(a) What will be the cost of the final product if testing and packaging adds \$1.60 to the completed product?

(b) The circuit design could be partitioned into two chips rather than one, but each die will increase in area by 15% in order to accommodate additional pads and I/O circuitry. If the testing and packaging cost remains the same, what is the cost of the two-chip set? Base your answers on Eq. (8.8). (See Problem 8.4 for the number of dice per wafer.)

8.7 **(a)** Repeat Problem 8.5 for a defect density of 5 defects/cm^2 and a wafer cost of \$150.

(b) Repeat Problem 8.5 for a defect density of 5 defects/cm^2 and a wafer cost of \$300.

8.8 A die has an area of 25 mm^2 and is being manufactured on a 100-mm-diameter wafer using a process rated at 2 defects/cm^2. A new process is being developed which allows the die area to be reduced by a factor of 2. However, because of the smaller feature sizes, the new process costs 30% more and is presently achieving only 10 defects/cm^2.

(a) Is it economical to switch to this new process?

(b) At what defect density does the cost of the new die equal the cost of the old die?

(c) Based on your judgment, would you recommend switching to the new process even if it is not now economical? Why?

(d) At what die size is the cost the same in either process? Use Eq. (8.8) for this problem.

8.9 What is the limit of the yield distribution in Eq. (8.9) as the clustering parameter approaches infinity?

8.10 Compare the predictions of yield equations 8.5 and 8.9 for D_0A ranging from 1 to 10 with $\alpha = 5$ and $\alpha = 5{,}000$.

8.11 Suppose $D_0 = 0.1/\text{cm}^2$. What is the average number of defects on 150 mm, 200 mm, and 300 mm wafers?

8.12 Suppose that going from 100-mm wafers to 150-mm wafers changes the wafer processing cost from \$150/wafer to \$250/wafer, and the defect density remains constant at 10 defects/cm^2. Assume a die cost of \$1.00 to find the die sizes. Use Eq. (8.9) with a cluster factor of 2. Use a calculator or computer to find the answer by iteration.

8.13 What would be the die yield in Fig. 8.16(b) if the defect positions were the same but the die pattern was rotated by 90°? How many good dice with four times the area of that in Fig. 8.16(a) would now exist?

8.14 A Gaussian probability density function for defect density is given by

$$f(D) = \frac{2}{D_0\sqrt{\pi}} \exp -\left[\frac{2(D-D_0)}{D_0}\right]^2, \text{ for } 0 \le D \le 2D_0, \text{ and 0 otherwise}$$

Calculate the yield Y for various values of D_0A and compare your results to those of the triangular distribution given in Eq. (8.7). (You may want to use a calculator or computer to perform the iteration.)

8.15 The wafers shown in Fig. 8.16 actually have 120 defects placed randomly on the wafer. Obviously, some chips must have several defects. Use Eq. (8.1) to predict how many dice will have exactly 1, 2, 3, 4, and 5 defects.

8.16 What defect density is required to achieve a yield of 70% for a 10×15 mm die if the process is characterized by a cluster parameter of 5? (b) Repeat for 80% yield. (c) Repeat for 90% yield.

8.17 What defect density is required to achieve a yield of 75% for a 20×20 mm die if the process is characterized by a cluster parameter of 6? (b) Repeat for 85% yield.

CHAPTER 9

MOS Process Integration

In Chapter 9, we explore a number of relationships between process and device design and circuit layout. Processes are usually developed to provide devices with the highest possible performance in a specific circuit application, and one must understand the circuit environment and its relation to device parameters and device layout.

In this chapter, we look at a number of basic concerns in MOS process design, including channel-length control; layout ground rules and ground-rule design; source–drain breakdown and punch-through voltages; and threshold-voltage adjustment. Metal-gate technology is discussed, and the important advantages of self-aligned silicon-gate technologies are presented. Discussions of CMOS and silicon-on-insulator technologies complete the chapter.

9.1 BASIC MOS DEVICE CONSIDERATIONS

To explore the relationship between MOS process design and basic device behavior, we begin by discussing the static current–voltage relationship for the MOS transistor, as developed in Volume IV of this series [1]. The cross section of two NMOS transistors is shown in Fig. 9.1. In the linear region of operation, the drain current is given by

$$I_D = \bar{\mu}_n C_O (w/L)(V_{GS} - V_{TN} - V_{DS}/2)V_{DS} \tag{9.1}$$

for $V_{GS} \ge V_{TN}$ and $V_{DS} \le V_{GS} - V_{TN}$. $C_O = K_O\varepsilon_O/X_O$ is the oxide capacitance per unit area, $\bar{\mu}_n$ is the average majority-carrier mobility in the inversion layer, and V_{TN} is the threshold voltage. W and L represent the width and length of the channel, respectively.

One of the first specifications required is the circuit power-supply voltages, which set the maximum value of V_{GS} and V_{DS} that the devices must withstand. Once this choice is made, the only variables in Eq. (9.1) that a circuit designer may adjust are the width and length of the transistor. Thus, the circuit designer varies the circuit topology and horizontal geometry to achieve the desired circuit function.

Other device parameters are fixed by the process designer, who must determine the process sequence, times, temperatures, etc., which ultimately determine the device

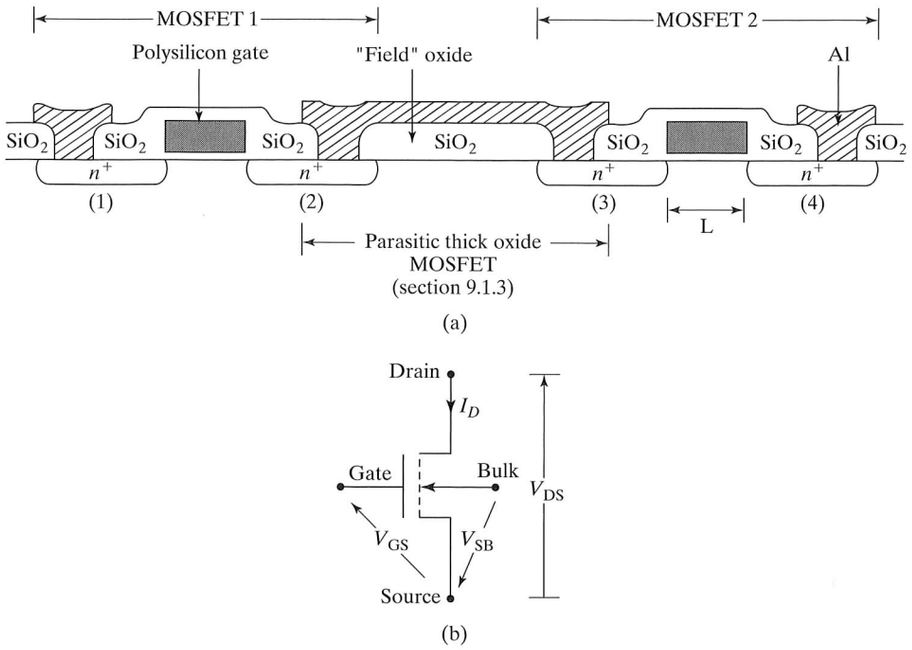


FIGURE 9.1

(a) Cross section of an integrated circuit showing two adjacent NMOS transistors. A parasitic NMOS device is formed by the aluminum interconnection over the field oxide with diffused regions (2) and (3) acting as source and drain. (b) An NMOS transistor with gate-to-source (V_{GS}), drain-to-source (V_{DS}), and source-to-bulk (V_{SB}) voltages defined.

structure and hence its characteristics. These include specifying the gate-oxide thickness, field-oxide thickness, substrate doping, and field and threshold-adjustment implantations. The process designer also supplies a set of design rules, or ground rules that must be obeyed during circuit layout. These include minimum channel length and width, spacings between features on the same and different mask levels, and overlaps between features on different mask levels. A mask alignment sequence and tolerances must also be developed for the process.

9.1.1 Gate-Oxide Thickness

Current flow in the MOS transistor, for a given set of terminal voltages, is inversely proportional to the gate-oxide thickness. The gate oxide will generally be made as thin as possible, commensurate with oxide breakdown and reliability considerations. High-quality silicon dioxide will typically break down at electric fields of 5 to 10 MV/cm, corresponding to 5 to 10 V across a 10-nm oxide. Present processes are using oxide thicknesses between 2 and 10 nm. Below 10 nm, current starts to flow by tunneling, and the oxide begins to lose its insulating qualities. The choice of oxide thickness is also related to hot electron injection into the oxide, a problem beyond the scope of this text [2–4]. Various alternative gate oxide materials are being investigated. Oxynitrides are formed by adding nitrogen to the silicon dioxide system either during or after formation of the gate oxide. Researchers are exploring a number of high-dielectric constant

gate materials that will permit the use of somewhat thicker oxides without reducing the oxide capacitance and transconductance of the transistor.

9.1.2 Substrate Doping and Threshold Voltage

Threshold voltage is an important parameter which determines the gate voltage necessary to initiate conduction in the MOS device. The threshold voltage [1] for a device with a uniformly doped substrate is given by

$$\text{NMOS: } V_{TN} = \Phi_M - \chi - \frac{E_g}{2q} + |\Phi_F| + \left[\sqrt{2K_s\epsilon_0 q N_B(2|\Phi_F| + V_{SB})}\right]/C_O - Q_{tot}/C_O \tag{9.2}$$

$$\text{PMOS: } V_{TP} = \Phi_M - \chi - \frac{E_g}{2q} - |\Phi_F| - \left[\sqrt{2K_s\epsilon_0 q N_B(2|\Phi_F| - V_{BS})}\right]/C_O - Q_{tot}/C_O$$

$$|\Phi_F| = (kT/q)\ \ln(N_B/n_i)$$

where N_B is the substrate doping. $\Phi_M - \chi = -0.11$ for an aluminum gate, $\Phi_M - \chi = 0$ for an n^+-doped polysilicon gate, and $\Phi_M - \chi = +1.12$ for a p^+-doped polysilicon gate.

Q_{tot}[1] represents the total oxide and interface charge per cm^2 and adds a parallel shift of the curves in Fig. 9.2 to more negative values of threshold. This charge contribution to the threshold voltage had an extremely important influence on early MOS device fabrication. Q_{tot} tends to be positive, which makes the MOS transistor threshold more negative; n-channel transistors become depletion-mode devices ($V_{TN} < 0$), whereas p-channel transistors remain enhancement-mode devices ($V_{TP} < 0$). During early days of MOS technology, Q_{tot} was high, and the only successful MOS processing was done using PMOS technology. After the industry gained an understanding of the origin of oxide and interface charges, and following the advent of ion implantation, NMOS technology became dominant, because of the mobility advantage of electrons over holes. Total charge levels have been reduced to less than 5×10^{10} charges/cm^2 in good MOS processes, and the oxide charge contribution to threshold voltage is minimal.

Substrate doping enters the threshold-voltage expression through both the $|\Phi_F|$ term and the square-root term. A plot of threshold voltage versus substrate doping for n- and p-channel, n^+ polysilicon-gate devices with 10-nm gate oxides is given in Fig. 9.2 for $Q_{tot} = 0$. The choice of substrate doping is complicated by other considerations, including drain-to-substrate breakdown voltage, drain-to-source punch-through voltage, source-to-substrate and drain-to-substrate capacitances, and substrate sensitivity or body effect.

[1] $Q_{tot} = Q_F + Q_{rr} + \gamma_M Q_M$

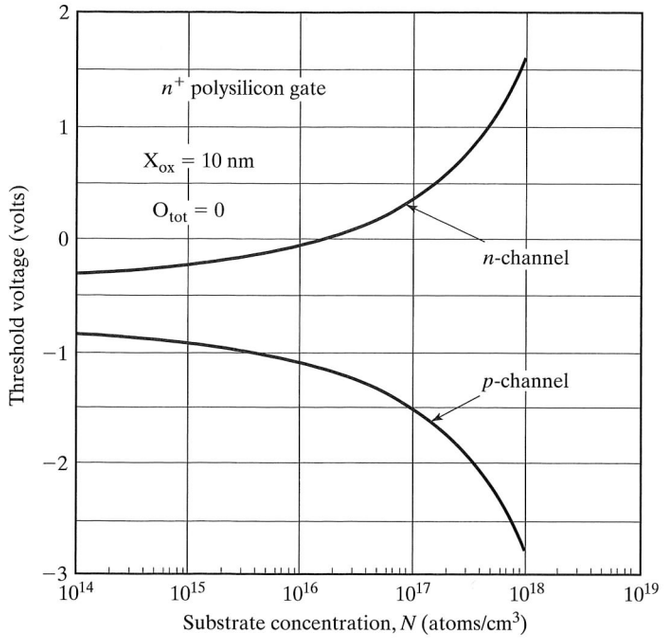


FIGURE 9.2

Threshold voltages for *n*- and *p*-channel polysilicon-gate transistors with 10-nm gate oxides, calculated from Eq. (9.2).

9.1.3 Junction Breakdown

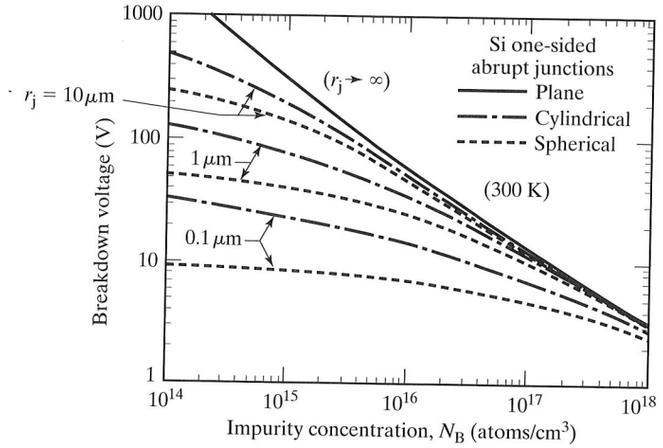
The source and drain regions are usually heavily doped to minimize their resistance and are essentially one-sided junctions in which the depletion region extends entirely into the substrate. Figure 9.3(a) gives the breakdown voltage of a one-sided *pn* junction as a function of the doping concentration on the lightly doped side of the junction [5]. Junction breakdown voltage decreases as doping level increases. Breakdown voltage is also a function of the radius of curvature of the junction space-charge region. Junction curvature enhances the electric field in the curved region of the depletion layer and reduces the breakdown voltage below that predicted by one-dimensional junction theory. A rectangular diffused area has regions with both cylindrical and spherical curvature, as shown in Fig. 9.3(b). It is worth noting that very shallow spherical junctions break down at voltages of less than 10 V, regardless of doping level.

9.1.4 Punch-through

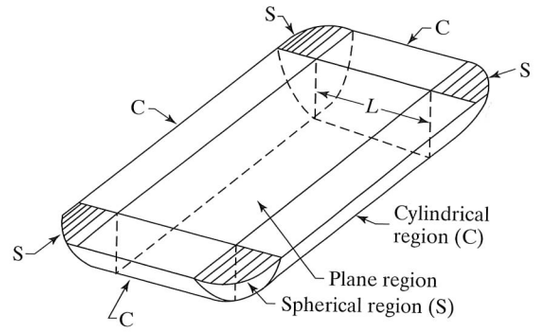
Punch-through occurs when the drain depletion region contacts the source depletion region, and substrate doping must be chosen to prevent the merging of these depletion regions when the MOSFET is off. Punch-through will not occur if the channel length exceeds the sum of the depletion-layer widths of the source-to-substrate and drain-to-substrate junctions. For a transistor used as a load device in a logic circuit, the source-to-substrate and drain-to-substrate junctions must both support a voltage equal to the drain supply voltage plus the substrate supply voltage. The depletion-layer widths can be estimated using the formula for the width of a one-sided step junction

$$W_d = \sqrt{(2K_s\epsilon_0(V_A + \Phi_{bi}))/qN_B},$$

$$\Phi_{bi} = 0.56 + (kT/q)\ \ln(N_B/n_i), \qquad (9.3)$$



(a)



(b)

FIGURE 9.3

(a) Abrupt *pn* junction breakdown voltage versus impurity concentration on the lightly doped side of the junction for both cylindrical and spherical structures. r_j is the radius of curvature. (b) Formation of cylindrical and spherical regions by diffusion through a rectangular window. Copyright 1985, John Wiley & Sons, Inc. Reprinted with permission from Ref. [5].

where V_A is the total applied voltage and Φ_{bi} is the built-in potential of the junction. If the channel length is greater than $2W_d$ punch-through should not occur. Figure 9.4 gives the depletion-layer width of *pn* junctions as a function of doping and applied voltage. Punch-through is not a limiting factor for most doping levels, except for very short-channel transistors. Ion implantation has been used to enhance the doping concentration below the channel region of short-channel devices to increase the punch-through voltage.

9.1.5 Junction Capacitance

The capacitance per unit area associated with a diffused junction is given by the parallel-plate capacitance formula with a plate spacing of W_d:

$$C_j = K_s\epsilon_0/W_d$$

The larger the doping, the larger the capacitance. Zero bias and a doping concentration of $10^{16}/cm^3$ result in a junction capacitance of approximately 10 nf/cm^2.

Eq. (9.2) shows that the threshold voltage depends on the source-to-substrate voltage, V_{SB}. This variation is known as "substrate sensitivity," or "body effect," and it becomes worse as the substrate doping level increases.

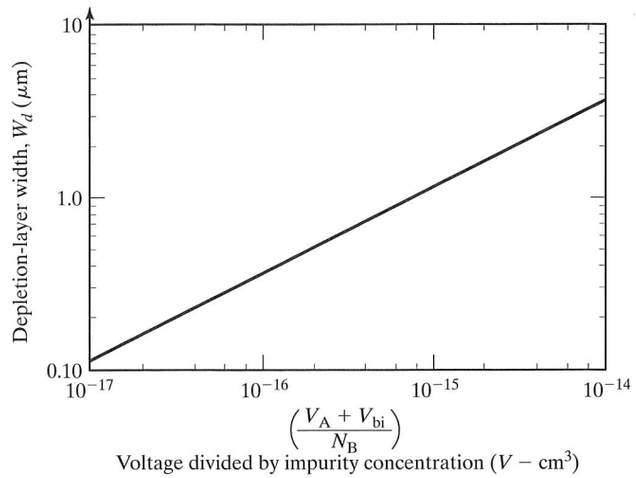


FIGURE 9.4

Depletion-layer width of a one-sided step junction as a function of doping and applied voltage calculated from eq. (9.3).

From the preceding discussion, one can see that there are trade-offs involved in the choice of substrate doping. Substrate doping is directly related to threshold voltage. It is desirable to reduce substrate doping to minimize junction capacitance and substrate sensitivity and to maximize breakdown voltage. Mobility also tends to be higher for lower doping levels. On the other hand, a heavily doped substrate will increase the punch-through voltage.

9.1.6 Threshold Adjustment

Ion implantation is routinely used to separate threshold-voltage design from the other factors involved in the choice of substrate doping. Substrate doping can be chosen based on a combination of breakdown, punch-through, capacitance, and substrate sensitivity considerations, and the threshold voltage is then adjusted to the desired value by adding a shallow ion-implantation step to the process. Figure 9.5 shows a step approximation to an implanted profile used to adjust the impurity concentration near the surface. These additional impurities cause a shift in threshold voltage given approximately by

$$\Delta V_{TN} = (1/C_O)\,(qQ_i)\,(1 - x_i/2x_d), \qquad x_i \ll x_d, \qquad x_d = \sqrt{qN_B/4K_s\epsilon_0|\Phi_F|} \tag{9.4}$$

where $Q_i = x_i N_i$ represents the implanted dose and x_d represents the depletion-layer width beneath the gate. For shallow implants, the threshold-voltage shift is proportional to the implanted dose. The threshold-voltage shift is positive for acceptor impurities and negative for donor impurities.

Example 9.1:

An NMOS transistor with an n^+ polysilicon gate is fabricated with a 10-nm gate oxide, a substrate doping of $10^{16}/cm^3$, and source–drain junction depths of 0.25 μm.

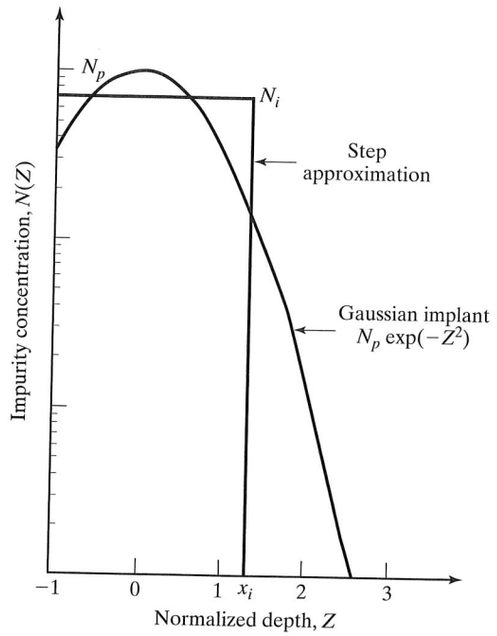


FIGURE 9.5

Step approximation to a Gaussian impurity profile used to estimate the threshold-voltage shift achieved using ion implantation.

Determine the threshold voltage and drain-to-substrate breakdown voltages for this device. What is the punch-through voltage for a channel length of 1 μm if the substrate bias is 0 V? A shallow boron implantation is to be used to adjust the threshold to 0.7 V. What is the dose of this implant? (Assume that $V_{SB} = 0$ and $Q_{tot} = 0$.)

Solution : For the n^+ polysilicon-gate transistor, $\Phi_M - \chi - E_g/2q = -0.56$ V and $|\Phi_F| = 0.36$ volts (for $n_i = 1 \times 10^{10}/cm^3$ and $kT/q = 0.026$ V). For $V_{SB} = 0$, the threshold voltage expression yields $V_{TN} = -0.56 + 0.36 + 0.14V = -0.06V$. Interpolating Fig. 9.3 for spherical breakdown with a substrate doping of $10^{16}/cm^3$ and a radius of curvature of 0.25 μm gives an estimated drain-to-substrate breakdown voltage of 20 V. To estimate the punch-through voltage, we use Eq. (9.3) with $2W = 1\mu m$ and $V_A = V_D$, where V_D is the drain voltage. Evaluating this expression yields $V_D = 1.01V$.

For a shallow implant, the threshold-voltage shift is approximately $\Delta V_T = qQ/C_O$. A voltage shift of 0.76V with an oxide thickness of 10 nm yields $\Delta Q = 1.64 \times 10^{12}/cm^2$.

Thin gate oxides mentioned earlier in this chapter also have potential problems with impurity diffusion from the polysilicon gates through the oxide and into the substrates. Any doping that makes it into the substrate is directly in the MOS channel region and will shift the threshold of the devices.

In the past, NMOS depletion-mode ($V_{TN} < 0$) transistors were routinely used in processes designed for high-performance logic applications. To reduce the NMOS threshold voltage, *n*-type impurities can be implanted to form a built-in channel connecting the source and drain regions of the transistor, as shown in Fig. 9.6. The device characteristics of a depletion-mode transistor are similar, although not identical, to those of an enhancement-mode NMOS transistor, and the dose needed to shift the threshold voltage may be estimated using Eq. (9.4).

9.1.7 Field-Region Considerations

The region between the two transistors in Fig. 9.1 is called the *field* region and must be designed to provide isolation between adjacent MOS devices. Several factors must be considered. The metal line over the field region can act as the gate of a "parasitic NMOS transistor" with diffused regions (2) and (3) acting as its source and drain. To ensure that this parasitic device is never turned on, the magnitude of the threshold voltage in this region must be much higher than that in the normal gate region. Referring to Eq. (9.2), we find that the threshold voltage may be made higher by increasing the oxide thickness in the field region and by increasing the doping below the field oxide. The field oxide is typically made three to ten times thicker than the gate oxide of the transistors.

Another problem occurs for NMOS transistors. The substrate for NMOS transistors is p-type, usually doped with boron. We know that thermal oxidation results in depletion of boron from the surface of the silicon, and looking at Eq. (9.2) we see that boron depletion will lower the threshold voltage of the transistors in the field region. A field implant step is often added to processes to increase the threshold voltage and compensate for the boron depletion during field-oxide growth.

For PMOS devices, the substrate is typically doped with phosphorus. During oxidation, phosphorus pileup at the surface tends to increase the threshold voltage in the field region. Thus, phosphorus pileup helps to keep the parasitic field devices turned off.

9.1.8 MOS Transistor Isolation

When two properly biased MOS transistors are placed near each other, they are isolated by reverse-biased source-substrate and drain-substrate junctions, as shown in Fig. 9.7. The MOS devices are referred to as *self-isolated*. No additional structure is required to achieve isolation, and this fact gives an inherent size advantage to MOS technology over the junction isolated bipolar structures discussed in the next chapter. However, to maintain this isolation, the depletion layers surrounding the various

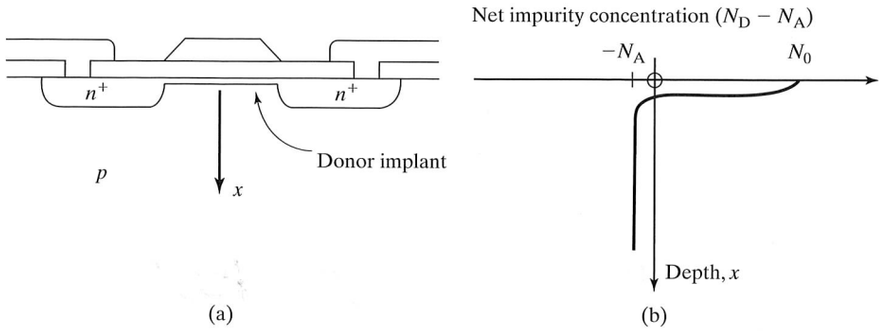


FIGURE 9.6

(a) Formation of a depletion-mode NMOS transistor using a shallow ion-implanted layer; (b) net impurity profile under the gate of the depletion-mode MOSFET.

source–drain diodes must not merge, and this requirement limits the minimum spacing between the devices. The spacing between adjacent transistors must be greater than twice the maximum depletion-layer width.

Example 9.2:

Use Eq. (9.3) to estimate the minimum spacing between the drains of two adjacent NMOS transistors if the substrate doping is $3 \times 10^{16}/\text{cm}^3$ and the maximum drain-substrate voltage is 5 V.

Solution: The n^+p drain-substrate junctions correspond to one-sided step junctions, so the use of Eq. (9.3) is appropriate. The built-in potential is equal to

$$V_{bi} = 0.56\ \text{V} + (0.0258\ \text{V})\ln(3 \times 10^{16}/10^{10}) = 0.94\ \text{V}$$

and the depletion layer width is

$$W_d = \sqrt{2(11.7)(8.854 \times 10^{-14}\ \text{F/cm})(5\ \text{V} + 0.94\ \text{V})/(1.6 \times 10^{-19}\text{C})(3{*}10^{16}/\text{cm}^3)}$$
$$= 0.51\ \mu\text{m}$$

Each transistor has a depletion layer around its drain, so the devices must be separated by at least twice this distance, and the minimum spacing between transistors (with no safety margin) is 1.02 μm.

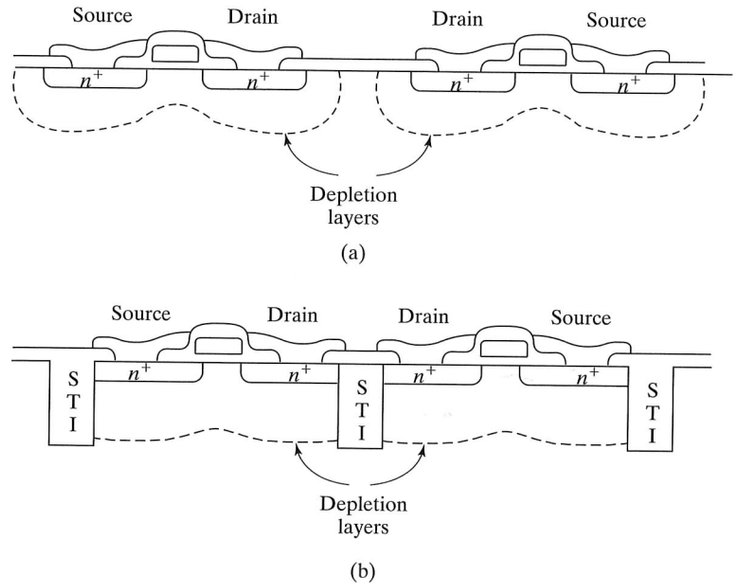


FIGURE 9.7

Isolation strategies (a) Intrinsic (b) Shallow trench isolation

In early technology based upon 2-μm or greater lithography, the spacing calculated in Example 9.2 does not represent a problem. However, for advanced processes with deep submicron feature sizes, this form of isolation is not satisfactory. This led first to the use of recessed oxide technology depicted in Fig. 3.12, and more recently to the pervasive utilization of shallow trench isolation (STI) in both MOS and bipolar technologies. For the STI depicted in Fig. 9.7(b), the depletion layers are effectively cut off and separated by the oxide. The minimum space between drain diffusions is now set by the minimum width of the STI region and can ideally approach a minimum feature size in the technology. Highly planar STI regions are produced using the CMP process described in Chapter 3. Note that the *pn* junction boundary intersects the STI oxide in Fig. 9.7(b). This interface represents a potential junction leakage site, but it is not a problem with well controlled processing. Note that *pn* junctions in MOS and bipolar transistors have always intersected the oxide at the surface of the silicon. (See Fig. 9.7(a).)

9.1.9 Lightly Doped Drain Structures

As devices are scaled to smaller dimensions, the substrate doping level tends to be increased, very shallow junctions with a high curvature are used, and the applied electric fields tend to increase. All of these factors tend to cause breakdown problems with the drain-substrate junction. A number of lightly doped drain (LDD) structures have evolved to control the breakdown problem. The concept is depicted in Fig. 9.8. After defining the polysilicon gate, we use an *n*-type implantation to form the LDD extension that ultimately defines the extent of the channel. An oxide or nitride "spacer" is formed on the edges of the gate by thermal oxidation, or CVD process, and then the highly doped source and drain contact regions are implanted. The reduced doping in the LDD region enhances the breakdown voltage of the transistor. A wide array of different process have been developed to achieve structures similar to that in Fig. 9.8.

9.1.10 MOS Transistor Scaling

The phenomenal increase in IC density and complexity has been driven by our ability to aggressively scale the physical dimensions (W, L, X_O, x_j, etc.) of the MOS transistor. A theoretical framework for MOSFET miniaturization was first provided by Dennard,

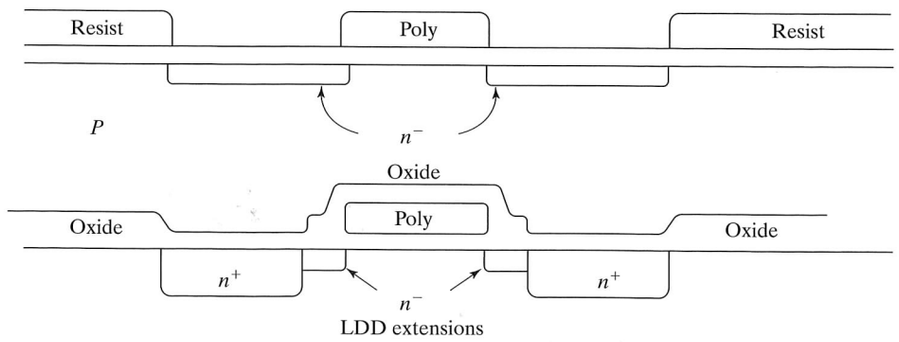


FIGURE 9.8

Self aligned polysilicon-gate transistor with lightly-doped source/drain regions

Gaensslen, Kuhn and Yu [22]. The basic tenant of the theory requires the electrical fields to be maintained constant within the device as the geometry is changed. Thus, if a physical dimension is reduced by a factor of α, then the voltage applied across that dimension must also be decreased by the same factor.

These rules are applied to the transconductance and linear region drain current for the MOSFET in Eq. (9.5) in which the three physical dimensions—W, L and X_O—are all reduced by the factor α, and each of the voltages including the threshold voltage is reduced by the same factor. For the *n*-channel MOSFET, we have

$$I_D^* = \bar{\mu}_n \frac{K_O \epsilon_O}{\left(\frac{X_O}{\alpha}\right)} \frac{\left(\frac{W}{\alpha}\right)}{\left(\frac{L}{\alpha}\right)} \left(\frac{V_{GS}}{\alpha} - \frac{V_{TN}}{\alpha} - \frac{V_{DS}}{2\alpha}\right) \frac{V_{DS}}{\alpha} = \frac{I_D}{\alpha}, \tag{9.5}$$

$$K_n^* = \bar{\mu}_n \frac{K_O \epsilon_O}{\left(\frac{X_O}{\alpha}\right)} \frac{\left(\frac{W}{\alpha}\right)}{\left(\frac{L}{\alpha}\right)} = \alpha \bar{\mu}_n \frac{K_O \epsilon_O}{X_O} \frac{W}{L} = \alpha K_n.$$

We see that the scaled drain current is actually reduced from the original value by the scale factor α, whereas the scaled transconductance parameter K_n^* is increased by the scale factor. In a similar manner, the total gate-channel capacitance of the device is also found to be reduced by α:

$$C_{GC}^* = (C_{OX}^*)^* W^* L^* = \frac{K_O \epsilon_O}{\left(\frac{X_O}{\alpha}\right)} \frac{\left(\frac{W}{\alpha}\right)}{\left(\frac{L}{\alpha}\right)} = \frac{C_{GC}}{\alpha} \tag{9.6}$$

We know that the delay of logic gates is limited by the transistor's ability to charge and discharge the capacitance associated with the circuit. Based upon $i = Cdv/dt$, an estimate of the delay of a scaled logic circuit is

$$\tau^* = C_{GC}^* \frac{\Delta V^*}{I_D^*} = \frac{C_{GC}}{\alpha} \frac{\frac{\Delta V}{\alpha}}{\frac{I_D}{\alpha}} = \frac{\tau}{\alpha} \tag{9.7}$$

We find that circuit delay is also improved by the scale factor α.

As we scale down the dimensions by α, the number of circuits in a given area will increase by a factor of α^2. An important concern in scaling is therefore what happens to the power per circuit, and hence the power per unit area (power density) as dimensions are reduced. The total power supplied to a transistor circuit will be equal to the product of the supply voltage and the transistor drain current:

$$P^* = V_{DD}^* I_D^* = \left(\frac{V_{DD}}{\alpha}\right)\left(\frac{I_D}{\alpha}\right) = \frac{P}{\alpha^2} \text{ and } \frac{P^*}{A^*} = \frac{P^*}{W^* L^*} = \frac{\frac{P}{\alpha^2}}{\left(\frac{W}{\alpha}\right)\left(\frac{L}{\alpha}\right)} = \frac{P}{WL} = \frac{P}{A} \tag{9.8}$$

This equation is extremely important. It indicates that the power per unit area remains constant if a technology is properly scaled. Even though we are increasing the number of circuits by α^2, the total power for a given size integrated circuit die will remain constant. Violation of the scaling theory over many years, by maintaining a constant 5-V power supply as dimensions were reduced, led to almost unmanagable power levels in many of today's integrated circuits. The problem could only be resolved by moving away from NMOS technology and into CMOS technology!

A useful figure of merit for comparing logic families is the power-delay product (PDP). The product of power and delay time represents energy, and the PDP represents a measure of the energy required to perform a simple logic operation:

$$PDP^* = P^* \tau^* = \frac{P}{\alpha^2}\frac{\tau}{\alpha} = \frac{PDP}{\alpha^3} \tag{9.9}$$

The PDP figure of merit shows the full power of technology scaling. The PDP is reduced by the cube of the scaling factor!

Each generation of lithography corresponds to a scale factor $\alpha = \sqrt{2}$, so each new technology generation increases the number of circuits by a factor of 2 and improves the PDP by a factor of almost 3. Table 9.1 summarizes the performance changes achieved with constant field scaling.

9.2 MOS TRANSISTOR LAYOUT AND DESIGN RULES

Design of the layout for transistors and circuits is constrained by a set of rules called the design rules, or ground rules. These rules are technology specific and specify minimum sizes, spacings, and overlaps for the various shapes that define transistors. Processes are designed around a *minimum feature size*, which is the width of the smallest line or space that can be reliably transferred to the surface of the wafer using a given generation of lithography.

To produce a basic set of ground rules, we must also know the maximum misalignment that can occur between two mask levels. Figure 9.9(a) shows the nominal position of a metal line aligned over a contact window. The metal overlaps the contact window by at least one *alignment tolerance* in all directions. During the fabrication process, the alignment will not be perfect, and the actual structure may have misalignment in both the *x*- and *y*-directions. Figures 9.9(b)–(d) show the result of worst-case misalignment of the patterns in the *x*-, *y*-, and both directions simultaneously. Our set of design rules will assume that this alignment tolerance is the same in both directions.

TABLE 9.1 Constant Electric Field Scaling Results

Performance Measure	Scale Factor
Area/Circuit	$1/\alpha^2$
Transconductance Parameter	α
Current	$1/\alpha$
Capacitance	$1/\alpha$
Circuit Delay	$1/\alpha$
Power/Circuit	$1/\alpha^2$
Power/Unit Area (Power Density)	1
Power-Delay Product (PDP)	$1/\alpha^3$

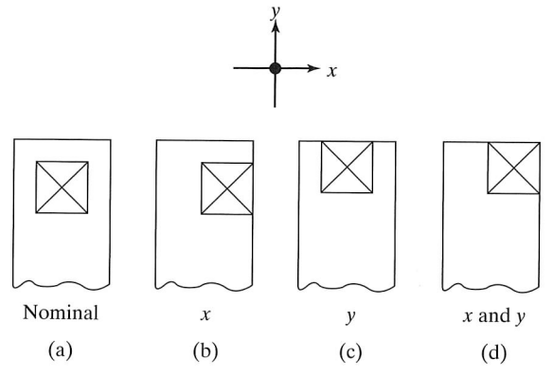


FIGURE 9.9
(a) Nominal alignment of the contact and metal masks; (b) worst-case misalignment in the *x*-direction, (c) in the *y*-direction, and (d) in both directions.

9.2.1 Metal-Gate Transistor Layout

The first successful MOS technologies utilized aluminum for the gate material. Although these metal-gate devices are seldom used in today's silicon processes, an understanding of their layout provides significant insight into the limitations of the metal gate process and the importance of the self-aligned silicon-gate technologies that replace them. The low melting temperature of aluminum greatly limits the type of processing steps that can be used following the metal deposition step. Refractory metals such as tungsten, which can withstand very high temperatures, have been used in experimental self-aligned metal-gate MOS processes.

Figure 9.10 shows the process sequence for a basic metal-gate process. The first mask defines the position of the source and drain diffusions. Following diffusion, the second mask is used to define a window for growth of the thin gate oxide. The third and fourth masks delineate the contact openings and metal pattern. The metal-gate mask sequence, omitting the final passivation layer mask, is as follows:

1. Source/drain diffusion mask	First mask
2. Thin oxide mask	Align to level 1
3. Contact window mask	Align to level 1
4. Metal mask	Align to level 2

An alignment sequence must be specified in order to properly account for alignment tolerances in the ground rules. In this metal-gate example, mask levels two and three are aligned to the first level, and level four is aligned to level two.

We will first look at a set of design rules for metal-gate transistors similar in concept to the rules developed by Mead and Conway [6]. These ground rules were designed to permit easy movement of a design from one generation of technology to another by simply changing the size of a single parameter λ. In order to achieve this goal, the rules are quite loose in terms of level-to-level alignment tolerance. We will explore tighter ground rules later in this chapter.

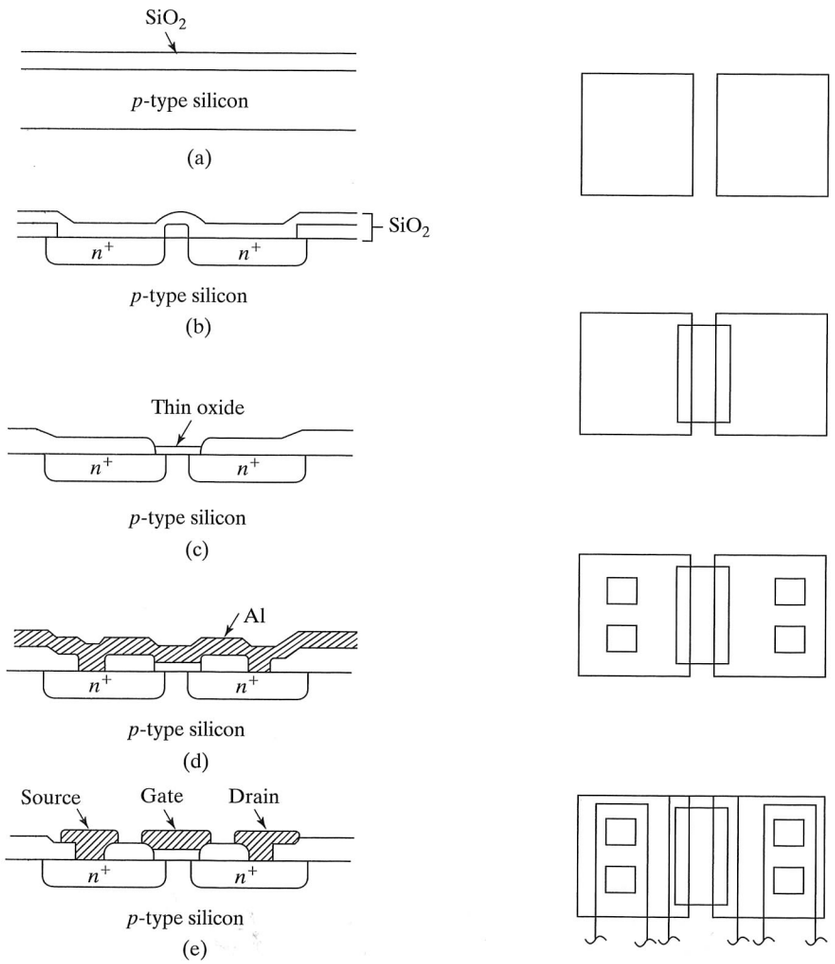


FIGURE 9.10

Mask steps and device cross sections in a metal-gate process. (a) Substrate ready for first mask step; (b) substrate following source/drain diffusion and oxide regrowth, (c) following gate-oxide growth, (d) following contact window mask and aluminum deposition, and (e) following metal delineation.

A set of metal-gate rules is shown in Fig. 9.11. The minimum feature size $F = 2\lambda$, and the alignment tolerance $T = \lambda$. The parameter λ could be 1 μm, .25 μm, or 0.1 μm, for example. Transistors designed using our ground rules will fail to operate properly if the misalignment exceeds the specified alignment tolerance T.

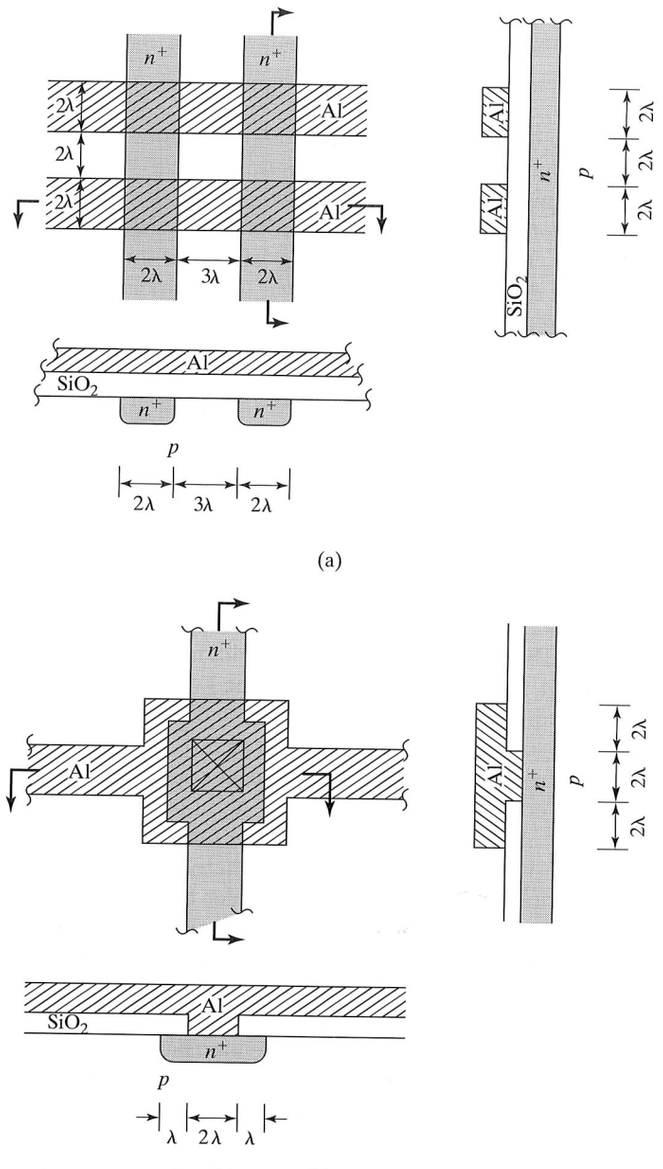


FIGURE 9.11

A simple "λ–based" set of "design rules" or "ground rules" based on an alignment sequence in which levels 2 and 3 are aligned to level 1 and level 4 is aligned to level 2. (a) Rules for metal and diffused interconnection lines; (b) rules for contacts between metal and diffusion.

On the metal level, minimum line widths and spaces are equal to 2λ. In some processes, the metal widths are made larger, because the metal level encounters the most mountainous topology of any level.

On the diffusion level, the minimum linewidth is 2λ. The minimum space between diffusions is increased to 3λ to ensure that the depletion layers of adjacent lines do not merge together. However, the spacing between the source–drain diffusions of a transistor may be 2λ.

In this set of rules, the alignment tolerance between two mask levels is assumed to be 1λ, which represents the maximum shift of one level away from its nominal position, relative to the level to which it is being aligned. A 1λ shift can occur in both the *x*- and *y*-directions.

Square contacts are a minimum feature size of 2λ in each dimension. It is normal practice to ensure that the contact is completely covered by metal even for worst-case alignment. Depending on the alignment sequence, a 1λ or 2λ metal border will be required around the contact window. Likewise, a contact window must be completely surrounded by a 1λ or 2λ border of the diffused region beneath the contact.

For our metal-gate transistors, the thin oxide region will be aligned to diffusion, so it requires a 1λ overlap over the source–drain diffusions in the length direction. The source–drain regions must also extend past the thin oxide by at least 1λ in the width direction. Contacts must be inside the diffusions by 1λ. The metal level is aligned to the thin oxide level, whereas the contacts are aligned to the diffusion level. A worst-case layout therefore requires a 2λ border of metal around contact windows, but only a 1λ border around the thin oxide regions.

Figure 9.12 shows the horizontal layout and vertical cross section of a minimum-size NMOS metal-gate transistor with $W/L = 10\lambda/2\lambda = 5/1$ at the mask level. The two diffusions are spaced by a minimum feature size of 2λ. Thin oxide must overlap the diffusions by 1λ in the length direction and underlap the diffusions by 1λ in the width direction. Metal must overlap thin oxide by 1λ. Accumulated alignment tolerances cause the minimum width of the gate metal to be 6λ. The spacing between metal lines must be 2λ. The metal over the contact holes must be 8λ wide, because of the alignment sequence used, and the contact hole must be 1λ inside the edge of the diffusion. The resulting minimum transistor is 26λ in the length direction and 16λ in the width direction.

A new design rule has been introduced into this layout. The gate metal is spaced 1λ from the diffusion to prevent the edge of a metal line from falling directly on top of the edge of the diffusion in the nominal layout.

Several observations can be made by looking at this structure. First, note that the transistor is $416\lambda^2$ in total area, whereas the active channel area of the device is $20\lambda^2$! The rest of the area is required in order to make contacts to the various regions, within the constraints of the minimum feature size and alignment tolerance rules. Second, there is a substantial area of thin and thick oxide in which the gate metal overlaps the source and drain regions of the transistor. This increases the gate-to-source and gate-to-drain capacitance of the transistor. In this metal-gate transistor layout, the channel is defined by the junction edges in the length direction and by the thin oxide region in the width direction.

It should also be noted that there are several small contact windows in the source and drain regions. The usual practice is to make all the contact windows the same size throughout the wafer. From a processing point of view, equal-size contact windows will all tend to open at the same time during the etching process. The uniform size of the contacts also facilitates modeling of the contact resistance as the area of the diffusion is changed.

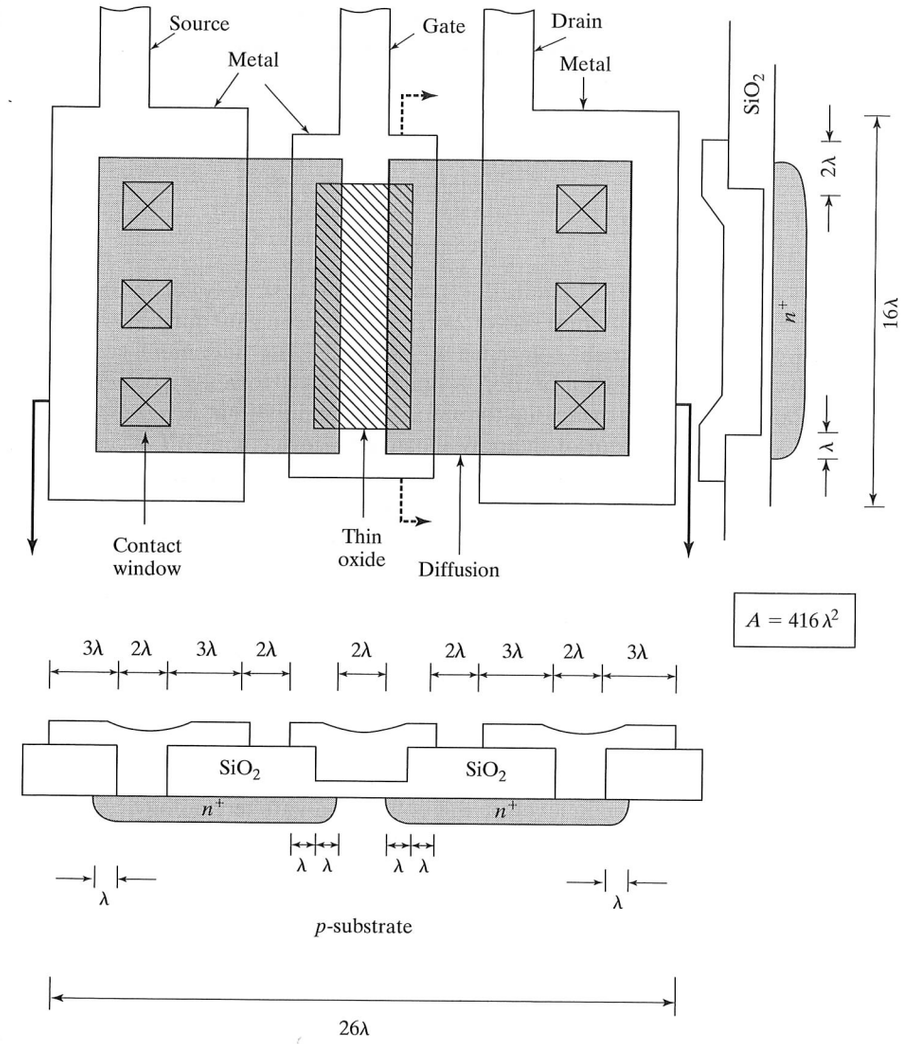


FIGURE 9.12

Minimum-size metal-gate transistor with *W/L* ratio of 5/1 using the design rules of Fig. 9.11. The active gate region is less than 5% of the total device area.

9.2.2 Polysilicon-Gate Transistor Layout

Transistors fabricated using polysilicon-gate technology have a number of important advantages over those built using metal-gate processes. The polysilicon gate can withstand high-temperature processing following its deposition, and this significantly improves the flexibility of the process. The silicon gate can be directly oxidized at high temperature to form an insulating layer over the gate. The heavily doped polysilicon represents an additional interconnect layer that other metal layers can easily cross, because of the oxide isolation. However, the most significant advantages are in layout

and parasitic capacitance reduction, and we will discover some of these advantages by looking at the layout and structure of the polysilicon-gate transistor.

The mask sequence for the basic polysilicon-gate process from Chapter 1 is (again without passivation layer) as follows:

1. Active region (thin oxide) mask	First mask
2. Polysilicon mask	Align to level 1
3. Contact window mask	Align to level 2
4. Metal mask	Align to level 3

Some new design rules must be introduced for this process. Polysilicon lines and spaces will both be a minimum feature size of 2λ. The polysilicon gate must overlap the thin oxide region by an alignment tolerance λ. The preceding alignment sequence requires 1λ polysilicon and 1λ metal borders around contacts. However, contact holes should have a 2λ border of thin oxide due to tolerance accumulation.

Figure 9.13 shows the layout of the polysilicon-gate device with $WL = 5/1$ using these design rules. The total area is $168\lambda^2$. The active channel region now represents 12% of the total area, compared with less than 5% for the metal-gate device. The polysilicon gate acts as a barrier material during source–drain implantation and results in the self-alignment of the edge of the gate to the edge of the source–drain regions. Self-alignment of the gate to the channel reduces the size of the transistor and eliminates the overlap region between the gate and the source–drain regions. In addition, the size of the transistor is reduced, because the source–drain metallization can be placed nearer to the gate. In the polysilicon-gate layout, the channel is defined by the polysilicon gate in the length direction and by the thin oxide in the width direction.

A very important side benefit resulting from this process is the third level of interconnection provided by the polysilicon. Circuit wiring may be accomplished on the diffusion, metal, and polysilicon levels in the polysilicon-gate technology.

A design rule concerning edges has again been introduced into this layout. Metal lines are spaced 1λ from the polysilicon gate to prevent the edge of the metal line from falling directly on top of the edge of the polysilicon line in the nominal layout.

9.2.3 More-Aggressive Design Rules

The design rules discussed so far have focused on minimum feature size and alignment tolerance. *F* and *T* are determined primarily by the type of lithography being practiced. However, linewidth expansion and shrinkage throughout the process also strongly affect the ground rules. Expansion or shrinkage may occur during mask fabrication, resist exposure, resist development, etching, or diffusion. These linewidth changes are normally factored into the design rules.

In addition, alignment variation is a statistical process. Worst-case misalignments occur only a very small percentage of the time. (For a Gaussian distribution, a 3σ misalignment occurs only 2% of the time.) Our set of rules based on worst-case alignment tolerances is very pessimistic. For example, assuming that contacts are misaligned by λ in one direction, at the same time that the metal level is misaligned in the opposite direction by λ, results in an accumulated tolerance of 2λ. However, this situation would most probably never occur.

Let us consider the impact of tightening two design rules in the polysilicon-gate process. First, we will let the edge of one layer align with the edge of another layer.

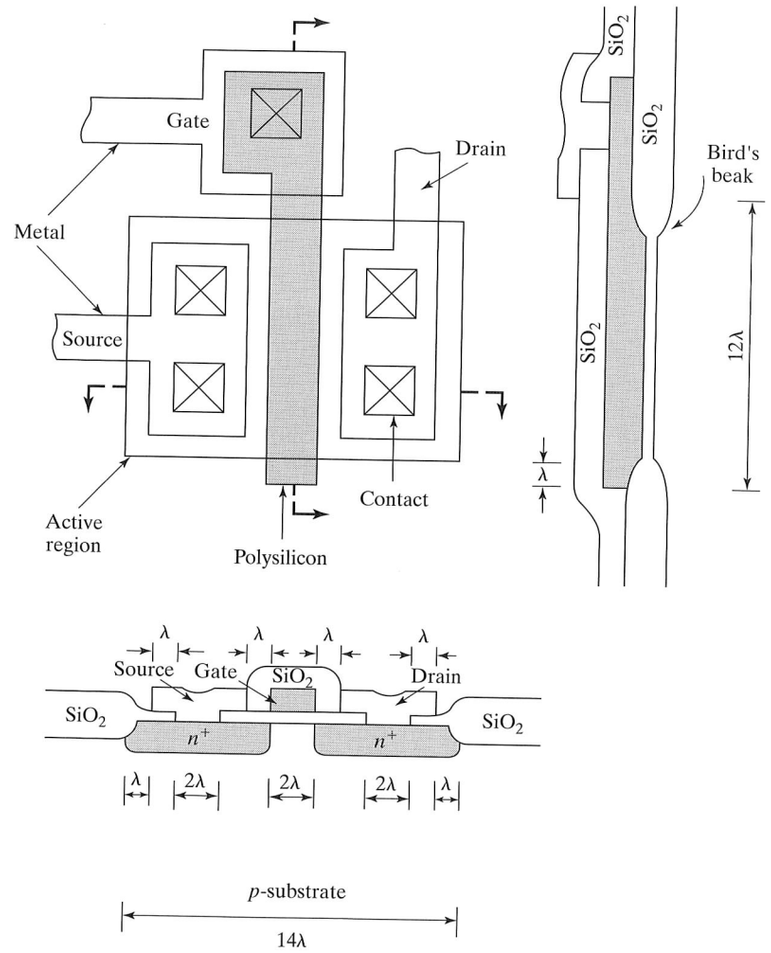


FIGURE 9.13

Minimum-size polysilicon-gate transistor layout for $W/L = 5/1$. The active gate region occupies 12% of the transistor area, and parasitic gate capacitance is minimized.

Second, a contact window will be allowed to run over onto the field oxide by 1λ. The resulting layout using our polysilicon-gate alignment sequence is shown in Fig. 9.14. The total area of the device has been reduced 25% to $120\lambda^2$, and the active channel region now represents 17% of the total transistor area. We see how ground rule changes can have a substantial effect on device area.

9.2.4 Channel Length and Width Biases

Figure 9.15 presents another example of the interaction of the process with design-rule definitions. Here we assume a polysilicon-gate process in which the source–drain junction depth is equal to λ/2 and lateral diffusion equals vertical diffusion. Since we know

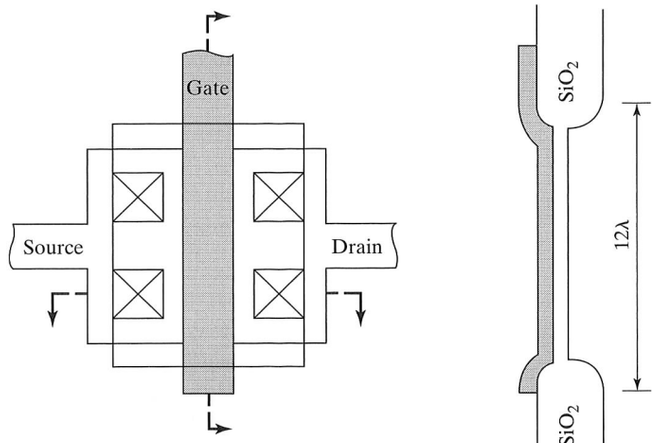


FIGURE 9.14

More aggressive layout of the polysilicon-gate transistor in which two ground rules have been relaxed. Active gate area is now 17% of total device area.

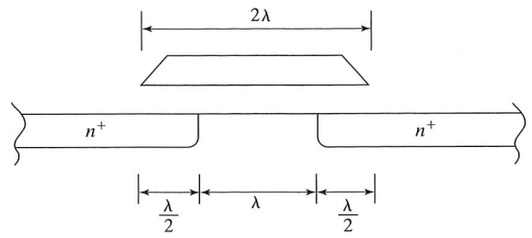


FIGURE 9.15

Channel-length bias in a polysilicon-gate NMOS device caused by lateral diffusion under the edge of the diffusion window. The transistor has $L = 2\lambda$ at the mask level but ends up with an actual $L = \lambda$ after the device is fabricated.

that the source–drain diffusions will move laterally under the edge of the oxide openings, the contact windows can be aligned within $\lambda/2$ of the edge of the diffusions at the mask level, but will still be 1λ within the border of the diffusion in the final structure.

However, lateral diffusion requires the length of channel at the mask level to be increased by λ to achieve the same electrical channel length in the device. The actual channel length $L = L_m - \Delta L$, where L_m is the channel length as originally drawn on the mask and ΔL is the channel-length shrinkage that occurs during processing. This is an important area where the process must be controlled. For devices with short channel lengths, ΔL may be so severe that the devices become unusable. For the layout of Fig. 9.15 width $W_m = 10\lambda$, $W_m/L_m = 10\lambda/2\lambda = 5/1$ at the mask level, whereas $W/L = 10\lambda/\lambda = 10/1$ in the fabricated transistor.

The development of self-aligned polysilicon-gate technology with ion-implanted source–drain regions was a major improvement. The polysilicon-gate process significantly reduces both the channel shrinkage caused by lateral diffusion and the overlap capacitance resulting from alignment tolerances in the metal-gate process.

In Fig.9.14, one can see another source of channel bias. The "bird's beak" reduces the size of the active region to below that defined by the active region mask, and it introduces a process bias into the channel width of the polysilicon-gate transistor. $W = W_m - \Delta W$, where W_m is the width at the mask level and ΔW is the channel-width shrinkage during processing.

In sets of very tight design rules developed for high-volume-production ICs such as dynamic memories, all critical dimensions are adjusted to account for the processing and alignment sequences. This often results in a layout that must conform to a set of 50 to 100 design rules [7]. Such a set of design rules is highly technology-specific and cannot be transferred from one generation of lithography to the next. The Mead-and-Conway-style rules [6] reach a compromise between a set of rules that is overly pessimistic and wastes a lot of silicon area, and one that is extremely complex, but squeezes out all excess area. The Mead-and-Conway-style design rules are used for ICs in which design time, and not silicon area, is of dominant importance.

9.3 COMPLEMENTARY MOS (CMOS) TECHNOLOGY

Complementary MOS (CMOS) technology is arguably the most commercially important silicon technology. It came to the forefront in the mid 1980s when its low-power benefits finally outweighed the perceived increase in process complexity. Today, scaling of CMOS to submicron dimensions has made the technology highly competitive not only in term of power, but also in raw speed.

9.3.1 The *n*-Well Process

The basic CMOS process of Fig. 1.8 requires a *n*-well diffusion and formation of both NMOS and PMOS transistors. Substrate resistivity is chosen to give the desired NMOS characteristics, and an additional implant step may be introduced to adjust the NMOS threshold separately. The *n*-well-to-substrate junction may range from a few microns to as much as 20 microns in depth. The net surface concentration of the *n*-well must be high enough above the substrate concentration to provide adequate process control without severely degrading the mobility and threshold voltage of the PMOS transistors. The surface concentration of the *n*-well typically ranges between 3 and 10 times the substrate impurity concentration. An additional implant step is often introduced to adjust the PMOS threshold voltage.

9.3.2 *p*-Well and Twin-Well Processes

The first successful CMOS technologies actually utilized *p*-well processes whose structures are simply a mirror image of Fig. 1.8. However, the drive toward ever higher performance led to the development of the *n*-well processes in which the NMOS transistors are placed in the lightly doped substrate region where the *n*-channel

mobility will be the highest. More recently, twin-well processes, such as in Fig. 9.16, have been developed that permit individual optimization of the characteristics of both the *n*- and *p*-channel devices [10].

A lightly doped *n*- or *p*-type epitaxial layer is grown on a heavily doped *n*- or *p*-type substrate. (Lightly doped *n*- and *p*-type regions are often referred to as *v* and π regions, respectively.) Separate implantations and diffusions are used to form wells for both the NMOS and PMOS transistors. The low-resistivity substrate substantially reduces the substrate resistance R_s and improves latchup resistance as discussed later.

9.3.3 Gate Doping

Early polysilicon gate CMOS processes used n^+ polysilicon gates for both transistors as assumed in the graph of Fig. 9.2. In many processes, it was found that use of an n^+ gate on a PMOS transistor led to formation of a buried channel rather than a surface channel device that causes problems with subthreshold turn-off of the device. With the push toward optimization of both devices with the advent of twin-well processes, p^+ doped polysilicon gates were introduced. With use of the p^+ gate, the PMOS device characteristics become more symmetrical to those of the NMOS devices, except for the inherent mobility differences. The threshold voltages also become symmetrical (See Prob. 9.4). Note that the twin-well process depicted in Fig. 9.16 produces p^+ and n^+ polysilicon gates. The p^+ gate is protected by the photoresist layer during the n^+ implant.

Example 9.3

An n^+ polysilicon gate CMOS process uses an *n*-type substrate with a doping of $10^{16}/\text{cm}^3$. An implant/drive-in schedule will be used to form a *p*-well with a net surface concentration of $10^{17}/\text{cm}^3$ and a junction depth of 3 μm. (**a**) What is the drive-in time at 1150 °C? (**b**) Solve for the implanted dose in silicon. (**c**) What are the threshold voltages of the *n*- and *p*-channel transistors, if the oxide thickness is 10 nm?

Solution: The 3-μm junction depth and low surface concentration suggest that the well has a Gaussian profile resulting from a two-step diffusion or implant/diffusion process. A final surface concentration of $1.1 \times 10^{17}/\text{cm}^3$ is required to produce a net concentration of $1 \times 10^{17}/\text{cm}^3$ at the surface. Solving for the *Dt* product yields

$$Dt = x_j^2/4 \ \ln(N_0/N_B) = 9.38 \times 10^{-9}\ \text{cm}^2.$$

At 1150 °C, $D = 8.87 \times 10^{-13}$ cm²/sec, which gives $t = 2.94$ h. The dose in silicon is given by $Q = N_O\sqrt{\pi Dt} = 1.89 \times 10^{13}$ /cm². The *p*-channel devices reside in the *n*-type substrate with a doping concentration of $10^{16}/\text{cm}^3$. From Fig. 9.2, the threshold voltage will be −1.1 V. The deep well diffusion will be almost constant near the surface with a value of $10^{17}/\text{cm}^3$. Figure 9.2 yields an *n*-channel threshold of 0.4 V. A threshold adjustment implant would be needed in this process to increase the *n*-channel threshold voltage.

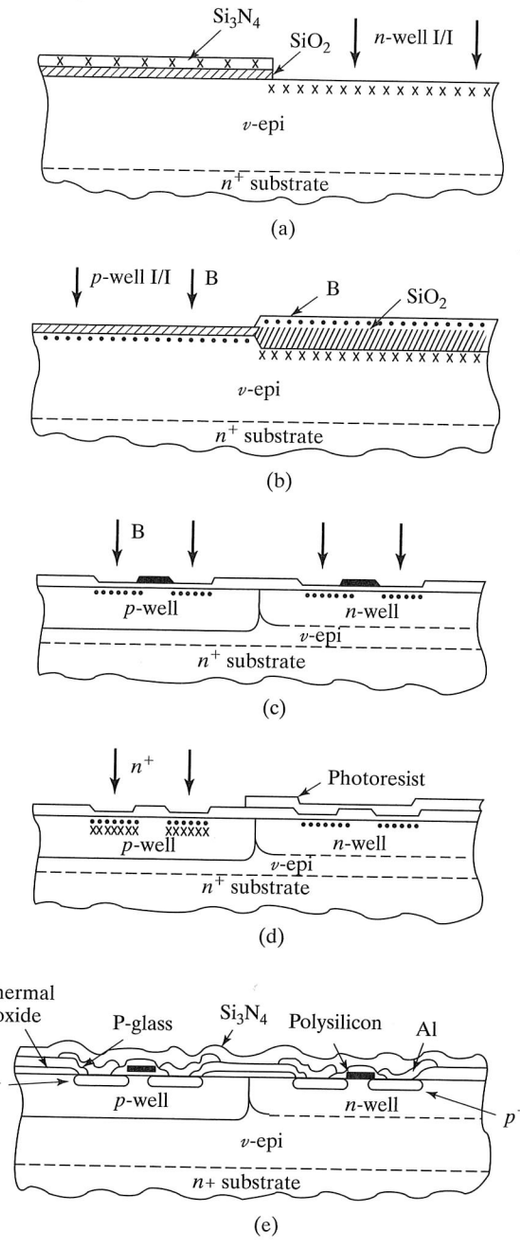


FIGURE 9.16

Twin-well CMOS structure at several stages of the process. (a) *n*-well ion implant; (b) *p*-well implant; (c) nonselective p^+ source/drain implant; (d) selective n^+ source/drain implant using photoresist mask; (e) final structure. Copyright 1980, IEEE. Reprinted with permission from Ref. [10].

9.3.4 CMOS Isolation

In order to maintain isolation in CMOS technology, the transistors must not only be separated from each other as described in Section 9.2.4, but they must also be separated from the edge of the well, as depicted in Fig. 9.17. The minimum spacing is equal to the sum of the depletion layers surrounding the drain diffusions plus the depletion layer of the well-substrate junction. An estimate of the minimum spacing is given in Example 9.4.

Example 9.4

Use Fig. 9.4 to estimate the minimum spacing between the drains of an adjacent NMOS and PMOS transistors in a CMOS process if the substrate doping is $3 \times 10^{16}/\text{cm}^3$, the well doping is $3 \times 10^{17}/\text{cm}^3$, and the maximum drain-substrate voltage is 3.3 V. Assume that the well is also reverse biased by 3.3 V.

Solution: For Fig. 9.4, we need the built-in potential that can be estimated from Eq. (9.3). For the NMOS device,

$$V_{bi} = 0.56\,\text{V} + (0.0258\,\text{V})\ln(3 \times 10^{17}/10^{10}) = 1.0\,\text{V},$$

and for the PMOS transistor and the well-substrate junction,

$$V_{bi} = 0.56\text{V} + (0.0258\,\text{V})\ln(5 \times 10^{16}/10^{10}) = 0.96\,\text{V}.$$

(Note that the well-substrate junction is not modeled as accurately as the other junctions by the step junction formula.) Using Fig. 9.4, we get the following depletion layer estimates:

$$\left(\frac{V_A + V_{bi}}{N_B}\right) = \frac{3.3\,\text{V} + 0.96\,\text{V}}{5 \times 10^{16}/\text{cm}^3} = 8.5 \times 10^{-17}\,\text{V}-\text{cm}^3 \rightarrow w_d = 0.33\ \mu\text{m},$$

$$\left(\frac{V_A + V_{bi}}{N_B}\right) = \frac{3.3\text{V} + 1.0\,\text{V}}{3 \times 10^{17}/\text{cm}^3} = 1.4 \times 10^{-17}\,\text{V}-\text{cm}^3 \rightarrow w_d = 0.13\ \mu\text{m}.$$

The depletion layers take up $(0.33 + 0.33 + 0.13) = 0.79$ μm. A safety margin must be added so that the minimum total spacing in this process might be 1.25 μm.

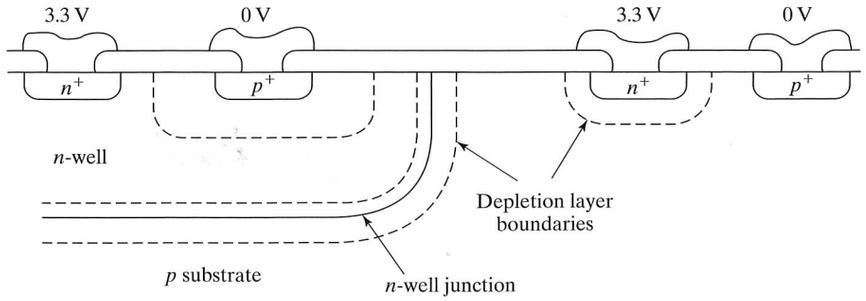


FIGURE 9.17

Minimum spacing requirements required to ensure isolation in an n-well CMOS technology.

In addition to the depletion layer extents, alignment tolerances must be added to the total spacing to ensure that minimum spacing is maintained under worst-case alignment errors. Also, diffusion of deep wells leads to significant lateral diffusion of the well boundary that must be taken into account in the CMOS layout.

9.3.5 CMOS Latchup

Parasitic bipolar devices are formed in the CMOS process in which merged *pnp* and *npn* transistors form a four-layer (*pnpn*) lateral SCR, as shown in Fig. 9.18. If this SCR is turned on, the device may destroy itself via a condition called *latchup* [8, 9]. The *n*-well depth and the spacings between the source–drain regions and the edge of the *n*-well must be carefully chosen to minimize the current gain of the bipolar transistors and the size of the shunting resistors R_s and R_w. A CMOS process will have a number of additional ground rules not present in an NMOS or PMOS process. A more detailed discussion of the design of bipolar transistors will be given in Chapter 10.

To reduce the resistance of the two shunting resistors, "guard ring" diffusions are sometimes added to the process, as shown in Fig. 9.18. Guard rings can be formed using the source–drain diffusions of the PMOS and NMOS transistors or can be added as separate diffusion steps.

9.3.6 Shallow Trench Isolation

Advanced processes with deep submicron feature sizes make use of shallow trench isolation, as depicted in Fig. 9.19, in which a twin-well process is shown [20]. Both the NMOS and PMOS devices are bounded by the STI oxide region. The STI in combina-

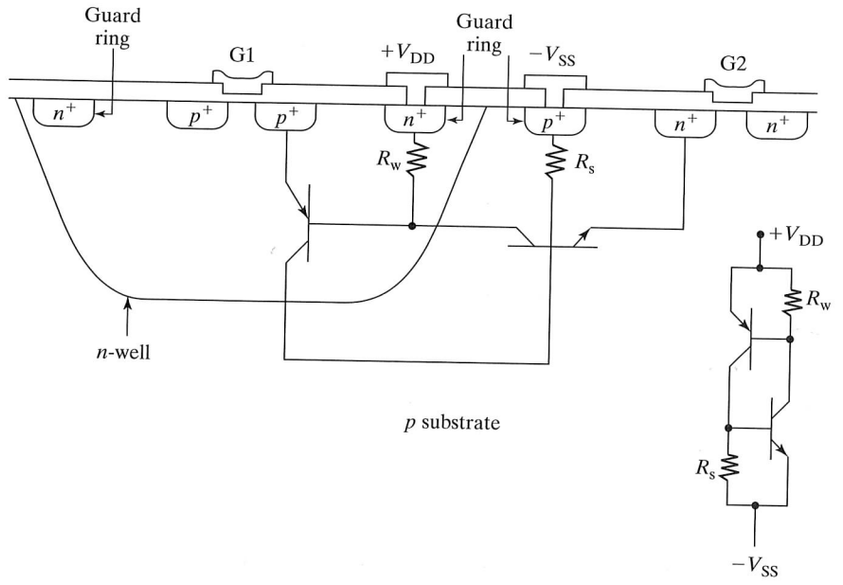


FIGURE 9.18

Cross-section of a CMOS structure, showing the existence of a parasitic lateral *pnpn* SCR and the use of guard rings to reduce the value of R_s and R_w.

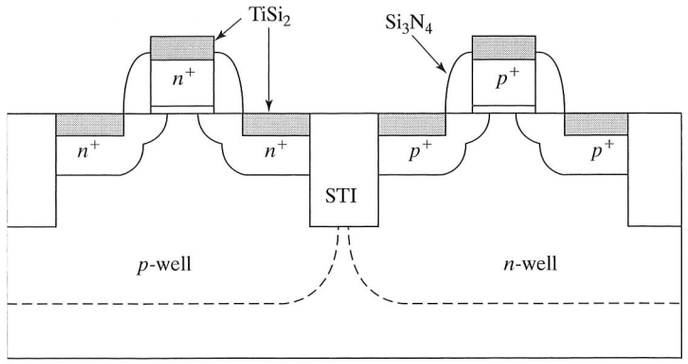


FIGURE 9.19

Application of shallow trench isolation to a twin-well CMOS technology. Copyright *IEEE* 1998. Reprinted with permission from Reference [20].

tion with a heavily doped substrate eliminates the need for guard rings by substantially reducing the current gain of the bipolar devices and decreasing the value of the shunt resistances. LDD extensions can be noted on both devices, as well as the silicon nitride spacers around the perimeter of the gate. Self-aligned tantalum silicide layers are used to reduce the effective sheet resistance of the polysilicon gate, as well as the source and drain regions. This type of CMOS process is being used for 0.18 μm devices and below.

9.4 SILICON ON INSULATOR

Insulating substrates provide the ultimate in device isolation and freedom from latchup problems. The earliest efforts to achieve an insulating substrate grew thin layers (< 10 μm) of single crystal silicon on a sapphire substrate that provides a reasonable match to the silicon crystal lattice. NMOS and PMOS devices were fabricated in the silicon film to produce a CMOS technology. This technology was termed silicon-on-sapphire (SOS) technology. The early attempts were plagued by problems at the silicon–sapphire interface, but the problems were eventually controlled well enough to produce a usable technology.

Our ability to produce a highly controlled silicon–silicon dioxide interface has led to newer forms of silicon-on-insulator (SOI) processes. High-energy ion-implantation can be used to place oxygen atoms in a layer well below the surface of a lightly-doped silicon wafer. Following implantation, the wafers are annealed at elevated temperature to produce a buried oxide layer well below the silicon surface, as depicted in Fig. 9.20. This technology is often referred to as Separation by Implanted Oxygen or

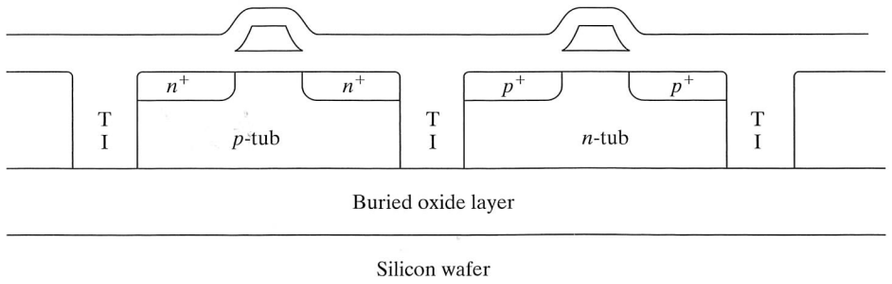


FIGURE 9.20

Trench isolated silicon-on-insulator technology.

SIMOX [19]. Twin-well tubs can then be formed in the lightly doped substrate and separated by trench isolation. NMOS and PMOS devices are then fabricated in the tubs to complete the SOI CMOS process.

Silicon wafer-to-wafer bonding, originally developed for use with MEMs, has also been used to fabricate SOI substrates. A silicon wafer is oxidized to form an insulating SiO_2 layer. A second silicon wafer is brought in contact with the oxidized surface and annealed at elevated temperature to form a bond between the two wafers. Obviously, surface cleanliness is of extreme importance to the success of this process, as indicated in Fig. 9.21. After the bonding is completed, the upper silicon layer is thinned by chemical etching until the desired silicon layer thickness is achieved. An alternative is to use mechanical lapping and polishing processes to thin the silicon wafer.

SUMMARY

In this chapter, we explored the interaction of process design with MOS device characteristics and transistor layout, including the relationships between processing parameters and breakdown voltage, punch-through voltage, threshold voltage, and junction capacitance. A low value of substrate doping is desired to minimize junction capacitance, substrate sensitivity, and junction breakdown voltage, whereas a high substrate doping is needed to maximize punch-through voltage. The use of ion implantation permits the designer to separately tailor the threshold voltage of the transistor.

We have developed basic ideas relating minimum feature size and alignment tolerances and have discussed simple sets of layout design rules. The strong relation between layout design rules and the size of transistors has been demonstrated. Polysilicon-gate technology has been shown to result in a much smaller device area than metal-gate technology for a given transistor W/L ratio, as well as to minimize the parasitic gate capacitance of the device. In addition, the polysilicon-gate process substantially reduces channel-length bias caused by lateral diffusion.

A combination of ion implantation and diffusion is commonly used to form the *p*- or *n*-well required for CMOS technology. VLSI CMOS often uses twin-well processes which permit separate optimization of both the *n*- and *p*-channel devices.

To achieve a high packing density for submicron processes, trench isolation, which provides excellent isolation between devices, is utilized. The source and drain regions of the transistors can be abutted with the oxide isolation regions. The combination of trench isolation and twin-well processes on a heavily doped substrate suppresses latchup and eliminates the need for guard ring diffusions. The ultimate in

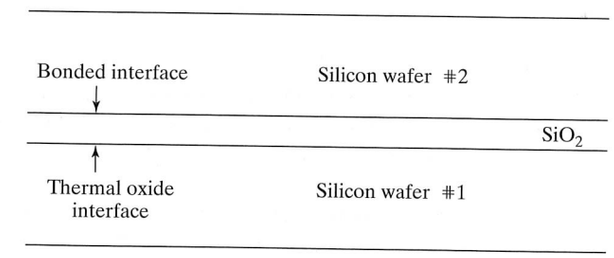


FIGURE 9.21

Formation of bonded wafer SOI

isolation and capacitance reduction is achieved with silicon-on-insulator or SOI substrates. The earliest versions, termed SOS, grew thin silicon layers on sapphire substrates. Today's SOI substrates are formed by high-energy implantation of oxygen or direct wafer-to-wafer bonding followed by chemical etching.

REFERENCES

[1] R.F. Pierret, *Field Effect Devices*, Volume IV in the Modular Series on Solid State Devices, Addison-Wesley, Reading, MA, 1983.

[2] S.A. Abbas and R.C. Dockerty, "*N*-channel Design Limitations due to Hot Electron Trapping," *IEEE IEDM Digest*, pp. 35–38, 1975.

[3] T.H. Ning, C.M. Osburn, and H.N. Yu, "Threshold Instability in IGFETs due to Emission of Leakage Electrons from Silicon Substrate into Silicon Dioxide," *Applied Physics Letters, 29*, 198–199, 1976.

[4] P.E. Cottrell and E.M. Buturla, "Steady State Analysis of Field Effect Transistors via the Finite Element Method," *IEEE IEDM Digest*, pp. 51–54, 1975.

[5] S.M. Sze, *Semiconductor Devices—Physics and Technology*, John Wiley & Sons, New York, 1985.

[6] C.A. Mead and L. Conway, *VLSI Design*, Addison-Wesley, Reading, MA, 1980.

[7] Brian Spinks, *Introduction to Integrated Circuit Layout*, Chapter 7, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[8] A. Ochoa, W. Dawes, and D. Estreich, "Latchup Control in CMOS Integrated Circuits," IEEE Transactions on Nuclear Science, *NS-26*, 5065–5068, December 1979.

[9] R.S. Payne, W.N. Grant, and W.J. Bertram, "The Elimination of Latchup in Bulk CMOS," *IEEE IEDM Digest*, p. 248–251, December 1980.

[10] L.C. Parrillo, R.S. Payne, R.E. Davis, G.W. Reutlinger, and R.L. Field, "Twin-Tub CMOS—A Technology for VLSI Circuits," *IEEE IEDM Digest*, p. 752–755, December 1980.

[11] B.J. Baliga and D.Y. Chen, *Power Transistors: Device Design and Applications*, IEEE Press, New York, 1984.

[12] K.P. Roenker and L.W. Linholm, "An NMOS Test Chip for a Course in Semiconductor Parameter Measurements," *National Bureau of Standards Internal Report* 84–2822, April 1984.

[13] T.J. Russell, T.F. Leedy, and R.L. Mattis, "A Comparison of Electrical and Visual Alignment Test Structures for Evaluating Photomask Alignment in Integrated Circuit Manufacturing," *IEEE IEDM Digest*, p. 7A–7F, December 1977.

[14] D.S. Perloff, "A Four-Point Electrical Measurement Technique for Characterizing Mask Superposition Errors on Semiconductor Wafers," *IEEE Journal of Solid-State Circuits*, SC–13, 436–444, August 1978.

[15] A. Hori et al., "High Speed 0.1 μm Dual Gate CMOS with Low Energy Phosphorus/Boron Implantation and Cobalt Silicide," *IEEE IEDM Technical Digest*, pp. 575–578, December 1996.

[16] H. Hwang, D-H Lee and J.M. Hwang, "Degradation of MOSFETs Drive Current Due to Halo Ion Implantation," *IEEE IEDM Technical Digest*, pp. 567–570, December 1996.

[17] A.J. Auberton-Hervé, "SOI: Materials to Systems," *IEEE IEDM Technical Digest*, pp. 3–10, December 1996.

[18] T. Hashimoto et al., "A 0.2-μm Bipolar-CMOS Technology on Bonded SOI with Copper Metallization for Ultra High-speed Processors," *IEEE IEDM Technical Digest*, pp. 209–212, December 1998.

[19] CMOS: From Bulk to SOI, IBIS Technology Corporation *Technical Note*, 1999, http://www.ibis.com.

[20] S. Yang et al., "A High Performance 180 nm Generation Logic Technology," *IEEE IEDM Digest*, pp. 197–200, December 1998.

[21] W. Maly, *Atlas of IC Technologies: An Introduction to VLSI Processes*, The Benjamin/Cummings Publishing Company, Inc., Menlo Park, CA: 1987.

[22] R. H. Dennard, F. H. Gaenssien, L. Kuhn and H. N. Yu, "Design of Micron MOS Switching Devices," *IEEE IEDM Digest*, pp. 168, December 1972.

PROBLEMS

9.1 What is the maximum gate-to-source voltage that a MOSFET with a 10-nm gate oxide can withstand? Assume that the oxide breaks down at 5 MV/cm and that the substrate voltage is zero.

9.2 Two n^+ diffused lines are running parallel in a substrate doped with 10^{15} boron atoms/cm^3. The substrate is biased to -2 V, and both lines are connected to $+3$ V. Using one-dimensional junction theory, calculate the minimum spacing needed between the lines to prevent their depletion regions from merging. (b) Repeat for a 3×10^{16}/cm^3 doping level.

9.3 Use one-dimensional junction theory to estimate the punch-through voltage of a MOSFET with a channel length of 1 μm. Assume a substrate doping of 3×10^{16}/cm^3 and a substrate bias of 0 V.

9.4 Plot a graph of thresholds similar to Fig. 9.2, but assume that a p^+ gate is used for the PMOS transistors.

9.5 What is the minimum substrate doping required to realize an enhancement-mode NMOS device ($V_{TN} > 0$) with a 10-nm gate oxide?

9.6 Calculate the threshold voltage for the NMOS transistor with the doping profile shown in Fig. P9.6 Assume an n^+ polysilicon-gate transistor with a gate-oxide thickness of 20 nm.

9.7 An implant with its peak concentration at the silicon surface is used to adjust the threshold of an NMOS transistor. We desire to model this implant by a rectangular approximation similar to that of Figure 9.5. Show that $N_i = N_p\pi/4$ and that $x_i = \Delta R_p\sqrt{8/\pi}$ by matching the first two moments of the two impurity distributions.

9.8 A MOS technology is scaled from a 1-μm feature size to 0.25 μm. What is the increase in the number of circuits/cm^2? What is the improvement in the power-delay product?

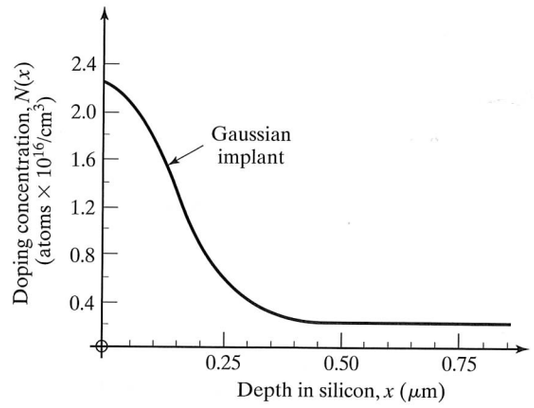


FIGURE P9.6

9.9 Suppose that the voltages are not scaled as the dimensions are reduced by a factor of α? How does the drain current of the transistor change? How do the power/circuit and power density scale?

9.10 High-performance NMOS logic processes used depletion-mode NMOS transistors for load devices. This requires a negative threshold, which can be obtained by implanting a shallow arsenic or phosphorus dose into the channel region. Calculate the arsenic dose needed to achieve a -3-V threshold in an n^+ polysilicon-gate NMOS transistor that has a substrate doping of $3\times10^{16}/\text{cm}^3$ and a gate-oxide thickness of 50 nm.

9.11 Draw the p-well version of the CMOS process in Fig. 1.8.

9.12 Our design rule examples used an alignment tolerance that was one-half the feature size. This ratio represents a very loose alignment capability. Develop a new set of design rules similar to those of Fig. 9.13 for $T=\alpha$ and $F=4\alpha$. Draw the new minimum-size polysilicon-gate transistor using your rules. Compare the area of your transistor with the area of the transistor of Fig. 9.13 if $\lambda=2\alpha$.

9.13 An n-well CMOS process starts with a substrate doping of $3\times10^{15}/\text{cm}^3$. The well doping near the surface is approximately constant at a level of $3\times10^{16}/\text{cm}^3$. The gate-oxide thicknesses are both 15 nm.

(a) Calculate the thresholds of the n- and p-channel transistors using Eqs. (9.2). Assume n^+ polysilicon gates.

(b) Calculate the boron doses needed to shift the NMOS threshold to $+1$ V and the PMOS threshold to -1 V. Assume that the threshold shifts are achieved through shallow ion implantations. Neglect oxide charge.

9.14 Early CMOS logic circuits operated from power supplies of 8 V or more. Estimate the minimum spacing between the drains of adjacent NMOS and PMOS transistors in a CMOS process if the substrate doping is $3\times10^{15}/\text{cm}^3$, the well doping is $5\times10^{16}/\text{cm}^3$, and the maximum drain-substrate voltage is 8 V. Assume that the well is also reverse biased by 8 V.

9.15 A twin-well process starts with a 3-μm-thick, 10-Ω-cm v epi layer on an n^+ substrate. (a) A p-well is to be formed by ion-implantation followed by a drive-in diffusion and is to have a surface concentration of $10^{16}/\text{cm}^3$ with a depth of 2 μm. What are the drive-in time at a temperature of 1075°C and impurity dose in silicon? What is the lateral diffusion distance of the well? (b) A phosphorus n-well is to be formed in the same substrate with a surface concentration of $5\times10^{16}/\text{cm}^3$ and a depth of 1.5 μm. What are the drive-in time at a temperature of 1075°C and impurity dose in silicon? What is the lateral diffusion distance of the n-well? (c) What is the total out-diffusion from the n^+ substrate following the formation of both wells if the substrate is arsenic doped with an arsenic concentration of $10^{20}/\text{cm}^3$?

9.16 Draw a composite view of the situation resulting from a worst-case misalignment of the masks for the MOSFET layout shown in Fig. 9.12. Assume that metal aligns to thin oxide and that thin oxide and contacts align to the diffusion.

9.17 Develop a new set of ground rules for the metal-gate transistor of Section 9.2, assuming that levels 2, 3, and 4 are all aligned to level 1. Redraw the transistor of Fig. 9.12 using your new rules. In what ways is this layout better or worse than that originally given in Fig. 9.12?

9.18 Draw a cross section of a metal-gate NMOS transistor and a composite view of its mask set, assuming an aggressive layout that takes into account all lateral diffusion. Assume a source–drain junction depth of 2.5 μm, and assume that lateral diffusion equals 80% of vertical diffusion. Assume λ is 2 μm and $W/L=10/1$.

9.19 Draw the layout of a three-input NMOS NOR-gate with the dimensions given on the circuit schematic in Fig. P9.19. Be sure to merge diffusions wherever possible. Use the more aggressive ground rules developed for polysilicon-gate devices.

9.20 Draw the layout (top view) of the CMOS inverter in Fig. P9.20 for an n-well technology using the λ-based ground rules from Fig 9.13 for the transistors. In addition, assume that source and drain regions must be a minimum of 8 λ from the edge of the well. What is the total area of the CMOS inverter (in λ^2)? What is the total gate area?

9.21 Repeat Problem 9.20, but this time assume shallow trench isolation with a minimum width of 4λ. Assume that the source and drain regions can butt against the oxide, as in Fig. 9.19.

9.22 A high energy (4 MeV) is used to implant oxygen well below the silicon surface in order to form a buried SiO_2 layer. Assume that the SiO_2 layer is desired to be 0.25 μm wide. (a) What is the oxygen dose required in silicon? (b) What beam current is required to achieve a throughput of five 200 mm wafers per hour? (c) How much power is being supplied to the ion beam?

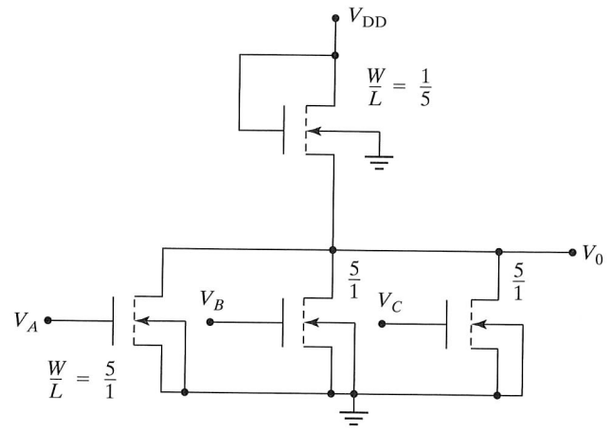


FIGURE P9.19

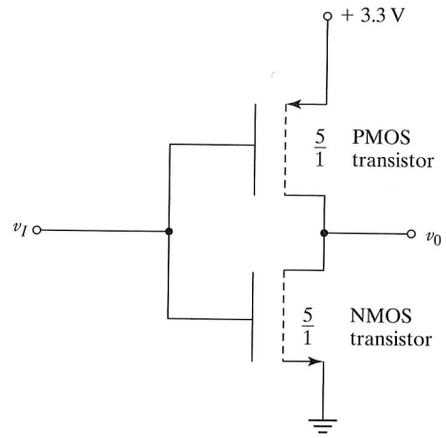


FIGURE P9.20

CMOS inverter with both W/L = 5/1

9.23 A number of types of alignment test structures have been developed [12, 13]. Figure P9.23 shows a simple test structure that can be used to measure the misregistration of the contact window mask relative to the diffusion mask [14]. Two linear potentiometers, one in the horizontal direction and one in the vertical direction, are fabricated using diffused resistors. The distance between contacts A and C is the same as that between C and E, and the contact from pad D is nominally one-half the distance between pads C and E. A current is injected between pads B and F, and the voltages between pads C–D and D–E are measured.

(a) Show that the misregistration in the y-direction is given by $\Delta Y = 1/2\ L\ (V_{DE} - V_{CD})/V_{AC}$.

(b) Derive a similar relationship for misregistration in the x-direction.

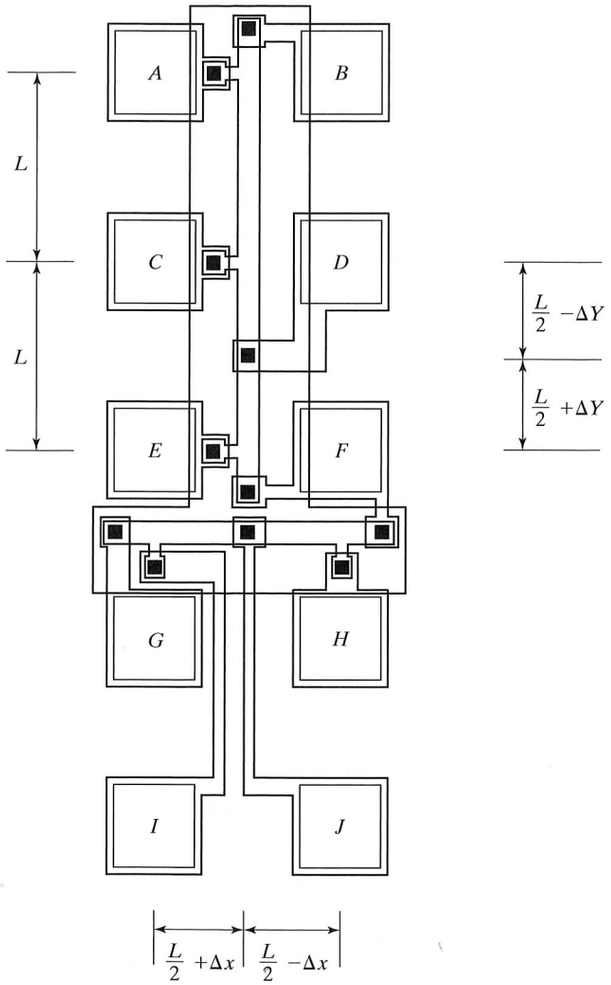


FIGURE P9.23

CHAPTER 10

Bipolar Process Integration

In this chapter, interactions between fabrication processes and bipolar device design and layout will be explored. In particular, we will look closely at relationships between impurity profiles and device parameters such as current gain, transit time, and breakdown voltage. Basic design rules for bipolar structures are introduced. The use of recessed oxidation, deep and shallow trenches, polysilicon electrodes and self-aligned processes in the formation of high-performance bipolar transistors will be presented. Dielectric and collector-diffused isolation processes are discussed, as well as silicon-germanium epitaxial-base transistors and advanced BiCMOS technologies, which provide bipolar and CMOS devices.

10.1 THE JUNCTION-ISOLATED STRUCTURE

The classic SBC process provides a backdrop for understanding the limitations of the basic bipolar transistor, as well as the structure of various other devices that are fabricated in bipolar IC processes. The basic junction-isolated bipolar process of Fig. 10.1 has been used throughout the IC industry for many years and has become known as the *standard buried collector* (SBC) process. In this junction-isolated process, adjoining devices are separated by back-to-back *pn* junction diodes that must be reverse biased to ensure isolation. (See Fig. 10.1(b).) The SBC process remains the primary bipolar process for analog and power circuit applications with power supplies exceeding 15 V. Although the SBC process was also originally used for logic circuits, most digital technologies have evolved to self-aligned, oxide-isolated processes using polysilicon and other technology advances first developed for MOS processes. Wafers with a <111> surface orientation were used specifically for bipolar fabrication for many years. However, in the past few years, it has become common to find bipolar processes also using <100> substrate material, which facilitates transfer of processes from MOS technology. Certainly, all BiCMOS technologies utilize <100> material.

The process flow for the SBC structure of Fig. 10.1(b) was discussed in Section 1.4 and will only be outlined here. An n^+ buried layer is formed by selective diffusion into a <111>-oriented p-type substrate and is followed by growth of an n-type epitaxial