO}}{\sqrt[n]{\beta}} \tag{2.26}$$

where BV_{CBO} is the plane breakdown voltage of the collector-base junction. Thus for a given epitaxial layer resistivity and corresponding collector-base breakdown voltage, an increase in beta gives a decrease in BV_{CEO}. As a result, using such a process modification to increase the beta of all the transistors in an operational amplifier is not possible because the modified transistors could not withstand the required operating voltage.

The problem of the trade-off between current gain and breakdown voltage can be avoided by fabricating two different types of devices on the same die. The standard device is similar to conventional transistors in structure. By inserting a second diffusion, however, high-beta devices also can be formed. A structure typical of such devices is shown in Fig. 2.48. These devices may be made by utilizing the same base diffusion for both devices and using separate emitter diffusions, or by using two different base diffusions and the same emitter diffusion. Both techniques are used. If the superbeta devices are used only as the input transistors in an operational amplifier, they are not required to have a breakdown voltage of more than about 1 V. Therefore, they can be diffused to extremely narrow base widths, giving current gain on the order of 2000 to 5000. At these base widths, the actual breakdown mechanism is often no longer collector multiplication at all but is due to the depletion layer of the collector-base junction depleting the whole base region and reaching the emitter-base depletion layer. This breakdown mechanism is called *punchthrough*. An application of these devices in op-amp design is described in Section 6.9.2.

MOS Transistors. MOS transistors are useful in bipolar integrated-circuit design because they provide high-performance analog switches and low-input-current amplifiers, and particularly because complex digital logic can be realized in a small area using MOS technology. The latter consideration is important since the partitioning of subsystems into analog and digital chips becomes more and more cumbersome as the complexity of the individual chips becomes greater.

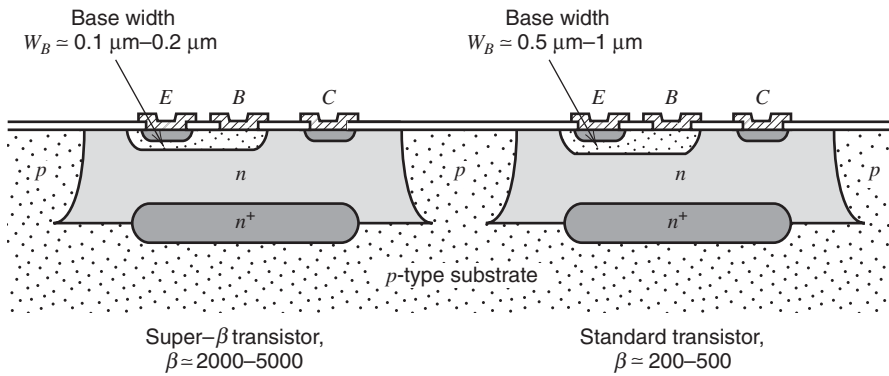


Figure 2.48 Superbeta device structure.

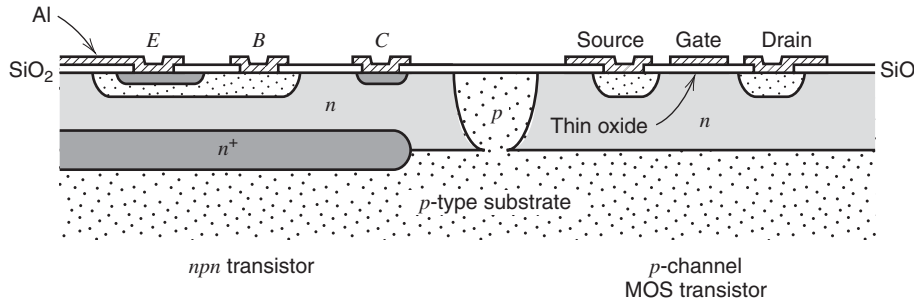


Figure 2.49 Compatible *p*-channel MOS transistor.

Metal-gate *p*-channel MOS transistors can be formed in a standard high-voltage bipolar analog IC process with one extra mask step.[19] If a capacitor mask is included in the original sequence, then no extra mask steps are required. As illustrated in Fig. 2.49, the source and drain are formed in the epi material using the base diffusion. The capacitor mask is used to define the oxide region over the channel and the aluminum metallization forms the metal gate.

A major development in IC processing in recent years has been the combination on the same chip of high-performance bipolar devices with CMOS devices in a BiCMOS process. This topic is considered in Section 2.11.

Double-Diffused *pnp* Transistors. The limited frequency response of the lateral *pnp* transistor places a limitation on the high-frequency performance attainable with certain types of analog circuits. While this problem can be circumvented by clever circuit design in many cases, the resulting circuit is often quite complex and costly. An alternative approach is to use a more complex process that produces a high-speed, double-diffused *pnp* transistor with properties comparable to those of the *npn* transistor.[20] The process usually utilizes three additional mask steps and diffusions: one to form a lightly doped *p*-type region, which will be the collector of the *pnp*; one *n*-type diffusion to form the base of the *pnp*; and one *p*-type diffusion to form the emitter of the *pnp*. A typical resulting structure is shown in Fig. 2.50. This process requires

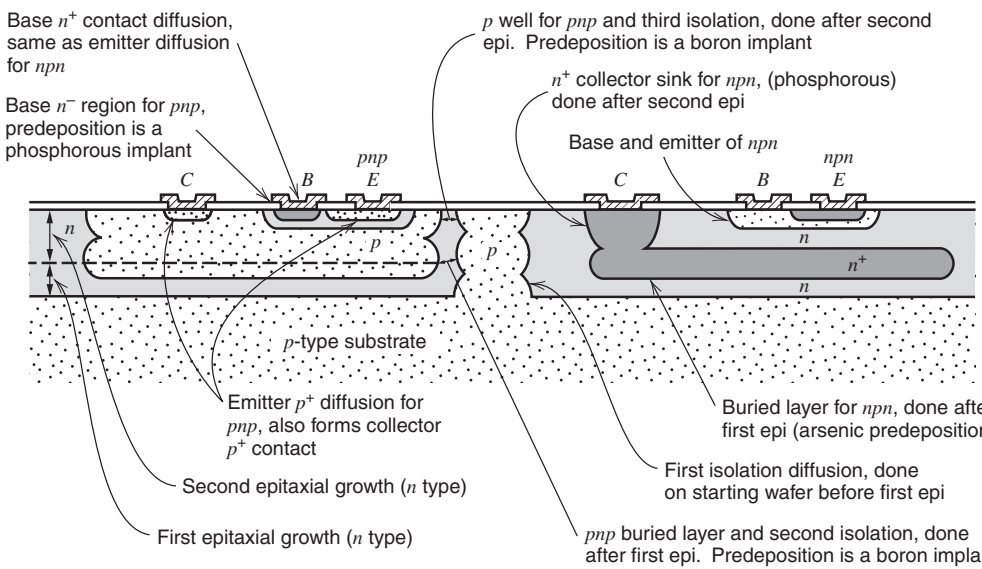


Figure 2.50 Compatible double-diffused *pnp* process.

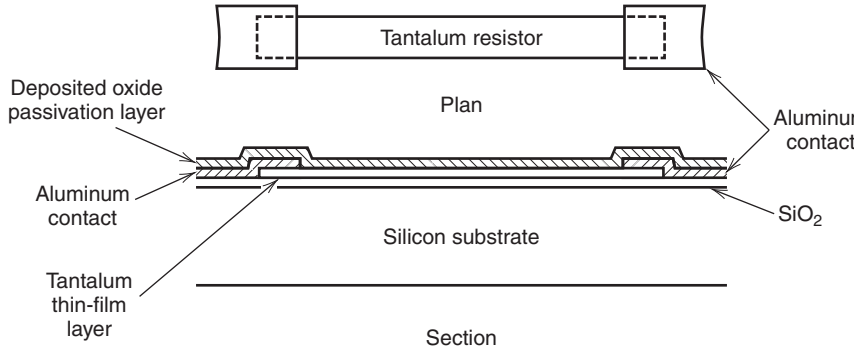


Figure 2.51 Typical thin-film resistor structure.

	Nichrome	Tantalum	Cermet (Cr-SiO)
Range of sheet resistance (Ω/□)	10 to 1000	10 to 1000	30 to 2500
Temperature coefficient (ppm/°C)	±10 to ±150	±5 to ±200	±50 to ±150

Figure 2.52 Properties of monolithic thin-film resistors.

ten masking steps and two epitaxial growth steps. Oxide isolation and poly-emitter technology have been incorporated into more advanced versions of this process.

2.7.3 High-Performance Passive Components

Diffused resistors have three drawbacks: They have high temperature coefficients, they have poor tolerance, and they are junction-isolated. The latter means that a parasitic capacitance is associated with each resistor, and exposure to radiation causes photocurrents to flow across the isolating junction. These drawbacks can be overcome by the use of thin-film resistors deposited on the top surface of the die over an insulating layer of oxide. After the resistor material itself is deposited, the individual resistors are defined in a conventional way using a masking step. They are then interconnected with the rest of the circuit using the standard aluminum interconnect process. The most common materials for the resistors are nichrome and tantalum, and a typical structure is shown in Fig. 2.51. The properties of the resulting resistors using these materials are summarized in Fig. 2.52.

2.8 MOS Integrated-Circuit Fabrication

Fabrication technologies for MOS integrated circuits span a considerably wider spectrum of complexity and performance than those for bipolar technology. CMOS technologies provide two basic types of transistors: enhancement-mode *n*-channel transistors (which have positive thresholds) and enhancement-mode *p*-channel transistors (which have negative thresholds). The magnitudes of the threshold voltages of these transistors are typically set to be 0.6 V to 0.8 V so that the drain current resulting from subthreshold conduction with zero gate-source voltage is very small. This property gives standard CMOS digital circuits high noise margins and essentially zero static power dissipation. However, such thresholds do not always minimize the *total* power dissipation because significant dynamic power is dissipated by

charging and discharging internal nodes during logical transitions, especially for high clock rates and power-supply voltages.[21] To reduce the minimum required supply voltage and the total power dissipation for some applications, low-threshold, enhancement-mode devices or depletion-mode devices are sometimes used instead of or along with the standard-threshold, enhancement-mode devices. For the sake of illustration, we will consider an example process that contains enhancement-mode *n*- and *p*-channel devices along with a depletion-mode *n*-channel device.

CMOS technologies can utilize either a *p*-type or *n*-type substrate, with the complementary device type formed in an implanted well of the opposite impurity type. We will take as an example a process in which the starting material is *p*-type. The starting material is a silicon wafer with a concentration in the range of 10^{14} to 10^{15} atoms/cm^3. In CMOS technology, the first step is the formation of a *well* of opposite impurity-type material where the complementary device will be formed. In this case, the well is *n*-type and is formed by a masking operation and ion implantation of a donor species, typically phosphorus. Subsequent diffusion results in the structure shown in Fig. 2.53. The surface concentration in the well following diffusion is typically between 10^{15} and 10^{16} atoms/cm^3.

Next, a layer of silicon nitride is deposited and defined with a masking operation so that nitride is left only in the areas that are to become active devices. After this masking operation, additional ion implantations are carried out, which increase the surface concentrations in the areas that are not covered by nitride, called the *field regions*. This often involves an extra masking operation so that the surface concentration in the well and that in the substrate areas can be independently controlled by means of separate implants. This increase in surface concentration in the field is necessary because the field regions themselves are MOS transistors with very thick gate oxide. To properly isolate the active devices from one another, the field devices must have a threshold voltage high enough that they never turn on. This can be accomplished by increasing the surface concentration in the field regions. Following the field implants, a local oxidation is performed, which results in the structure shown in Fig. 2.54.

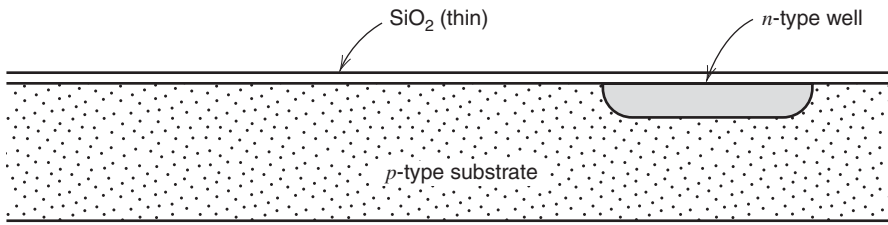


Figure 2.53 Cross section of sample following implantation and diffusion of the *n*-type well. Subsequent processing will result in formation of an *n*-channel device in the unimplanted *p*-type portions of the substrate and a *p*-type transistor in the *n*-type well region.

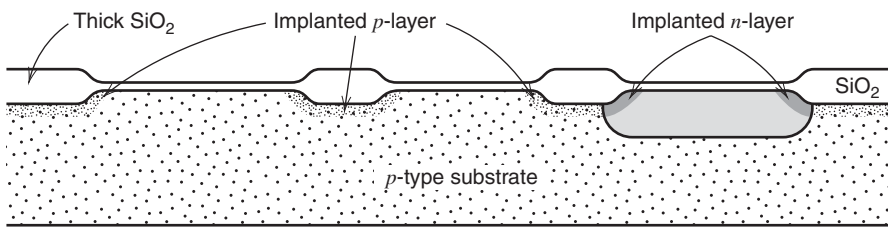


Figure 2.54 Cross section of the sample following field implant steps and field oxidation.

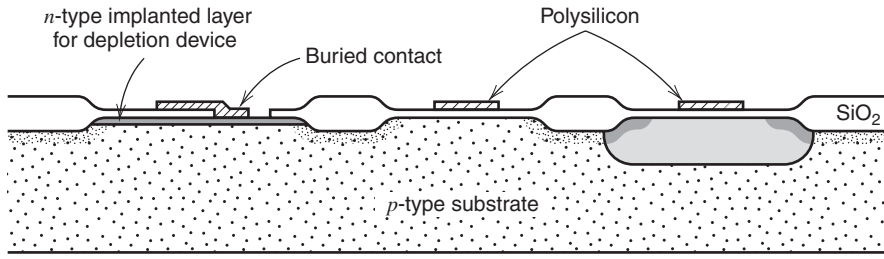


Figure 2.55 Cross section of the sample following deposition and definition of the polysilicon gate layer. Ion implantations have been performed in the thin-oxide regions to adjust the thresholds of the devices.

After field-oxide growth, the nitride is removed from the active areas, and implantation steps are carried out, which adjust the surface concentrations in what will become the channel of the MOS transistors. Equation 1.139, applied to the doping levels usually found in the active-device areas, gives an *n*-channel threshold of within a few hundred millivolts of zero, and *p*-channel threshold of about -2 V. To shift the magnitudes of the device threshold voltages to 0.6 V to 0.8 V, an implantation step that changes the impurity concentration at the surface in the channel regions of the two transistor types is usually included. This shift in threshold can sometimes be accomplished by using a single sheet implant over the entire wafer, which simultaneously shifts the thresholds of both types of devices. More typically, however, two separate masked implants are used, one for each device type. Also, if a depletion-mode *n*-channel device is included in the process, it is defined at this point by a masking operation and subsequent implant to shift the threshold of the selected devices to a negative value so that they are normally on.

Next, a layer of polysilicon is deposited, and the gates of the various devices are defined with a masking operation. The resulting structure is shown in Fig. 2.55. Silicon-gate MOS technology provides three materials that can be used for interconnection: polysilicon, diffusion, and several layers of metal. Unless special provision is made in the process, connections between polysilicon and diffusion layers require a metallization bridge, since the polysilicon layer acts as a mask for the diffused layers. To provide a direct electrical connection between polysilicon and diffusion layers, a buried contact can be included just prior to the polysilicon deposition. This masking operation opens a window in the silicon dioxide under the polysilicon, allowing it to touch the bare silicon surface when it is deposited, forming a direct polysilicon-silicon contact. The depletion device shown in Fig. 2.55 has such a buried contact connecting its source to its gate.

Next, a masking operation is performed such that photoresist covers the *p*-channel devices, and the wafer is etched to remove the oxide from the source and drain areas of the *n*-channel devices. Arsenic or phosphorus is then introduced into these areas, using either diffusion or ion implantation. After a short oxidation, the process is repeated for the *p*-channel source and drain areas, where boron is used. The resulting structure is shown in Fig. 2.56.

At this point in the process, a layer of silicon dioxide is usually deposited on the wafer, using chemical vapor deposition or some other similar technique. This layer is required to reduce the parasitic capacitance of the interconnect metallization and cannot be thermally grown because of the redistribution of the impurities within the device structures that would result during the growth. Following the oxide deposition, the contact windows are formed with a masking operation, and metallization is deposited and defined with a second masking operation. The final structure is shown in Fig. 2.57. A microscope photograph of such a device is shown in Fig. 2.58. Subsequent fabrication steps are as described in Section 2.3 for bipolar technology.

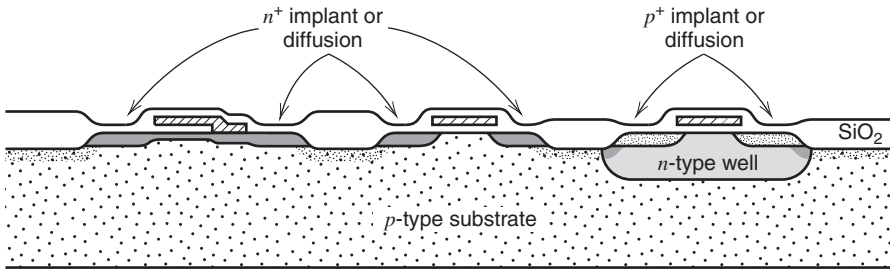


Figure 2.56 Cross section of the sample following the source drain masking and diffusion operations.

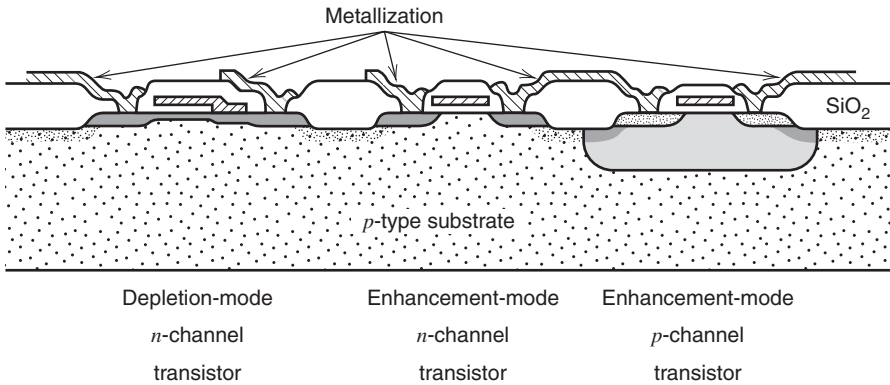


Figure 2.57 Cross section of the sample after final process step. The enhancement and depletion n-channel devices are distinguished from each other by the fact that the depletion device has received a channel implantation of donor impurities to lower its threshold voltage, usually to the range of -1.5 V to -3 V.

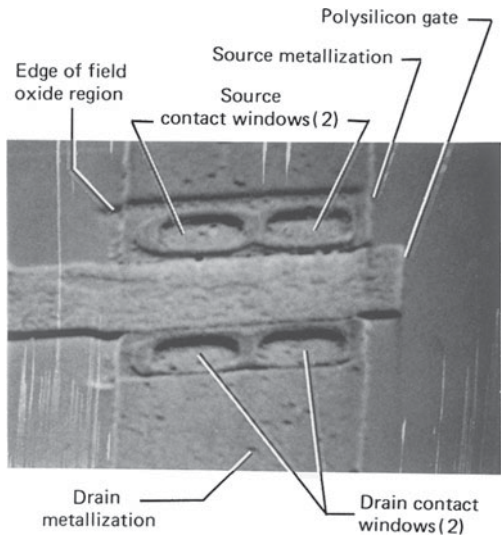


Figure 2.58 Photomicrograph of a silicon-gate MOS transistor. Visible in this picture are the polysilicon gate, field-oxide region boundary, source and drain metallization, and contact windows. In this particular device, the contact windows have been broken into two smaller rectangular openings rather than a single long one as shown in Fig. 2.59. Large contact windows are frequently implemented with an array of small openings so that all individual contact holes in the integrated circuit have the same nominal geometry. This results in better uniformity of the etch rate of the contact windows and better matching.

2.9 Active Devices in MOS Integrated Circuits

The process sequence described in the previous section results in a variety of device types having different threshold voltages, channel mobilities, and parasitic capacitances. In addition, the sequence allows the fabrication of a bipolar emitter follower, using the well as a base. In this section, we explore the properties of these different types of devices.

2.9.1 *n*-Channel Transistors

A typical layout of an *n*-channel MOS transistor is shown in Fig. 2.59. The electrically active portion of the device is the region under the gate; the remainder of the device area simply provides electrical contact to the terminals. As in the case of integrated bipolar transistors, these areas contribute additional parasitic capacitance and resistance.

In the case of MOS technology, the circuit designer has even greater flexibility than in the bipolar case to tailor the properties of each device to the role it is to play in the individual circuit application. Both the channel width (analogous to the emitter area in bipolar) and the channel length can be defined by the designer. The latter is analogous to the base width of a bipolar device, which is not under the control of the bipolar circuit designer since it is a process parameter and not a mask parameter. In contrast to a bipolar transistor, the transconductance

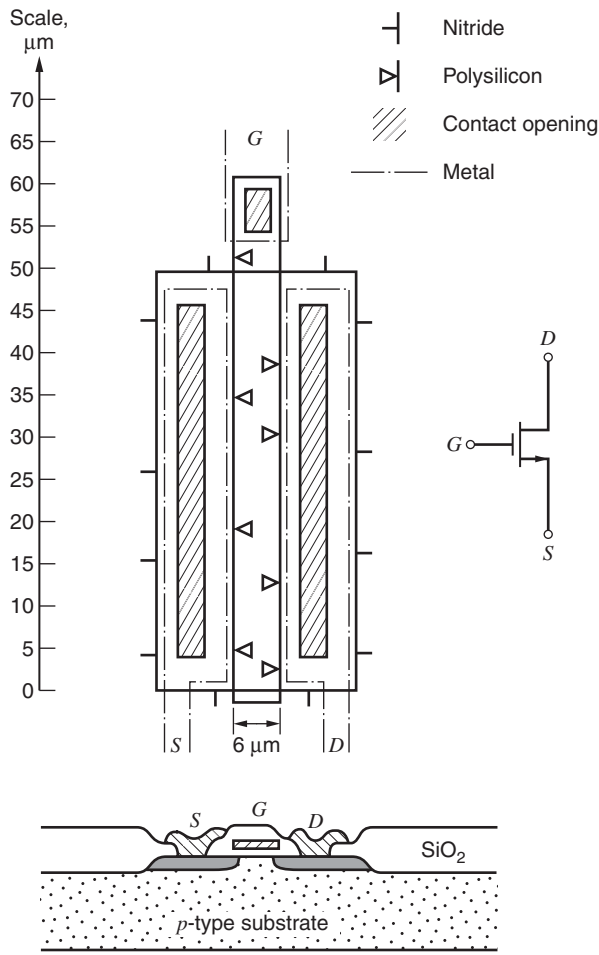


Figure 2.59 Example layout of an *n*-channel silicon-gate MOS transistor. The mask layers are coded as shown.

of an MOS device can be made to vary over a wide range at a fixed drain current by simply changing the device geometry. The same is true of the gate-source voltage. In making these design choices, the designer must be able to relate changes in device geometry to changes in the electrical properties of the device. To illustrate this procedure, we will calculate the model parameters of the device shown in Fig. 2.59. This device has a drawn channel length of 6 μm and channel width of 50 μm. We will assume the process has the parameters that are summarized in Table 2.1. This is typical of processes with minimum allowed gate lengths of 3 μm. Parameters for more advanced processes are given in Tables 2.2, 2.3, 2.4, 2.5, and 2.6.

Table 2.1 Summary of Process Parameters for a Typical Silicon-Gate *n*-Well CMOS Process with 3 μm Minimum Allowed Gate Length

Parameter	Symbol	Value *n*-Channel Transistor	Value *p*-Channel Transistor	Units
Substrate doping	N_A, N_D	1×10^{15}	1×10^{16}	Atoms/cm^3
Gate oxide thickness	t_{ox}	400	400	Å
Metal-silicon work function	ϕ_{ms}	−0.6	−0.1	V
Channel mobility	μ_n, μ_p	700	350	cm^2/V-s
Minimum drawn channel length	L_{drwn}	3	3	μm
Source, drain junction depth	X_j	0.6	0.6	μm
Source, drain side diffusion	L_d	0.3	0.3	μm
Overlap capacitance per unit gate width	C_{ol}	0.35	0.35	fF/μm
Threshold adjust implant (box dist)				
impurity type		P	P	
effective depth	X_i	0.3	0.3	μm
effective surface concentration	N_{si}	2×10^{16}	0.9×10^{16}	Atoms/cm^3
Nominal threshold voltage	V_t	0.7	−0.7	V
Polysilicon gate doping concentration	N_{dpoly}	10^{20}	10^{20}	Atoms/cm^3
Poly gate sheet resistance	R_s	20	20	Ω/□
Source, drain-bulk junction capacitances (zero bias)	C_{j0}	0.08	0.20	fF/μm^2
Source, drain-bulk junction capacitance grading coefficient	n	0.5	0.5	
Source, drain periphery capacitance (zero bias)	C_{jsw0}	0.5	1.5	fF/μm
Source, drain periphery capacitance grading coefficient	n	0.5	0.5	
Source, drain junction built-in potential	ψ_0	0.65	0.65	V
Surface-state density	$\frac{Q_{SS}}{q}$	10^{11}	10^{11}	Atoms/cm^2
Channel-length modulation parameter	$\left\lvert \frac{dX_d}{dV_{DS}} \right\rvert$	0.2	0.1	μm/V

Table 2.2 Summary of Process Parameters for a Typical Silicon-Gate *n*-Well CMOS Process with 1.5 μm Minimum Allowed Gate Length

Parameter	Symbol	Value *n*-Channel Transistor	Value *p*-Channel Transistor	Units
Substrate doping	N_A, N_D	2×10^{15}	1.5×10^{16}	Atoms/cm^3
Gate oxide thickness	t_{ox}	250	250	Å
Metal-silicon work function	ϕ_{ms}	−0.6	−0.1	V
Channel mobility	μ_n, μ_p	650	300	cm^2/V-s
Minimum drawn channel length	L_{drwn}	1.5	1.5	μm
Source, drain junction depth	X_j	0.35	0.4	μm
Source, drain side diffusion	L_d	0.2	0.3	μm
Overlap capacitance per unit gate width	C_{ol}	0.18	0.26	fF/μm
Threshold adjust implant (box dist)				
impurity type		P	P	
effective depth	X_i	0.3	0.3	μm
effective surface concentration	N_{si}	2×10^{16}	0.9×10^{16}	Atoms/cm^3
Nominal threshold voltage	V_t	0.7	−0.7	V
Polysilicon gate doping concentration	N_{dpoly}	10^{20}	10^{20}	Atoms/cm^3
Poly gate sheet resistance	R_s	20	20	Ω/□
Source, drain-bulk junction capacitances (zero bias)	C_{j0}	0.14	0.25	fF/μm^2
Source, drain-bulk junction capacitance grading coefficient	n	0.5	0.5	
Source, drain periphery capacitance (zero bias)	C_{jsw0}	0.8	1.8	fF/μm
Source, drain periphery capacitance grading coefficient	n	0.5	0.5	
Source, drain junction built-in potential	ψ_0	0.65	0.65	V
Surface-state density	$\frac{Q_{SS}}{q}$	10^{11}	10^{11}	Atoms/cm^2
Channel-length modulation parameter	$\left\lvert \frac{dX_d}{dV_{DS}} \right\rvert$	0.12	0.06	μm/V

Threshold Voltage. In Chapter 1, an MOS transistor was shown to have a threshold voltage of

$$V_t = \phi_{ms} + 2\phi_f + \frac{Q_b}{C_{ox}} - \frac{Q_{ss}}{C_{ox}} \tag{2.27}$$

where ϕ_{ms} is the metal-silicon work function, ϕ_f is the Fermi level in the bulk silicon, Q_b is the bulk depletion layer charge, C_{ox} is the oxide capacitance per unit area, and Q_{ss} is the

Table 2.3 Summary of Process Parameters for a Typical Silicon-Gate *n*-Well CMOS Process with 0.8μm Minimum Allowed Gate Length

Parameter	Symbol	Value *n*-Channel Transistor	Value *p*-Channel Transistor	Units
Substrate doping	N_A, N_D	4×10^{15}	3×10^{16}	Atoms/cm^3
Gate oxide thickness	t_{ox}	150	150	Å
Metal-silicon work function	ϕ_{ms}	−0.6	−0.1	V
Channel mobility	μ_n, μ_p	550	250	cm^2/V-s
Minimum drawn channel length	L_{drwn}	0.8	0.8	μm
Source, drain junction depth	X_j	0.2	0.3	μm
Source, drain side diffusion	L_d	0.12	0.18	μm
Overlap capacitance per unit gate width	C_{ol}	0.12	0.18	fF/μm
Threshold adjust implant (box dist)				
impurity type		P	P	
effective depth	X_i	0.2	0.2	μm
effective surface concentration	N_{si}	3×10^{16}	2×10^{16}	Atoms/cm^3
Nominal threshold voltage	V_t	0.7	−0.7	V
Polysilicon gate doping concentration	N_{dpoly}	10^{20}	10^{20}	Atoms/cm^3
Poly gate sheet resistance	R_s	10	10	Ω/□
Source, drain-bulk junction capacitances (zero bias)	C_{j0}	0.18	0.30	fF/μm^2
Source, drain-bulk junction capacitance grading coefficient	n	0.5	0.5	
Source, drain periphery capacitance (zero bias)	C_{jsw0}	1.0	2.2	fF/μm
Source, drain periphery capacitance grading coefficient	n	0.5	0.5	
Source, drain junction built-in potential	ψ_0	0.65	0.65	V
Surface-state density	$\frac{Q_{SS}}{q}$	10^{11}	10^{11}	Atoms/cm^2
Channel-length modulation parameter	$\left\lvert \frac{dX_d}{dV_{DS}} \right\rvert$	0.08	0.04	μm/V

concentration of surface-state charge. An actual calculation of the threshold is illustrated in the following example.

Often the threshold voltage must be deduced from measurements, and a useful approach to doing this is to plot the square root of the drain current as a function of V_{GS}, as shown in Fig. 2.60. The threshold voltage can be determined as the extrapolation of the straight portion

Table 2.4 Summary of Process Parameters for a Typical Silicon-Gate *n*-Well CMOS Process with 0.4 μm Minimum Allowed Gate Length

Parameter	Symbol	Value *n*-Channel Transistor	Value *p*-Channel Transistor	Units
Substrate doping	N_A, N_D	5×10^{15}	4×10^{16}	Atoms/cm^3
Gate oxide thickness	t_{ox}	80	80	Å
Metal-silicon work function	ϕ_{ms}	−0.6	−0.1	V
Channel mobility	μ_n, μ_p	450	150	cm^2/V-s
Minimum drawn channel length	L_{drwn}	0.4	0.4	μm
Source, drain junction depth	X_j	0.15	0.18	μm
Source, drain side diffusion	L_d	0.09	0.09	μm
Overlap capacitance per unit gate width	C_{ol}	0.35	0.35	fF/μm
Threshold adjust implant (box dist)				
impurity type		P	P	
effective depth	X_i	0.16	0.16	μm
effective surface concentration	N_{si}	4×10^{16}	3×10^{16}	Atoms/cm^3
Nominal threshold voltage	V_t	0.6	−0.8	V
Polysilicon gate doping concentration	N_{dpoly}	10^{20}	10^{20}	Atoms/cm^3
Poly gate sheet resistance	R_s	5	5	Ω/□
Source, drain-bulk junction capacitances (zero bias)	C_{j0}	0.2	0.4	fF/μm^2
Source, drain-bulk junction capacitance grading coefficient	n	0.5	0.4	
Source, drain periphery capacitance (zero bias)	C_{jsw0}	1.2	2.4	fF/μm
Source, drain periphery capacitance grading coefficient	n	0.4	0.3	
Source, drain junction built-in potential	ψ_0	0.7	0.7	V
Surface-state density	$\frac{Q_{SS}}{q}$	10^{11}	10^{11}	Atoms/cm^2
Channel-length modulation parameter	$\left\lvert \frac{dX_d}{dV_{DS}} \right\rvert$	0.02	0.04	μm/V

of the curve to zero current. The slope of the curve also yields a direct measure of the quantity $\mu_n C_{ox} W/L_{\text{eff}}$ for the device at the particular drain-source voltage at which the measurement is made. The measured curve deviates from a straight line at low currents because of subthreshold conduction and at high currents because of mobility degradation in the channel as the carriers approach scattering-limited velocity.

Table 2.5 Summary of Process Parameters for a Typical CMOS Process with 0.2 μm Minimum Allowed Gate Length

Parameter	Symbol	Value *n*-Channel Transistor	Value *p*-Channel Transistor	Units
Substrate doping	N_A, N_D	8×10^{16}	8×10^{16}	Atoms/cm^3
Gate oxide thickness	t_{ox}	42	42	Angstroms
Metal-silicon work function	ϕ_{ms}	−0.6	−0.1	V
Channel mobility	μ_n, μ_p	300	80	cm^2/V-s
Minimum drawn channel length	L_{drwn}	0.2	0.2	μm
Source, drain junction depth	X_j	0.16	0.16	μm
Source, drain side diffusion	L_d	0.01	0.015	μm
Overlap capacitance per unit gate width	C_{ol}	0.36	0.33	fF/μm
Threshold adjust implant (box dist.)				
impurity type		P	P	
effective depth	X_i	0.12	0.12	μm
effective surface concentration	N_{si}	2×10^{17}	2×10^{17}	Atoms/cm^3
Nominal threshold voltage	V_t	0.5	−0.45	V
Polysilicon gate doping concentration	N_{dpoly}	10^{20}	10^{20}	Atoms/cm^3
Poly gate sheet resistance	R_s	7	7	Ω/□
Source, drain-bulk junction capacitances (zero bias)	C_{j0}	1.0	1.1	fF/μm^2
Source, drain-bulk junction capacitance grading coefficient	n	0.36	0.45	
Source, drain periphery capacitance (zero bias)	C_{jsw0}	0.2	0.25	fF/μm
Source, drain periphery capacitance grading coefficient	n	0.2	0.24	
Source, drain junction built-in potential	ψ_0	0.68	0.74	V
Surface-state density	$\frac{Q_{SS}}{q}$	10^{11}	10^{11}	Atoms/cm^2
Channel-length modulation parameter	$\left\lvert \frac{dX_d}{dV_{DS}} \right\rvert$	0.028	0.023	μm/V

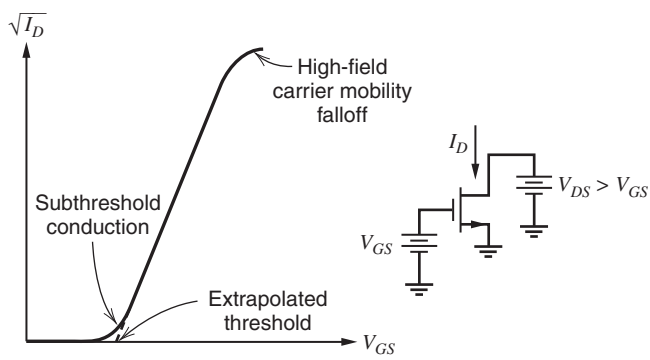


Figure 2.60 Typical experimental variation of drain current as a function of the square root of gate-source voltage in the active region.

Table 2.6 Summary of Process Parameters for a Typical CMOS Process with 0.1 μm Minimum Allowed Gate Length

Parameter	Symbol	Value *n*-Channel Transistor	Value *p*-Channel Transistor	Units
Substrate doping	N_A, N_D	1×10^{17}	1×10^{17}	Atoms/cm^3
Gate oxide thickness	t_{ox}	25	25	Angstroms
Gate leakage current density	J_G	1.2	0.4	nA/μm^2
Metal-silicon work function	ϕ_{ms}	−0.6	−0.1	V
Channel mobility	μ_n, μ_p	390	100	cm^2/V-s
Minimum drawn channel length	L_{drwn}	0.1	0.1	μm
Source, drain junction depth	X_j	0.15	0.16	μm
Source, drain side diffusion	L_d	0.005	0.005	μm
Overlap capacitance per unit gate width	C_{ol}	0.10	0.07	fF/μm
Threshold adjust implant (box dist.)				
impurity type		P	P	
effective depth	X_i	0.1	0.1	μm
effective surface concentration	N_{si}	5×10^{17}	5×10^{17}	Atoms/cm^3
Nominal threshold voltage	V_t	0.27	−0.28	V
Polysilicon gate doping concentration	N_{dpoly}	10^{20}	10^{20}	Atoms/cm^3
Poly gate sheet resistance	R_s	10	10	Ω/□
Source, drain-bulk junction capacitances (zero bias)	C_{j0}	1.0	1.1	fF/μm^2
Source, drain-bulk junction capacitance grading coefficient	n	0.25	0.35	
Source, drain periphery capacitance (zero bias)	C_{jsw0}	0.05	0.06	fF/μm
Source, drain periphery capacitance grading coefficient	n	0.05	0.05	
Source, drain junction built-in potential	ψ_0	0.6	0.65	V
Surface-state density	$\frac{Q_{SS}}{q}$	10^{11}	10^{11}	Atoms/cm^2
Channel-length modulation parameter	$\left\lvert \frac{dX_d}{dV_{DS}} \right\rvert$	0.06	0.05	μm/V

■ **EXAMPLE**

Calculate the zero-bias threshold voltage of the unimplanted and implanted NMOS transistors for the process given in Table 2.1.

Each of the four components in the threshold voltage expression (2.27) must be calculated. The first term is the metal-silicon work function. For an *n*-channel transistor with an *n*-type polysilicon gate electrode, this has a value equal to the difference in the Fermi potentials in the two regions, or approximately −0.6 V.

The second term in the threshold voltage equation represents the band bending in the semiconductor that is required to strongly invert the surface. To produce a surface concentration of electrons that is approximately equal to the bulk concentration of holes, the surface potential

must be increased by approximately twice the bulk Fermi potential. The Fermi potential in the bulk is given by

$$\phi_f = \frac{kT}{q} \ln\left(\frac{N_A}{n_i}\right) \tag{2.28}$$

For the unimplanted transistor with the substrate doping given in Table 2.1, this value is 0.27 V. Thus the second term in (2.27) takes on a value of 0.54 V. The value of this term will be the same for the implanted transistor since we are defining the threshold voltage as the voltage that causes the surface concentration of electrons to be the same as that of holes in the bulk material beneath the channel implant. Thus the potential difference between the surface and the bulk silicon beneath the channel implant region that is required to bring this condition about is still given by (2.30), independent of the details of the channel implant.

The third term in (2.27) is related to the charge in the depletion layer under the channel. We first consider the unimplanted device. Using (1.137), with a value of N_A of 10^{15} atoms/cm^3,

$$\begin{aligned} Q_{b0} &= \sqrt{2\, qN_A \epsilon 2\phi_f} = \sqrt{2\,(1.6 \times 10^{-19})(10^{15})(11.6 \times 8.86 \times 10^{-14})(0.54)} \\ &= 1.34 \times 10^{-8}\ \text{C/cm}^2 \end{aligned} \tag{2.29}$$

Also, the capacitance per unit area of the 400-Å gate oxide is

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{3.9 \times 8.86 \times 10^{-14}\ \text{F/cm}}{400 \times 10^{-8}\ \text{cm}} = 8.6 \times 10^{-8}\ \frac{\text{F}}{\text{cm}^2} = 0.86\ \frac{\text{fF}}{\mu\text{m}^2} \tag{2.30}$$

The resulting magnitude of the third term is 0.16 V.

The fourth term in (2.27) is the threshold shift due to the surface-state charge. This positive charge has a value equal to the charge of one electron multiplied by the density of surface states, 10^{11} atoms/cm^2, from Table 2.1. The value of the surface-state charge term is then

$$\frac{Q_{ss}}{C_{ox}} = \frac{1.6 \times 10^{-19} \times 10^{11}}{8.6 \times 10^{-8}} = 0.19\ \text{V} \tag{2.31}$$

Using these calculations, the threshold voltage for the unimplanted transistor is

$$V_t = -0.6\ \text{V} + 0.54\ \text{V} + 0.16\ \text{V} - 0.19\ \text{V} = -0.09\ \text{V} \tag{2.32}$$

For the implanted transistor, the calculation of the threshold voltage is complicated by the fact that the depletion layer under the channel spans a region of nonuniform doping. A precise calculation of device threshold voltage would require consideration of this nonuniform profile. The threshold voltage can be approximated, however, by considering the implanted layer to be approximated by a box distribution of impurities with a depth X_i and a specified impurity concentration N_i. If the impurity profile resulting from the threshold-adjustment implant and subsequent process steps is sufficiently deep so that the channel-substrate depletion layer lies entirely within it, then the effect of the implant is simply to raise the effective substrate doping. For the implant specified in Table 2.1, the average doping in the layer is the sum of the implant doping and the background concentration, or 2.1×10^{16} atoms/cm^3. This increases the Q_{b0} term in the threshold voltage to 0.71 V and gives device threshold voltage of 0.47 V. The validity of the assumption regarding the boundary of the channel-substrate depletion layer can be checked by using Fig. 2.29. For a doping level of 2.1×10^{16} atoms/cm^3, a one-sided step junction displays a depletion region width of approximately 0.2 μm. Since the depth of the layer is 0.3 μm in this case, the assumption is valid.

Alternatively, if the implantation and subsequent diffusion had resulted in a layer that was very shallow, and was contained entirely within the depletion layer, the effect of the implanted layer would be simply to increase the effective value of Q_{ss} by an amount equal to the effective

implant dose over and above that of the unimplanted transistor. The total active impurity dose for the implant given in Table 2.1 is the product of the depth and the impurity concentration, or 6×10^{11} atoms/cm^2. For this case, the increase in threshold voltage would have been 1.11 V, giving a threshold voltage of 1.02 V. ■

Body-Effect Parameter. For an unimplanted, uniform-channel transistor, the body-effect parameter is given by (1.141).

$$\gamma = \frac{1}{C_{ox}}\sqrt{2q\epsilon N_A} \tag{2.33}$$

The application of this expression is illustrated in the following example.

■ **EXAMPLE**

Calculate the body-effect parameter for the unimplanted n-channel transistor in Table 2.1.

Utilizing in (2.33) the parameters given in Table 2.1, we obtain

$$\gamma = \frac{\sqrt{2\,(1.6 \times 10^{-19})(11.6 \times 8.86 \times 10^{-14})(10^{15})}}{8.6 \times 10^{-8}} = 0.21\ \text{V}^{1/2} \tag{2.34}$$

■

The calculation of body effect in an implanted transistor is complicated by the fact that the channel is not uniformly doped and the preceding simple expression does not apply. The threshold voltage as a function of body bias voltage can be approximated again by considering the implanted layer to be approximated by a box distribution of impurity of depth X_i and concentration N_i. For small values of body bias where the channel-substrate depletion layer resides entirely within the implanted layer, the body effect is that corresponding to a transistor with channel doping $(N_i + N_A)$. For larger values of body bias for which the depletion layer extends into the substrate beyond the implanted distribution, the incremental body effect corresponds to a transistor with substrate doping N_A. A typical variation of threshold voltage as a function of substrate bias for this type of device is illustrated in Fig. 2.61.

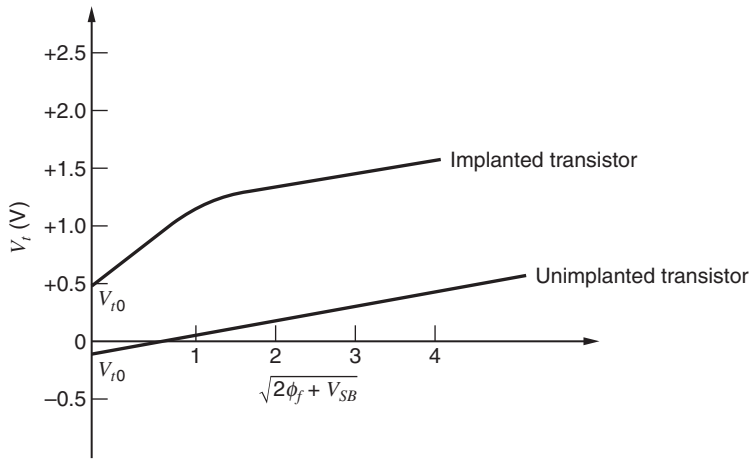


Figure 2.61 Typical variation of threshold voltage as a function of substrate bias for n-channel devices with uniform channel doping (no channel implant) and with nonuniform channel doping resulting from threshold adjustment channel implant.

Effective Channel Length. The gate dimension parallel to current flow that is actually drawn on the mask is called the drawn channel length L_{drwn}. This is the length referred to on circuit schematics. Because of exposure variations and other effects, the physical length of the polysilicon strip defining the gate may be somewhat larger or smaller than this value. The actual channel length of the device is the physical length of the polysilicon gate electrode minus the side or lateral diffusions of the source and the drain under the gate. This length will be termed the *metallurgical channel length* and is the distance between the metallurgical source and drain junctions. Assuming that the lateral diffusion of the source and drain are each equal to L_d, the metallurgical channel length is $L = (L_{\text{drwn}} - 2L_d)$.

When the transistor is biased in the active or saturation region, a depletion region exists between the drain region and the end of the channel. In Chapter 1, the width of this region was defined as X_d. Thus for a transistor operating in the active region, the actual effective channel length L_{eff} is given by

$$L_{\text{eff}} = L_{\text{drwn}} - 2L_d - X_d \tag{2.35}$$

A precise determination of X_d is complicated by the fact that the field distribution in the drain region is two-dimensional and quite complex. The drain depletion width X_d can be approximated by assuming that the electric field in the drain region is one-dimensional and that the depletion width is that of a one-sided step junction with an applied voltage of $V_{DS} - V_{ov}$, where $V_{ov} = V_{GS} - V_t$ is the potential at the drain end of the channel with respect to the source. This assumption is used in the following example.

As shown in Chapter 1, the small-signal output resistance of the transistor is inversely proportional to the effective channel length. Because the performance of analog circuits often depends strongly on the transistor small-signal output resistance, analog circuits often use channel lengths that are longer than the minimum channel length for digital circuits. This statement is particularly true for unimplanted transistors.

■ **EXAMPLE**

Estimate the effective channel length for the unimplanted and implanted transistors for the process shown in Table 2.1 and the device geometry shown in Fig. 2.59. Assume the device is biased at a drain-source voltage of 5 V and a drain current of 10 μA. Calculate the transconductance and the output resistance. For the calculation of X_d, assume that the depletion region between the drain and the end of the channel behaves like a step junction. At the given drain bias voltage, assume that the values of dX_d/dV_{DS} have been deduced from other measurements to be 0.1 μm/V for the unimplanted device and 0.02 μm/V for the implanted device.

The metallurgical channel length is given by

$$L = L_{\text{drwn}} - 2L_d = 6\ \mu\text{m} - (2 \times 0.3\ \mu\text{m}) = 5.4\ \mu\text{m} \tag{2.36}$$

The effective channel length is this length minus the width of the depletion region at the drain X_d. In the active region, the voltage at the drain end of the channel is approximately $(V_{GS} - V_t)$. From (1.166),

$$V_{GS} - V_t = \sqrt{\frac{2I_D}{\mu_n C_{ox} W/L}} = V_{ov} \tag{2.37}$$

If we ignore X_d at first and assume that $L \simeq L_{\text{eff}}$, we obtain a V_{ov} of 0.16 V using the data from Table 2.1. Thus the voltage across the drain depletion region is approximately 4.84 V. To estimate the depletion-region width, assume it is a one-sided step junction that mainly exists in the lightly doped side. Since the channel and the drain are both *n*-type regions,

the built-in potential of the junction is near zero. The width of the depletion layer can be calculated using (1.14) or the nomograph in Fig. 2.29. Using (1.14), and assuming $N_D \gg N_A$,

$$X_d = \sqrt{\frac{2\epsilon\,(V_{DS} - V_{ov})}{qN_A}} \tag{2.38}$$

For the unimplanted device, this equation gives a depletion width of 2.4 μm. For the implanted device, the result is 0.5 μm, assuming an effective constant channel doping of 2.1×10^{16} atoms/cm^3. Thus the effective channel lengths of the two devices would be approximately 3.0 μm and 4.9 μm, respectively.

From (1.180), the device transconductance is given by

$$g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D} \tag{2.39}$$

Assuming that $\mu_n = 700$ cm^2/V-s, we find

$$g_m = \sqrt{2\,(700)(8.6 \times 10^{-8})(50/3.0)(10 \times 10^{-6})} = 141\ \mu\text{A/V} \tag{2.40}$$

for the unimplanted transistor and

$$g_m = \sqrt{2\,(700)(8.6 \times 10^{-8})(50/4.9)(10 \times 10^{-6})} = 111\ \mu\text{A/V} \tag{2.41}$$

for the implanted transistor.

The output resistance can be calculated by using (1.163) and (1.194). For the unimplanted device,

$$r_o = \frac{L_{\text{eff}}}{I_D}\left(\frac{dX_d}{dV_{DS}}\right)^{-1} = \left(\frac{3.0\ \mu\text{m}}{10\ \mu\text{A}}\right)\frac{1}{0.1\ \mu\text{m/V}} = 3.0\ \text{M}\Omega \tag{2.42}$$

For the implanted device,

$$r_o = \left(\frac{4.9\ \mu\text{m}}{10\ \mu\text{A}}\right)\frac{1}{0.02\ \mu\text{m/V}} = 25\ \text{M}\Omega \tag{2.43}$$

■

Because the depletion region for unimplanted devices is much wider than for implanted devices, the channel length of unimplanted devices must be made longer than for implanted devices to achieve comparable punchthrough voltages and small-signal output resistances under identical bias conditions.

Effective Channel Width. The effective channel width of an MOS transistor is determined by the gate dimension parallel to the surface and perpendicular to the channel length over which the gate oxide is thin. Thick field oxide regions are grown at the edges of each transistor by using the local-oxidation process described in Sections 2.2.7 and 2.8. Before the field oxide is grown, nitride is deposited and patterned so that it remains only in areas that should become transistors. Therefore, the width of a nitride region corresponds to the the drawn width of a transistor. To minimize the width variation, the field oxide should grow only vertically; that is, the oxide thickness should increase only in regions where nitride does not cover the oxide. In practice, however, some lateral growth of oxide also occurs near the edges of the nitride during field-oxide growth. As a result, the edges of the field oxide are not vertical, as shown in Figures 2.9 and 2.54. This lateral growth of the oxide reduces the effective width of MOS transistors compared to their drawn widths. It is commonly referred to as the *bird's beak* because the gradually decreasing oxide thickness in the cross sections of Figures 2.9 and 2.54 resembles the corresponding portion of the profile of a bird.

As a result, both the effective lengths and the effective widths of transistors differ from the corresponding drawn dimensions. In analog design, the change in the effective length is usually much more important than the change in the effective width because transistors usually have drawn lengths much less than their drawn widths. As a result, the difference between the drawn and effective width is often ignored. However, this difference is sometimes important, especially when the matching between two ratioed transistors limits the accuracy of a given circuit. This topic is considered in Section 4.2.

Intrinsic Gate-Source Capacitance. As described in Chapter 1, the intrinsic gate-source capacitance of the transistor in the active region of operation is given by

$$C_{gs} = \frac{2}{3} W L_{\text{eff}} C_{ox} \tag{2.44}$$

The calculation of this parameter is illustrated in the next example.

Overlap Capacitance. Assuming that the source and drain regions each diffuse under the gate by L_d after implantation, the gate-source and gate-drain overlap capacitances are given by

$$C_{ol} = W L_d C_{ox} \tag{2.45}$$

This parasitic capacitance adds directly to the intrinsic gate-source capacitance. It constitutes the entire drain-gate capacitance in the active region of operation.

Junction Capacitances. Source-substrate and drain-substrate capacitances result from the junction-depletion capacitance between the source and drain diffusions and the substrate. A complicating factor in calculating these capacitances is the fact that the substrate doping around the source and drain regions is not constant. In the region of the periphery of the source and drain diffusions that border on the field regions, a relatively high surface concentration exists on the field side of the junction because of the field threshold adjustment implant. Although approximate calculations can be carried out, the zero-bias value and grading parameter of the periphery capacitance are often characterized experimentally by using test structures. The bulk-junction capacitance can be calculated directly by using (1.21) or can be read from the nomograph in Fig. 2.29.

An additional capacitance that must be accounted for is the depletion capacitance between the channel and the substrate under the gate, which we will term C_{cs}. Calculation of this capacitance is complicated by the fact that the channel-substrate voltage is not constant but varies along the channel. Also, the allocation of this capacitance to the source and drain varies with operating conditions in the same way as the allocation of C_{gs}. A reasonable approach is to develop an approximate total value for this junction capacitance under the gate and allocate it to source and drain in the same ratio as the gate capacitance is allocated. For example, in the active region, a capacitance of two-thirds of C_{cs} would appear in parallel with the source-substrate capacitance and none would appear in parallel with the drain-substrate capacitance.

■ **EXAMPLE**

Calculate the capacitances of an implanted device with the geometry shown in Fig. 2.59. Use the process parameters given in Table 2.1 and assume a drain-source voltage of 5 V, drain current of 10 μA, and no substrate bias voltage. Neglect the capacitance between the channel and the substrate. Assume that X_d is negligibly small.

From (2.44), the intrinsic gate-source capacitance is

$$C_{gs} = \frac{2}{3} W L_{\text{eff}} C_{ox} = \left(\frac{2}{3}\right) 50\ \mu\text{m} \times 5.4\ \mu\text{m} \times 0.86\ \text{fF}/\mu\text{m}^2 = 155\ \text{fF} \tag{2.46}$$

From (2.45), the overlap capacitance is given by

$$C_{ol} = W L_d C_{ox} = 50\ \mu\text{m} \times 0.3\ \mu\text{m} \times 0.86\ \text{fF}/\mu\text{m}^2 = 12.9\ \text{fF} \tag{2.47}$$

Thus the total gate-source capacitance is $(C_{gs} + C_{ol})$ or 168 fF. The gate-drain capacitance is equal to the overlap capacitance, or 12.9 fF.

The source- and drain-to-substrate capacitances consist of two portions. The periphery or sidewall part C_{jsw} is associated with that portion of the edge of the diffusion area that is adjacent to the field region. The second portion C_j is the depletion capacitance between the diffused junction and the bulk silicon under the source and drain. For the bias conditions given, the source-substrate junction operates with zero bias and the drain-substrate junction has a reverse bias of 5 V. Using Table 2.1, the periphery portion for the source-substrate capacitance is

$$C_{jsw}(\text{source}) = (50\ \mu\text{m} + 9\ \mu\text{m} + 9\ \mu\text{m})(0.5\ \text{fF}/\mu\text{m}) = 34\ \text{fF} \tag{2.48}$$

Here, the perimeter is set equal to $W + 2L$ because that is the distance on the surface of the silicon around the part of the source and drain regions that border on field-oxide regions. Since the substrate doping is high along this perimeter to increase the magnitude of the threshold voltage in the field regions, the sidewall capacitance here is dominant. The bulk capacitance is simply the source-diffusion area multiplied by the capacitance per unit area from Table 2.1.

$$C_j(\text{source}) = (50\ \mu\text{m})(9\ \mu\text{m})(0.08\ \ \text{fF}/\mu\text{m}^2) = 36\ \text{fF} \tag{2.49}$$

The total capacitance from source to bulk is the sum of these two, or

$$C_{sb} = 70\ \text{fF} \tag{2.50}$$

For the geometry given for this example, the transistor is symmetrical, and the source and drain areas and peripheries are the same. From Table 2.1, both the bulk and periphery capacitances have a grading coefficient of 0.5. As a result, the drain-bulk capacitance is the same as the source-bulk capacitance modified to account for the 5 V reverse bias on the junction. Assuming $\psi_0 = 0.65$ V,

$$C_{db} = \frac{(70\ \text{fF})}{\sqrt{1 + V_{DB}/\psi_0}} = \frac{(70\ \text{fF})}{\sqrt{1 + 5/0.65}} = 24\ \text{fF} \tag{2.51}$$

■

As the minimum channel length decreases, second-order effects cause the operation of short-channel MOS transistors to deviate significantly from the simple square-law models in Chapters 1 and 2.[22] Equations that include these second-order effects are complicated and make hand calculations difficult. Therefore, simple models and equations that ignore these effects are often used as a design aid and to develop intuition. SPICE simulations with highly accurate device models are used to verify circuit peformance and to refine a design.

For processes with minimum allowed channel length less than 0.2μm, the gate-oxide thickness can fall below 30 Angstroms (for example, see Table 2.6). With such thin gate oxide, enough carriers in the channel can tunnel through the gate oxide and create nonzero dc gate current that is sometimes important. [23] This current is referred to as gate-leakage current and is a complicated function of the operating point and oxide thickness.[24,25] The gate-leakage current I_G is the product of the gate-leakage-current density J_G and the gate area. SPICE models are available that include gate-leakage current and accurately predict short-channel-device operation.[26,27]

2.9.2 *p*-Channel Transistors

The *p*-channel transistor in most CMOS technologies displays dc and ac properties that are comparable to the *n*-channel transistor. One difference is that the transconductance parameter k' of a *p*-channel device is about one-half to one-third that of an *n*-channel device because holes have correspondingly lower mobility than electrons. As shown in (1.209), this difference in mobility also reduces the f_T of *p*-channel devices by the same factor. Another difference is that for a CMOS technology with a *p*-type substrate and *n*-type wells, the substrate terminal of the *p*-channel transistors can be electrically isolated since the devices are made in an implanted well. Good use can be made of this fact in analog circuits to alleviate the impact of the high body effect in these devices. For a CMOS process made on an *n*-type substrate with *p*-type wells, the *p*-channel devices are made in the substrate material, which is connected to the highest power-supply voltage, but the *n*-channel devices can have electrically isolated substrate terminals.

The calculation of device parameters for *p*-channel devices proceeds exactly as for *n*-channel devices. An important difference is the fact that for the *p*-channel transistors the threshold voltage that results if no threshold adjustment implant is used is relatively high, usually in the range of 1 to 3 V. This occurs because the polarities of the Q_{ss} term and the work-function term are such that they tend to increase the *p*-channel threshold voltages while decreasing the *n*-channel threshold voltages. Thus the *p*-type threshold adjustment implant is used to *reduce* the surface concentration by partially compensating the doping of the *n*-type well or substrate. Thus in contrast to the *n*-channel device, the *p*-channel transistor has an effective surface concentration in the channel that is lower than the bulk concentration, and as a result, often displays a smaller incremental body effect for low values of substrate bias and a larger incremental body effect for larger values of substrate bias.

Nota: enhancement == canal inducido || depletion == canal preformado

2.9.3 Depletion Devices

The properties of depletion devices are similar to those of the enhancement device already considered, except that an implant has been added in the channel to make the threshold negative (for an *n*-channel device). In most respects a depletion device closely resembles an enhancement device with a voltage source in series with the gate lead of value $(V_{tD} - V_{tE})$, where V_{tD} is the threshold voltage of the depletion-mode transistor and V_{tE} is the threshold voltage of the enhancement-mode transistor. Depletion transistors are most frequently used with the gate tied to the source. Because the device is on with $V_{GS} = 0$, if it operates in the active region, it operates like a current source with a drain current of

$$I_{DSS} = I_D|_{V_{GS}=0} = \frac{\mu_n C_{ox}}{2}\frac{W}{L}V_{tD}^2 \tag{2.52}$$

An important aspect of depletion-device performance is the variation of I_{DSS} with process variations. These variations stem primarily from the fact that the threshold voltage varies substantially from its nominal value due to processing variations. Since the transistor I_{DSS} varies as the square of the threshold voltage, large variations in I_{DSS} due to process variations often occur. Tolerances of ± 40 percent or more from nominal due to process variations are common. Because I_{DSS} determines circuit bias current and power dissipation, the magnitude of this variation is an important factor. Another important aspect of the behavior of depletion devices stems from the body effect. Because the threshold voltage varies with body bias, a depletion device with $V_{GS} = 0$ and $v_{sb} \neq 0$ displays a finite conductance in the active region even if the effect of channel-length modulation is ignored. In turn, this finite conductance has a strong effect on the performance of analog circuits that use depletion devices as load elements.

2.9.4 Bipolar Transistors

Standard CMOS technologies include process steps that can be used to form a bipolar transistor whose collector is tied to the substrate. The substrate, in turn, is tied to one of the power supplies. Fig. 2.62*a* shows a cross section of such a device. The well region forms the base of the transistor, and the source/drain diffusion of the device in the well forms the emitter. Since the current flow through the base is perpendicular to the surface of the silicon, the device is a vertical bipolar transistor. It is a *pnp* transistor in processes that utilize *p*-type substrates as in Fig. 2.62*a* and an *npn* transistor in processes that use an *n*-type substrate. The device is particularly useful in band-gap references, described in Chapter 4, and in output stages, considered in Chapter 5. The performance of the device is a strong function of well depth and doping but is generally similar to the substrate *pnp* transistor in bipolar technology, described in Section 2.5.2.

The main limitation of such a vertical bipolar transistor is that its collector is the substrate and is connected to a power supply to keep the substrate *p*-*n* junctions reverse biased. Standard CMOS processes also provide another bipolar transistor for which the collector need not be connected to a power supply.[28] Figure 2.62*b* shows a cross section of such a device. As in the vertical transistor, the well region forms the base and a source/drain diffusion forms the emitter. In this case, however, another source/drain diffusion forms the collector C_1. Since the

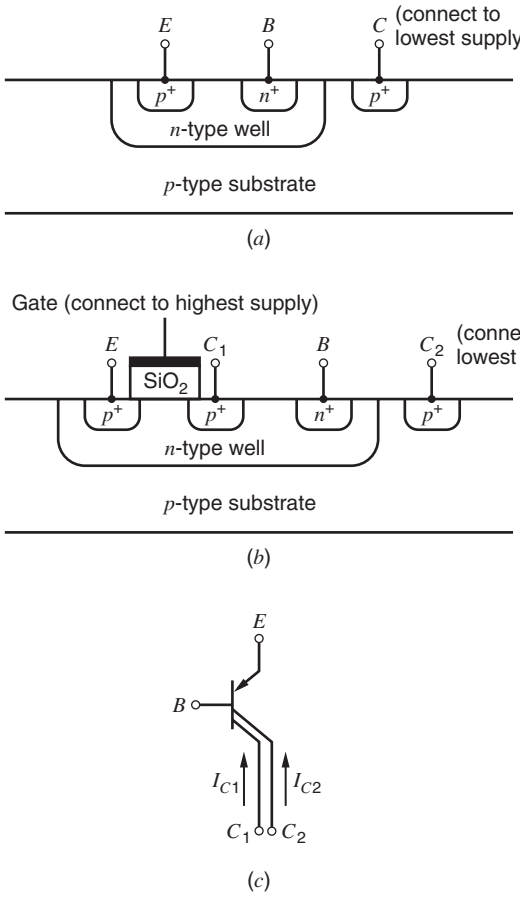


Figure 2.62 (*a*) Cross section of a vertical *pnp* transistor in an *n*-well CMOS process. (*b*) Cross section of lateral and vertical *pnp* transistors in an *n*-well CMOS process. (*c*) Schematic of the bipolar transistors in (*b*).

current flow through the base is parallel to the surface of the silicon, this device is a lateral bipolar transistor. Again, it is a *pnp* transistor in processes that utilize *n*-type wells and an *npn* transistor in processes that use *p*-type wells. The emitter and collector of this lateral device correspond to the source and drain of an MOS transistor. Since the goal here is to build a bipolar transistor, the MOS transistor is deliberately biased to operate in the cutoff region. In Fig. 2.62*b*, for example, the gate of the *p*-channel transistor must be connected to a voltage sufficient to bias it in the cutoff region. A key point here is that the base width of the lateral bipolar device corresponds to the channel length of the MOS device.

One limitation of this structure is that when a lateral bipolar transistor is intentionally formed, a vertical bipolar transistor is also formed. In Fig. 2.62*b*, the emitter and base connections of the vertical transistor are the same as for the lateral transistor, but the collector is the substrate, which is connected to the lowest supply voltage. When the emitter injects minority carriers into the base, some flow parallel to the surface and are collected by the collector of the lateral transistor C_1. However, others flow perpendicular to the surface and are collected by the substrate C_2. Figure 2.62*c* models this behavior by showing a transistor symbol with one emitter and one base but two collectors. The current I_{C1} is the collector current of the lateral transistor, and I_{C2} is the collector current of the vertical transistor. Although the base current is small because little recombination and reverse injection occur, the undesired current I_{C2} is comparable to the desired current I_{C1}. To minimize the ratio, the collector of the lateral transistor usually surrounds the emitter, and the emitter area as well as the lateral base width are minimized. Even with these techniques, however, the ratio of I_{C2}/I_{C1} is poorly controlled in practice.[28,29] If the total emitter current is held constant as in many conventional circuits, variation of I_{C2}/I_{C1} changes the desired collector current and associated small-signal parameters such as the transconductance. To overcome this problem, the emitter current can be adjusted by negative feedback so that the desired collector current is insensitive to variations in I_{C2}/I_{C1}.[30]

Some important properties of the lateral bipolar transistor, including its β_F and f_T, improve as the base width is reduced. Since the base width corresponds to the channel length of an MOS transistor, the steady reduction in the minimum channel length of scaled MOS technologies is improving the performance and increasing the importance of the available lateral bipolar transistor.

2.10 Passive Components in MOS Technology

In this section, we describe the various passive components that are available in CMOS technologies. Resistors include diffused, poly-silicon, and well resistors. Capacitors include poly-poly, metal-poly, metal-silicon, silicon-silicon, and vertical and lateral metal-metal.

2.10.1 Resistors

Diffused Resistors. The diffused layer used to form the source and drain of the *n*-channel and *p*-channel devices can be used to form a diffused resistor. The resulting resistor structure and properties are very similar to the resistors described in Section 2.6.1 on diffused resistors in bipolar technology. The sheet resistances, layout geometries, and parasitic capacitances are similar.

Polysilicon Resistors. At least one layer of polysilicon is required in silicon-gate MOS technologies to form the gates of the transistors, and this layer is often used to form resistors. The geometries employed are similar to those used for diffused resistors, and the resistor exhibits a

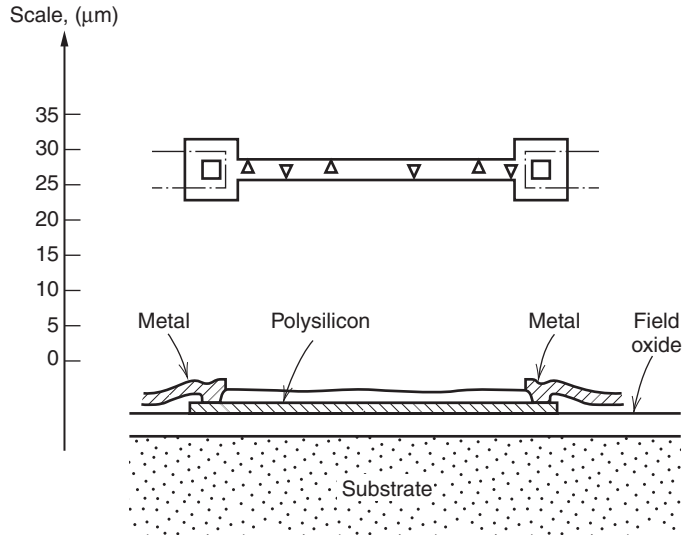


Figure 2.63 (*a*) Plan view and cross section of polysilicon resistor.

parasitic capacitance to the underlying layer much like a diffused resistor. In this case, however, the capacitance stems from the oxide layer under the polysilicon instead of from a reverse-biased *pn* junction. The nominal sheet resistance of most polysilicon layers that are utilized in MOS processes is on the order of 20 Ω/□ to 80 Ω/□ and typically displays a relatively large variation around the nominal value due to process variations. The matching properties of polysilicon resistors are similar to those of diffused resistors. A cross section and plan view of a typical polysilicon resistor are shown in Fig. 2.63*a*.

The sheet resistance of polysilicon can limit the speed of interconnections, especially in submicron technologies. To reduce the sheet resistance, a silicide layer is sometimes deposited on top of the polysilicon. Silicide is a compound of silicon and a metal, such as tungsten, that can withstand subsequent high-temperature processing with little movement. Silicide reduces the sheet resistance by about an order of magnitude. Also, it has little effect on the oxidation rate of polysilicon and is therefore compatible with conventional CMOS process technologies.[31] Finally, silicide can be used on the source/drain diffusions as well as on the polysilicon.

Well Resistors. In CMOS technologies the well region can be used as the body of a resistor. It is a relatively lightly doped region and when used as a resistor provides a sheet resistance on the order of 10 kΩ/□. Its properties and geometrical layout are much like the epitaxial resistor described in Section 2.6.2 and shown in Fig. 2.42. It displays large tolerance, high voltage coefficient, and high temperature coefficient relative to other types of resistors. Higher sheet resistance can be achieved by the addition of the pinching diffusion just as in the bipolar technology case.

MOS Devices as Resistors. The MOS transistor biased in the triode region can be used in many circuits to perform the function of a resistor. The drain-source resistance can be calculated by differentiating the equation for the drain current in the triode region with respect to the drain-source voltage. From (1.152),

$$R = \left(\frac{\partial I_D}{\partial V_{DS}}\right)^{-1} = \frac{L}{W}\frac{1}{k'(V_{GS} - V_t - V_{DS})} \tag{2.53}$$

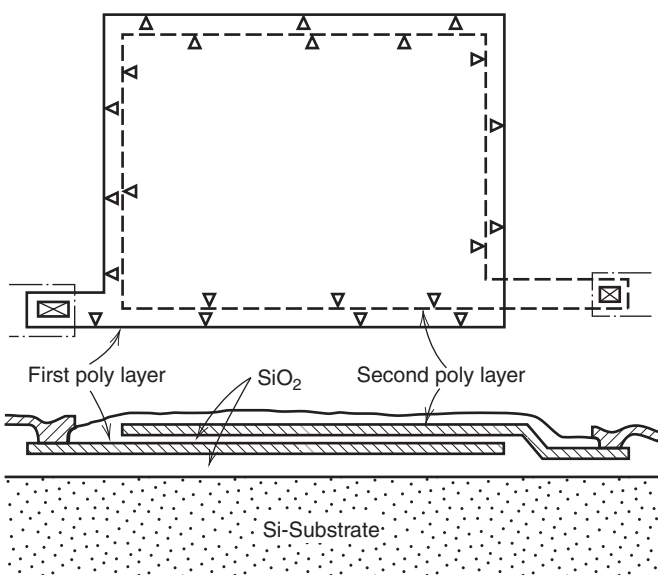
Since L/W gives the number of squares, the second term on the right side of this equation gives the sheet resistance. This equation shows that the effective sheet resistance is a function of the applied gate bias. In practice, this sheet resistance can be much higher than polysilicon or diffused resistors, allowing large amounts of resistance to be implemented in a small area. Also, the resistance can be made to track the transconductance of an MOS transistor operating in the active region, allowing circuits to be designed with properties insensitive to variations in process, supply, and temperature. An example of such a circuit is considered in Section 9.4.3. The principal drawback of this form of resistor is the high degree of nonlinearity of the resulting resistor element; that is, the drain-source resistance is not constant but depends on the drain-source voltage. Nevertheless, it can be used very effectively in many applications.

2.10.2 Capacitors in MOS Technology

As a passive component, capacitors play a much more important role in MOS technology than they do in bipolar technology. Because of the fact that MOS transistors have virtually infinite input resistance, voltages stored on capacitors can be sensed with little leakage using MOS amplifiers. As a result, capacitors can be used to perform many functions that are traditionally performed by resistors in bipolar technology.

Poly-Poly Capacitors. Many MOS technologies that are used to implement analog functions have two layers of polysilicon. The additional layer provides an efficient capacitor structure, an extra layer of interconnect, and can also be used to implement floating-gate memory cells that are electrically programmable and optically erasable with UV light (EPROM). A typical poly-poly capacitor structure is shown in cross section and plan view in Fig. 2.63*b*. The plate separation is usually comparable to the gate oxide thickness of the MOS transistors.

An important aspect of the capacitor structure is the parasitic capacitance associated with each plate. The largest parasitic capacitance exists from the bottom plate to the underlying layer, which could be either the substrate or a well diffusion whose terminal is electrically isolated. This bottom-plate parasitic capacitance is proportional to the bottom-plate area and typically has a value from 10 to 30 percent of the capacitor itself.



(*b*)

Figure 2.63 (*b*) Plan view and cross section of typical poly-poly capacitor.

The top-plate parasitic is contributed by the interconnect metallization or polysilicon that connects the top plate to the rest of the circuit, plus the parasitic capacitance of the transistor to which it is connected. In the structure shown in Fig. 2.63*b*, the drain-substrate capacitance of an associated MOS transistor contributes to the top-plate parasitic capacitance. The minimum value of this parasitic is technology dependent but is typically on the order of 5 fF to 50 fF.

Other important parameters of monolithic capacitor structures are the tolerance, voltage coefficient, and temperature coefficient of the capacitance value. The tolerance on the absolute value of the capacitor value is primarily a function of oxide-thickness variations and is usually in the 10 percent to 30 percent range. Within the same die, however, the matching of one capacitor to another identical structure is much more precise and can typically be in the range of 0.05 percent to 1 percent, depending on the geometry. Because the plates of the capacitor are a heavily doped semiconductor rather than an ideal conductor, some variation in surface potential relative to the bulk material of the plate occurs as voltage is applied to the capacitor.[32] This effect is analogous to the variation in surface potential that occurs in an MOS transistor when a voltage is applied to the gate. However, since the impurity concentration in the plate is usually relatively high, the variations in surface potential are small. The result of these surface potential variations is a slight variation in capacitance with applied voltage. Increasing the doping in the capacitor plates reduces the voltage coefficient. For the impurity concentrations that are typically used in polysilicon layers, the voltage coefficient is usually less than 50 ppm/V, [32,33] a level small enough to be neglected in most applications.

A variation in the capacitance value also occurs with temperature variations. This variation stems primarily from the temperature variation of the surface potential in the plates previously described.[32] Also, secondary effects include the temperature variation of the dielectric constant and the expansion and contraction of the dielectric. For heavily doped polysilicon plates, this temperature variation is usually less than 50 ppm/°C.[32,33]

MOS Transistors as Capacitors. The MOS transistor itself can be used as a capacitor when biased in the triode region, the gate forming one plate and the source, drain, and channel forming another. Unfortunately, because the underlying substrate is lightly doped, a large amount of surface potential variation occurs with changes in applied voltage and the capacitor displays a high voltage coefficient. In noncritical applications, however, it can be used effectively under two conditions. The circuit must be designed in such a way that the device is biased in the triode region when a high capacitance value is desired, and the high sheet resistance of the bottom plate formed by the channel must be taken into account.

Other Vertical Capacitor Structures. In processes with only one layer of polysilicon, alternative structures must be used to implement capacitive elements. One approach involves the insertion of an extra mask to reduce the thickness of the oxide on top of the polysilicon layer so that when the interconnect metallization is applied, a thin-oxide layer exists between the metal layer and the polysilicon layer in selected areas. Such a capacitor has properties that are similar to poly-poly capacitors.

Another capacitor implementation in single-layer polysilicon processing involves the insertion of an extra masking and diffusion operation such that a diffused layer with low sheet resistance can be formed underneath the polysilicon layer in a thin-oxide area. This is not possible in conventional silicon-gate processes because the polysilicon layer is deposited before the source-drain implants or diffusions are performed. The properties of such capacitors are similar to the poly-poly structure, except that the bottom-plate parasitic capacitance

is that of a *pn* junction, which is voltage dependent and is usually larger than in the poly-poly case. Also, the bottom plate has a junction leakage current that is associated with it, which is important in some applications.

To avoid the need for extra processing steps, capacitors can also be constructed using the metal and poly layers with standard oxide thicknesses between layers. For example, in a process with one layer of polysilicon and two layers of metal, the top metal and the poly can be connected together to form one plate of a capacitor, and the bottom metal can be used to form the other plate. A key disadvantage of such structures, however, is that the capacitance per unit area is small because the oxide used to isolate one layer from another is thick. Therefore, such capacitors usually occupy large areas. Furthermore, the thickness of this oxide changes little as CMOS processes evolve with reduced minimum channel length. As a result, the area required by analog circuits using such capacitors undergoes a much smaller reduction than that of digital circuits in new technologies. This characteristic is important because reducing the area of an integrated circuit reduces its cost.

Lateral Capacitor Structures. To reduce the capacitor area, and to avoid the need for extra processing steps, lateral capacitors can be used.[34] A lateral capacitor can be formed in one layer of metal by separating one plate from another by spacing s, as shown in Fig. 2.64*a*. If w is the width of the metal and t is the metal thickness, the capacitance is $(wt\epsilon/s)$, where ϵ is the dielectric constant. As technologies evolve to reduced feature sizes, the minimum metal spacing shrinks but the thickness changes little; therefore, the die area required for a given lateral capacitance decreases in scaled technologies.[35] Note that the lateral capacitance is proportional to the perimeter of each plate that is adjacent to the other in a horizontal plane. Geometries to increase this perimeter in a given die area have been proposed.[35]

Lateral capacitors can be used in conjunction with vertical capacitors, as shown in Fig. 2.64*b*.[34] The key point here is that each metal layer is composed of multiple pieces, and each capacitor node is connected in an alternating manner to the pieces in each layer.

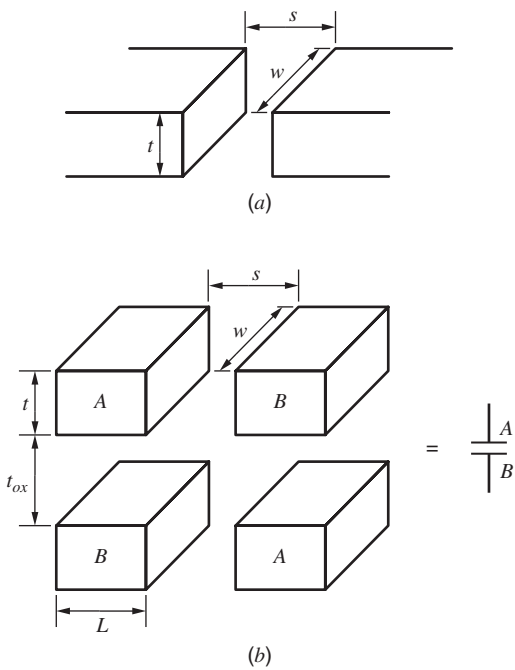


Figure 2.64 (*a*) Lateral capacitor in one level of metal. (*b*) Capacitor using two levels of metal in which both lateral and vertical capacitance contribute to the desired capacitance.

As a result, the total capacitance includes vertical and lateral components arising between all adjacent pieces. If the vertical and lateral dielectric constants are equal, the total capacitance is increased compared to the case in which the same die area is used to construct only a vertical capacitor when the minimum spacing $s < \sqrt{2t\ (t_{ox})}$, where t is the metal thickness and t_{ox} is the oxide thickness between metal layers. This concept can be extended to additional pieces in each layer and additional layers.

2.10.3 Latchup in CMOS Technology

The device structures that are present in standard CMOS technology inherently comprise a *pnpn* sandwich of layers. For example, consider the typical circuit shown in Fig. 2.65*a*. It uses one *n*-channel and one *p*-channel transistor and operates as an inverter if the two gates are connected together as the inverter input. Figure 2.65*b* shows the cross section in an *n*-well process. When the two MOS transistors are fabricated, two parasitic bipolar transistors are also formed: a lateral *npn* and a vertical *pnp*. In this example, the source of the *n*-channel transistor forms the emitter of the parasitic lateral *npn* transistor, the substrate forms the base, and the *n*-well forms the collector. The source of the *p*-channel transistor forms the emitter of a parasitic vertical *pnp* transistor, the *n*-well forms the base, and the *p*-type substrate forms the collector. The electrical connection of these bipolar transistors that results from the layout shown is illustrated in Fig. 2.65*c*. In normal operation, all the *pn* junctions in the structure are reverse biased. If the two bipolar transistors enter the active region for some reason, however, the circuit can display a large amount of positive feedback, causing both transistors to conduct heavily. This device structure is similar to that of a silicon-controlled rectifier (SCR), a widely used component in

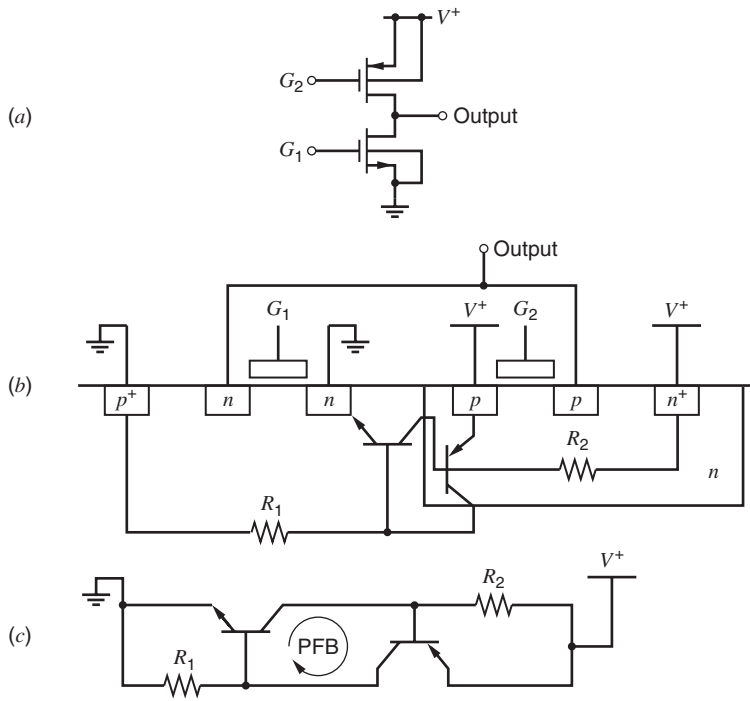


Figure 2.65 (*a*) Schematic of a typical CMOS device pair. (*b*) Cross section illustrating the parasitic bipolar transistors. (*c*) Schematic of the parasitic bipolar transistors.

power-control applications. In power-control applications, the property of the *pnpn* sandwich to remain in the *on* state with no externally supplied signal is a great advantage. However, the result of this behavior here is usually a destructive breakdown phenomenon called *latchup.*

The positive feedback loop is labeled in Fig. 2.65*c*. Feedback is studied in detail in Chapters 8 and 9. To explain why the feedback around this loop is positive, assume that both transistors are active and that the base current of the *npn* transistor increases by i for some reason. Then the collector current of the *npn* transistor increases by $\beta_{npn}i$. This current is pulled out of the base of the *pnp* transistor if R_2 is ignored. As a result, the current flowing out of the collector of the *pnp* transistor increases by $\beta_{npn}\beta_{pnp}i$. Finally, this current flows into the base of the *npn* transistor if R_1 is ignored. This analysis shows that the circuit generates a current that flows in the same direction as the initial disturbance; therefore, the feedback is positive. If the gain around the loop is more than unity, the response of the circuit to the initial disturbance continues to grow until one or both of the bipolar transistors saturate. In this case, a large current flows from the positive supply to ground until the power supply is turned off or the circuit burns out. This condition is called *latchup.* If R_1 and R_2 are large enough that base currents are large compared to the currents in these resistors, the gain around the loop is $\beta_{npn}\beta_{pnp}$. Therefore, latchup can occur if the product of the betas is greater than unity.

For latchup to occur, one of the junctions in the sandwich must become forward biased. In the configuration illustrated in Fig. 2.65, current must flow in one of the resistors between the emitter and the base of one of the two transistors in order for this to occur. This current can come from a variety of causes. Examples are an application of a voltage that is larger than the power-supply voltage to an input or output terminal, improper sequencing of the power supplies, the presence of large dc currents in the substrate or *p*- or *n*-well, or the flow of displacement current in the substrate or well due to fast-changing internal nodes. Latchup is more likely to occur in circuits as the substrate and well concentration is made lighter, as the well is made thinner, and as the device geometries are made smaller. All these trends in process technology tend to increase R_1 and R_2 in Fig. 2.65*b*. Also, they tend to increase the betas of the two bipolar transistors. These changes increase the likelihood of the occurrence of latchup.

The layout of CMOS-integrated circuits must be carried out with careful attention paid to the prevention of latchup. Although the exact rules followed depend on the specifics of the technology, the usual steps are to keep R_1 and R_2, as well as the product of the betas, small enough to avoid this problem. The beta of the vertical bipolar transistor is determined by process characteristics, such as the well depth, that are outside the control of circuit designers. However, the beta of the lateral bipolar transistor can be decreased by increasing its base width, which is the distance between the source of the *n*-channel transistor and the *n*-type well. To reduce R_1 and R_2, many substrate and well contacts are usually used instead of just one each, as shown in the simple example of Fig. 2.65. In particular, *guard rings* of substrate and well contacts are often used just outside and inside the well regions. These rings are formed by using the source/drain diffusion and provide low-resistance connections in the substrate and well to reduce series resistance. Also, special protection structures at each input and output pad are usually included so that excessive currents flowing into or out of the chip are safely shunted.

2.11 BiCMOS Technology

In Section 2.3, we showed that to achieve a high collector-base breakdown voltage in a bipolar transistor structure, a thick epitaxial layer is used (17 μm of 5 Ω-cm material for 36-V operation). This in turn requires a deep *p*-type diffusion to isolate transistors and other devices.

On the other hand, if a low breakdown voltage (say about 7 V to allow 5-V supply operation) can be tolerated, then a much more heavily doped (on the order of 0.5 Ω-cm) collector region can be used that is also much thinner (on the order of 1 μm). Under these conditions, the bipolar devices can be isolated by using the same local-oxidation technique used for CMOS, as described in Section 2.4. This approach has the advantage of greatly reducing the bipolar transistor collector-substrate parasitic capacitance because the heavily doped high-capacitance regions near the surface are now replaced by low-capacitance oxide isolation. The devices can also be packed much more densely on the chip. In addition, CMOS and bipolar fabrication technologies begin to look rather similar, and the combination of high-speed, shallow, ion-implanted bipolar transistors with CMOS devices in a BiCMOS technology becomes feasible (at the expense of several extra processing steps).[36] This technology has performance advantages in digital applications because the high current-drive capability of the bipolar transistors greatly facilitates driving large capacitive loads. Such processes are also attractive for analog applications because they allow the designer to take advantage of the unique characteristics of both types of devices.

We now describe the structure of a typical high-frequency, low-voltage, oxide-isolated BiCMOS process. A simplified cross section of a high-performance process[37] is shown in Fig. 2.66. The process begins with masking steps and the implantation of *n*-type antimony buried layers into a *p*-type substrate wherever an *npn* bipolar transistor or PMOS device is to be formed. A second implant of *p*-type boron impurities forms a *p*-well wherever an NMOS device is to be formed. This is followed by the growth of about 1 μm of n^- epi, which forms the collectors of the *npn* bipolar devices and the channel regions of the PMOS devices. During this and subsequent heat cycles, the more mobile boron atoms out-diffuse and the *p*-well extends to the surface, whereas the antimony buried layers remain essentially fixed.

A masking step defines regions where thick field oxide is to be grown and these regions are etched down into the epi layer. Field-oxide growth is then carried out, followed by a planarization step where the field oxide that has grown above the plane of the surface is etched back level with the other regions. This eliminates the lumpy surface shown in Fig. 2.57 and helps to overcome problems of ensuring reliable metal connections over the oxide steps (so-called *step coverage*). Finally, a series of masking steps and *p*- and *n*-type implants are carried out to form bipolar base and emitter regions, low-resistance bipolar collector contact, and source and drain regions for the MOSFETs. In this sequence, gate oxide is grown, polysilicon gates and emitters are formed, and threshold-adjusting implants are made for the MOS devices. Metal contacts are then made to the desired regions, and the chip is coated with a layer of deposited SiO_2. A second layer of metal interconnects is formed on top of this oxide with connections where necessary to the first layer of metal below. A further deposited layer of SiO_2 is then added with a third layer of metal interconnect and vias to give even more connection flexibility and thus to improve the density of the layout.

2.12 Heterojunction Bipolar Transistors

A *heterojunction* is a *pn* junction made of two different materials. Until this point, all the junctions we have considered have been *homojunctions* because the same material (silicon) has been used to form both the *n*-type and the *p*-type regions. In contrast, a junction between an *n*-type region of silicon and a *p*-type region of germanium or a compound of silicon and germanium forms a heterojunction.

In homojunction bipolar transistors, the emitter doping is selected to be much greater than the base doping to give an emitter injection efficiency γ of about unity, as shown by

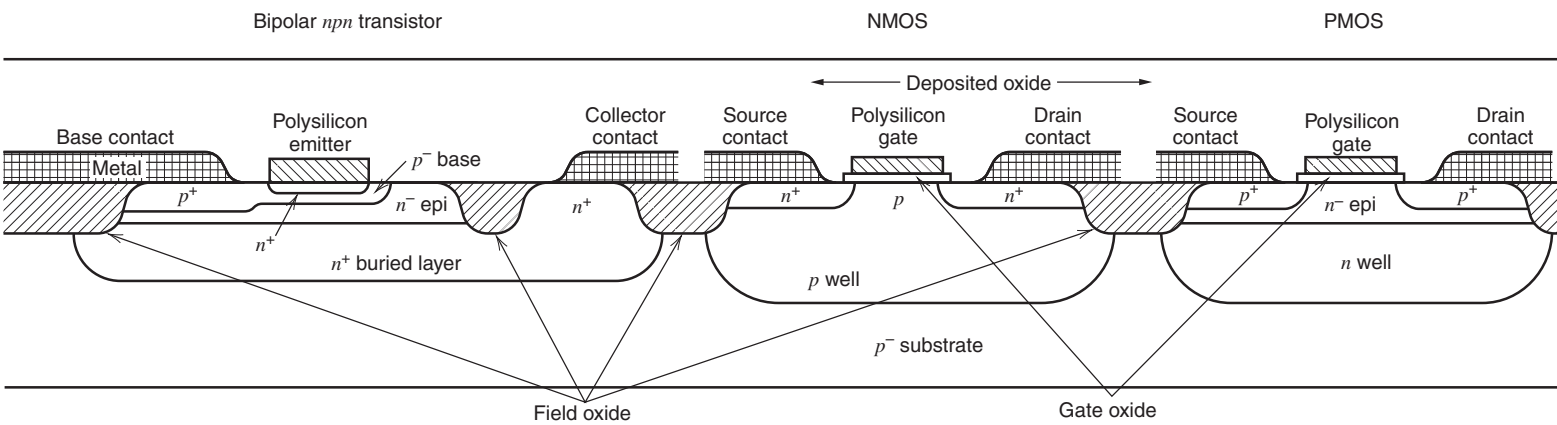


Figure 2.66 Cross section of a high-performance BiCMOS process.

(1.51b). As a result, the base is relatively lightly doped while the emitter is heavily doped in practice. Section 1.4.8 shows that the f_T of bipolar devices is limited in part by τ_F, which is the time required for minority carriers to cross the base. Maximizing f_T is important in some applications such as radio-frequency electronics. To increase f_T, the base width can be reduced. If the base doping is fixed to maintain a constant γ, however, this approach increases the base resistance r_b. In turn, this base resistance limits speed because it forms a time constant with capacitance attached to the base node. As a result, a tradeoff exists in standard bipolar technology between high f_T on the one hand and low r_b on the other, and both extremes limit the speed that can be attained in practice.

One way to overcome this tradeoff is to add some germanium to the base of bipolar transistors to form heterojunction transistors. The key idea is that the different materials on the two sides of the junction have different band gaps. In particular, the band gap of silicon is greater than for germanium, and forming a SiGe compound in the base reduces the band gap there. The relatively large band gap in the emitter can be used to increase the potential barrier to holes that can be injected from the base back to the emitter. Therefore, this structure does not require that the emitter doping be much greater than the base doping to give $\gamma \simeq 1$. As a result, the emitter doping can be decreased and the base doping can be increased in a heterojunction bipolar transistor compared to its homojunction counterpart. Increasing the base doping allows r_b to be constant even when the base width is reduced to increase f_T. Furthermore, this change also reduces the width of the base-collector depletion region in the base when the transistor operates in the forward active region, thus decreasing the effect of base-width modulation and increasing the early voltage V_A. Not only does increasing the base doping have a beneficial effect on performance, but also decreasing the emitter doping increases the width of the base-emitter, space-charge region in the emitter, reducing the C_{je} capacitance and further increasing the maximum speed.

The base region of the heterojunction bipolar transistors can be formed by growing a thin epitaxial layer of SiGe using ultra-high vacuum chemical vapor deposition (UHV/CVD).[38] Since this is an epi layer, it takes on the crystal structure of the silicon in the substrate. Because the lattice constant for germanium is greater than that for silicon, the SiGe layer forms under a compressive strain, limiting the concentration of germanium and the thickness of the layer to avoid defect formation after subsequent high-temperature processing used at the back end of conventional technologies.[39] In practice, with a base thickness of 0.1 μm, the concentration of germanium is limited to about 15 percent so that the layer is unconditionally stable.[40] With only a small concentration of germanium, the change in the band gap and the resulting shift in the potential barrier that limits reverse injection of holes into the emitter is small. However, the reverse injection is an exponential function of this barrier; therefore, even a small change in the barrier greatly reduces the reverse injection and results in these benefits.

In practice, the concentration of germanium in the base need not be constant. In particular, the UHV/CVD process is capable of increasing the concentration of germanium in the base from the emitter end to the collector end. This grading of the germanium concentration results in an electric field that helps electrons move across the base, further reducing τ_F and increasing f_T.

The heterojunction bipolar transistors described above can be included as the bipolar transistors in otherwise conventional BiCMOS processes. The key point is that the device processing sequence retains the well-established properties of silicon integrated-circuit processing because the average concentration of germanium in the base is small.[39] This characteristic is important because it allows the new processing steps to be included as a simple addition to an existing process, reducing the cost of the new technology. For example, a BiCMOS process with a minimum drawn CMOS channel length of 0.3 μm and heterojunction bipolar transistors

with a f_T of 50 GHz has been reported.[40] The use of the heterojunction technology increases the f_T by about a factor of two compared to a comparable homojunction technology.

2.13 Interconnect Delay

As the minimum feature size allowed in integrated-circuit technologies is reduced, the maximum operating speed and bandwidth have steadily increased. This trend stems partly from the reductions in the minimum base width of bipolar transistors and the minimum channel length of MOS transistors, which in turn increase the f_T of these devices. While scaling has increased the speed of the transistors, however, it is also increasing the delay introduced by the interconnections to the point where it could soon limit the maximum speed of integrated circuits.[41] This delay is increasing as the minimum feature size is reduced because the width of metal lines and spacing between them are both being reduced to increase the allowed density of interconnections. Decreasing the width of the lines increases the number of squares for a fixed length, increasing the resistance. Decreasing the spacing between the lines increases the lateral capacitance between lines. The delay is proportional to the product of the resistance and capacitance. To reduce the delay, alternative materials are being studied for use in integrated circuits.

First, copper is replacing aluminum in metal layers because copper reduces the resistivity of the interconnection by about 40 percent and is less susceptible to electromigration and stress migration than aluminum. Electromigration and stress migration are processes in which the material of a conductor moves slightly while it conducts current and is under tension, respectively. These processes can cause open circuits to appear in metal interconnects and are important failure mechanisms in integrated circuits. Unfortunately, however, copper can not simply be substituted for aluminum with the same fabrication process. Two key problems are that copper diffuses through silicon and silicon dioxide more quickly than aluminum, and copper is difficult to plasma etch.[42] To overcome the diffusion problem, copper must be surrounded by a thin film of another metal that can endure high-temperature processing with little movement. To overcome the etch problem, a damascene process has been developed.[43] In this process, a layer of interconnection is formed by first depositing a layer of oxide. Then the interconnect pattern is etched into the oxide, and the wafer is uniformly coated by a thin diffusion-resistant layer and then copper. The wafer is then polished by a chemical-mechanical process until the surface of the oxide is reached, which leaves the copper in the cavities etched into the oxide. A key advantage of this process is that it results in a planar structure after each level of metalization.

Also, low-permittivity dielectrics are being studied to replace silicon dioxide to reduce the interconnect capacitance. The dielectric constant of silicon dioxide is 3.9 times more than for air. For relative dielectric constants between about 2.5 and 3.0, polymers have been studied. For relative dielectric constants below about 2.0, the proposed materials include foams and gels, which include air.[42] Other important requirements of low-permittivity dielectric materials include low leakage, high breakdown voltage, high thermal conductivity, stability under high-temperature processing, and adhesion to the metal layers.[41] The search for a replacement for silicon dioxide is difficult because it is an excellent dielectric in all these ways.

2.14 Economics of Integrated-Circuit Fabrication

The principal reason for the growing pervasiveness of integrated circuits in systems of all types is the reduction in cost attainable through integrated-circuit fabrication. Proper utilization of the technology to achieve this cost reduction requires an understanding of the factors influencing

the cost of an integrated circuit in completed, packaged form. In this section, we consider these factors.

2.14.1 Yield Considerations in Integrated-Circuit Fabrication

As pointed out earlier in this chapter, integrated circuits are batch-fabricated on single wafers, each containing up to several thousand separate but identical circuits. At the end of the processing sequence, the individual circuits on the wafer are probed and tested prior to the breaking up of the wafer into individual dice. The percentage of the circuits that are electrically functional and within specifications at this point is termed the wafer-sort yield Y_{ws} and is usually in the range of 10 percent to 90 percent. The nonfunctional units can result from a number of factors, but one major source of yield loss is point defects of various kinds that occur during the photoresist and diffusion operations. These defects can result from mask defects, pinholes in the photoresist, airborne particles that fall on the surface of the wafer, crystalline defects in the epitaxial layer, and so on. If such a defect occurs in the active region on one of the transistors or resistors making up the circuit, a nonfunctional unit usually results. The frequency of occurrence per unit of wafer area of such defects is usually dependent primarily on the particular fabrication process used and not on the particular circuit being fabricated. Generally speaking, the more mask steps and diffusion operations that the wafer is subjected to, the higher will be the density of defects on the surface of the finished wafer.

The existence of these defects limits the size of the circuit that can be economically fabricated on a single die. Consider the two cases illustrated in Fig. 2.67, where two identical wafers with the same defect locations have been used to fabricate circuits of different area. Although the defect locations in both cases are the same, the wafer-sort yield of the large die would be zero. When the die size is cut to one-fourth of the original size, the wafer sort yield is 62 percent. This conceptual example illustrates the effect of die size on wafer-sort yield. Quantitatively, the expected yield for a given die size is a strong function of the complexity of the process, the nature of the individual steps in the process, and perhaps, most importantly, the maturity and degree of development of the process as a whole and the individual steps within it. Since the inception of the planar process, a steady reduction in defect densities has occurred as a result of improved lithography, increased use of low-temperature processing steps such as ion implantation, improved manufacturing environmental control, and so forth. Three typical curves derived from yield data on bipolar and MOS processes are shown in Fig. 2.68. These are representative of yields for processes ranging from a very complex process with many yield-reducing steps to a very simple process carried out in an advanced VLSI fabrication facility. Also, the yield curves can be raised or lowered by more conservative design rules, and other factors. Uncontrolled factors such as testing problems and design problems in the circuit can cause results for a particular integrated circuit to deviate widely from these curves, but still the overall trend is useful.

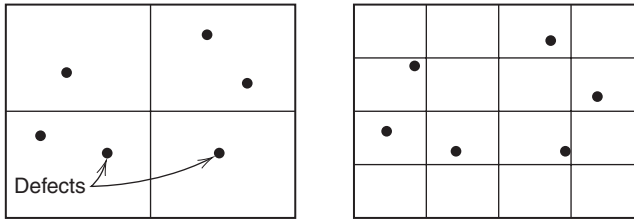


Figure 2.67 Conceptual example of the effect of die size on yield.

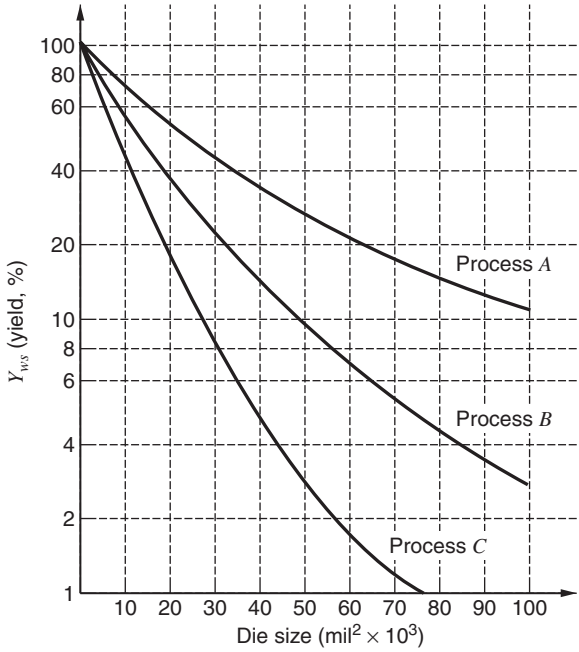


Figure 2.68 Typically observed yield versus die size for the three different processes, ranging from a very simple, well-developed process (curve *A*), to a very complex process with many yield-reducing steps (curve *C*).

In addition to affecting yield, the die size also affects the total number of dice that can be fabricated on a wafer of a given size. The total number of usable dice on the wafer, called the gross die per wafer *N*, is plotted in Fig. 2.69 as a function of die size for several wafer sizes. The product of the gross die per wafer and the wafer-sort yield gives the net good die per wafer, plotted in Fig. 2.70 for the yield curve of Fig. 2.68, assuming a 4-inch wafer.

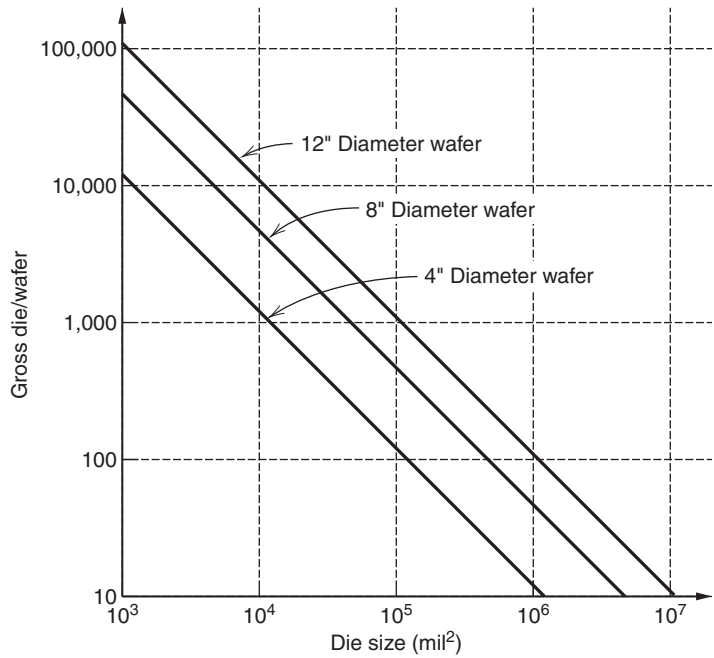


Figure 2.69 Gross die per wafer for 4-in., 8-in., and 12-in. wafers.

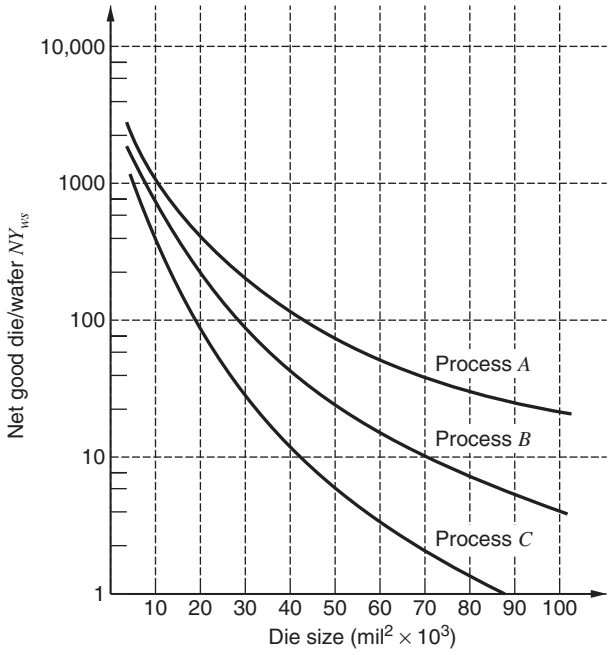


Figure 2.70 Net good die per wafer for the three processes in Fig. 2.66, assuming a 4-in. wafer. The same curve can be obtained approximately for other wafer sizes by simply scaling the vertical axis by a factor equal to the wafer area.

Once the wafer has undergone the wafer-probe test, it is separated into individual dice by sawing or scribing and breaking. The dice are visually inspected, sorted, and readied for assembly into packages. This step is termed *die fab,* and some loss of good dice occurs in the process. Of the original electrically good dice on the wafer, some will be lost in the die fab process due to breakage and scratching of the surface. The ratio of the electrically good dice following die fab to the number of electrically good dice on the wafer before die fab is called the *die fab yield* Y_{df}. The good dice are then inserted in a package, and the electrical connections to each die are made with bonding wires to the pins on the package. The packaged circuits then undergo a final test, and some loss of functional units usually occurs because of improper bonding and handling losses. The ratio of the number of good units at final test to the number of good dice into assembly is called the *final test yield* Y_{ft}.

2.14.2 Cost Considerations in Integrated-Circuit Fabrication

The principal direct costs to the manufacturer can be divided into two categories: those associated with fabricating and testing the wafer, called the *wafer fab cost* C_w, and those associated with packaging and final testing the individual dice, called the *packaging cost* C_p. If we consider the costs incurred by the complete fabrication of one wafer of dice, we first have the wafer cost itself C_w. The number of electrically good dice that are packaged from the wafer is $NY_{ws}Y_{df}$. The total cost C_t incurred once these units have been packaged and tested is

$$C_t = C_w + C_p NY_{ws}Y_{df} \tag{2.54}$$

The total number of good finished units N_g is

$$N_g = NY_{ws}Y_{df}Y_{ft} \tag{2.55}$$

Thus the cost per unit is

$$C = \frac{C_t}{N_g} = \frac{C_w}{NY_{ws}Y_{df}Y_{ft}} + \frac{C_p}{Y_{ft}} \tag{2.56}$$

The first term in the cost expression is wafer fab cost, while the second is associated with assembly and final testing. This expression can be used to calculate the direct cost of the finished product to the manufacturer as shown in the following example.

■ **EXAMPLE**

Plot the direct fabrication cost as a function of die size for the following two sets of assumptions.

(a) Wafer-fab cost of \$75.00, packaging and testing costs per die of \$0.06, a die-fab yield of 0.9, and a final-test yield of 0.9. Assume yield curve *B* in Fig. 2.68. This set of conditions might characterize an operational amplifier manufactured on a medium-complexity bipolar process and packaged in an inexpensive 8 or 14 lead package.

From (2.56),

$$C = \frac{\$75.00}{(NY_{ws})(0.81)} + \frac{0.06}{0.9} = \frac{\$92.59}{NY_{ws}} + 0.066 \tag{2.57}$$

This cost is plotted versus die size in Fig. 2.71*a*.

(b) A wafer-fab cost of \$100.00, packaging and testing costs of \$1.00, die-fab yield of 0.9, and final-test yield of 0.8. Assume yield curve *A* in Fig. 2.68. This might characterize a complex analog/digital integrated circuit, utilizing an advanced CMOS process and packaged in a large, multilead package. Again, from (2.56),

$$C = \frac{\$100.00}{(NY_{ws})(0.72)} + \frac{\$1.00}{0.8} = \frac{\$138.89}{NY_{ws}} + \$1.25 \tag{2.58}$$

■ This cost is plotted versus die size in Fig. 2.71*b*.

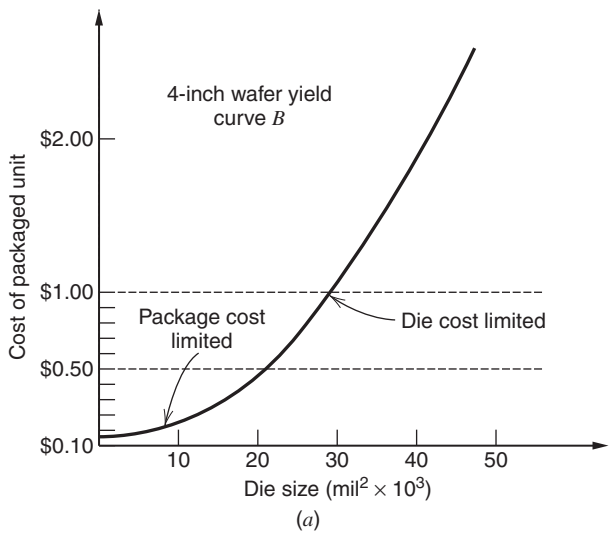


Figure 2.71 (*a*) Cost curve for example *a*.

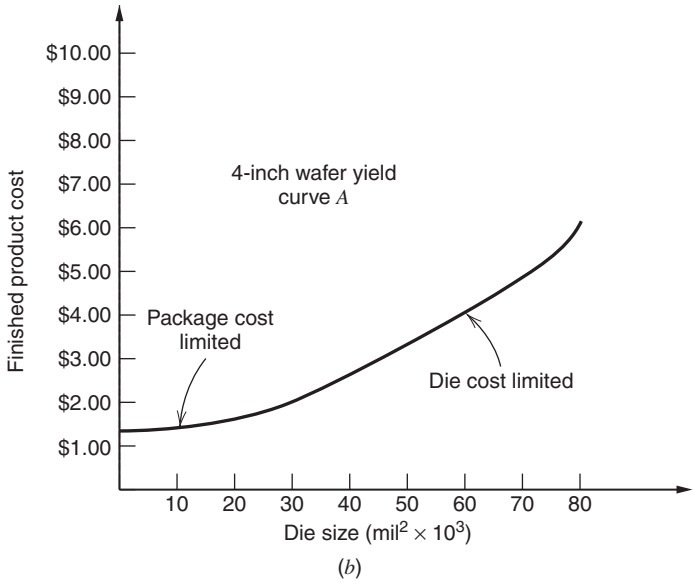


Figure 2.71 (*b*) Cost curve for example *b*.

This example shows that most of the cost comes from packaging and testing for small die sizes, whereas most of the cost comes from wafer-fab costs for large die sizes. This relationship is made clearer by considering the cost of the integrated circuit in terms of cost per unit area of silicon in the finished product, as illustrated in Fig. 2.72 for the examples previously given. These curves plot the ratio of the finished-product cost to the number of square mils of silicon on the die. The minimum cost per unit area of silicon results midway between the package-cost and die-cost limited regions for each example. Thus the fabrication of excessively large

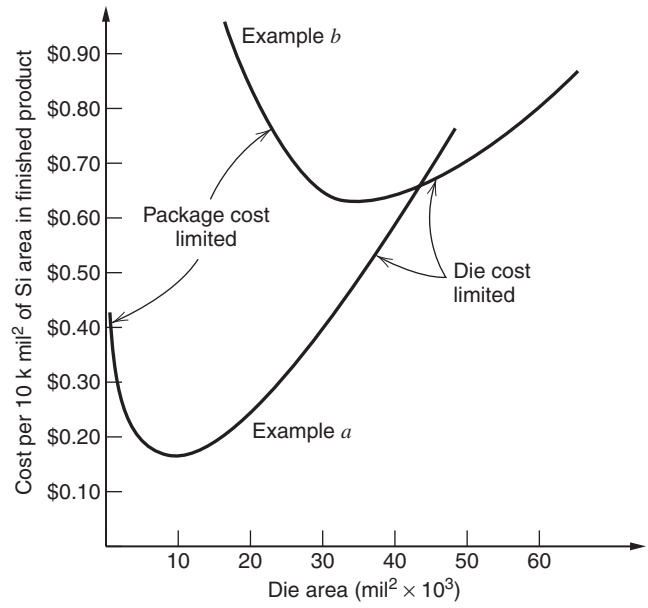


Figure 2.72 Cost of finished product in terms of cost per unit of silicon area for the two examples. Because the package and testing costs are lower in example *a*, the minimum cost point falls at a much smaller die size. The cost per unit of silicon area at large die sizes is smaller for example *b* because process *A* gives higher yield at large die sizes.

or small dice is uneconomical in terms of utilizing the silicon die area at minimum cost. The significance of these curves is that, for example, if a complex analog/digital system, characterized by example *b* in Fig. 2.72 with a total silicon area of 80,000 square mils is to be fabricated in silicon, it probably would be most economical to build the system on two chips rather than on a single chip. This decision would also be strongly affected by other factors such as the increase in the number of total package pins required for the two chips to be interconnected, the effect on performance of the required interconnections, and the additional printed circuit board space required for additional packages. The shape of the cost curves is also a strong function of the package cost, test cost of the individual product, yield curve for the particular process, and so forth.

The preceding analysis concerned only the direct costs to the manufacturer of the fabrication of the finished product; the actual selling price is much higher and reflects additional research and development, engineering, and selling costs. Many of these costs are fixed, however, so that the selling price of a particular integrated circuit tends to vary inversely with the quantity of the circuits sold by the manufacturer.

APPENDIX

A.2.1 SPICE MODEL-PARAMETER FILES

In this section, SPICE model-parameter symbols are compared with the symbols employed in the text for commonly used quantities.

Bipolar Transistor Parameters

SPICE Symbol	Text Symbol	Description
IS	I_S	Transport saturation current
BF	β_F	Maximum forward current gain
BR	β_R	Maximum reverse current gain
VAF	V_A	Forward Early voltage
RB	r_b	Base series resistance
RE	r_{ex}	Emitter series resistance
RC	r_c	Collector series resistance
TF	τ_F	Forward transit time
TR	τ_R	Reverse transit time
CJE	C_{je0}	Zero-bias base-emitter depletion capacitance
VJE	ψ_{0e}	Base-emitter junction built-in potential
MJE	n_e	Base-emitter junction-capacitance exponent
CJC	$C_{\mu 0}$	Zero-bias base-collector depletion capacitance
VJC	ψ_{0c}	Base-collector junction built-in potential
MJC	n_c	Base-collector junction-capacitance exponent
CJS	C_{CS0}	Zero-bias collector-substrate depletion capacitance
VJS	ψ_{0s}	Collector-substrate junction built-in potential
MJS	n_s	Collector-substrate junction-capacitance exponent

Note: Depending on which version of SPICE is used, a separate diode may have to be included to model base-substrate capacitance in a lateral *pnp* transistor.

MOSFET Parameters

SPICE Symbol	Text Symbol	Description
VTO	V_i	Threshold voltage with zero source-substrate voltage
KP	$k' = \mu C_{ox}$	Transconductance parameter
GAMMA	$\gamma = \frac{\sqrt{2q\epsilon N_A}}{C_{ox}}$	Threshold voltage parameter
PHI	$2\phi_f$	Surface potential
LAMBDA	$\lambda = \frac{1}{L_{\text{eff}}}\frac{dX_d}{dV_{DS}}$	Channel-length modulation parameter
CGSO	C_{ol}	Gate-source overlap capacitance per unit channel width
CGDO	C_{ol}	Gate-drain overlap capacitance per unit channel width
CJ	C_{j0}	Zero-bias junction capacitance per unit area from source and drain bottom to bulk (substrate)
MJ	n	Source-bulk and drain-bulk junction capacitance exponent (grading coefficient)
CJSW	C_{jsw0}	Zero-bias junction capacitance per unit junction perimeter from source and drain sidewall (periphery) to bulk
MJSW	n	Source-bulk and drain-bulk sidewall junction capacitance exponent
PB	ψ_0	Source-bulk and drain-bulk junction built-in potential
TOX	t_{ox}	Oxide thickness
NSUB	N_A, N_D	Substrate doping
NSS	Q_{ss}/q	Surface-state density
XJ	X_j	Source, drain junction depth
LD	L_d	Source, drain lateral diffusion

PROBLEMS

2.1 What impurity concentration corresponds to a 1 Ω-cm resistivity in *p*-type silicon? In *n*-type silicon?

2.2 What is the sheet resistance of a layer of 1 Ω-cm material that is 5 μm thick?

2.3 Consider a hypothetical layer of silicon that has an *n*-type impurity concentration of 10^{17} cm^{-3} at the top surface, and in which the impurity concentration decreases exponentially with distance into the silicon. Assume that the concentration has decreased to $1/e$ of its surface value at a depth of 0.5 μm, and that the impurity concentration in the sample before the insertion of the *n*-type impurities was 10^{15} cm^{-3} *p*-type. Determine the depth below the surface of the *pn* junction that results and determine the sheet resistance of the *n*-type layer. Assume a constant electron mobility of 800 cm^2/V-s. Assume that the width of the depletion layer is negligible.

2.4 A diffused resistor has a length of 200 μm and a width of 5 μm. The sheet resistance of the base diffusion is 100 Ω/□ and the emitter diffusion is 5 Ω/□. The base pinched layer has a sheet resistance of 5 kΩ/□. Determine the resistance of the resistor if it is an emitter-diffused, base-diffused, or pinch resistor.

2.5 A base-emitter voltage of from 520 mV to 580 mV is measured on a test *npn* transistor structure with 10 μA collector current. The emitter dimensions on the test transistor are 100 μm × 100 μm. Determine the range of values of Q_B implied by this data. Use this information to calculate the range of values of sheet resistance that will be observed in the pinch resistors in the circuit. Assume a constant electron diffusivity, $\overline{D}_n$, of 13 cm^2/s, and a constant hole mobility of 150 cm^2/V-s. Assume that the width of the depletion layer is negligible.

2.6 Estimate the series base resistance, series collector resistance r_c, base-emitter capacitance, base-collector capacitance, and collector-substrate

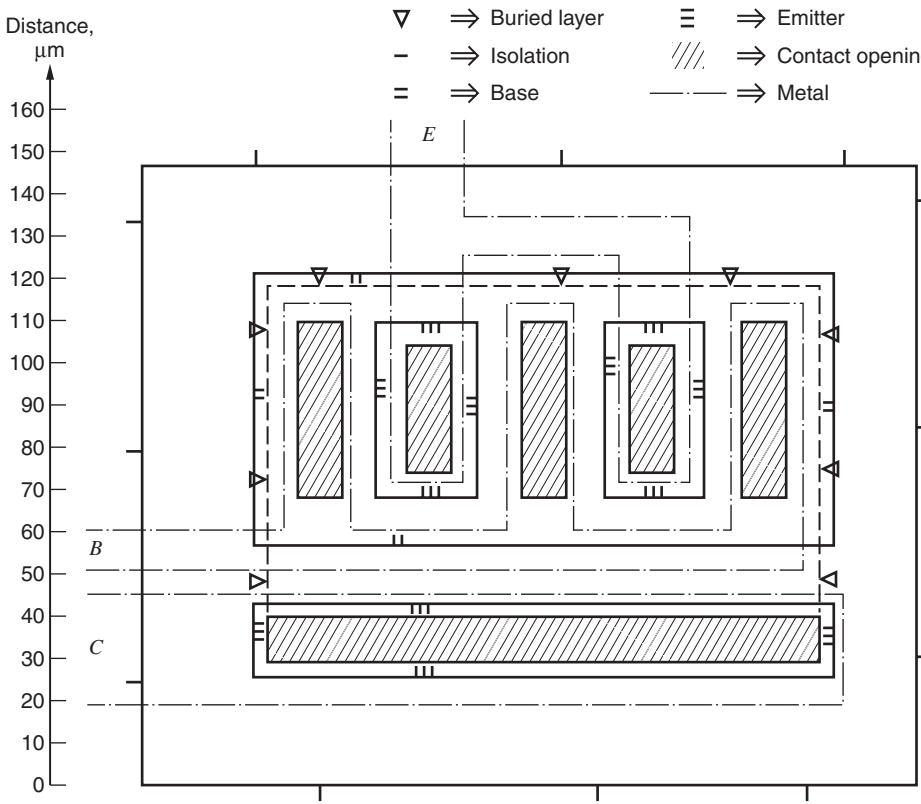


Figure 2.73 Device structure for Problem 2.6.

capacitance of the high-current *npn* transistor structure shown in Fig. 2.73. This structure is typical of those used as the output transistor in operational amplifiers that must supply up to about 20 mA. Assume a doping profile as shown in Fig. 2.17.

2.7 If the lateral *pnp* structure of Fig. 2.33*a* is fabricated with an epi layer resistivity of 0.5 Ω-cm, determine the value of collector current at which the current gain begins to fall off. Assume a diffusivity for holes of: $\overline{D}_p = 10$ cm^2/s. Assume a base width of 8 μm.

2.8 The substrate *pnp* of Fig. 2.36*a* is to be used as a test device to monitor epitaxial layer thickness. Assume that the flow of minority carriers across the base is vertical, and that the width of the emitter-base and collector-base depletion layers is negligible. Assume that the epi layer resistivity is known to be 2 Ω-cm by independent measurement. The base-emitter voltage is observed to vary from 525 mV to 560 mV over several wafers at a collector current of 10 μA. What range of epitaxial layer thickness does this imply? What is the corresponding range of sheet resistance that will be observed in the epitaxial pinch resistors? Assume a hole diffusivity of 10 cm^2/s, and an electron mobility of 800 cm^2/V-s. Neglect the depletion layer thickness. Assume a junction depth of 3 μm for the base diffusion.

2.9 Calculate the total parasitic junction capacitance associated with a 10-kΩ base-diffused resistor if the base sheet resistance is 100 Ω/□ and the resistor width is 6 μm. Repeat for a resistor width of 12 μm. Assume the doping profiles are as shown in Fig. 2.17. Assume the clubheads are 26 μm × 26 μm, and that the junction depth is 3 μm. Account for sidewall effects.

2.10 For the substrate *pnp* structure shown in Fig. 2.36*a*, calculate I_S, C_{je}, C_μ, and τ_F. Assume the doping profiles are as shown in Fig. 2.17.

2.11 A base-emitter voltage of 480 mV is measured on a super-β test transistor with a 100 μm × 100 μm emitter area at a collector current of 10 μA. Calculate the Q_B and the sheet resistance of the

base region. Estimate the punch-through voltage in the following way. When the base depletion region includes the entire base, charge neutrality requires that the number of ionized acceptors in the depletion region in the base be equal to the number of ionized donors in the depletion region on the collector side of the base. [See (1.2).] Therefore, when enough voltage is applied that the depletion region in the base region includes the whole base, the depletion region in the collector must include a number of ionized atoms equal to Q_B. Since the density of these atoms is known (equal to N_D), the width of the depletion layer in the collector region at punch-through can be determined. If we assume that the doping in the base N_A is much larger than that in the collector N_D, then (1.15) can be used to find the voltage that will result in this depletion layer width. Repeat this problem for the standard device, assuming a V_{BE} measured at 560 mV. Assume an electron diffusivity $\overline{D}_n$ of 13 cm^2/s, and a hole mobility $\overline{\mu}_p$ of 150 cm^2/V-s. Assume the epi doping is 10^{15} cm^{-3}. Use $\epsilon = 1.04 \times 10^{-12}$F/cm for the permittivity of silicon. Also, assume ψ_o for the collector-base junction is 0.55 V.

2.12 An MOS transistor biased in the active region displays a drain current of 100 μA at a V_{GS} of 1.5 V and a drain current of 10 μA at a V_{GS} of 0.8 V. Determine the threshold voltage and $\mu_n C_{ox}$(W/L). Neglect subthreshold conduction and assume that the mobility is constant.

2.13 Calculate the threshold voltage of the *p*-channel transistors for the process given in Table 2.1. First do the calculation for the unimplanted transistor, then for the case in which the device receives the channel implant specified. Note that this is a *p*-type implant, so that the effective surface concentration is the difference between the background substrate concentration and the effective concentration in the implant layer.

2.14 An *n*-channel implanted transistor from the process described in Table 2.1 displays a measured output resistance of 5 MΩ at a drain current of 10 μA, biased in the active region at a V_{DS} of 5 V. The drawn dimensions of the device are 100 μm by 7 μm. Find the output resistance of a second device on the same technology that has drawn dimensions of 50 μm by 12 μm and is operated at a drain current of 30 μA and a V_{DS} of 5 V.

2.15 Calculate the small-signal model parameters of the device shown in Fig. 2.74, including g_m, g_{mb}, r_o, C_{gs}, C_{gd}, C_{sb}, and C_{db}. Assume the transistor is biased at a drain-source voltage of 2 V and a drain current of 20 μA. Use the process parameters that are specified in Table 2.4. Assume $V_{SB} = 1$ V.

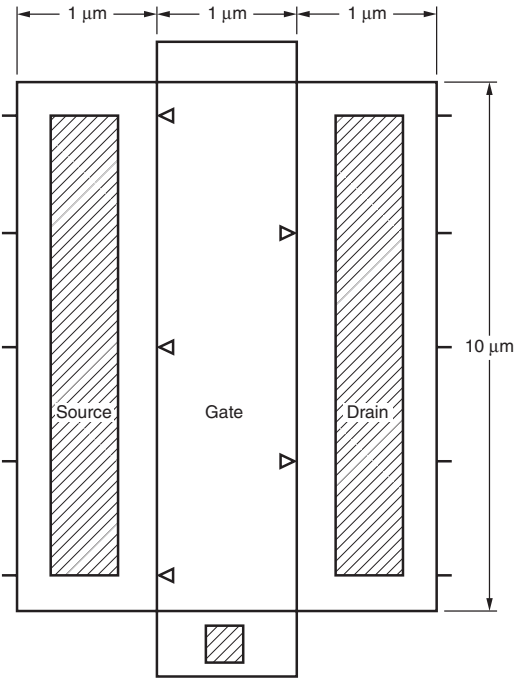


Figure 2.74 Transistor for Problem 2.15.

2.16 The transistor shown in Fig. 2.74 is connected in the circuit shown in Fig. 2.75. The gate is grounded, the substrate is connected to −1.5 V, and the drain is open circuited. An ideal current source is tied to the source, and this source has a value of zero for $t < 0$ and 10 μA for $t > 0$. The source and drain are at an initial voltage of +1.5 V at $t = 0$. Sketch the voltage at the source and drain from $t = 0$ until the drain voltage reaches −1.5 V. For simplicity, assume that the source-substrate and drain-substrate capacitances are constant at their zero-bias values. Assume the transistor has a threshold voltage of 0.6 V.

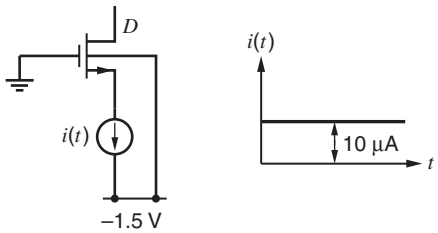


Figure 2.75 Circuit for Problem 2.16.

2.17 Show that two MOS transistors connected in parallel with channel widths of W_1 and W_2 and identical channel lengths of L can be modeled as one equivalent MOS transistor whose width is $W_1 + W_2$ and whose length is L, as shown in Fig. 2.76. Assume the transistors are identical except for their channel widths.

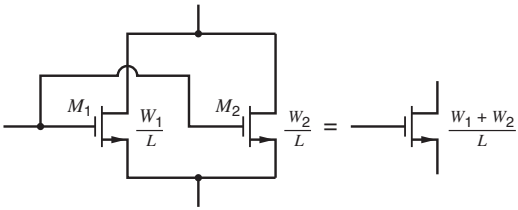


Figure 2.76 Circuit for Problem 2.17.

2.18 Show that two MOS transistors connected in series with channel lengths of L_1 and L_2 and identical channel widths of W can be modeled as one equivalent MOS transistor whose width is W and whose length is $L_1 + L_2$, as shown in Fig. 2.77. Assume the transistors are identical except for their channel lengths. Ignore the body effect and channel-length modulation.

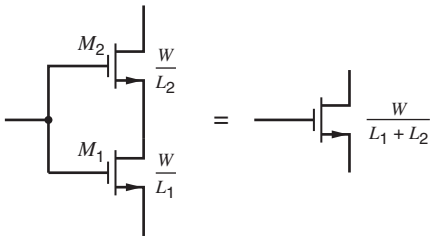


Figure 2.77 Circuit for Problem 2.18.

2.19 An integrated electronic subsystem is to be fabricated, which requires 40,000 square mils of silicon area. Determine whether the system should be put on one or two chips, assuming that the fabrication cost of the two chips is the only consideration. Assume that the wafer-fab cost is \$100.00, the packaging and testing costs are \$0.60, the die-fab yield is 0.9, and the final-test yield is 0.8. Assume the process used follows curve *B* in Fig. 2.68. Repeat the problem assuming yield curve *A*, and then yield curve *C*. Assume a 4-inch wafer.

2.20 Determine the direct fabrication cost of an integrated circuit that is 150 mils on a side in size. Assume a wafer-fab cost of \$130.00, a package and testing cost of \$0.40, a die-fab yield of 0.8, and a final-test yield of 0.8. Work the problem for yield curves *A*, *B*, and *C* in Fig. 2.68. Assume a 4-inch wafer.

2.21(a) A frequently used empirical approximation for the yield of an IC process as a function of die size is

$$Y_{ws} = \exp(-A/A_0)$$

where A is the die area and A_0 is a constant. Using Fig. 2.68, determine approximate values of A_0 for each of the three processes shown. Use the point on the curve at which the yield is e^{-1} to determine A_0. Plot the yield predicted by this expression and compare with the curves shown in Fig. 2.68.

(b) Use the expression derived in (a), together with the gross-die-per-wafer curves shown in Fig. 2.69, to develop an analytical expression for the cost of silicon per unit area as a function of die size, Y_{df}, Y_{ft}, C_p, and C_w for each of the three processes *A*, *B*, and *C*.

2.22 Calculate the small-signal model parameters g_m, r_o, C_{gs}, and C_{gd} for a NMOS transistor. Assume the transistor operates in the active region with $I_D = 100$ μA, $V_{DS} = 1$ V, $V_{BS} = 0$ V, $W = 0.9$ μm, and $L = 0.2$ μm. Use the transistor model data in Table 2.5.

2.23 Calculate the small-signal model parameters g_m, r_o, C_{gs}, and C_{gd} for a NMOS transistor. Also calculate the gate-leakage current I_G. Assume the transistor operates in the active region with $I_D = 100$ μA, $V_{DS} = 1$ V, $V_{BS} = 0$ V, $W = 0.5$ μm, and $L = 0.1$ μm. Use the transistor model data in Table 2.6.

REFERENCES

1. A. S. Grove. *Physics and Technology of Semiconductor Devices.* Wiley, New York, 1967.

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

3. E. M. Conwell. "Properties of Silicon and Germanium," *Proc. IRE,* Vol. 46, pp. 1281–1300, June 1958.

4. J. C. Irvin. "Resistivity of Bulk Silicon and of Diffused Layers in Silicon," *Bell System Tech. Journal,* Vol. 41, pp. 387–410, March 1962.

5. R. W. Russell and D. D. Culmer. "Ion-Implanted JFET-Bipolar Monolithic Analog Circuits," *Digest of Technical Papers, 1974 International Solid-State Circuits Conference,* Philadelphia, PA, pp. 140–141, February 1974.

6. D. J. Hamilton and W. G. Howard. *Basic Integrated Circuit Engineering.* McGraw-Hill, New York, 1975.

7. Y. Tamaki, T. Shiba, I. Ogiwara, T. Kure, K. Ohyu, and T. Nakamura. "Advanced Device Process Technology for 0.3 μm Self-Aligned Bipolar LSI," *Proceedings of the IEEE Bipolar Circuits and Technology Meeting,* pp. 166–168, September 1990.

8. M. Kurisu, Y. Sasyama, M. Ohuchi, A. Sawairi, M. Sigiyama, H. Takemura, and T. Tashiro.

"A Si Bipolar 21 GHz Static Frequency Divider," *Digest of Technical Papers, 1991 International Solid-State Circuits Conference,* pp. 158–159, February 1991.

9. R. M. Burger and R. P. Donovan. *Fundamentals of Silicon Integrated Device Technology.* Vol. 2, pp. 134–136. Prentice-Hall, Englewood Cliffs, NJ, 1968.

10. R. J. Whittier and D. A. Tremere. "Current Gain and Cutoff Frequency Falloff at High Currents," *IEEE Transactions Electron Devices,* Vol. ED-16, pp. 39–57, January 1969.

11. H. R. Camenzind. *Electronic Integrated Systems Design.* Van Nostrand Reinhold, New York, 1972. Copyright © 1972 Litton Educational Publishing, Inc. Reprinted by permission of Van Nostrand Reinhold Company.

12. H. J. DeMan. "The Influence of Heavy Doping on the Emitter Efficiency of a Bipolar Transistor," *IEEE Transactions on Electron Devices,* Vol. ED-18, pp. 833–835, October 1971.

13. H. C. Lin. *Integrated Electronics.* Holden-Day, San Francisco, 1967.

14. N. M. Nguyen and R. G. Meyer. "Si IC-Compatible Inductors and *LC* Passive Filters," *IEEE Journal of Solid-State Circuits,* Vol. 25, pp. 1028–1031, August 1990.

15. J. Y.-C. Chang, A. A. Abidi, and M. Gaitan. "Large Suspended Inductors on Silicon and Their Use in a 2 μm CMOS RF Amplifier," *IEEE Electron Device Letters,* Vol. 14, pp. 246–248, May 1993.

16. K. Negus, B. Koupal, J. Wholey, K. Carter, D. Millicker, C. Snapp, and N. Marion. "Highly Integrated Transmitter RFIC with Monolithic Narrowband Tuning for Digital Cellular Handsets," *Digest of Technical Papers, 1994 International Solid-State Circuits Conference,* San Francisco, CA, pp. 38–39, February 1994.

17. P. R. Gray and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits,* Third Edition, Wiley, New York, 1993.

18. R. J. Widlar. "Design Techniques for Monolithic Operational Amplifiers," *IEEE Journal of Solid-State Circuits,* Vol. SC-4, pp. 184–191, August 1969.

19. K. R. Stafford, P. R. Gray, and R. A. Blanchard. "A Complete Monolithic Sample/Hold Amplifier," *IEEE Journal of Solid-State Circuits,* Vol. SC-9, pp. 381–387, December 1974.

20. P. C. Davis, S. F. Moyer, and V. R. Saari. "High Slew Rate Monolithic Operational Amplifier Using Compatible Complementary *pnp*'s," *IEEE Journal of Solid-State Circuits,* Vol. SC-9, pp. 340–346, December 1974.

21. A. P. Chandrakasan, S. Sheng, and R. W. Brodersen. "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits,* Vol. 27, pp. 473–484, April 1992.

22. Y. P. Tsividis, *Operation and Modeling of the MOS Transistor*. McGraw-Hill, 1987.

23. S.-H. Lo, D. A. Buchanan, Y. Taur and W. Wang, "Quantum-Mechanical Modeling of Electron Tunneling from the Inversion Layer of Ultra-Thin-Oxide nMOSFET's," *IEEE Electron Device Letters*, pp. 209–211, May 1997.

24. K. F. Schuegraf, C. C. King, and C. Hu, "Ultra-thin Silicon Dioxide Leakage Current and Scaling Limits," *Symp. on VLSI Technology, Digest of Technical Papers*, pp. 18–19, 1992.

25. W-C. Lee, C. Hu, "Modeling Gate and Substrate Currents due to Conduction- and Valence-Band Electron and Hole Tunneling," *Symp. on VLSI Technology, Digest of Technical Papers*, pp. 198–199, 2000.

26. www-device.eecs.berkeley.edu/~bsim3/.

27. D. P. Foty, *MOSFET Modeling with SPICE*, Prentice-Hall, Upper Saddle River, NJ, 1997.

28. E. A. Vittoz. "MOS Transistors Operated in the Lateral Bipolar Mode and Their Application in CMOS Technology," *IEEE Journal of Solid-State Circuits,* Vol. SC-18, pp. 273–279, June 1983.

29. W. T. Holman and J. A. Connelly. "A Compact Low-Noise Operational Amplifier for a 1.2 μm Digital CMOS Technology," *IEEE Journal of Solid-State Circuits,* Vol. 30, pp. 710–714, June 1995.

30. C. A. Laber, C. F. Rahim, S. F. Dreyer, G. T. Uehara, P. T. Kwok, and P. R. Gray. "Design Considerations for a High-Performance 3-μm CMOS Analog Standard-Cell Library," *IEEE Journal of Solid-State Circuits,* Vol. SC-22, pp. 181–189, April 1987.

31. B. L. Crowder and S. Zirinsky. "1 μm MOSFET VLSI Technology: Part VII—Metal Silicide Interconnection Technology—A Future Perspective," *IEEE Journal of Solid-State Circuits,* Vol. SC-14, pp. 291–293, April 1979.

32. J. L. McCreary. "Matching Properties, and Voltage and Temperature Dependence of MOS Capacitors," *IEEE Journal of Solid-State Circuits,* Vol. SC-16, pp. 608–616, December 1981.

33. D. J. Allstot and W. C. Black, Jr. "Technological Design Considerations for Monolithic MOS Switched-Capacitor Filtering Systems," *Proceedings of the IEEE,* Vol. 71, pp. 967–986, August 1983.

34. O. E. Akcasu. "High Capacitance Structure in a Semiconductor Device," *U.S. Patent 5,208,725,* May 1993.

35. H. Samavati, A. Hajimiri, A. R. Shahani, G. N. Nasserbakht, and T. H. Lee. "Fractal Capacitors,"

IEEE Journal of Solid-State Circuits, Vol. 33, pp. 2035–2041, December 1998.

36. A. R. Alvarez. *BiCMOS Technology and Applications.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.

37. J. L. de Jong, R. Lane, B. van Schravendijk, and G. Conner. "Single Polysilicon Layer Advanced Super High-speed BiCMOS Technology," *Proceedings of the IEEE Bipolar Circuits and Technology Meeting,* pp. 182–185, September 1989.

38. D. L. Harame, J. H. Comfort, J. D. Cressler, E. F. Crabbé, J. Y.-C. Sun, B. S. Meyerson, and T. Tice. "Si/SiGe Epitaxial-Base Transistors–Part I: Materials, Physics, and Circuits," *IEEE Transactions on Electron Devices,* Vol. 42, pp. 455–468, March 1995.

39. J. D. Cressler, D. L. Harame, J. H. Comfort, J. M. C. Stork, B. S. Meyerson, and T. E. Tice. "Silicon-Germanium Heterojunction Bipolar Technology: The Next Leap in Silicon?" *Digest of Technical Papers, 1994 International Solid-State Circuits Conference,* San Francisco, CA, pp. 24–27, February 1994.

40. D. L. Harame, J. H. Comfort, J. D. Cressler, E. F. Crabbé, J. Y.-C. Sun, B. S. Meyerson, and T. Tice. "Si/SiGe Epitaxial-Base Transistors—Part II: Process Integration and Analog Applications," *IEEE Transactions on Electron Devices,* Vol. 42, pp. 469–482, March 1995.

41. M. T. Bohr. "Interconnect Scaling—The Real Limiter to High Performance ULSI," *Technical Digest, International Electron Devices Meeting,* pp. 241–244, December 1995.

42. C. S. Chang, K. A. Monnig, and M. Melliar-Smith. "Interconnection Challenges and the National Technology Roadmap for Semiconductors," *IEEE International Interconnect Technology Conference,* pp. 3–6, June 1998.

43. D. Edelstein, J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, P. Roper, T. McDevitt, W. Motsiff, A. Simon, J. Dukovic, R. Wachnik, H. Rathore, R. Schulz, L. Su, S. Luce, and J. Slattery. "Full Copper Wiring in a Sub-0.25 μm CMOS ULSI Technology," *IEEE International Electron Devices Meeting,* pp. 773–776, December 1997.