11 MOS Transistors

The *field-effect transistor,* or FET, has a long and complicated history. Its initial invention preceded that of the bipolar transistor by some seventeen years, but all of the early attempts to manufacture field-effect transistors failed because of processing problems.[1] Many of these problems were associated with the growth of thin, high-quality dielectric films. By the time these problems were finally overcome, Bardeen and Brattain had already developed the bipolar transistor.

Since the growth of thin dielectric films proved to be so difficult, the first practical field-effect transistors used reverse-biased junctions in place of dielectrics. The resulting devices were called *junction field-effect transistors* (JFETs). Although JFETs were relatively cumbersome devices, they offered much lower input currents than bipolar transistors could achieve. Certain types of operational amplifiers were designed with JFET input stages to reduce their input currents. These devices have become quite successful and are still being produced today.

Thin insulating films suitable for gate dielectrics were finally produced in 1960.[2] This achievement made possible the manufacture of the *metal-oxide-semiconductor field-effect transistor* (MOSFET), often simply called the *MOS transistor.* The early MOS devices still had their share of problems. Their threshold voltages were notoriously unstable, and their thin gate oxides were exceedingly vulnerable to electrostatic discharge (ESD). Once these problems were overcome, MOS transistors began to seriously challenge established bipolar technologies. MOS integrated circuits proved especially useful for low-power digital devices such as digital watches and pocket calculators.

The earliest MOS processes only offered PMOS transistors. These were soon superseded by processes that could produce both enhancement and depletion NMOS transistors. Demands for lower current consumption and greater design flexibility led to the introduction of processes that could simultaneously fabricate both NMOS and PMOS transistors. Although originally intended for digital applications, these *complimentary metal-oxide-semiconductor* (CMOS) processes could also be

[1] D. Kahng, "A Historical Perspective on the Development of MOS Transistors and Related Devices," *IEEE Trans. on Electron Devices,* Vol. ED-23, #7, 1976, pp. 655–657.

[2] D. Kahng and M. M. Atalla, "Silicon-silicon dioxide field-induced devices," Solid-State Device Research Conference, Pittsburgh, June 1960.

used to design a variety of analog integrated circuits. These soon began to replace bipolar integrated circuits in selected applications, but CMOS transistors were not able to duplicate all of the capabilities of bipolar. Many newer processes now merge both bipolar and CMOS transistors onto a common substrate.

This chapter describes the operation and construction of MOS transistors, particularly those employing self-aligned polysilicon gate technology. Chapter 12 covers a variety of more specialized devices, including high-voltage transistors, power transistors, and JFETs.

11.1 TOPICS IN MOS TRANSISTOR OPERATION

MOS transistors are relatively easy to understand and apply. Unlike their bipolar counterparts, they do not experience thermal runaway or secondary breakdown, and they have relatively few troublesome parasitics. Although most textbooks offer a relatively complete discussion of MOS transistor operation, a few topics are slighted because they rarely concern circuit designers. This section discusses several of these neglected topics, including the effects of process and layout on device parameters and the behavior of MOS transistors operating in breakdown.

11.1.1. Modeling the MOS Transistor

Figure 11.1 shows a simplified three-terminal circuit model of an NMOS transistor. No DC current flows through the gate terminal because of the insulating layer between it and the rest of the transistor. Capacitors C_{GS} and C_{GD} represent the gate-to-source and gate-to-drain capacitances produced by this *gate dielectric*. The slashes through these capacitors signify that their values depend on biasing. Voltage-controlled current source I_1 models the current flowing from drain to source through the channel beneath the gate.

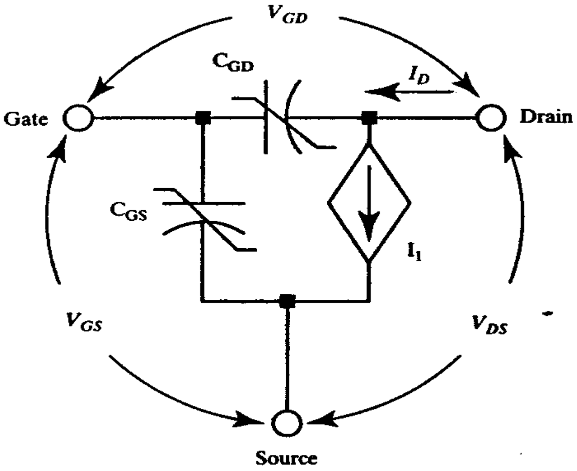


FIGURE 11.1 Simplified three-terminal model of an NMOS transistor.

The magnitude of the drain current I_D depends on the gate-to-source voltage V_{GS} and the gate-to-drain voltage V_{DS}. If the gate-to-source voltage is less than the threshold voltage V_t then no channel forms and the transistor remains in *cutoff* and conducts little or no current. A channel begins to form as soon as the gate-to-source voltage V_{GS} exceeds the *threshold voltage* V_t. The difference between these two quantities is sometimes called the *effective gate voltage* V_{gst}:

$$V_{gst} = V_{GS} - V_t \qquad [11.1]$$

A larger V_{gst} generates a stronger channel that can conduct more current. The drain current also depends on the drain-to-source voltage V_{DS}. If the drain-to-source voltage is less than the effective gate voltage, then the drain current varies linearly with drain-to-source voltage and the transistor is said to operate in the *linear region* (also called the *triode region*). If the drain-to-source voltage exceeds the threshold voltage, then the drain current becomes essentially independent of drain-to-source voltage and the transistor is said to operate in the *saturation region.* The relationship between I_D, V_{GS}, and V_{DS} can be described by a pair of equations—one for the linear region and the other for saturation:

$$If\ 0 \le V_{DS} < V_{gst}, I_D = k\left(V_{gst} - \frac{V_{DS}}{2}\right)V_{DS} \qquad [11.2A]$$

$$If\ V_{DS} \ge V_{gst}, I_D = \frac{k}{2}V_{gst}^2 \qquad [11.2B]$$

These are the *Shichman-Hodges equations* for the NMOS transistor.[3] The parameter, *k,* is called the *device transconductance.* which is something of a misnomer because it has units of A/V^2 rather than A/V. The equations for the PMOS transistor differ in only one respect: equation 11.2A applies when $0 \ge V_{DS} > V_{gst}$, and equation 11.2B applies when $V_{DS} \le V_{gst}$. For an enhancement-mode PMOS, V_{DS}, V_{GS}, V_t, k, and I_D must all be negative quantities for the equations to yield the proper results.

The Shichman-Hodges equations for the NMOS do not cover the case where $V_{DS} < 0$. Strictly speaking, this condition cannot occur because the source and drain of a MOS transistor are determined by electrical biasing rather than by arbitrary terminal markings. If one attempts to make the drain-to-source voltage less than zero, then the source and drain simply swap roles. The drain terminal now plays the role of the source, and *vice versa.* If one insists on retaining terminal names that no longer correspond to the roles that the terminals actually play, then the terminal conditions must be transformed into actual biasing conditions before applying the Shichman-Hodges equations (see Exercise 11.2).

Device Transconductance

The device transconductance, *k,* determines the amount of drain current that flows through a MOS transistor in response to a given V_{gst}. The device transconductance thus specifies the size of a MOS transistor in much the same way that the emitter saturation current I_S quantifies the size of a bipolar transistor. The device transconductance has units of A/V^2, or (more commonly) μA/V^2. It is related to the layout dimensions of the transistor by the following equation

$$k = k'\left(\frac{W}{L}\right) \qquad [11.3]$$

where W and L represent the width and length of the MOS channel and k' is a constant called the *process transconductance,* which equals

$$k' = \frac{\mu\varepsilon_o\varepsilon_r}{t_{ox}} \qquad [11.4]$$

where the quantity μ represents the *effective mobility* of the carriers (electrons in an NMOS, holes in a PMOS). Surface scattering reduces the mobility of carriers

[3] The equations given in the text ignore channel length modulation and the body effect. For further details see H. Shichman and D. A. Hodges, "Modeling and Simulation of Insulated-Gate Field-Effect Transistor Switching Circuits." *IEEE J. Solid-State Circuits,* SC-3, 1968.

confined within a MOS channel, so the effective mobilities appearing in equation 11.4 are considerably smaller than the bulk mobilities discussed in Section 1.1.1. The effective mobility of electrons and holes in silicon are about $675\text{cm}^2/\text{V}\cdot\text{s}$ and $240\text{cm}^2/\text{V}\cdot\text{s}$, respectively. The constant ε_o denotes a universal physical constant called the *permittivity of free space,* which equals $8.85\cdot10^{-12}$ F/m. The constant ; ε_r represents the relative permittivity of the gate dielectric, which for pure silicon dioxide equals about 3.9. The actual permittivity of oxide may vary slightly from this theoretical value (see Table 6.1). The quantity t_{ox} represents the thickness of the gate dielectric. Substituting these values into the previous equation yields the following simplified formulas for the process transconductance of an NMOS transistor k'_n and of a PMOS transistor k'_p

$$k'_n \cong \frac{23000}{t_{ox}}\mu A/V^2 \qquad [11.5A]$$

$$k'_p \cong \frac{8200}{t_{ox}}\mu A/V^2 \qquad [11.5B]$$

where t_{ox} is measured in Angstroms (Å). MOS processes use the thinnest possible gate oxides to produce the largest possible device transconductances. The dielectric strength of gate oxide equals about 10^7V/cm, or about 0.1V/Å. In practice, the gate dielectric is restricted to substantially lower field intensities to prevent a delayed breakdown mechanism called *time-dependent dielectric breakdown* (TDDB). Transistors with gate oxides thicker than 500Å are generally restricted to field intensities of no more than $3\cdot10^6$V/cm (30mV/Å). Thinner gate oxides can withstand somewhat higher electric field intensities,[4] so transistors with a gate oxide only 100 to 200Å thick can safely operate at fields in excess of $5\cdot10^6$V/cm (50mV/Å). Assuming a conservative limit of 30mV/Å, then the maximum possible process transconductance for an operating voltage V_{op} equals

$$k'_n \cong \frac{690}{V_{op}}\mu A/V^2 \qquad [11.6A]$$

$$k'_p \cong \frac{240}{V_{op}}\mu A/V^2 \qquad [11.6B]$$

These formulas indicate that a 5V CMOS process can achieve an NMOS transconductance of about $138\mu\text{A/V}^2$ and a PMOS transconductance of about $48\mu\text{A/V}^2$. In practice, short-channel transistors often experience additional transconductance reductions caused by velocity saturation and other high-field effects. In order to construct a PMOS transistor with the same transconductance as a given NMOS, the W/L ratio of the PMOS must equal almost three times the W/L ratio of the NMOS. The PMOS transistor will thus require nearly three times the area of the NMOS. This disparity is most noticeable in power transistors, but even minimum-size logic gates often increase the size of PMOS transistors to compensate for their low process transconductances.

The device transconductance of MOS transistors decreases with temperature. This variation is primarily due to the temperature coefficient of the carrier mobility. As the temperature rises, lattice vibrations become more energetic and cause increased carrier scattering. Consequently, carrier mobilities are approximately pro-

[4] C. M. Osburn and D. W. Ormond, "Dielectric Breakdown in Silicon Dioxide Films on Silicon, II. Influence of Processing and Materials," *J. Electrochem. Soc.*, Vol. 119, #5, 1972, pp. 597–603.

portional to the inverse square of absolute temperature.[5] The device transconductance at 150°C equals about half of its value at 25°C. The drain current for a given gate-to-source voltage scales somewhat similarly. Since an increase in temperature causes a decrease in drain current, MOS transistors generally do not exhibit thermal runaway. Power MOS transistors do not require ballasting, which not only simplifies their construction but also enables them to achieve extremely low on-resistances (Section 12.2.1).

The device transconductance given in Equation 11.4 corresponds to that found in most engineering texts,[6] as well as that used in the level-1 MOS model of the simulation program SPICE. Since SPICE was written at the University of California at Berkeley, this definition of device transconductance is often called the *Berkeley k.* A few authors use an alternative definition equal to one-half of the Berkeley k and adjust the Shichman-Hodges equations accordingly.

Threshold Voltage

The ***threshold voltage*** V_t equals the gate-to-source voltage required to just establish a channel beneath the gate dielectric when the backgate is connected to the source. An ***enhancement-mode*** MOS transistor requires the application of a non-zero gate-to-source voltage in order to form a channel. The channel of an enhancement-mode NMOS consists of electrons attracted to the surface of the P-type backgate by the positively charged gate electrode (Figure 11.2A). The threshold voltage of the enhancement NMOS is therefore positive. The channel of an enhancement PMOS consists of holes attracted to the surface of an N-type backgate by the negatively charged gate electrode (Figure 11.2B). The threshold voltage of an enhancement PMOS is therefore negative. Another type of MOS transistor exhibits a channel even at a gate-to-source voltage of zero. These ***depletion-mode*** transistors are normally conducting and require the application of an external gate-to-source voltage in order to turn them off (Figures 11.2C, D). The threshold voltage of a depletion NMOS is negative and the threshold of a depletion PMOS is positive.

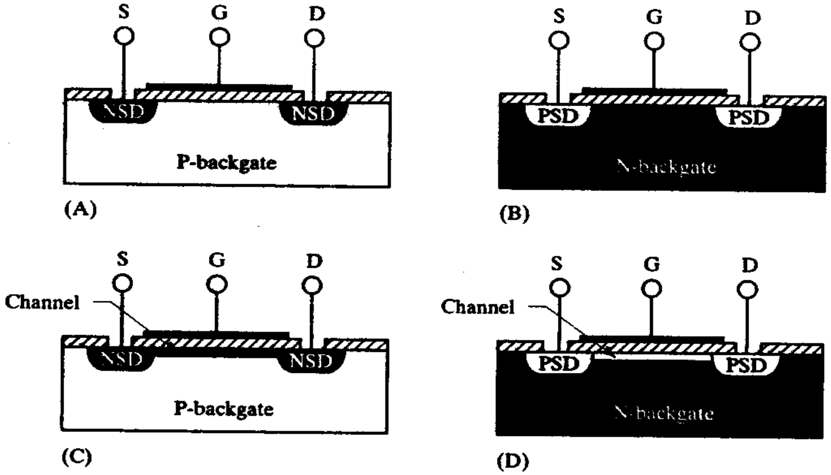


FIGURE 11.2 The four types of MOS transistors:
(A) enhancement NMOS,
(B) enhancement PMOS,
(C) depletion NMOS, and
(D) depletion PMOS.

[5] Electron mobility varies as $T^{-2.42}$ and hole mobility varies as $T^{-2.20}$; see S.M. Sze, *Physics of Semiconductor Devices,* 2nd ed. (New York: John Wiley and Sons, 1981), p. 29.

[6] For instance, see R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits,* 2nd ed. (New York: John Wiley & Sons, 1986), p. 430.

MOS transistors are often described as electrically actuated switches. An enhancement MOS resembles a normally open switch because it is normally off and it requires the application of an external gate bias to turn it on. A depletion MOS resembles a normally closed switch because it is on by default and it requires the application of an external gate bias to turn it off. Most processes are optimized to fabricate enhancement-mode transistors because these devices are much more convenient to use than their depletion-mode counterparts. Some processes also offer depletion-mode devices as process extensions.

The threshold voltage of a MOS transistor depends on several factors, including the gate electrode material, the doping of the backgate, the thickness of the gate oxide, the surface state charge density, and the oxide charge density (both fixed and mobile). Each of these factors will be considered in turn.

If the materials used to form the gate and the backgate are not identical, then a nonzero contact potential develops between them. This contact potential represents a net voltage difference between the two materials even if they are separated by an insulating layer. Any change in the contact potential produces a corresponding change in the threshold voltage. Modern MOS transistors almost always use heavily doped polysilicon gate electrodes. There are two choices of gate material: N+ poly or P+ poly. The substitution of a P+ gate for an N+ gate causes the contact potential to increase by about 1.2V, regardless of backgate doping (Table 11.1). Similarly, the substitution of an N+ gate for a P+ gate causes the contact potential to decrease by about 1.2V. These variations in contact potential will produce equal variations in threshold voltage.

TABLE 11.1 Computed poly gate-to-backgate contact potentials (10^{20}cm^{-3} poly).

Backgate Material	N+ Poly Gate	P+ Poly Gate
N-type, $N_D = 10^{14}$cm^{-3}	–0.36	0.82
N-type, $N_D = 10^{16}$cm^{-3}	–0.24	0.94
N-type, $N_D = 10^{18}$cm^{-3}	–0.12	1.06
P-type, $N_A = 10^{14}$cm^{-3}	–0.82	0.36
P-type, $N_A = 10^{16}$cm^{-3}	–0.94	0.24
P-type, $N_A = 10^{18}$cm^{-3}	–1.06	0.12

A few examples may help to explain the effects of swapping gate materials. Suppose that a certain NMOS transistor develops a threshold voltage of 0.7V using an N+ gate. If this same transistor were to use a P+ gate, it would have a threshold of about 1.9V. Suppose that the corresponding enhancement PMOS develops a threshold voltage of –0.7V using a P+ gate. If this transistor were to use an N+ gate, it would become a depletion device with a threshold of about +0.5V.

The backgate doping concentration also has a strong impact on the threshold voltage. In order to form a channel, enough carriers must be attracted to the surface to invert the silicon. A heavily doped backgate is difficult to invert and the gate must therefore exert a stronger electric field to muster enough carriers to create a channel. The magnitude of the threshold voltage therefore increases with backgate doping. The effect is small at low doping concentrations, but at higher doping levels it actually dominates the expression for the threshold voltage.

The thickness of the gate oxide can also play an important role in determining the threshold voltage of a transistor. A given gate-to-source voltage produces a weaker electric field across a thick gate oxide than across a thin one. Transistors with thick gate oxides are more difficult to invert than ones with thin gate oxides, so increasing the thickness of the gate oxide increases the magnitude of the threshold voltage. For example, the oxide in the field regions is made as thick as possible to raise the thick-field threshold. Unfortunately, thickness alone does not guarantee an adequate thick-field threshold for most processes. Consider the entries in Table 11.2 for a 10kÅ oxide. If the background doping equals 10^{15}cm^{-3}, then the thick-field threshold only equals about 4V. This is obviously inadequate. Most processes use channel stop implants to raise the doping in the field regions.[8] If the channel stop implant raises the doping under a 10kÅ field oxide to 10^{17}cm^{-3}, then the thick-field threshold will approach fifty volts. Since both the NMOS and the PMOS backgate doping levels are relatively low, most processes must employ a combination of boron and phosphorus channel stops to ensure that both thick-field thresholds lie well above the nominal operating voltages.

Backgate Doping	100Å	250Å	10kÅ
10^{14}cm^{-3}	–0.23	–0.21	0.89
10^{15}cm^{-3}	–0.08	–0.02	3.91
10^{16}cm^{-3}	0.14	0.35	13.9
10^{17}cm^{-3}	0.60	1.31	47.9
10^{18}cm^{-3}	1.86	4.28	162
10^{20}cm^{-3}	7.94	19.1	747

TABLE 11.2 Typical NMOS threshold voltages as a function of backgate doping and gate oxide thickness (N-type poly gate).[7]

The threshold voltage is also affected by the presence of residual charges within the gate oxide and along the oxide-silicon interface. These residual charges can be divided into three types: fixed oxide charge, mobile oxide charge, and surface state charge. The *fixed oxide charge* Q_f consists of defect sites scattered randomly throughout the oxide film. Gate oxides grown in dry oxygen at relatively low temperatures have very small fixed oxide charges. The fixed oxide charge can drastically increase if holes are injected into the oxide, as happens during oxide breakdown, hot carrier injection, and exposure to ionizing radiation.

The *mobile oxide charge* Q_m consists of positively charged mobile ions such as sodium and potassium.[9] The threshold voltage shift that they produce also depends on their location within the oxide film, and this in turn depends on gate biasing. Consider the case of an NMOS transistor whose gate oxide is contaminated by sodium. When a positive gate-to-source voltage is applied, the mobile ions move away from the positively charged gate electrode and move closer to the negatively charged backgate. As the mobile ions shift closer to the channel region, they exert an increasing effect on it. Thus the movement of the mobile ions causes the NMOS threshold to decrease. Any shift in the threshold voltage of MOS transistors can cause offsets. In order to obtain accurate matching, mobile ions must either be excluded from the process, or they must somehow be immobilized (Section 4.2.2). Modern processing techniques have reduced the magnitude of the mobile oxide charge to negligible proportions.

The *surface state charge* Q_{ss} is concentrated in a thin layer near the oxide/silicon interface. It is generally positive, but its magnitude depends on silicon crystal orientation and annealing conditions. The precise mechanisms responsible for generating the surface state charge are not fully understood, but they are believed to involve mismatches between the molecular structure of the silicon lattice and that of the

[7] The values in this table assume $Q_{ss} = 0$ and $\Phi_{MS} = -0.7\text{V}$.

[8] J. D. Sansbury, "MOS Field Threshold Increase by Phosphorus-Implanted Field," *IEEE Trans. on Electron Devices,* Vol. ED-20, #5, 1973, pp. 473–476.

[9] Sodium is the primary mobile ion encountered in silicon processing; lesser contributors include potassium and hydrogen ions. See B. E. Deal, "The Current Understanding of Charges in the Thermally Oxidized Silicon Structure," *J. Electrochem. Soc.,* Vol. 121, # 6, 1974, pp. 198C–205C.

oxide macromolecule.[10] Oxides grown on (100) silicon have less than half the surface state charge density of oxides grown on (111) silicon.[11] All modern MOS processes employ (100) silicon to minimize the impact of surface state charge on the threshold voltages of MOS transistors. Standard bipolar uses (111) silicon to deliberately raise the NMOS thick-field threshold. The surface state charge can also be reduced by annealing the oxide in a reducing atmosphere such as hydrogen or forming gas (a mixture of nitrogen and hydrogen). This anneal presumably offers opportunities for the molecular structure of the oxide to adjust to match that of the silicon. Although it is not possible to entirely eliminate the surface state charge, the use of (100) silicon in combination with proper annealing can minimize threshold voltage variation.

Each of the mechanisms discussed above contributes a small amount of variability to the threshold voltage. With care, threshold voltages can be held to within ±15%. This translates into a variation of ±0.1V in a threshold voltage of 0.7V. Threshold voltages also vary with temperature. The magnitude of this variation depends on the backgate doping and oxide thickness, but it typically ranges from –2mV/°C to –4mV/°C.[12,13] A temperature variation of –2mV/°C over a temperature range of –55 to 125°C produces a threshold variation of about ±0.2V. Combining this with the process variation yields a total variation of about ±0.3V. A transistor with a nominal threshold voltage of 0.7V might actually have a V_t as low as 0.4V or as high as 1.0V. Although it might seem that the minimum threshold could safely be decreased to 0.3V or less, this is actually not the case. MOS transistors continue to conduct small amounts of current when the gate-to-source voltage is less than the threshold voltage due to a mechanism called *subthreshold conduction*. The magnitude of the subthreshold current decreases exponentially, and the gate-to-source voltage must drop at least 0.3 to 0.4V below the threshold in order to reduce the drain current to negligible levels. The magnitude of the nominal threshold voltage must therefore equal at least 0.6 to 0.7V. Transistors with smaller threshold voltages are useful in certain applications, but they cannot be used as switching devices. Transistors used in very low current applications may require nominal threshold voltages of 0.8 to 1.0V to prevent objectionable subthreshold conduction. Alternatively, special circuit design techniques may be used to apply a reverse gate-to-source voltage to ensure proper cutoff.

11.1.2. Parasitics of MOS Transistors

A real-world MOS transistor contains a number of parasitic elements that affect its operation. Perhaps the most important of these are the junctions that isolate the source and drain regions from the backgate. These junctions remain reverse-biased during normal operation, but either or both of them may begin to conduct under certain circumstances. The forward-biased junctions will inject minority carriers into the backgate, at best causing unexpected leakage and at worst triggering latchup.

A complete model of all of the parasitics contained in an MOS transistor would include a number of distributed effects, the discussion of which lies beyond the

[10] This discussion is somewhat oversimplified, as several types of charges reside along the interface, including a component of the fixed oxide charge and a variety of interface traps (the so-called *fast surface states* and *slow surface states*). The component due to fixed oxide charge is always positive, while the interface traps may contribute either positive or negative charges (see Muller and Kamins, p. 152ff).

[11] Deal's data shows Q_{SS} values on (100) silicon equal to 20 to 30% of those on (111) silicon: Deal, *ibid.*

[12] R. Wang, J. Dunkley, T. A. DeMassa. and L. F. Jelsma. "Threshold Voltage Variations with Temperature in MOS Transistors," *IEEE Trans. on Electron Devices*, Vol. ED-18, #6, 1971, pp. 386–388.

[13] F. M. Klaasen and W. Hes, "On the Temperature Coefficient of the MOSFET Threshold Voltage," *Solid-state Electronics*, Vol. 29, #8, 1986, pp. 787–789.

scope of this text. Figure 11.3 shows a simplified parasitic model of an NMOS transistor constructed in an N-well CMOS process. The circuit contains a three-terminal NMOS transistor, M_1, that models the intended functionality of the device. Capacitors C_{GD}, C_{GS}, and C_{GB} represent the gate-to-drain, gate-to-source, and gate-to-backgate capacitances, respectively. The gate-to-backgate capacitance, C_{GB}, models the capacitance across the thin gate oxide separating the gate electrode from the backgate diffusion. This capacitance decreases as the transistor approaches inversion, and drops to zero when a channel forms. The gate-to-drain and gate-to-source capacitances, C_{GD} and C_{GS}, consist primarily of the overlap capacitances between the gate electrode and the respective diffusions. These have been greatly reduced by the introduction of self-aligned polysilicon gates, and they now consist mainly of fringing capacitances. The gate-to-source and gate-to-drain capacitances abruptly increase when a channel forms because of the sudden addition of the capacitance between the gate electrode and the channel.[14] The slashes drawn through C_{GD}, C_{GS}, and C_{GB} denote the voltage dependence of these devices.

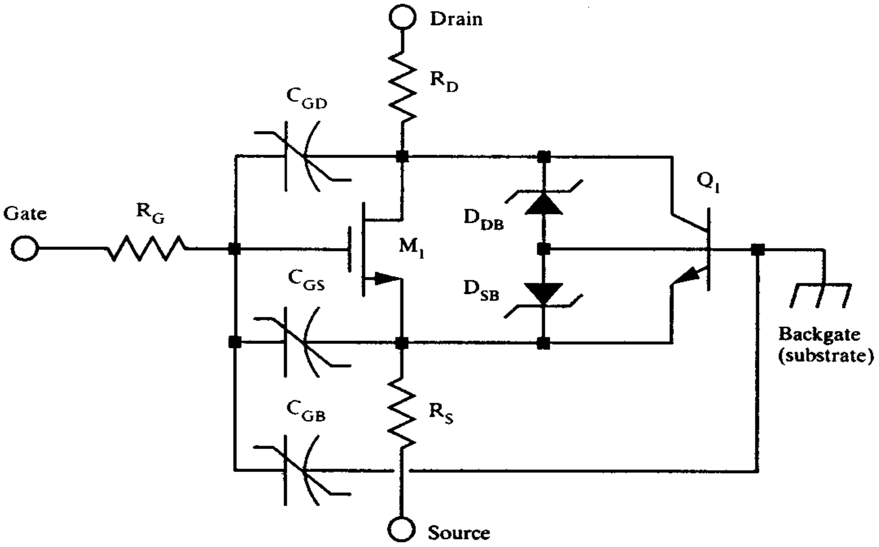


FIGURE 11.3 Simplified parasitic model of an NMOS transistor constructed in N-well CMOS.

Resistors R_G, R_S, and R_D represent the resistance of the gate, source, and drain terminals, respectively. The gate resistance actually forms a distributed network with the three gate capacitances, but this simplified model depicts it as a lumped quantity. The gate resistance has no effect on the DC performance of the transistor, but it does slow the switching speed because it limits the current that is available to charge and discharge capacitances C_{GS}, C_{GD}, and C_{GB}. The gate poly is often silicided to minimize R_G. The drain and source resistances R_D and R_S consist of the Ohmic resistances between the contact and the edge of the source/drain diffusions abutting the channel. These resistances can be minimized by siliciding the surfaces of the source/drain diffusions. A silicided poly gate electrode is sometimes called a *clad gate*, and silicided source/drain regions are also called *clad moats*. Many processes use clad gates, but generally only submicron processes have sufficient transconductance to merit clad moats.

[14] J. E. Meyer, MOS Models and Circuit Simulation. *RCA Rev.*, Vol. 32, March 1971, pp. 42–63.

Diodes D_{DB} and D_{SB} represent the drain-backgate and source-backgate junctions, respectively. The diodes model the junction capacitance added to the drain and source terminals by these junctions, and they also model their avalanche breakdown characteristics. Lateral NPN transistor Q_1 represents one possible path for minority carriers to travel from drain to source (or *vice versa*). Since the NMOS transistor resides in the epi, the minority carriers can also flow to adjacent NMOS source/drain regions, or to adjacent wells. The flow of carriers to adjacent wells can lead to CMOS latchup, as discussed below.

Figure 11.4 shows a simplified parasitic model for a PMOS transistor constructed in an N-well process. Capacitors C_{GS}, C_{GD}, and C_{GB} play the same roles in this device as their counterparts do in an NMOS. Similarly, resistors R_G, R_S, and R_D represent the same terminal resistances in the PMOS as they do in the NMOS. Diodes D_{SB} and D_{DB} represent the junctions between the source/drain regions and the backgate, in this case consisting of an N-well diffusion contacted through a fourth terminal. If either of these junctions forward-biases into the well, then the minority carriers will travel either to the other source/drain diffusion or to the substrate. Lateral PNP Q_1 represents the minority carrier conduction path from drain to source (or *vice versa*). Lateral PNP transistors Q_2 and Q_3 represent the minority carrier paths from the source/drain diffusions to the substrate.

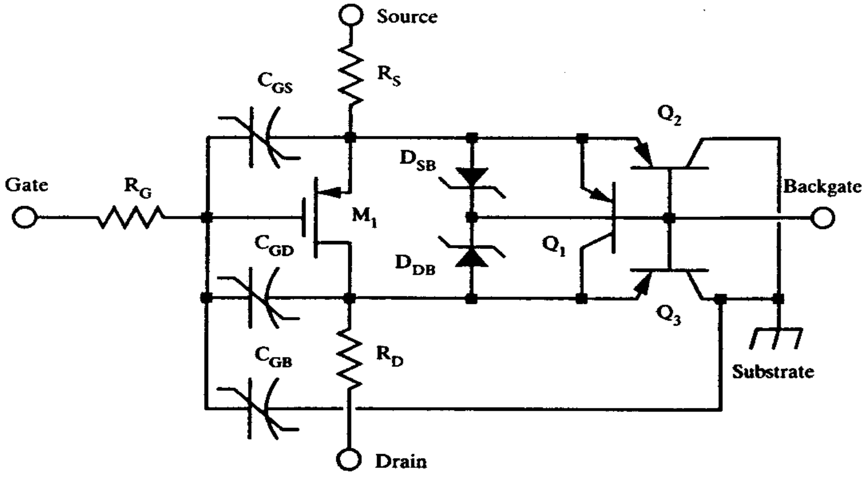


FIGURE 11.4 Simplified parasitic model of a PMOS transistor constructed in N-well CMOS.

Breakdown Mechanisms

Several different mechanisms limit the operating voltage of MOS transistors. One of these corresponds to the V_{CER} breakdown mechanism seen in bipolar transistors (Section 8.1.2). For purposes of discussion, assume that the gate and backgate electrodes both connect to the source. As the drain-to-source voltage rises, it eventually reaches a point where the drain-backgate junction begins to break down. Avalanche multiplication injects large numbers of majority carriers into the lightly doped backgate, causing it to debias. The source-backgate junction begins to inject minority carriers into the backgate as soon as it forward-biases. Most of these minority carriers flow across to the drain, where they stimulate further avalanche multiplication. This beta multiplication process causes the breakdown voltage of the MOS transistor to snap back from the initial *trigger voltage* to a lower *sustain voltage* (Figure 11.5). Short-channel transistors have somewhat lower breakdown volt-

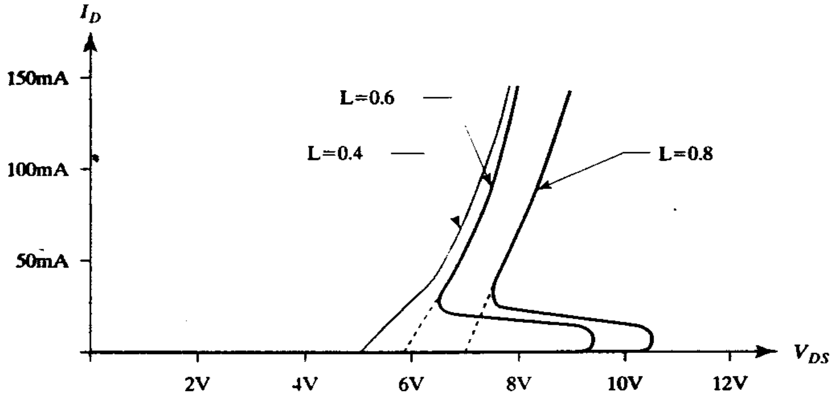


FIGURE 11.5 Breakdown characteristics of short-channel NMOS transistors. The solid curves show the breakdown characteristics in cutoff, and the dotted lines show the change in characteristics when in conduction.

ages because the narrower basewidth of their parasitic lateral bipolar transistor raises its gain and enhances the beta multiplication process.

Once an MOS transistor avalanches, most of its drain current flows through the parasitic bipolar transistor rather than through the MOS channel.[15] The device therefore becomes vulnerable to hot spot formation and current focusing. These mechanisms can severely restrict the transistor's ability to withstand transient overloading. Localized conduction is often triggered by the presence of sharp corners in the channel that cause electric field intensification and premature avalanche breakdown. The designer should attempt to eliminate as many of these corners as possible. Circular annular MOS devices (Section 11.2.6) are particularly robust because their channels have no corners at all. The conventional finger-style transistor has corners at either end of each channel, but these are not as vulnerable as one might expect because they lie along the periphery of the device where the temperatures are not as extreme. Switching transistors should not incorporate bends into their channels, because the resulting corners lie inside the device where localized conduction is more likely to result in thermal runaway.

The resistances of the source and drain diffusions help ballast the avalanching MOS transistor. High source/drain sheet resistances and large spacings between the contacts and the adjacent channel enhance the effectiveness of the ballasting. Silicidation of the source/drain regions eliminates most of the ballasting and renders the transistors more vulnerable to transient overloads. One can sometimes employ a silicide block mask to prevent the silicide from abutting the gate in order to introduce at least a small amount of ballasting. Additional ballasting can be generated by spacing the source/drain contacts further away from the channel.

Short-channel transistors sometimes experience another form of breakdown called *punchthrough*. The source/drain regions are heavily doped, so the depletion regions surrounding them extend primarily into the lightly doped backgate. The drain depletion region widens as the drain-to-source voltage increases. If the drain/backgate junction does not avalanche first, the drain depletion region will eventually extend entirely across the channel to touch the source depletion region. Punchthrough breakdown does not exhibit snapback because it does not activate the parasitic lateral bipolar transistor. The curves of Figure 11.5 highlight the differences between punchthrough (in the 0.4μm device) and avalanche breakdown (in

[15] F.-C. Hsu, P.-K. Ko. S. Tam, C. Hu, and R. S. Muller, "An Analytical Breakdown Model for Short-Channel MOSFETs," *IEEE Trans. on Electron Devices*, ED-29, #11. 1982, pp. 1735–1740.

the 0.6 and 0.8μm devices). The snapback characteristic of avalanche breakdown only appears if the transistor is in cutoff. Even low levels of subthreshold conduction produce enough beta multiplication to obscure the snapback characteristic (as shown by the dotted lines in Figure 11.5). Low-voltage isolated MOS transistors may also exhibit punchthrough breakdown between the drain and the substrate. In N-well CMOS processes, only the PMOS transistor normally experiences this form of punchthrough. This problem can be solved by using a deeper well diffusion, but this will result in larger spacings between components. Some low-voltage processes use high-energy implants called *punchthrough stops* to increase the doping at the bottom of the well to prevent vertical punchthrough. Alternatively, the well can be formed using a high-energy implant to produce a peak doping concentration deep beneath the surface of the silicon. Such a *retrograde well* has characteristics similar to those of a regular well augmented by a punchthrough stop.[16]

The thin gate dielectric is also vulnerable to a third form of breakdown. If the voltage across the gate oxide rises beyond a certain point, then avalanche generation begins to occur within the oxide itself. Holes generated by the avalanche process become trapped in the oxide, producing a positive charge that increases the electric field across the oxide. Below a certain field intensity, electron-hole recombination stabilizes the magnitude of the positive charge. Above this critical field intensity, the positive charge continues to grow and runaway conduction overheats and destroys the gate oxide.[17] This mechanism is called *dielectric breakdown*, or more commonly, *oxide rupture*. Unlike avalanche and punchthrough, oxide rupture is catastrophic and irreversible. The oxide rupture voltage places an effective limit on the gate-to-source voltage that an MOS transistor can withstand. Unless special precautions are taken in the design of the transistor, it also limits the achievable gate-to-drain voltage rating (see Section 12.1).

Oxide dielectrics may eventually be destroyed by electric field intensities that are somewhat lower than those required to trigger immediate rupture. Avalanche generation produces hot carriers that damage the integrity of the gate oxide. The damaged gate oxide allows larger currents to flow, producing an ever-increasing flow of hot carriers. At some point, the generation rate becomes so great that electron-hole recombination cannot keep pace, and catastrophic oxide rupture quickly ensues. The total amount of charge required to induce catastrophic failure remains roughly constant regardless of the magnitude of the currents involved, although it decreases at higher temperatures.[18] This mechanism is called *time-dependent dielectric breakdown* (TDDB). If the avalanche currents are relatively high, TDDB can occur in a matter of seconds. On the other hand, if the avalanche currents are very low, TDDB may require months or even years. The existence of TDDB explains why the voltage across the gate oxide must not exceed a small fraction of the actual rupture voltage.

MOS transistors are also subject to a fourth breakdown mechanism. The electric field across the pinched-off portion of an MOS channel can become very intense. The carriers flowing across the pinched-off region accelerate to very high velocities and become so-called *hot carriers*. Some of these carriers collide with the lattice and recoil out of the channel. Most of them travel into the backgate, and eventually contribute to the backgate current. Some of the hot carriers also travel into the gate

[16] R. D. Rung, C. J. Dell'oca, and L. G. Walker, "A Retrograde p-well for Higher Density CMOS," *IEEE Trans. on Electron Devices*, Vol. ED-28, #10, 1981, pp. 1115–1119.

[17] P. Soloman, "Breakdown in silicon oxide—A review," *J. Vac. Sci. Technol.*, Vol. 14, #5, 1977, pp. 1122–1130.

[18] G. A. Swartz, "Gate Oxide Integrity of NMOS Transistor Arrays," *IEEE Trans. on Electron Devices*, Vol. ED-33, #11, 1986, pp. 1826–1829.

oxide. Although most of these eventually return to the silicon, a few become permanently trapped and contribute to a slowly increasing fixed oxide charge (Section 4.3.1). As the transistor continues to operate, the accumulating fixed oxide charge causes the threshold voltage to gradually shift. This effect can easily disrupt matching between MOS transistors operating under different biases. If the threshold voltage shifts too far, the transistor may not even be able to switch on and off. Short-channel transistors are especially susceptible to hot-carrier generation because the high backgate doping levels required to prevent punchthrough shorten the pinched-off portion of the channel. The resulting electric fields produce hot carriers at lower voltages than would occur in a transistor with a lightly doped backgate. Transistors acting as switches are less susceptible to hot carrier generation because they normally operate either in the linear region or in cutoff. MOS transistors used in analog circuitry are more vulnerable, as these devices often operate continuously in the saturation region.

Any given MOS transistor will not necessarily experience all of these breakdown mechanisms. Long-channel transistors are usually limited by avalanche breakdown, oxide rupture, and TDDB. Short-channel transistors are usually limited by punchthrough, oxide rupture, and TDDB. Transistors operating for long periods of time under high drain-to-source voltages may also experience hot carrier-induced threshold voltage shifts.

CMOS Latchup

When a source/drain diffusion forward-biases into the backgate, it injects minority carriers that can flow to the reverse-biased junctions of adjacent devices. The exchange of minority carriers between adjacent NMOS and PMOS transistors can trigger *CMOS latchup* (Section 4.4.2). Minority-carrier guard rings can prevent latchup, but they are not necessarily easy to construct in CMOS processes.

The absence of NBL and deep-N+ makes it impossible to construct effective blocking guard rings in a pure CMOS process. One can still construct hole- and electron-collecting guard rings using PMoat and NMoat, although the collection efficiency of these shallow diffusions usually leaves much to be desired. Any PMOS transistor that can inject minority carriers into its well should be surrounded by a hole-collecting guard ring constructed of PMoat (Figure 11.6A). This guard ring should connect to substrate potential to reverse-bias the PMoat/N-well junction as

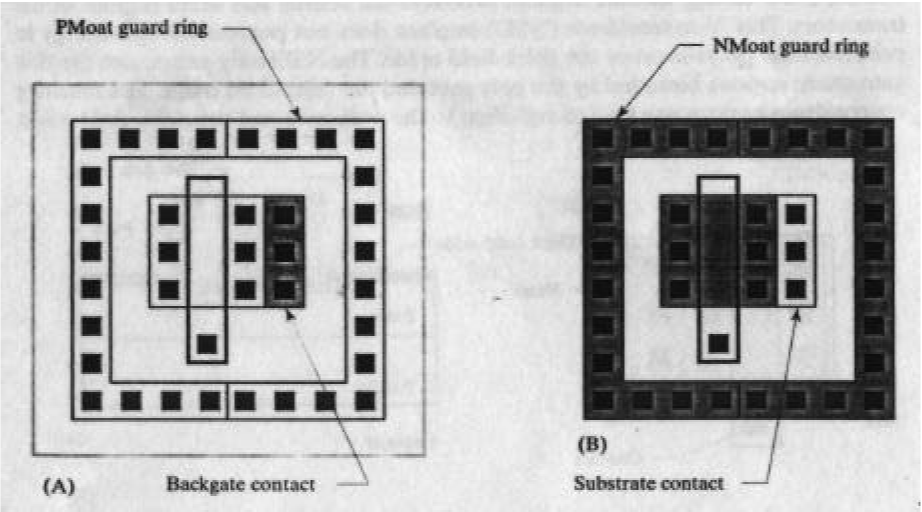


FIGURE 11.6 Examples of guard rings for protecting N-well CMOS transistors: (A) PMOS with a hole-collecting guard ring and (B) NMOS with an electron-collecting guard ring.

strongly as possible. The guard ring collects a percentage of the minority carriers injected laterally by the enclosed PMOS transistor. Although this reduces the lateral flow of carriers toward adjacent devices, it does not stop holes from traveling downward to the substrate. CMOS processes usually employ a P+ substrate to minimize debiasing.

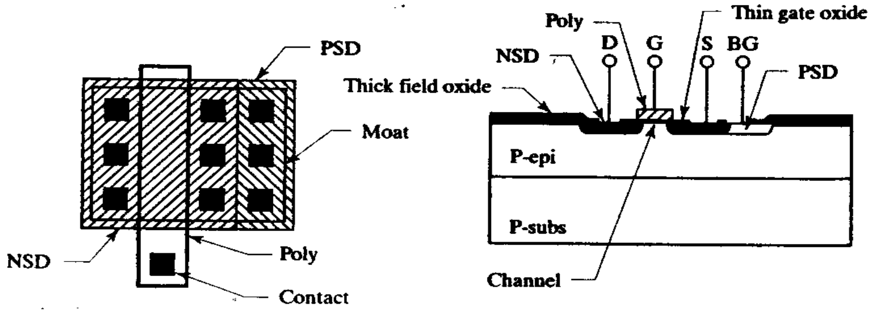
Any NMOS transistor that can forward-bias into the substrate should be surrounded by an electron-collecting guard ring. These guard rings are usually constructed from NMoat rather than from N-well (Figure 11.6B). The deeper well diffusion intercepts a larger fraction of the carriers, but its large vertical resistance makes it prone to debiasing unless it connects to a relatively high-voltage supply. NMoat has a lower collection efficiency, but it is virtually immune to debiasing. NMoat guard rings are sometimes connected to a positive power supply to help drive the depletion region deeper into the substrate to enhance collection efficiency. In low-voltage processes, NMoat guard rings should be connected to substrate potential to minimize hot-carrier generation in their depletion regions (Section 13.2.3).

Guard rings, by themselves, cannot provide total latchup immunity. The flow of even a few minority carriers around the guard rings will trigger parasitic bipolar conduction if not for the presence of backgate contacts. Backgate contacts remove the collected carriers and prevent them from biasing the parasitic lateral bipolar transistors into conduction. Section 11.2.7 discusses the design of backgate contacts in greater detail.

11.2 SELF-ALIGNED POLY-GATE CMOS TRANSISTORS

Most modern CMOS and BiCMOS processes are designed to produce self-aligned poly-gate transistors. Figure 11.7 shows a layout and cross section of a simple self-aligned poly-gate NMOS. The backgate of this transistor consists of a P– epitaxial layer grown on a P+ substrate. The areas between adjacent transistors are called *field regions*. LOCOS oxidation covers these with a *thick-field oxide* that helps suppress parasitic channel formation. The nitride oxidation mask prevents thick oxide from growing in the *moat regions* where transistors will eventually reside. After the removal of the nitride, the moat regions are re-oxidized to form the *thin gate oxide* of the MOS transistors. Doped polysilicon is then deposited on top of the gate oxide to form the gate electrodes of the MOS transistors. After the poly has been patterned, a low-energy arsenic implant produces the source and drain regions of the transistors. This *N-source/drain* (NSD) implant does not possess enough energy to penetrate the polysilicon or the thick-field oxide. The NSD only penetrates the thin gate oxide regions bounded by the poly gate and the thick-field oxide. The resulting source/drain regions are said to *self-align* to the poly gate and the thick-field oxide.

FIGURE 11.7 Layout and cross section of a simple self-aligned poly-gate NMOS transistor.



Next, a *P-source/drain* (PSD) implant is performed. This implant gains its name from the role it plays in the construction of PMOS transistors. The NMOS transistor uses PSD to contact the lightly doped P-epi backgate. A brief anneal activates the source/drain implants and completes the formation of the transistors.

Early MOS processes used aluminum to form the gate electrodes. Aluminum-gate processes are inferior to polysilicon-gate processes in several respects. Aluminum cannot withstand the temperatures required to anneal the source/drain implants, so it must be deposited after implantation. This precludes the self-alignment of the source/drain diffusions, so these implants must overlap the gate by an amount sufficient to account for misalignment. These overlaps greatly increase the gate-to-source capacitance, C_{GS}, and the gate-to-drain capacitance, C_{GD}, which in turn greatly reduces the switching speed of the transistor. The overlap capacitances of a poly-gate transistor are much smaller because the source and drain regions self-align to the gate.

Some attempts have been made to construct gate electrodes from refractory metals such as tungsten. These materials facilitate self-alignment because they can withstand the temperatures required to anneal the source/drain implants. Despite this advantage, refractory-metal gates have not enjoyed widespread success because they still exhibit threshold voltage variations caused by mobile ions and by variations in contact potential. Poly gates are preferred because they provide much more stable and reproducible threshold voltages (Section 4.2.2).

11.2.1. Coding the MOS Transistor

A simple N-well CMOS process requires a total of seven masks: N-well, moat, poly, NSD, PSD, contact, metal, and protective overcoat. The layout database contains the geometric information required to construct each of these seven masks. In the simplest database, the geometries for each mask are drawn upon a different layer. Figure 11.8A shows the layout of an NMOS transistor following this approach.

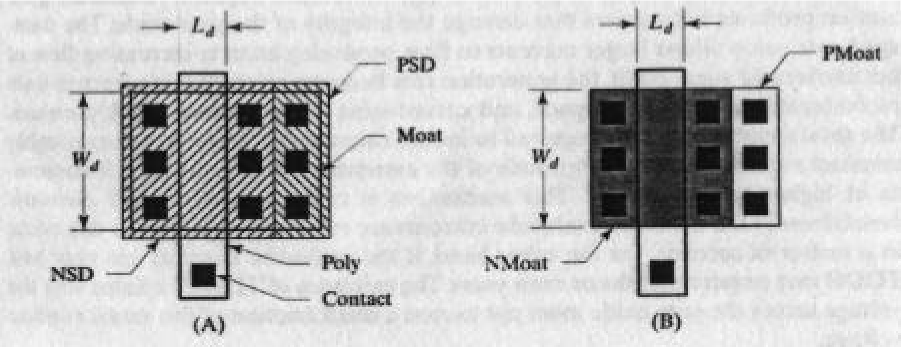


FIGURE 11.8 An NMOS transistor can be coded (A) using NSD, PSD, and moat mask layers or (B) using NMoat and PMoat coding layers.

Figure 11.8B shows another way to draw the same transistor. Two new layers called *NMoat* and *PMoat* are used to generate the NSD, PSD, and moat masks. A geometry placed on the NMoat layer produces corresponding geometries on both the moat and the NSD masks. The NSD geometry is automatically oversized to account for misalignment.[19] A geometry drawn on the PMoat layer creates corresponding geometries on the moat and PSD masks, and the PSD geometries are again automatically oversized.

[19] Although the moat geometries could be undersized instead, this would affect the drawn width of the devices and is therefore a less intuitive option.

The NSD, PSD, and moat layers used to construct the layout in Figure 11.8A are called *mask layers* because the information they contain is transferred directly to the corresponding masks without any intermediate processing. NMoat and PMoat are called *drawing layers* or *coding layers* because they are used only during the drawing (or *coding*) of the layout. The data on the coding layers must pass through a series of geometric transformations to generate the actual mask data.

Coding layers simplify the layout in a number of ways. Data entry takes less time because the layout contains fewer geometries, displays and plots become less cluttered, verification programs run more quickly, and databases consume less space. Each of these advantages may seem relatively insignificant, but together they make a strong argument for the use of coding layers.

The process of transforming coding data into mask data sometimes produces unexpected results on complicated geometries. After the NSD and PSD geometries are oversized, they must be trimmed so that the two implants do not overlap along abutting edges (Figure 11.8A). The trimming algorithm is relatively simple and straightforward as long as the NMoat and PMoat geometries do not contain bends or notches. The algorithm becomes much more complex if it must handle these special cases. The more complex the algorithm becomes, the more likely it is to produce unexpected results under circumstances that are not anticipated by its designer. The only way to entirely eliminate problems of this sort is to avoid the use of coding layers.

The choice between mask layers and coding layers is by no means an easy one. Many designers appreciate the simplifications introduced by coding layers, but they do not always understand the accompanying problems. The people responsible for writing verification and pattern generation programs must ultimately decide whether to use coding layers or to avoid them. This text uses NMoat and PMoat coding layers because they significantly simplify the illustrations.

Width and Length

The length of the transistor equals the distance between the source diffusion and the drain diffusion. The *drawn length* L_d of a self-aligned transistor equals the distance across the poly gate from source to drain as measured in the layout database. The *effective length* L_{eff} of the transistor may be slightly larger or smaller than the drawn length due to overetching, underetching, straggle, outdiffusion, and other factors.[20] These corrections remain relatively constant regardless of the dimensions of the gate, so L_{eff} can be approximated by

$$L_{eff} \cong L_d + \delta L \qquad [11.7]$$

where δL is constant for any given process. The value of δL is usually less than 1 μm, so it primarily affects short-channel devices. Submicron transistors, in particular, exhibit substantial differences between drawn and effective channel lengths. In such cases, the channel length, L, used in the Shichman-Hodges equations 11.2 must equal the effective channel length L_{eff} and not the drawn channel length L_d.

The poly gate must overhang both ends of the source/drain region to prevent the source and drain from shorting together. The width of a self-aligned MOS transistor is therefore set by the moat mask rather than the poly mask. The *drawn width* W_d equals the width of the moat geometry in the layout database or, equivalently, the width of the NMoat or PMoat geometry (Figure 11.8). The *effective width* W_{eff} varies slightly due to straggle, outdiffusion, the presence of the bird's beak, and

[20] G. Massobrio and P. Antognetti, *Semiconductor Device Modeling with SPICE*, 2nd ed. (New York: McGraw-Hill, 1993), pp. 279–283.

other factors. These corrections remain relatively constant regardless of the moat dimensions, so the effective width W_{eff} can be approximated by

$$W_{eff} \cong W_d + \delta W \qquad [11.8]$$

where δW is a constant for any given process. The value of δW is also usually less than 1μm.

11.2.2. N-well and P-well Processes

The NMOS transistors in Figure 11.8 are fabricated in a P– epitaxial layer deposited on a P+ substrate. The heavily doped substrate improves latchup immunity, but it introduces an additional process step. Providing that other measures have been taken to prevent latchup, the epitaxial layer can be eliminated and the transistors built directly into the substrate. Many early processes used this approach to minimize fabrication costs, and some digital processes continue to do so today. Most analog CMOS processes use an epitaxial layer because the epi doping can be controlled very accurately, and therefore the threshold voltages of transistors constructed in the epi vary less than the thresholds of those constructed in the substrate.

A P-epi allows the construction of NMOS transistors, and an N-epi allows the construction of PMOS transistors, but neither allows the construction of both simultaneously. In order to build complementary transistors, another diffusion must be added to counterdope the backgate region of one transistor or the other. If a P-epi is used, then a deep, lightly doped N-type diffusion must be added for PMOS transistors (Figure 11.9A). If an N-epi is used, then a deep, lightly doped P-type diffusion must be added for NMOS transistors (Figure 11.9B). These deep diffusions are commonly called *wells*. An N-type well is called an *N-well* and a P-type well is called a *P-well*. Most processes use either an N-well or a P-well, but not both. In these *single-well* processes, one type of transistor or the other resides in the epi. In an N-well process, the NMOS occupies the epi and the PMOS the N-well. In a P-well process,

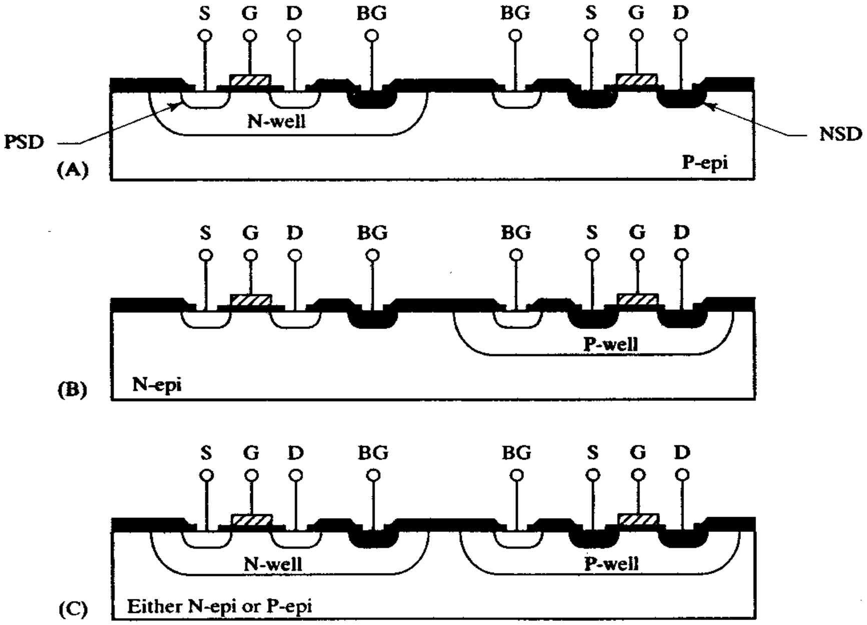


FIGURE 11.9 Three types of CMOS processes: (A) N-well, (B) P-well, and (C) twin-well.

the PMOS occupies the epi and the NMOS the P-well. Some processes include both an N-well and a P-well (Figure 11.9C). In such a *twin-well process*, the NMOS is formed in the P-well and the PMOS is formed in the N-well.

Single-well processes are simpler and cheaper than twin-well processes, but submicron processes sometimes require two wells. As the channel length of the transistor decreases, the backgate doping must increase to prevent punchthrough breakdown. The counterdoping mechanism that creates the well becomes difficult to control on heavily doped substrates. Heavy counterdoping also causes a slight reduction in carrier mobility and a more significant reduction in well-substrate breakdown voltages. These considerations force most submicron processes to use twin wells driven into a lightly doped epi.

The choice of epi also has several consequences. In a single-well process, transistors formed in the epi share a common backgate connection, while transistors formed in wells can be isolated from one another. Although the separate wells consume additional die area, isolation offers an extra degree of design flexibility. An N-well process produces isolated PMOS transistors, while a P-well process produces isolated PMOS transistors. A similar consideration affects the choice of the epi type for a twin-well process. If a P-epi is chosen, then this epitaxial layer shorts all of the P-wells on the die together, and all of the NMOS transistors share a common backgate connection. Similarly, if an N-epi is chosen, the epi shorts the N-wells together and all of the PMOS transistors share a common backgate connection.

N-well processes are favored over P-well processes for several reasons. Most schematics reference their power supplies to a common ground potential. If all of the power supplies deliver positive voltages with respect to ground, as is often the case, then ground becomes the most negative node in the circuit. The substrate of an N-well process can connect to this common ground, but the substrate of a P-well process must connect to the highest-voltage power supply. In multiple-supply systems, it is difficult to ensure that one power supply will always generate a higher voltage than the others, especially during start-up and shut-down. P-well processes are thus poorly suited to multiple-supply applications. One could theoretically reference multiple negative voltages to a positive ground, but this is rarely done in practice.

The mobility of carriers in the counterdoped well will be slightly less than the mobility of carriers in the epi. Since electrons are more mobile than holes, the NMOS transistor has a higher transconductance than the PMOS transistor. Many circuit designers prefer to degrade the performance of the already-inferior PMOS rather than reduce the superior transconductance of the NMOS. This consideration also favors the use of an N-well process.

BiCMOS processes generally employ a P-epi on a P-substrate because this combination simplifies the isolation of the bipolar transistors. The NPN transistor uses the lightly doped N-well as a collector region and the P-epi as isolation, a practice called *collector-diffused isolation* (CDI). Most analog BiCMOS processes are either N-well processes or twin-well processes built on a P-type epi.

N-well BiCMOS processes can construct isolated NMOS transistors as well as isolated PMOS transistors. The isolated NMOS uses a combination of NBL and deep-N+ (or N-well) to isolate the section of P-epi forming the backgate of the transistor (Figure 11.10). The NBL severs the isolated P-epi tank from the P-substrate beneath, and the deep-N+ (or N-well) ring isolates it from adjacent P-epi regions.[21] In order to ensure complete isolation, the ring must contain no gaps and the NBL

[21] E. Bayer, W. Bucksch, K. Scoones, K. Wagensohner, J. Erdeljac, and L. Hutter, "A 1.0μm Linear BiCMOS Technology with Power DMOS Capability," *BCTM Proceedings*, 1995, pp. 137–141.

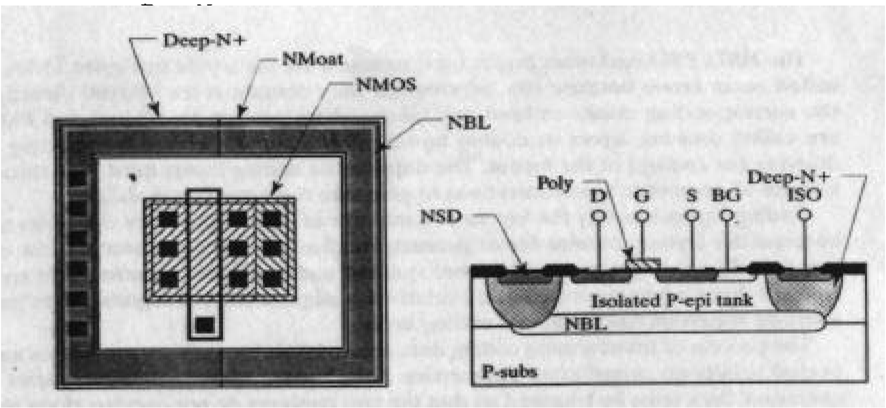


FIGURE 11.10 Layout and cross section of an isolated NMOS in an N-well analog BiCMOS process.

must overlap it sufficiently to allow for misalignment. If deep-N+ is available, then it is usually used instead of N-well because it produces a lower-resistance connection to the NBL. The N-well/epi junction usually has a higher breakdown voltage than the deep-N+/epi junction, so devices that must operate at a high voltage relative to the substrate will normally use N-well isolation rings, either alone or surrounding a deep-N+ isolation ring.

The isolation ring must connect to a voltage equal to or greater than that applied to the isolated P-epi tank. The source/drain regions easily punch through the lightly doped tank, so most isolated NMOS transistors cannot withstand the application of more than a few volts drain-to-isolation or source-to-isolation. These operating voltages increase slightly if the isolation ring connects to a potential midway between that of the backgate and that of the source/drain diffusions. This configuration allows a portion of the depletion region surrounding the isolation/NBL region to intrude into the lightly doped outer fringes of the NBL. Because the NBL dopant diffuses very slowly, the degree of field relief offered is slight and the improvement in operating voltage amounts to only a few volts.

The backgate region of the isolated NMOS consists of a thin, lightly doped P-epi layer. This layer has much more lateral resistance than the P-epi/substrate sandwich comprising the backgate of the nonisolated NMOS. Rapidly slewing signals can momentarily forward-bias the source/drain regions of an isolated NMOS into its backgate. Most of the injected minority carriers flow to the isolation ring, but a few travel from source to drain (or *vice versa*). If the transistor operates at a relatively large drain-to-source potential, then minority carrier injection can trigger V_{CER} breakdown and snapback. Adequate backgate contact minimizes the magnitude of the snapback and the likelihood of triggering it (Section 11.2.7).

11.2.3. Channel Stops

Self-aligned poly-gate MOS transistors form wherever poly intersects PMoat or NMoat geometries. Under certain circumstances, MOS transistors can also form underneath the thick-field oxide. These unwanted *parasitic transistors* interfere with the operation of the integrated circuit unless they are somehow suppressed.

The threshold voltages of the parasitic transistors can be raised by implanting the field regions with a suitable dopant before growing the thick-field oxide. A doping concentration of 10^{17} atoms/cm^3 beneath a 10kÅ field oxide will provide a thick-field threshold of nearly 50V (Table 11.2). This thick-field threshold will provide a comfortable safety margin for a 30V process. Implants used to raise the doping of field regions are called *channel-stop implants*.

The thick-field thresholds can also be raised by deliberately introducing surface state charges. Standard bipolar processes produce large surface state charges by using (111)-oriented silicon in combination with a final oxidizing anneal. The resulting positive charge raises the magnitude of the PMOS thick-field threshold and lowers the magnitude of the NMOS thick-field threshold. Standard bipolar uses a heavily doped P+ isolation system to suppress NMOS parasitic channel formation, and relies on the surface state charge to elevate the PMOS thick-field threshold. Thick-field thresholds of 40V can routinely be achieved by this means.

CMOS processes cannot tolerate the introduction of excess surface state charge because its effects are not limited to field regions. The magnitude of the surface state charge varies with processing conditions, and this, in turn, causes threshold voltage fluctuations. CMOS processes use (100)-oriented silicon and conduct an inert anneal to minimize the residual surface state charge. This anneal usually occurs in conjunction with the deposition of the protective overcoat. Many designers do not appreciate the importance of proper annealing. If a wafer is removed from processing before nitride deposition, then an anneal must be conducted in order to stabilize the threshold voltages and to sinter the contacts. Because this anneal does not necessarily duplicate the conditions of nitride deposition, the threshold voltages of no-nitride wafers do not always correspond to those of the finished product.

Most CMOS processes use two complementary channel-stop implants to suppress both NMOS and PMOS parasitic channels. All P-type field regions receive the P-type channel-stop implant to increase the magnitude of the PMOS thick-field threshold. Similarly, all N-type field regions receive the N-type channel stop implant to increase the magnitude of the NMOS thick-field threshold. Several methods have been devised to ensure the proper alignment of these channel-stop implants. The most common techniques involve either a blanket boron channel-stop implant and a patterned phosphorus channel-stop implant, or *vice versa*. Figure 3.23 shows the steps required to produce a blanket boron channel-stop implant and a patterned phosphorus channel-stop implant in an N-well CMOS process. Figure 11.11 shows the results.

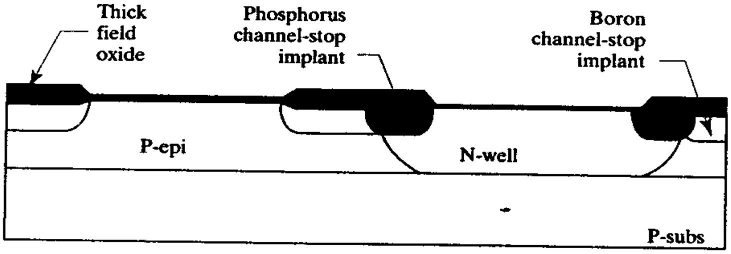


FIGURE 11.11 N-well CMOS wafer with boron and phosphorus channel-stop implants.

The channel-stop implants diffuse downwards during the long, high-temperature field oxidation. Lateral outdiffusion causes the two implants to intersect along the edges of the N-well. This zone of intersection limits the breakdown voltage of the N-well/P-epi junction. Fortunately, the channel-stop implants are sufficiently deep and lightly doped that the breakdown voltage lies well above normal operating voltages, and a 15V CMOS process can generally obtain an N-well/P-epi breakdown voltage in excess of 30V.

The patterned channel-stop implant requires a masking step. The locations receiving it depend on the type of process selected and which of the two implants is patterned. A patterned phosphorus channel stop in N-well CMOS goes into all N-well regions not inside moat. Although it is possible to draw the geometries for the channel-stop mask, it is much easier to generate the mask from existing coding layers. For example, the phosphorus channel-stop mask can be generated from the data present on the N-well, PMoat, and NMoat coding layers.

Submicron processes can sometimes dispense with one or both channel-stop implants. As the channel length shrinks, the backgate doping concentration rises and the operating voltage drops. The wells of a submicron process may therefore contain enough dopant to raise the thick-field threshold above the relatively low operating voltage of the process.

The channel-stop implants are designed to raise the NMOS and PMOS thick-field thresholds above the maximum operating voltage of the process. Many layout designers assume that this provides unconditional protection against parasitic channel formation, but in practice it does not. The maximum operating voltage of a process is usually set by either the gate oxide rupture voltage or by the breakdown voltage of the source/drain implants into their respective backgates. Certain components can operate at much higher voltages: for example a poly resistor is limited only by the breakdown of the thick-field oxide, which can easily reach several hundred volts. Similarly, the well-epi junctions can usually withstand several times the operating voltage of the process. Section 4.3.2 discusses the techniques used to suppress parasitic channel formation in circuits operating at or above the thick-field threshold.

11.2.4. Threshold Adjust Implants

Ideally, the threshold voltages of enhancement transistors should lie between 0.6 and 0.8V. The *native*, or *natural*, thresholds are determined by the doping of the gate and backgate and by the thickness of the gate oxide. Most processes dope the gate poly with phosphorus, reducing the magnitude of the NMOS threshold and increasing that of the PMOS. The natural NMOS threshold usually lies well below 0.6V and the magnitude of the natural PMOS threshold well above 0.8V. Over extremes of process and temperature, the NMOS goes into depletion, and the magnitude of the PMOS threshold exceeds 1.5V (Table 11.3). These thresholds are completely unacceptable for most applications.

TABLE 11.3 Worst-case natural and adjusted threshold voltages for a typical 10V N-well CMOS process.[22]

Worst-case Corner	Natural NMOS	Adjusted NMOS	Natural PMOS	Adjusted PMOS
Minimum	–0.10	0.50	–1.75	–1.15
Nominal	0.20	0.80	–1.40	–0.80
Maximum	0.55	1.15	–1.10	–0.50

The threshold voltage of an MOS transistor can be altered by implanting its channel region. A P-type implant produces a positive threshold shift and an N-type implant a negative one. NMOS and PMOS transistors using phosphorus-doped gate poly both require a positive threshold shift. Providing that the initial backgate dopant concentrations have been properly chosen, a single boron implant can adjust the thresholds of both types of transistors. This boron implant is called the *threshold adjust implant*, or simply the *threshold adjust*. Transistors receiving this implant are called *adjusted* transistors, while those not receiving it are called *native*, or *natural*,

[22] These values assume the listed natural V_t targets, ±0.15V threshold control and a –2mV/°C temperature coefficient.

transistors. The threshold adjust implant does not necessarily require a photomask. If the implant is performed across the entire wafer immediately after stripping the LOCOS nitride, then it appears in every moat region. This blanket implant simultaneously adjusts the threshold voltage of every MOS transistor to the targeted value. This practice precludes the fabrication of natural devices.

Circuit designers can often improve the performance of their circuits if they have access to both natural and adjusted transistors. Many processes therefore offer natural transistors as a process option. This option requires a single mask, properly called the *threshold adjust implant mask,* but more often referred to as a *natural* V_t *mask.* The associated coding layer has been given many names; in this text it is called *NatVT.*[23] This layer must be coded around the gate region of each natural transistor (Figure 11.12). The NatVT figure should slightly overlap the channel region to allow for misalignment and lateral outdiffusion. If a design does not use any natural transistors, then the NatVT mask can usually be omitted. A few processes may use the NatVT mask to fabricate certain other devices, such as Schottky diodes.

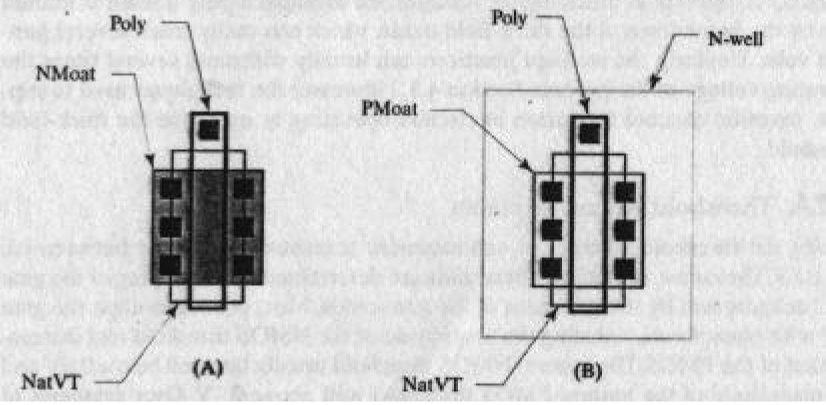


FIGURE 11.12 Layout of natural transistors using NatVT: (A) natural NMOS and (B) natural PMOS.

Although many processes have successfully used a single boron threshold adjust implant, submicron processes often require a different strategy. The boron implant reduces the magnitude of the PMOS threshold, but it also has the undesired effect of producing a *buried channel.* In order to obtain a large threshold shift, so much boron must be implanted that it actually inverts a thin layer of the backgate. The inversion region appears beneath the surface because this is where the peak doping concentration from the implant occurs. The buried channel lies so close to the surface that the electric field produced by the contact potential of the gate electrode inverts it, and it does not interfere with the normal operation of the transistor. This situation changes in a submicron transistor because the backgate doping increases as the channel length decreases. The increased doping partially shields the buried channel from the influence of the gate electrode. The gate can no longer fully invert the channel, so the buried channel begins to conduct current. Submicron buried-channel PMOS transistors are therefore somewhat leaky.

The buried channel can be eliminated by using a phosphorus channel-stop implant for the PMOS transistors. Since phosphorus induces a negative threshold shift, the PMOS transistor must begin with a relatively low threshold voltage, which can be achieved by using a boron-doped gate poly (Figure 11.13).

[23] The natural V_t mask is also called *NVT,* but this name is also used for the N-type threshold adjust mask used in a dual-doped poly process.

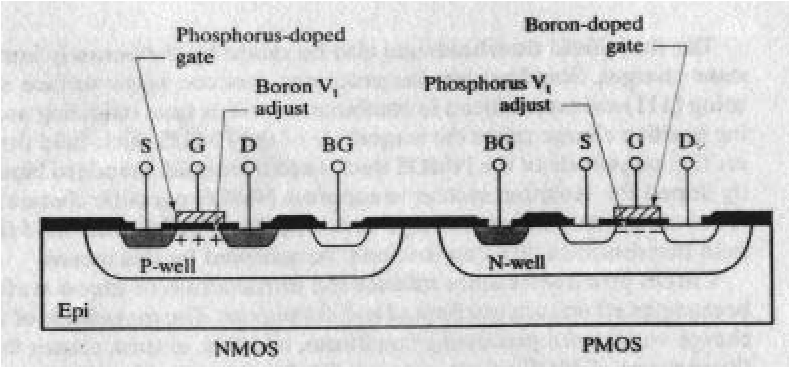


FIGURE 11.13 Cross section of dual-doped poly CMOS transistors.

The processing required to produce *dual-doped poly* CMOS transistors is as follows. After the LOCOS nitride has been stripped, the wafer is patterned using the mask for the *P-type V_t adjust* (PVT). This low-energy boron implant adjusts the NMOS threshold. Next, the wafer is patterned using the mask for the *N-type V_t adjust* (NVT). This low-energy phosphorus implant adjusts the PMOS threshold. After gate oxidation, the gate poly is deposited in a near-intrinsic state. Before etching, it is doped with a patterned boron implant using the *P-type gate poly* (PPoly) mask, followed by a patterned phosphorus implant using the *N-type gate poly* (NPoly) mask.

The most elaborate version of this process requires four new masks (PVT, NVT, PPoly, and NPoly). As long as no natural transistors are required, the P-well mask can be reused for the PVT step, and the N-well mask can be reused for the NVT step. If natural transistors are required, then the well masks cannot be reused in this manner and separate PVT and NVT masks become necessary. A well mask can also be used to define either the PPoly or the NPoly, but not both. Suppose that the N-well mask is used to pattern the NPoly. Both the P-well and the epi regions must receive the PPoly implant, and a special mask must be produced for this purpose. This process therefore requires at least one new mask, and possibly as many as four.

The number of masking steps can be reduced at the cost of compromising performance. For example, the gate poly can be doped with a blanket boron implant and a patterned phosphorus implant. The resulting poly is not quite as heavily doped as that formed by using two separate masking steps. The process can also use a blanket V_t implant followed by a patterned V_t implant of the opposite polarity in order to save one masking step. The transistor that receives both implants has somewhat more threshold voltage variation than it would have had using one implant. If both of these modifications are adopted, then only two masking steps are required instead of four. Even so, the additional steps add cost and complexity, so they are only used for submicron processes that would otherwise exhibit unacceptable leakage due to buried channel formation. In general, operating voltages of 5V or more require well dopings compatible with single-doped poly, while lower operating voltages require dual-doped gate poly.

Selective gate doping produces unwanted poly diodes unless the gate poly is silicided. PN junctions appear wherever the PPoly and the NPoly abut one another. They can be shorted by metal jumpers or, more conveniently, by silicidation. As long as the designer takes care not to block the silicide from locations where PPoly and NPoly abut one another, silicidation automatically shorts all of the poly diodes. Silicidation greatly increases the rate at which dopants diffuse through the poly, so the intersections between PPoly and NPoly must be spaced well away from the gate regions of adjacent MOS transistors.[24] Although the PPoly-NPoly junction does exhibit rectification, poly diodes are not recommended as circuit components because the presence of grain boundaries within the depletion regions causes substantial leakage.

Almost all CMOS processes adjust the NMOS and PMOS threshold voltages. The majority of analog processes offer natural NMOS and PMOS transistors either as part of the baseline process or as process extensions. A few processes offer additional threshold voltage options, such as depletion-mode transistors or low-V_t PMOS transistors. Each such option requires its own threshold adjust implant, formed through an additional masking step. Transistors using these special implants are coded much like natural transistors, except that NatVT is replaced by the layer that codes for the special implant.

11.2.5. Scaling the Transistor

Integrated circuits have become vastly more complex over the past thirty years. The first digital integrated circuits contained ten or twenty transistors; their modern equivalents contain tens of millions. This remarkable increase in complexity has largely been made possible by corresponding reductions in the size of individual transistors. From 1973 to 2000, minimum channel lengths went from 8μm to about 0.2μm.[25] These reductions in size have also improved the performance of the transistors. A set of guidelines called *scaling laws* have been developed that dictate how the various dimensions of an MOS transistor should be reduced to obtain the best performance.

Scaling laws fall into two general categories, both of which presume that width and length are multiplied by a *scaling factor S. Constant-voltage scaling* holds the operating voltage of the transistor constant while scaling its dimensions. As the transistor shrinks further and further, it becomes increasingly difficult to avoid hot-carrier generation and punchthrough breakdown. *Constant-field scaling* avoids these problems by reducing the supply voltage to keep the electric fields in the transistor constant regardless of scale. Most modern processes use some variant of constant-field scaling. Table 11.3 shows simplified rules for both constant-voltage and constant-field scaling laws.[26]

As an example, consider a process producing 5V transistors with minimum dimensions of 1μm long by 2.5μm wide using a 250Å gate oxide and a backgate doping concentration of 10^{16}cm^{-3}. Suppose the channel length of this process is reduced to 0.8μm using constant-field scaling. The scaling factor S equals 0.8μm/1.0μm, or 80%. According to Table 11.4, the scaled transistor should have a minimum width of 2.0μm, a 200Å gate-oxide, and a backgate doping of $1.25 \cdot 10^{16}\text{cm}^{-3}$. Since processes are generally identified by gate length, the original (100%) process would be considered a 1μm process, while the 80% shrink would be an 0.8μm process.

Shrinking a transistor actually improves its performance. The smaller dimensions reduce parasitic capacitances and increase switching speeds. The *gate delay* of a CMOS process equals the time required for a digital signal to propagate through a representative CMOS gate. As the transistors scale down, the gate delay decreases and the circuit can handle faster switching speeds. Early microprocessors operated

[24] Y. P. Tsividis, *Operation and Modeling of the MOS Transistor* (New York: McGraw-Hill, 1988), p. 439.

[25] The 8μm figure is from D. A. Pucknell and K. Eshraghian, *Basic VLSI Design*, 3rd ed. (Sydney: Prentice-Hall Australia, 1994), p. 7. Both figures are approximations of industry practice; much smaller dimensions are possible in a research environment.

[26] These laws have been adapted from Pucknell *et al.*, p. 129.

TABLE 11.4 Constant-voltage and constant-field scaling laws.

Quantity	Constant-voltage	Constant-field
Supply voltage	1	S
Minimum channel width	S	S
Minimum channel length	S	S
Gate oxide thickness	1	S
Backgate doping	$1/S^2$	$1/S$
Gate delay	S^2	S
Power-delay product	S^2	S^3

at clock speeds of 1 to 10MHz; their modern equivalents operate at 100 to 1000MHz.

Not only does a smaller transistor switch faster, but it requires less power to do so. CMOS logic gates require pulses of power to charge and discharge gate capacitances each time they switch. The faster the gate switches, the more transitions occur per second, and the larger the current consumption becomes. The supply current required by a gate can be reduced at the expense of increasing its gate delay. The product of gate delay and power consumption remains approximately constant for any given process. This *power-delay product* decreases as the size of the transistor shrinks. For example, 80% constant-field scaling reduces the power-delay product to about half of its initial value. As this example suggests, even relatively minor decreases in size significantly reduce power consumption. This is fortuitous, since otherwise a microprocessor running at several hundred megahertz would literally melt from its own waste heat.

Scaling laws are frequently applied to existing digital layouts to convert them for use with newer processes. Rather than laboriously recoding the layout, the designer simply runs a program that scales all of the data by a specified amount. This type of scaling is called an *optical shrink* because it produces the same results as photoreducing the existing mask set. Optical shrinks are denoted by the percentage scaling factor used to transform the data from their original, or *drawn*, dimensions to their final, or *shrunk*, dimensions. A 100% shrink indicates that the final dimensions equal the drawn dimensions, while an 80% shrink indicates that they equal 4/5 of the drawn dimensions. Figure 11.14A shows a 1μm transistor drawn at 100%. Figure 11.14B shows the same transistor optically shrunk to 80%. The optical shrink scales both the width and the length of the device in a manner consistent with either constant-voltage or constant-field scaling. The process engineers will adjust gate oxide thickness, backgate doping, and other parameters according to the desired type of scaling.

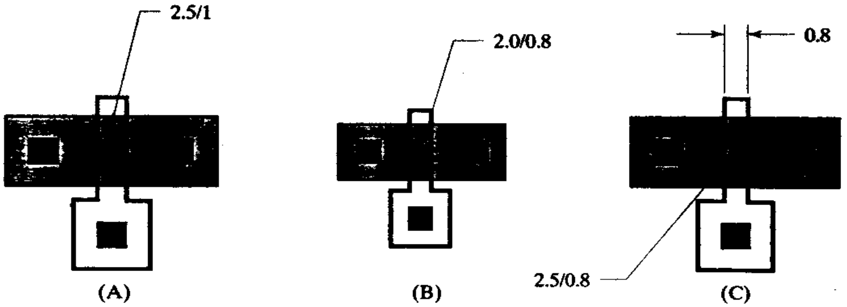


FIGURE 11.14 Examples of scaled MOS transistors: (A) drawn at 100%, (B) optically shrunk to 80%, (C) selectively shrunk to 80% of drawn gate length. The wells have been omitted for clarity.

An optical shrink affects all dimensions equally, but some are more difficult to scale than others. Multilayer metal systems have proved especially difficult to scale to submicron dimensions. Although fine-line metal systems certainly exist, they are very expensive. Many processes selectively scale channel length while retaining the previous dimensions for all other layout rules. Figure 11.14C shows a 1μm transistor whose gate has been shrunk to 0.8μm. The selective gate shrink requires a more complicated set of geometric transformations than a simple optical shrink, but it is still far simpler and quicker than a full relayout. The benefits of a selective gate shrink are somewhat less than for a full optical shrink (Table 11.5), but they are still sufficient to justify selective gate shrinks for many processes.

TABLE 11.5 Scaling laws for selective gate shrinks.

Quantity	Constant-voltage	Constant-field
Supply voltage	1	S
Minimum channel width	1	1
Minimum channel length	S	S
Gate oxide thickness	1	S
Backgate doping	$1/S^2$	$1/S$
Gate delay	S	1
Power-delay product	S	S^2

The scaling laws were originally developed for digital processes. CMOS logic circuits respond quite predictably to scaling. but the same is not true of analog or mixed-signal circuits. No set of predetermined scaling laws can comprehend the full complexity of analog circuit design. Indiscriminate scaling usually causes analog circuits to fail parametric specifications. and in some cases it may cause outright malfunctions. For example, an 80% optical shrink reduces all capacitors to 64% of their former values. Since analog designs rely on capacitors to stabilize feedback loops, a reduction in capacitance can actually destabilize the circuit. Constant-field scaling only makes matters worse because it simultaneously increases transconductance and reduces capacitance. Selective gate shrinks do not change capacitance values, but they are still risky because they can introduce unforeseen parametric changes in short-channel transistors that frequently prove more significant for analog circuits than for digital ones. To summarize. analog and mixed-signal circuits should not be scaled without re-evaluating the performance of the resulting circuit to ensure that it still meets functional and parametric specifications.

11.2.6. Variant Structures

The simplest type of self-aligned poly-gate transistor consists of a rectangle of NMoat or PMoat bisected by a strip of poly. This type of structure serves admirably for width-to-length ratios of less than ten. Transistors with larger W/L ratios become increasingly unwieldy unless they are divided into multiple identical sections connected in parallel. Figure 11.15A shows the layout of a three-section transistor. The paralleled fingers not only produce a more convenient aspect ratio, but they also save area because adjacent sections share source and drain fingers. The merger of adjacent source/drain fingers can also reduce parasitic junction capacitance by up to 50%.

The division of a transistor into sections can affect its matching, so circuit designers often specify the number of sections for critical transistors. The most common notation for a sectioned transistor is $N(W/L)$, which denotes N sections, each with a drawn width of W and a drawn length of L. Transistors specified in this manner should be laid out exactly as requested. If the transistor is specified as having dimen-

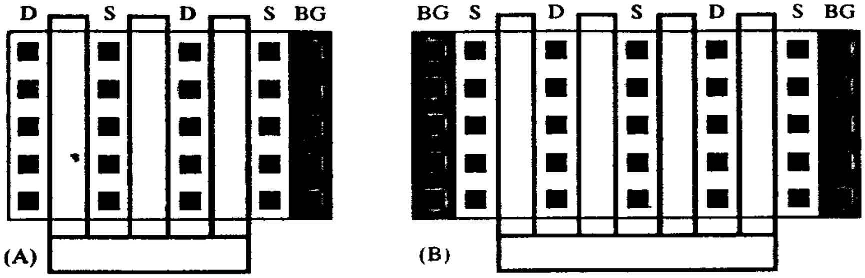


FIGURE 11.15 Sectioned transistors of (A) three and (B) four sections. The fingers are marked S (source), D (drain), and BG (backgate). The wells have been omitted for clarity.

sions *W/L*, this usually indicates that it can have any number of segments desired. If this transistor must match one having dimensions *N(W/L)*, then the former device should be laid out as a single section.

Transistors with even numbers of sections always contain odd numbers of source/drain fingers (Figure 11.15B). Such transistors are usually constructed with source fingers at either end. Not only does this allow the use of abutting backgate contacts on either or both ends, but it also reduces the number of drain fingers by one. This arrangement minimizes parasitic drain junction capacitance at the expense of source capacitance. Drain capacitance usually has more effect upon circuit performance than source capacitance, so a reduction in drain capacitance at the expense of source capacitance usually improves circuit performance.

Transistors sharing common source or drain connections are frequently merged to save space or to minimize parasitic junction capacitance. The merger is a relatively simple matter so long as both transistors contain sections of the same width. Differing widths require the use of a notched moat (Figure 11.16). The layout rules usually prohibit the placement of polysilicon immediately adjacent to a moat edge due to the large oxide step present at this location. The spacing between poly and moat S_{PM} forces a slight increase in the area of the shared source/drain finger, but one shared finger still consumes less area than two separate fingers.

Transistors M_1 and M_2 in Figure 11.16 share a common source region, so the drain fingers must occupy the ends of the array. This arrangement precludes the use of an abutting backgate contact, so the backgate contact is placed some distance away from the devices. The spacing between the backgate contact and the merged transistors may seem to eliminate any area benefit produced by the merger, but this backgate contact can also serve several other devices. Transistors sharing a common drain have source fingers on either end of the array and can therefore use abutting backgate contacts.

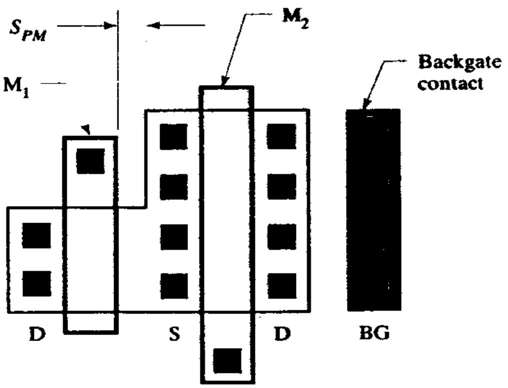


FIGURE 11.16 Merged transistors M_1 and M_2 share a common source (well omitted for clarity).

CMOS layout makes extensive use of merged devices to save space and to minimize capacitance. Figure 11.17 shows a simple layout of a two-input NAND gate that illustrates many of the techniques in common use. PMOS transistors M_1 and M_2 occupy a common well placed at the top of the layout. These transistors share a common drain region that not only reduces the width of the cell but also minimizes the drain capacitance on the output node Z. The two PMOS transistors also share a single backgate contact at the right end of the well. NMOS transistors M_3 and M_4 reside next to one another near the bottom of the layout. These transistors have been placed in series—the drain of M_3 simultaneously acts as the source of M_4. No contacts are necessary, because the current simply flows from one channel to the next. One strip of poly forms the gates of transistors M_2 and M_3, and a second strip of poly forms the gates of M_1 and M_4. The spacing between M_3 and M_4 is slightly larger than minimum. If desired, this spacing can be minimized by angling the gate leads toward one another.

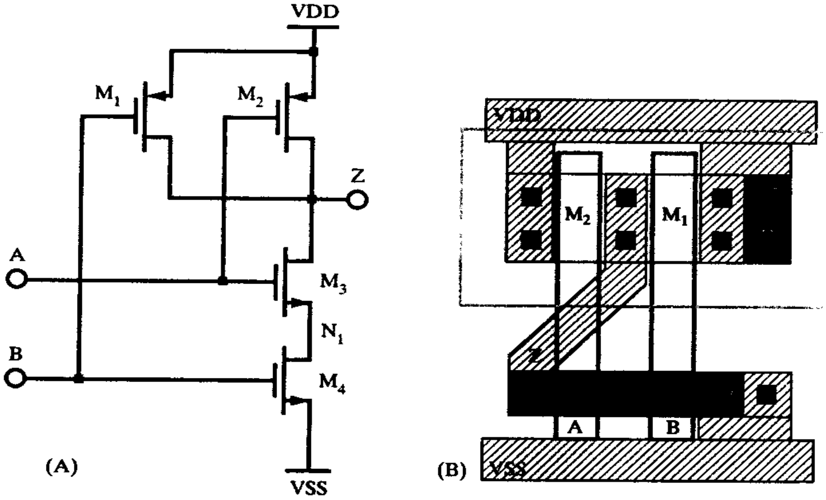


FIGURE 11.17 (A) Schematic and (B) layout of a two-input NAND gate.

The layout in Figure 11.17 follows the general guidelines of digital standard cell design. The power and ground rails run across the top and bottom of the cell, respectively. The width and spacing of these leads should be the same for all logic cells so they can stack end-to-end. The PMOS transistors occupy a common well spanning the top of the cell. When multiple logic cells stack end-to-end, their wells overlap to form a single contiguous region running the entire length of the assembly. This arrangement avoids the well-to-well spacings that would otherwise appear between adjacent cells. The NMOS transistors reside near the bottom of the cell, either in the epi or in another common well. Each cell contains at least one substrate and one backgate contact. Larger cells should contain additional substrate and backgate contacts wherever possible. The input and output connections exit from either the top or the bottom of the cell, whichever is more convenient for a given layout. Digital standard cells frequently contain special elements called *ports* and *prels* required by autorouting software. Since most autorouters cannot handle analog routing, there is little point in adding ports and prels to analog cells. The designer may still wish to employ concepts such as standard cell heights and consistent power and ground rail placement to allow analog cells to stack together. The height of analog cells is usually much greater than that of digital cells to accommodate their larger components and greater interconnection complexity. Additional rails are sometimes necessary to accommodate separate analog and digital supplies or to distribute several different supplies throughout a multisupply system.

Serpentine Transistors

Some designs require transistors with very long channels. The most convenient layout for such devices consists of a strip of NMoat or PMoat placed underneath a plate of polysilicon. A very compact layout results if one folds the moat into a serpentine pattern (Figure 11.18). The total channel length is computed by a procedure analogous to that used for serpentine resistors. Each 90° bend in the channel adds one-half of the transistor's width to its total length. The channel length of the transistor in Figure 11.18 therefore equals $2L_X + L_Y + W$. Serpentine transistors will not match precisely unless they have identical geometries, but most designs do not require especially precise matching from long-channel devices.

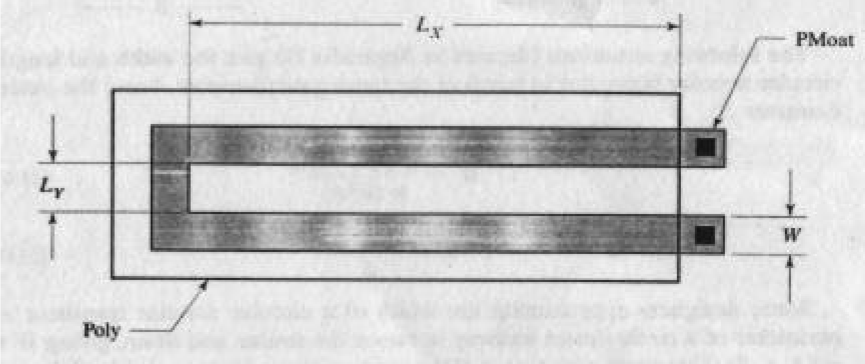


FIGURE 11.18 Serpentine PMOS transistor (well omitted for clarity).

Annular Transistors

The drain capacitance of an MOS transistor limits its switching speed and frequency response. Many circuits, analog as well as digital, can benefit from reduced drain capacitances. A smaller transistor has less capacitance, but it also provides less transconductance. These factors offset one another, so the smaller transistor is generally no faster than its larger counterparts. In order to actually increase switching speed, one must reduce the ratio of drain capacitance to transistor width C_D/W. Interdigitation reduces the C_D/W ratio by half because it surrounds each drain with two gates. This same principle can be carried still further by surrounding the drain on all four sides by an annular gate (Figure 11.19). An annular transistor will provide the smallest possible C_D/W ratio, but the decreased drain capacitance comes at the expense of increased source capacitance. The increased source capacitance is not necessarily injurious because the source often connects to a low-impedance node such as a power supply rail.

Two basic types of annular transistors exist: those that use a square gate geometry (Figure 11.19A) and those that use a circular gate (Figure 11.19B). The circular gate theoretically provides the highest C_D/W ratio because it minimizes the area-to-periphery ratio of the drain. The current flow through a circular gate is quite symmetric and the width of the transistor is easily computed. The current flow through square gates is less uniform and the effective width of the transistor is less easily computed. Square gates also have sharp corners that can induce premature avalanche breakdown due to electric field intensification.

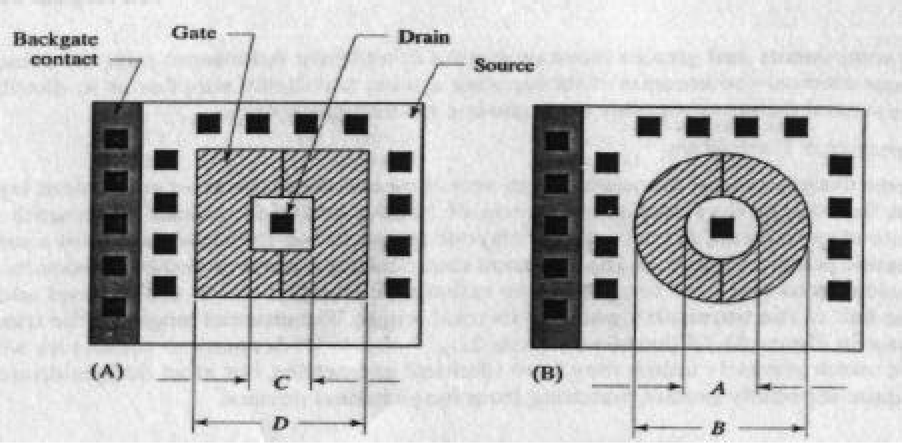


FIGURE 11.19 Annular MOS structures: (A) square and (B) circular.

The following equations (derived in Appendix D) give the width and length of a circular annular transistor in terms of the inner gate diameter A and the outer gate diameter B:

$$W = \frac{\pi\,(B - A)}{\ln\,(B/A)} \qquad [11.9A]$$

$$L = \frac{B - A}{2} \qquad [11.9B]$$

Some designers approximate the width of a circular annular transistor as the perimeter of a circle drawn halfway between the source and drain, giving $W \cong 1/2\ \pi\ (A + B)$. This approximation slightly overestimates the true width of the transistor. The errors caused by the approximation have little impact because precision circuits always rely on matching between identical devices rather than the properties of any one device.

The width and length of a square annular transistor are given by the following approximations, which do not correct for corner effects:

$$W \cong 2(C + D) \qquad [11.10A]$$

$$L \cong \frac{D - C}{2} \qquad [11.10B]$$

Annular transistors are often elongated to produce gate geometries similar to those in Figure 11.20. The C_D/W ratio of the elongated annular transistor is not much smaller than that of a conventional interdigitated transistor, so elongated structures are not recommended for minimizing drain capacitance. They are still sometimes used to produce an enclosed channel (Section 12.1.2). The W and L of an elongated circular annular transistor (Figure 11.20A) are approximately

$$W = \pi\,\frac{(B - A)}{\ln(B/A)} + 2U \qquad [11.11A]$$

$$L = \frac{B - A}{2} \qquad [11.11B]$$

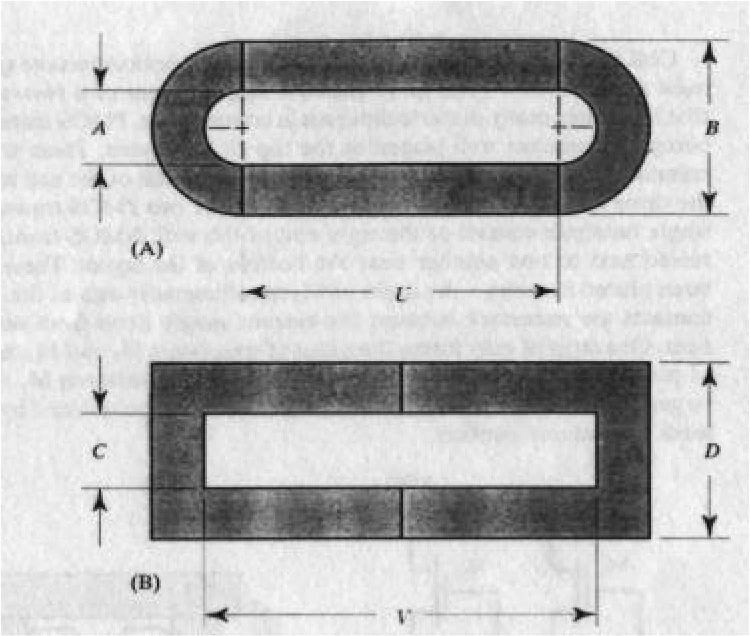


FIGURE 11.20 Gate geometries for elongated annular transistors: (A) circular and (B) square.

Similarly, the *W* and *L* of an elongated square annular transistor (Figure 11.20B) are approximately

$$W \cong 2V + C - D \qquad [11.12A]$$

$$L \cong \frac{D - C}{2} \qquad [11.12B]$$

11.2.7. Backgate Contacts

All MOS transistors require electrical connection to their backgates even though no current normally flows through them. MOS transistors that lack backgate contacts or have excessive backgate resistance are particularly prone to latchup. Every PMOS transistor contains a parasitic lateral PNP and every NMOS contains a parasitic lateral NPN. Together these form a parasitic SCR (Figure 4.20). The backgate contacts short the base-emitter junctions of these parasitic transistors, and the associated backgate resistances become base turn-off resistors (R_1 and R_2 in Figure 4.20). The SCR will remain off as long as the voltage across both of these resistors remains less than the forward voltage of the respective base-emitter junctions. The voltage necessary to trigger the SCR into conduction equals about 0.65 to 0.7V at 25°C, but this falls to 0.4 to 0.45V at 150°C due to the temperature coefficient of V_{BE}. Not only does the trigger voltage drop at high temperatures, but the betas of the parasitic transistors actually increase. Thus. CMOS latchup is most likely to occur at high temperature.

Most CMOS products must pass a standardized test that measures their latchup susceptibility. Positive and negative test current pulses are applied to each pin, while power is applied to the part. Depending on specifications, the magnitude of these test pulses may range from as little as ±100mA to as much as ±250mA. The supply current is measured both before and after the application of each test pulse. If these two currents are not approximately the same. then the part fails the test.

This latchup test can be modeled mathematically. Suppose that a test current I_T flows through the source/drain region of an MOS transistor M_1. In order to prevent latchup from occurring between M_1 and a complementary MOS transistor M_2, at least one of the following inequalities must be true:

$$\beta_{12}\beta_{21}(1-\eta_{c12})(1-\eta_{c21}) < 1 \qquad [11.13A]$$

$$I_T R_{B2}(1-\eta_{c12})\left(\frac{\beta_{12}}{\beta_{12}+1}\right) < V_{trig} \qquad [11.13B]$$

β_{12} represents the beta of the parasitic bipolar formed by minority carriers flowing from the source/drain region of M_1 to the backgate of M_2 in the absence of guard rings. β_{21} represents the beta of the parasitic bipolar formed by minority carriers flowing from the source/drain region of M_2 to the backgate of M_1, again in the absence of guard rings. η_{c12} represents the fraction of minority carriers flowing from M_1 to M_2 intercepted by guard rings. Similarly, η_{c21} represents the fraction of minority carriers flowing from M_2 to M_1 intercepted by guard rings. I_T equals the test current, R_{B2} equals the backgate resistance of M_2, and V_{trig} equals the trigger voltage of the SCR (about 0.4V at 150°C).

These equations provide some insight into the roles of guard rings and backgate contacts in suppressing latchup. Equation 11.13A represents the condition required to avoid sustained feedback. Minimizing parasitic betas β_{12} and β_{21} and adding guard rings to improve collector efficiencies η_{c12} and η_{c21} can help prevent sustained conduction. Any device meeting this criterion is invulnerable to CMOS latchup regardless of the magnitude of the test currents applied. Unfortunately, few CMOS processes can satisfy equation 11.13A because their transistors lie too close together and their guard rings are too inefficient. CMOS devices can still achieve conditional latchup immunity by satisfying the conditions of equation 11.13B. The four terms in this inequality represent the magnitude of the test current and the backgate resistance, the effectiveness of the guard rings, and the magnitude of the parasitic beta, respectively. The contributions of guard rings and backgate contacts multiply one another, producing a synergistic relationship between the two. Even if neither guard rings nor backgate contacts alone can stop latchup, a combination of the two frequently can. Guard rings require so much room that they can only be placed around a few devices—usually those that may potentially inject minority carriers into the die. Backgate contacts require much less area, so each transistor can have its own backgate contact or can at least share a backgate contact with another transistor.

The backgate of an NMOS must connect to a voltage less than or equal to its source, and the backgate of a PMOS must connect to a voltage greater than or equal to its source. In many applications, the backgate can be connected to the source. A few transistors operate under conditions in which it is difficult or impossible to distinguish source from drain. A few circuits connect the backgate to a voltage that is different from the source to increase the threshold voltage using the body effect. Some high-speed circuits also avoid connecting the source and backgate in order to minimize the capacitance appearing at the source node. All of these circuits require an independent backgate contact like that in Figure 11.21B. Transistors whose source and backgate operate at the same potential can use an abutting backgate contact (Figure 11.21A). This contact saves considerable area by eliminating the spacing between source and backgate diffusions. The intersection of two heavily doped diffusions produces a leaky and unreliable junction, but this defect can be tolerated as long as the two diffusions are connected together by metal or silicide.

Relatively small transistors such as those in Figure 11.21 require only one backgate contact. The distance across the transistor grows larger as additional segments

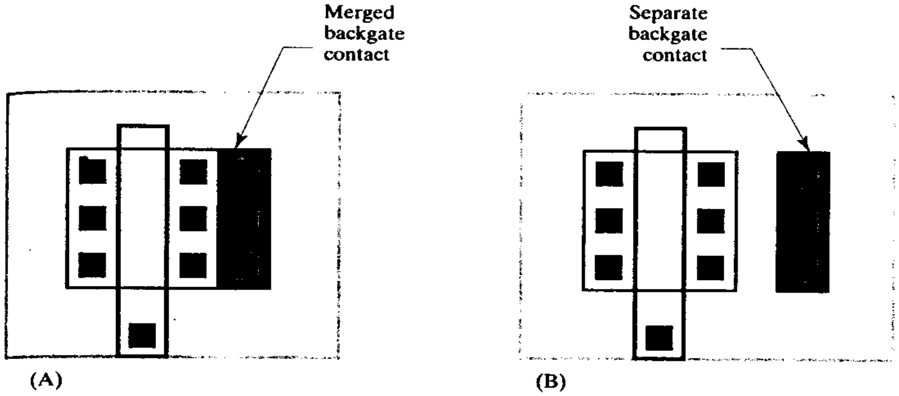


FIGURE 11.21 Examples of (A) abutting and (B) separate backgate contacts.

are added, and at some point this distance becomes so great that it produces an unacceptably large backgate resistance. A second backgate contact on the opposite side of the transistor reduces the distance to the backgate contacts at the cost of slightly increasing the device area (Figure 11.15B). The point at which a second backgate contact becomes necessary varies depending on the sheet resistance of the backgate. An NMOS transistor constructed in the P-epi above a P+ substrate would have a lower backgate resistance than a PMOS constructed in a shallow, lightly doped N-well. Adding a buried layer to a well likewise decreases the backgate resistance of transistors occupying it. Factors of this sort make it difficult to provide quantitative rules for backgate contact spacing. Some processes specify a maximum distance between any portion of a transistor and the nearest backgate contact. The maximum allowed distance becomes shorter as the backgate resistance increases. Typical spacings range from 25μm to 250μm. Transistors subject to large transients should use a smaller distance to provide additional latchup immunity. These include ones whose source/drain regions connect to pins, and those residing next to transistors whose source/drain regions connect to pins.

Large transistors with many fingers may require substrate contacts embedded within the body of the transistor itself. This is usually achieved by placing strips of backgate contact through the transistor at regular intervals (Figure 11.22A). Although these *interdigitated backgate contacts* reduce the distance to the nearest

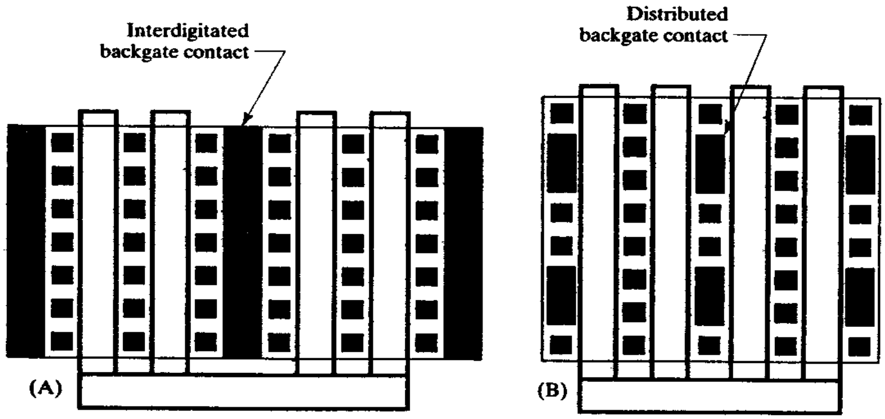


FIGURE 11.22 Additional styles of backgate contacts: (A) interdigitated backgate contact and (B) distributed backgate contact.

substrate contact, they also substantially increase the size of the transistor. Some processes allow another type of backgate contact, consisting of small plugs of backgate diffusion placed in holes within the source fingers of the transistor (Figure 11.22B). These *distributed backgate contacts* slightly increase the source resistance of the overall transistor, but they greatly reduce the area required by the backgate contacts. A substantial area savings can be obtained even if the transistor must be enlarged to compensate for increased source resistance. Distributed backgate contacts can be placed on every source finger, as shown, or they may be placed on only a few source fingers distributed at regular intervals across the transistor. A larger number of distributed backgate contacts further reduces backgate resistance, but not all applications necessarily require the same degree of latchup protection. Distributed backgate contacts always connect to the source of the transistor, so some applications may still require the use of interdigitated contacts.

Analog BiCMOS processes often contain additional diffusions that can help reduce the latchup susceptibility of MOS transistors. For example, many analog BiCMOS processes fabricate NPN transistors in the same N-well as PMOS transistors. The NBL that is used to reduce the collector resistance of the CDI NPN can also reduce the backgate resistance of the PMOS. If NBL is available, it should be placed in all MOS transistors that use the same well as the CDI NPN. Very efficient hole-blocking guard rings can be produced in many processes that offer deep-N+ and NBL. Some transistors may actually operate under conditions in which their source/drain regions regularly forward-bias into the backgate. Deep-N+ guard rings can help ensure that substrate injection from these PMOS transistors does not disrupt the operation of the rest of the circuit.

NMOS transistors are somewhat more difficult to protect from latchup than PMOS transistors. An isolated NMOS structure (Section 11.2.2) offers total immunity to CMOS latchup, but at the price of greatly increased backgate resistance. Although these transistors do not require a low backgate resistance in order to suppress CMOS latchup, parasitic lateral NPN action remains a concern. If the transistors must operate at relatively high voltages, they may become susceptible to snapback breakdown due to parasitic NPN action. Even if they operate well below the $V_{CEO(sus)}$ of the NPN, minority carrier injection still causes sluggish switching due to charge storage. A system of distributed backgate contacts can minimize these problems. An NMOS transistor fabricated above a P+ substrate and surrounded by a deep-N+ electron-collecting guard ring will also prove quite resistant to latchup. This style of transistor is suited to applications where the transistor must withstand severe transients, but where its source/drain regions do not routinely forward-bias into the backgate. Transistors of the latter sort are best constructed as isolated NMOS transistors.

11.3 SUMMARY

This chapter has covered the construction of conventional small-signal poly-gate CMOS transistors. The next chapter covers a variety of more specialized types of transistors, including extended-voltage transistors, power transistors, DMOS transistors, and JFETs. These transistors can fill a very wide range of applications, including many that are traditionally filled by bipolar transistors.

11.4 EXERCISES

Refer to Appendix C for layout rules and process specifications.

11.1. Suppose an enhancement NMOS has a threshold voltage of 0.7V and a transconductance of 220μA/V^2. Determine the region of operation and compute the drain current for each of the following biasing conditions:

a. $V_{GS} = 1.2V, V_{DS} = 2.3V$.
b. $V_{GS} = 1.2V, V_{DS} = 0.2V$.
c. $V_{GS} = -1.0V, V_{DS} = 4.4V$.

11.2. Suppose the enhancement NMOS in Exercise 11.1 is subjected to the following terminal voltages: $V_{GS} = 1.2V, V_{DS} = -2.3V$. Recognizing that the source and drain have swapped roles, determine the true electrical biasing conditions, the mode of operation, and the drain terminal current.

11.3. What is the process transconductance of an NMOS transistor having a composite gate dielectric consisting of 150Å of nitride ($\varepsilon_r = 6.8$) sandwiched between two layers of oxide, each 50Å thick ($\varepsilon_r = 3.9$)? *Hint:* See Section 6.1.

11.4. Estimate the process transconductances of NMOS and PMOS transistors having a maximum operating voltage of 15V.

11.5. Suppose an enhancement PMOS transistor with an N+ poly gate electrode has a nominal threshold voltage of –0.95V. What would the nominal threshold voltage become if the transistor used a P+ poly gate?

11.6. Would an enhancement PMOS transistor with a nominal threshold voltage of –0.4V serve as a useful device for constructing digital logic gates? Explain.

11.7. Lay out the inverter shown in Figure 11.23 using the poly-gate CMOS rules listed in Appendix C. Place a hole-collecting guard ring around the PMOS transistor and an electron-collecting guard ring around the NMOS. Connect the guard rings to provide the best possible protection.

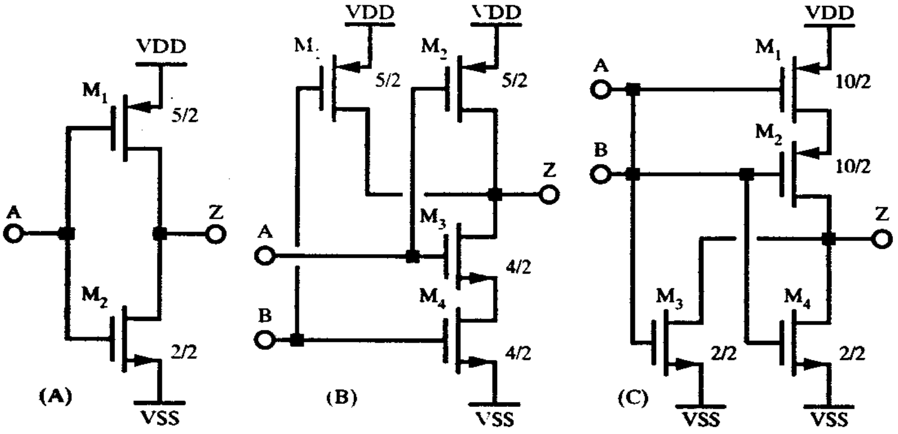


FIGURE 11.23 Three standard-cell logic gates: (A) Inverter. (B) NAND, and (C) NOR.

11.8. Construct an isolated NMOS transistor using the analog BiCMOS rules in Appendix C. The transistor should have a W/L ratio of 10/5. Include a single abutting substrate contact.

11.9. A 3μm CMOS process can withstand a maximum operating voltage of 12V and offers a gate delay of 2.3nS. Suppose constant-field scaling is applied to produce a 2μm process. Predict the maximum operating voltage and gate delay of the scaled process.

11.10. Lay out the following transistors using poly-gate CMOS rules. Include abutting backgate contacts, gate interconnections, and well geometries as necessary.
a. NMOS. 3(5/15)
b. NMOS. 12(20/5)
c. PMOS. 7(10/25)
d. PMOS. 4(10/3)

11.11. Lay out a natural NMOS transistor with dimensions of 3(10/3) and a natural PMOS transistor with dimensions of 25/25. Include abutting backgate contacts, gate interconnections, and well geometries as necessary. The layout rules for the NATVT layer are as follows:

1.	**NATVT** width	4μm
2.	**NATVT** overlap of **GATE**	2μm
3.	**NATVT** spacing to **TOX**	4μm
4.	**NATVT** spacing to **POLY**	4μm

Note: GATE is defined as the intersection of POLY with either NMOAT or PMOAT.

11.12. Construct standard-cell layouts of each of the three logic gates in Figure 11.23. The NAND gate should resemble the layout in Figure 11.17B. The VDD and VSS leads should be 4μm wide, and each cell should have at least one substrate contact and one well contact. Design the cells so they can be stacked together by abutting their VDD and VSS leads. No violations of layout rules should occur regardless of the order in which the cells are stacked.

11.13. Using the poly-gate CMOS rules of Appendix C, construct a serpentine PMOS with a nominal device transconductance of 0.01μA/V^2. Fold the gate as many times as necessary to produce an approximately square layout.

11.14. Lay out the following annular transistors using poly-gate CMOS rules. Include abutting backgate contacts, gate interconnections, and well geometries as necessary.

a. Circular NMOS, 31.4/4

b. Square PMOS, 48/4

c. Elongated circular PMOS, 51.4/4

11.15. Construct a 5000/2 PMOS transistor using poly-gate CMOS rules. Divide the transistor into as many sections as required to produce an approximately square layout. Include enough interdigitated backgate contacts to ensure that no part of the transistor is more than 50μm from the nearest backgate contact.