12 Applications of MOS Transistors

Some designs require MOS transistors to operate at high voltages without experiencing breakdown or parametric shifts. Although breakdown voltages can be increased by using lower dopant concentrations and thicker gate oxides, hot-carrier generation remains a concern. A variety of specialized transistor structures have been developed to minimize hot-carrier generation at high voltages. This chapter examines several structures that are compatible with ordinary CMOS processing, as well as several that require process extensions.

MOS power transistors are particularly useful for high-current, low-resistance switching. Conventional CMOS transistors can conduct large currents, but have operating voltage limitations. Many processes offer *double-diffused MOS*, or DMOS, transistors as a process extension. These devices allow the construction of compact, high-voltage, high-current transistors.

MOS transistors are only one example of a wider category of devices called *field-effect transistors*, or FETs. Some processes also offer *junction field-effect transistors*, or JFETs. This chapter summarizes the properties of the JFET and presents a selection of typical layouts.

Finally, this chapter also discusses the matching of MOS transistors. MOS transistors used in analog circuits frequently require a high degree of matching. The techniques used to match MOS transistors differ quite markedly from those used for bipolar transistors.

12.1 EXTENDED-VOLTAGE TRANSISTORS

Early MOS processes produced relatively long-channel transistors with lightly doped backgates. Avalanche breakdown of the source/drain regions limited these transistors to operating voltages of 10 to 15V. Processing improvements have enabled channel lengths to decrease from about 8μm to less than 0.3μm. If backgate doping and operating voltages remained constant, the pinched-off portion of the channel would represent an ever-increasing percentage of the channel length. The resulting transistors would exhibit increased channel length modulation and premature punchthrough breakdown.

Channel length modulation can be minimized and punchthrough averted either by reducing the operating voltage or by increasing the backgate doping. Since the operating voltages of analog circuits are not easily reduced below 3 to 5V, most processes have opted to increase backgate doping. This minimizes the width of the pinched-off region at the cost of intensifying the lateral electric field across it. Intense electric fields generate hot carriers that in turn produce undesirable long-term parametric drifts (Section 4.3.1). Holes are more difficult to accelerate than electrons, so PMOS transistors are less prone to hot-carrier generation than are NMOS transistors.

The *operating voltage* of a MOS transistor equals the maximum drain-to-source voltage allowed during saturation, while the *blocking voltage* equals the maximum drain-to-source voltage allowed during cutoff. Since the pinched-off region disappears in cutoff, hot-carrier generation ceases and the blocking voltage is limited only by drain-backgate avalanche and gate oxide rupture. Hot-carrier generation may restrict the operating voltage to a lower value than the blocking voltage. Higher voltages do not necessarily cause instant failure, but they produce gradual shifts in both threshold voltage and device transconductance.[1] Analog circuits are particularly sensitive to hot carrier degradation because they frequently contain matched devices operating in saturation. MOS transistors in digital circuits enter saturation only during brief switching transients. Thus, digital circuits operating at low clock speeds can often ignore the voltage limitations imposed by hot-carrier generation.

Hot carriers become an increasingly serious problem as channel lengths diminish. Specialized transistor structures offer extended operating voltage ranges at the cost of additional process complexity and larger spacings. Virtually all submicron CMOS processes use some form of these *extended-voltage transistors.*

12.1.1. LDD and DDD Transistors

All extended-voltage transistors incorporate some form of specialized drain structure that absorbs a portion of the electric field that would otherwise appear across their channel. Figure 12.1A shows a plot of the lateral electric field intensity across the drain end of a saturated MOS transistor. This example assumes constant backgate and drain doping concentrations and abrupt junctions. The field intensity rises linearly across the pinched-off region and reaches a maximum at the drain metallurgical junction. The field then drops linearly across the depletion region inside the drain. The widths of the pinched-off region x_p and the drain depletion region x_d are proportional to the inverse square root of the doping of the respective regions. The voltages sustained across the pinched-off region V_p and across the drain depletion region V_d equal the areas of the respective triangles (Figure 12.1A).

The total drain-to-source voltage V_{DS} equals the sum of the areas of both triangles:

$$V_{DS} = \frac{E_{max}}{2}(x_p + x_d) \qquad [12.1]$$

The maximum electric field intensity E_{max} and the width of the pinched-off region x_p are limited by hot-carrier generation and channel length modulation, respectively. No such hard-and-fast limit exists on the width of the drain depletion region x_d. The

[1] C. Duvvury and S. Aur, "Hot-Carrier Degradation Effects in CMOS Technologies," *TI Technical Journal,* Vol. 8, #1, 1991, pp. 56–66.

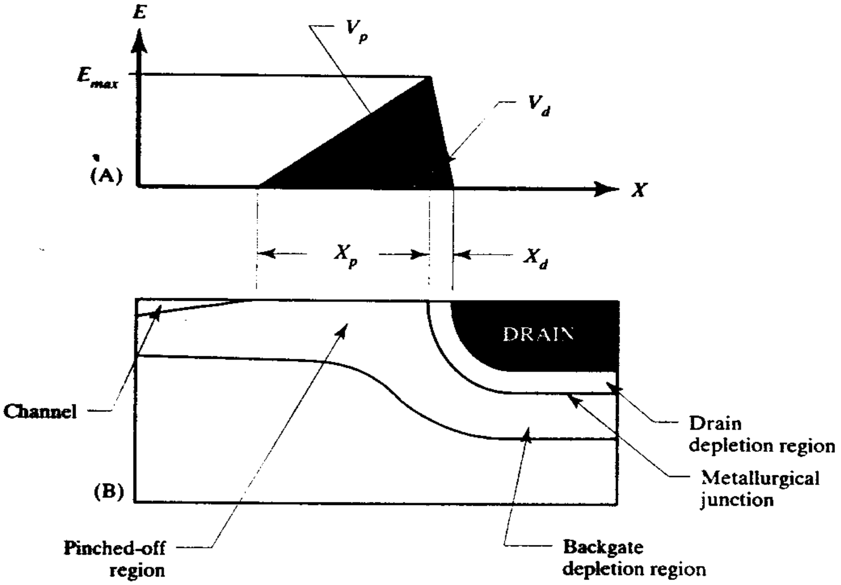

FIGURE 12.1 Diagram showing the lateral electric field across the drain end of a saturated MOS transistor.

only way to increase the operating voltage V_{DS} is to increase the width of the depletion region x_d. This in turn requires a more lightly doped drain.

In order to provide significant benefit, the width of the drain depletion region x_d must equal a significant fraction of the width of the pinched-off region x_p. This dictates that the drain doping not greatly exceed the backgate doping. Unfortunately, a lightly doped drain is also a highly resistive drain. Most modern processes minimize drain resistance by using a composite structure consisting of a lightly doped fringe surrounding a heavily doped core. The fringe depletes at relatively low voltages, forming a *drift region* bounded by the metallurgical junction on one side and by the heavily doped core, or *extrinsic drain,* on the other. The width of the drift region sets the width of the drain depletion region x_d. The drift region should be made just wide enough to support the desired operating voltage, and no wider. Any additional width would increase drain resistance without providing any corresponding benefit. The optimal width of the drift region usually represents no more than a small fraction of the channel length.

Several device structures have been developed that control the drift region width through various forms of self-alignment. The drift region must self-align to both the extrinsic drain and to the poly gate in order to minimize overlap capacitance. Figure 12.2

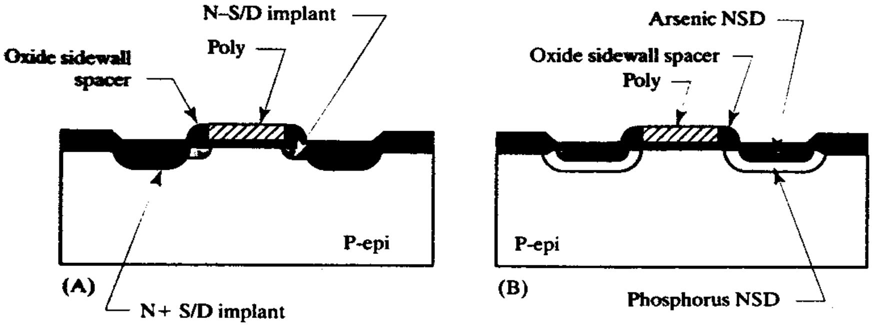

FIGURE 12.2 Extended-voltage transistors: (A) lightly doped drain and (B) double-diffused drain.

shows two structures that fulfill these requirements. Both employ a special feature called an *oxide sidewall spacer* formed by isotropically depositing and anisotropically etching an oxide layer (Section 3.2.4). By its very nature, the oxide sidewall spacer self-aligns to the poly gate. The sidewall spacer is approximately as wide as the poly is thick. A limited range of spacer widths can be fabricated by adjusting poly thickness and etch conditions.

A *lightly doped drain,* or LDD, uses oxide sidewall spacers to define the width of the drift region (Figure 12.2A). The LDD structure requires two separate source/drain implants: one performed before spacer formation and one afterward. The first implant forms a lightly doped drift region aligned to the polysilicon gate, while the second forms a heavily doped extrinsic drain aligned to the oxide sidewall spacer.[2] This technique requires no additional masking steps, but it does require an oxide deposition, an oxide removal, and a second drain implant. The performance advantages gained from higher operating voltages offset the cost of the additional processing.

The process that forms the LDD structure does not discriminate between source and drain. The source resistance can be reduced if the LDD structure appears only on the drain end of the device. The resulting transistor is said to be *asymmetric* because its source and drain terminals are not interchangeable. Asymmetric LDD transistors require additional masking and processing steps to remove the sidewall spacer from the source end of the transistor. The benefits produced by this additional processing are rarely sufficient to justify the additional cost.

The *double-diffused drain,* or DDD, uses two implants driven through the same oxide opening to form a composite drain structure (Figure 12.2B).[3] These two implants require dopants of widely differing diffusivities, most commonly arsenic and phosphorus. A brief drive causes the phosphorus to diffuse outside the boundaries of the arsenic implant to form a lightly doped drift region. An oxide sidewall spacer minimizes overlap capacitances by preventing the drift region from diffusing underneath the poly gate.

The double-diffused drain is not readily applicable to PMOS transistors because of the absence of a slow-diffusing acceptor for silicon. NMOS transistors sometimes use the DDD structure in preference to the LDD structure because it can fabricate very narrow drift regions with great precision. The DDD structure also increases the avalanche voltage of the source/drain implants by grading their junctions. It is difficult to construct wide DDD drift regions because the drive required to force the phosphorus under a wide sidewall spacer also interferes with threshold voltage control. The LDD structure is therefore favored for wider drift regions and the DDD structure for narrower ones. The cost and complexity of the two techniques are comparable, as both require oxide sidewall spacers and both employ two drain implants.

The electric field intensity required to generate hot holes is two or three times larger than that required to generate hot electrons, so many applications that require LDD or DDD NMOS transistors can still use ordinary PMOS transistors. These PMOS devices only require a single source/drain implant, so they are called *single-diffused drain* (SDD) transistors. If the process includes an LDD or DDD NMOS, then the SDD PMOS also receives oxide sidewall spacers. The spacers can actually transform a buried-channel PMOS transistor into an LDD device. The contact po-

[2] S. Ogura, P. J. Tsang, W. W. Walker, D. L. Critchlow, and J. F. Shepard, "Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistor," *IEEE Trans. on Electron Devices,* Vol. ED-27, #8, 1980, pp. 1359–1367.

[3] E. Takeda, H. Kume, T. Toyabe, and S. Asai, "Submicrometer MOSFET Structure for Minimizing Hot-carrier Generation," *IEEE Trans. on Electron Devices,* Vol. ED-29, #4, 1982, pp. 611–618.

tential of the doped poly gate inverts the portions of the buried channel underneath it. The portions of the buried channel protruding under the oxide sidewall spacers do not invert because they do not experience the full electric field generated by the gate electrode. The stubs of the buried channel therefore form two lightly doped P-type regions that act as lightly doped drains (Figure 12.3). This type of structure is sometimes called a *buried-channel lightly doped drain* (BCLDD).[4]

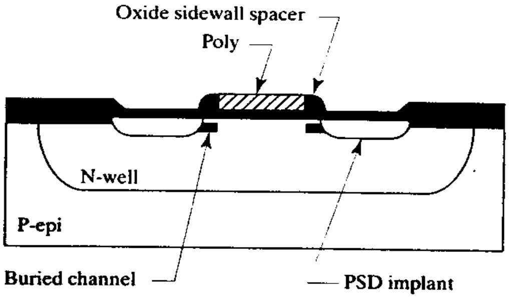

FIGURE 12.3 PMOS buried-channel lightly doped drain (BCLDD).

12.1.2 Extended-drain Transistors

Higher-voltage LDD and DDD transistors require thicker oxide sidewall spacers. The difficulties inherent in constructing wide spacers limit these structures to voltages of about 15 to 20V. Much higher voltages can be achieved using non-self-aligned composite drains that do not employ sidewall spacers to delimit the drift regions. The simplest structure of this sort is called an *extended drain.* It consists of a shallow, heavily doped diffusion entirely contained within a deeper, more lightly doped one. The inner diffusion forms the extrinsic drain, while the fringes of the outer diffusion act as a drift region. For example, an NMOS extended drain might consist of NSD within N-well; the NSD implant then acts as the extrinsic drain, and the N-well as the drift region.

One can often create extended-drain transistors from existing diffusions. The resulting devices are usually larger than purpose-built devices such as lateral DMOS transistors (Section 12.2.2). On the other hand, the extended-drain devices do not require additional processing steps or masks. If an integrated circuit only requires a few small high-voltage transistors, then the most economical solution probably consists of extended-drain transistors constructed from existing masks. On the other hand, large, low-resistance devices are better constructed using purpose-built devices that can achieve lower specific on resistances and overlap capacitances.

Extended-drain NMOS Transistors

Figure 12.4A shows the cross section of a typical extended-drain NMOS transistor constructed in an N-well CMOS process. The extended drain consists of an NSD plug contained inside a larger N-well geometry. The N-well diffuses outward to produce a very lightly doped drain capable of withstanding high voltages. These voltages will rupture the thin gate oxide unless the drain incorporates a special *field-relief structure.* This structure consists of a section of thick-field oxide placed just inside the metallurgical drain-backgate junction. As the depletion region intrudes further into the

[4] R. H. Eklund, R. A. Haken, R. H. Havemann, and L. N. Hutter, "BiCMOS Process Technology," in *BiCMOS Technology and Applications,* 2nd ed., A. R. Alvarez, ed. (Boston: Kluwer Academic Publishers, 1993, pp. 93–95.

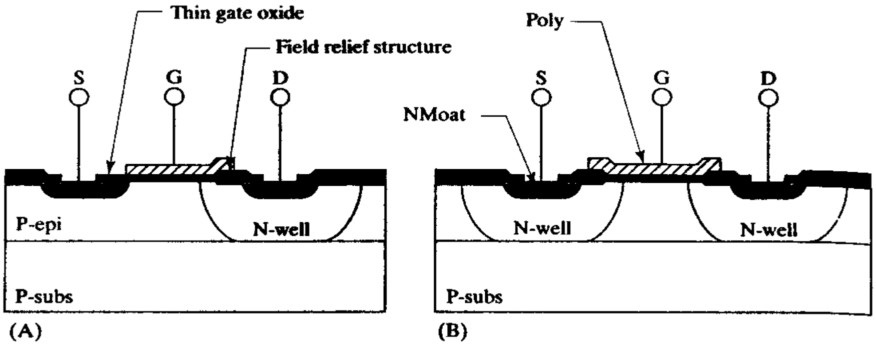

FIGURE 12.4 Cross sections of (A) asymmetric and (B) symmetric extended-drain NMOS transistors.

drain, it passes underneath the bird's beak. The highest drain-to-source voltages appear across the thick-field oxide over the drift region. As long as the field-relief structure is properly positioned relative to the metallurgical junction, the thin gate oxide experiences the full drain-to-source voltage. The field-relief structure has no effect on the transconductance or the threshold voltage of the transistor because these depend solely on the channel, all of which remains underneath the thin gate oxide.

Figure 12.4A illustrates an *asymmetric extended-drain NMOS* transistor. Only one end of this transistor receives an extended-drain structure. This produces a relatively compact layout, but one that is not suitable for applications where either end of the transistor may see high voltages. The *symmetric extended-drain NMOS* in Figure 12.4B equips both ends of the transistor with extended drains. The symmetric transistor can withstand large drain-to-source voltages regardless of which end of the transistor acts as its drain. It cannot simultaneously withstand large voltage differentials between the source and both source/drain regions, because one of these must serve as its source. Transistors that must withstand large gate-to-source voltages require thicker-gate oxides (Section 12.1.3).

Figure 12.5 shows the layout of asymmetric and symmetric extended-drain NMOS transistors. In both cases, the drawn gate length L_d, equals the distance across the moat beneath the gate. The minimum drawn gate lengths of these transistors are relatively large (typically 4 to 6μm), but the effective gate lengths are much smaller because of outdiffusion of the wells. Asymmetric transistors with multiple gate fingers lend themselves to a compact layout in which source and drain fingers alternate; this allows efficient use of the N-well strips, which form the extended drains.

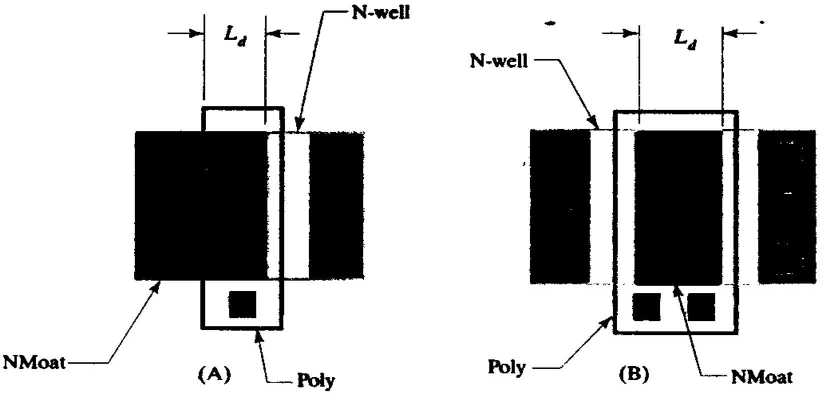

FIGURE 12.5 Layouts of (A) asymmetric and (B) symmetric extended-drain NMOS transistors.

The extremely light doping of the N-well suppresses hot electron generation, so the operating voltage ratings of properly designed extended-drain NMOS transistors are limited only by the avalanche voltage of the well-substrate junction and by the effectiveness of the field-relief structures. It is not particularly difficult to design extended-drain transistors that can withstand operating voltages two or three times larger than those of regular NMOS and PMOS transistors. Such voltages generally exceed the thick-field threshold of the process, requiring the addition of field plates or channel stops (Section 4.3.2).

Extended-drain PMOS Transistors

Figure 12.6A shows the cross section of an *asymmetric extended-drain PMOS* transistor constructed in an N-well CMOS process. The drift region of the extended drain consists of channel-stop implant. The CMOS process flow presented in Section 3.2 uses a patterned phosphorus channel-stop implant to counterdope a blanket boron channel-stop implant. If the channel-stop mask is modified to block the phosphorus implant from the vicinity of the extended drain, then this region receives only the boron implant. The boron outdiffuses during the field oxidation to form a deep, lightly doped P-type diffusion suitable for use as a drift region. Figure 12.6B shows a *symmetric extended-drain PMOS* that uses the channel-stop implant to form drift regions for both source/drain terminations.

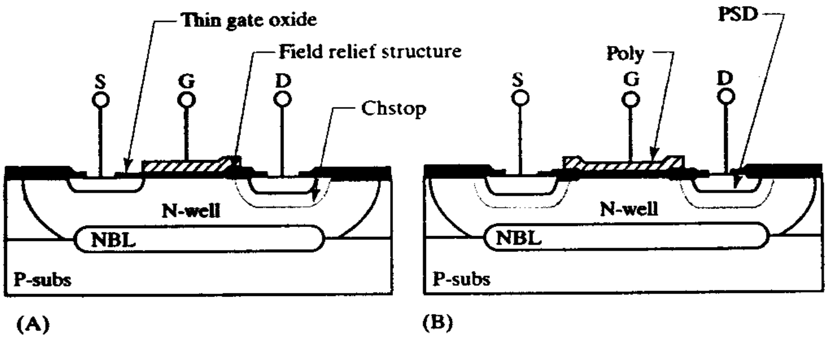

FIGURE 12.6 Cross sections of (A) asymmetric and (B) symmetric extended-drain PMOS transistors.

Extended-drain PMOS transistors use a special coding layer called *Chstop* to block the patterned phosphorus channel-stop implant. The geometries on the Chstop layer are added to the channel-stop mask during mask generation. Different processes may employ other coding techniques, but the principles remain broadly the same.

Figure 12.7 shows the layout of asymmetric and symmetric extended-drain PMOS transistors. In both cases, the drawn gate length equals the distance across the moat region beneath the gate. The transistors must contain NBL to stop vertical depletion from punching through the lightly doped bottom of the N-well and shorting the drain to the substrate.

12.1.3. Multiple Gate Oxides

The previous two sections have described transistors that can operate at large drain-to-source voltages. These transistors can also operate at large drain-to-gate voltages as long as they incorporate suitable field-relief structures. If they must operate simultaneously at large gate-to-source voltages, then no remedy exists but to increase the thickness of the gate oxide. Any increase in gate oxide thickness also reduces device transconductance. Circuit designers take an understandably dim

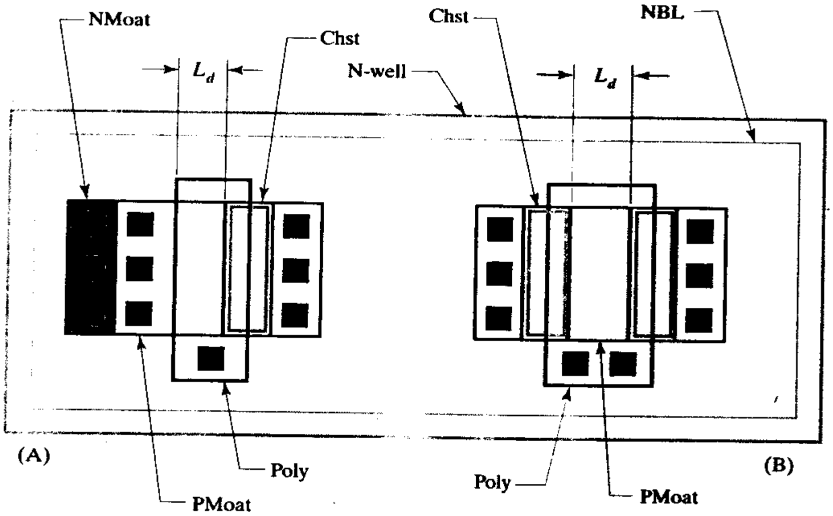

FIGURE 12.7 Layouts of (A) asymmetric and (B) symmetric extended-drain PMOS transistors.

view of increasing the gate oxide thickness of all transistors merely to accommodate increased voltages on a few. This conflict can be resolved by producing two separate thicknesses of gate oxide. The thinner gate oxide provides high transconductance for low-voltage applications, while the thicker gate oxide can withstand higher voltages. The circuit designer chooses which gate oxide each transistor receives based on expected operating conditions.

Multiple gate oxides are fabricated using either staged oxidation or etch-and-regrowth techniques. *Staged oxidation* requires a separate polysilicon deposition for each gate oxide. The thinnest gate oxide is grown first, followed by the deposition of the first polysilicon layer (Figure 12.8A). Once patterned, the poly acts as an oxidation mask for a continuation of the gate oxidation (Figûre 12.8B). After the gate oxidation is complete, the second poly layer is deposited and patterned (Figure 12.8C). Any transistor with a poly-1 gate receives the thin gate oxide, and any transistor with a poly-2 gate receives the thick gate oxide.

Processes with only one layer of poly can use the *etch-and-regrowth* technique instead of staged oxidation. The etch-and-regrowth process requires an additional masking step instead of an additional poly deposition. The extra mask patterns a layer of photoresist spun on top of a thin-gate oxide. The exposed oxide regions are etched away (Figure 12.8D), after which the gate oxidation is resumed. A thin-gate oxide forms over the areas that were etched back, while a thicker layer of oxide forms over the areas where the initial oxide was left undisturbed (Figure 12.8E). A single layer of polysilicon can now form the gates of both thin-oxide and thick-oxide transistors (Figure 12.8F).

Some processes use staged oxidation, while others use etch-and-regrowth. Insofar as the layout designer is concerned, the main difference between the two lies in the number of polysilicon layers required. Figure 12.9 shows a comparison of layouts required by staged oxidation and etch-and-regrowth. Both processes use a *Moat-2* geometry coded around the gate region, but this geometry performs different functions for each process. In the staged oxidation process, Moat-2 defines the region receiving the threshold adjusts for the thick-oxide devices. In the etch-and-regrowth process, the Moat-2 geometry defines both the regions protected from etchback and the regions receiving the thick-oxide threshold adjusts.

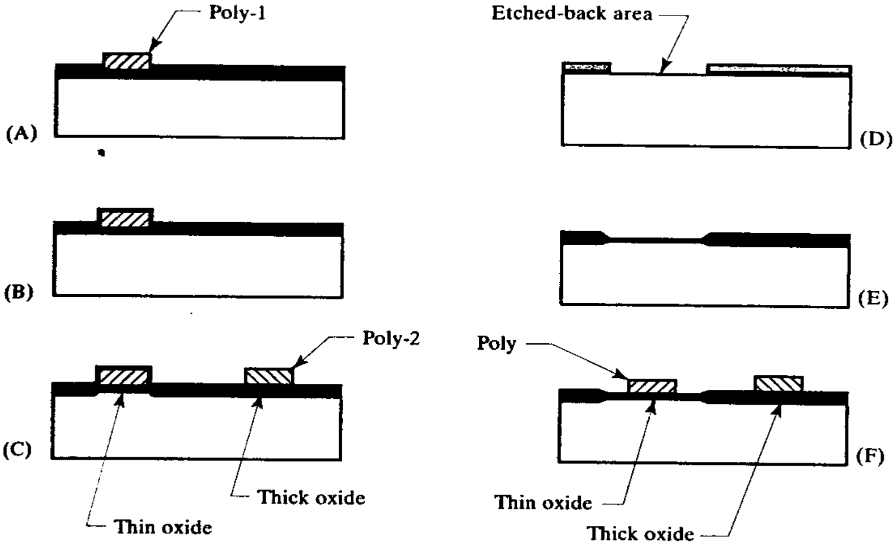

FIGURE 12.8 Process steps for growing multiple thicknesses of oxide using staged oxidation (A-B-C) and etch-and-regrowth (D-E-F) techniques.

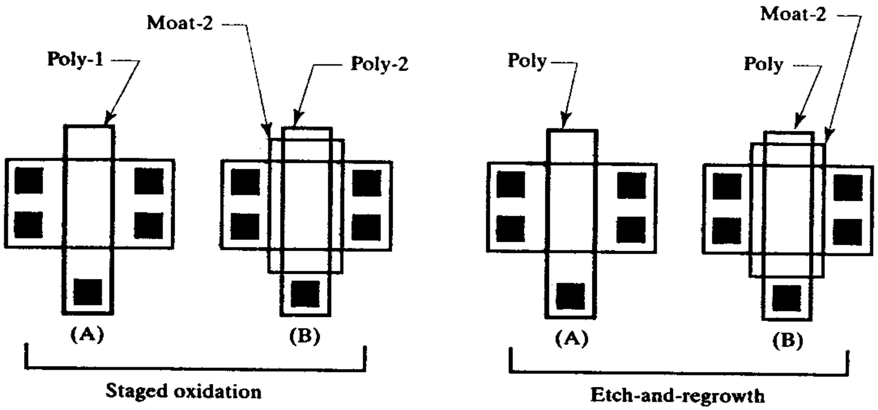

FIGURE 12.9 Comparison of layouts of (A) thin-oxide and (B) thick-oxide transistors constructed for a staged-oxidation process and an etch-and-regrowth process.

12.2 POWER MOS TRANSISTORS

MOS transistors can switch or regulate large amounts of power. Devices specifically designed for such applications are called *power transistors* to distinguish them from low-power, or *small-signal,* devices. MOS power transistors have several advantages over their bipolar counterparts. The forward-bias safe operating area of an MOS transistor is not constrained by thermal runaway or secondary breakdown, as is the safe operating area of a bipolar transistor. MOS transistors also make superior switches because they are not subject to the saturation effects that plague bipolar transistors. MOS transistors also have several significant limitations, most notably low transconductance and poor transient power handling capability.

Thermal Runaway

The collector current of a bipolar transistor increases with temperature, making these devices vulnerable to a positive feedback mechanism called *thermal runaway* (Section 8.1.3). MOS transistors are almost immune to thermal runaway because their drain currents actually decrease with temperature. The positive temperature coefficient of the bipolar collector current stems from the temperature coefficient of V_{BE} (–2mV/°C). Threshold voltages also decrease with temperature, but the low transconductance of the MOS transistor prevents this temperature coefficient from having much effect on drain current. The negative temperature coefficient of the device transconductance causes the drain current to decrease with temperature, as stated above.

Secondary Breakdown

Secondary breakdown (Section 8.1.3) clips off a portion of the forward-biased safe operating area (FBSOA) curve of a bipolar transistor (Figure 12.10A). Since MOS transistors do not experience secondary breakdown, the MOS FBSOA curve is limited only by breakdown voltage, power dissipation, and current-handling capability (Figure 12.10B). The safe operating area of the MOS transistor not only exceeds that of an equivalent bipolar transistor, but it is also more predictable because it is limited only by packaging, metallization, and breakdown voltages. Bipolar transistors may have more robust characteristics than equivalent MOS transistors under reverse breakdown or when handling transient power pulses.

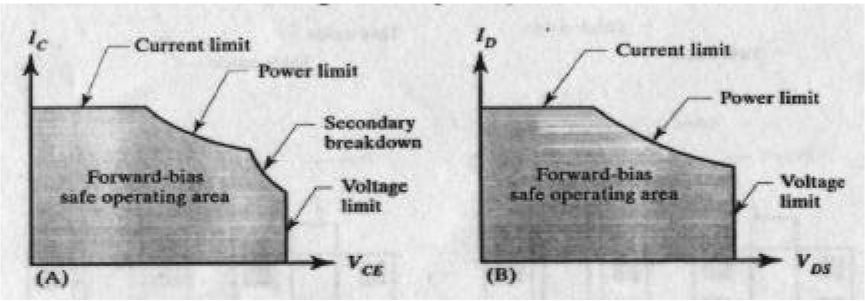

FIGURE 12.10 Forward-biased safe operating area (FBSOA) curves for (A) a power bipolar transistor and (B) a power MOS transistor.

Rapid Transient Overload

Although MOS transistors generally do not experience thermal runaway or secondary breakdown, they do sometimes experience current focusing problems caused by debiasing in long, narrow polysilicon gate fingers. The resistance of these fingers can become quite large, especially if the poly is not silicided. This resistance forms a distributed RC network with the gate capacitance. When the gate voltage slews rapidly, the ends of the transistor nearest the gate connection turn on and off before the rest of the device. This progressive turn-on characteristic sometimes leads to localized overheating and device failure.

Figure 12.11 shows a simplified model of a single gate finger of an MOS power device. M_1 represents the portion of the gate finger nearest the gate connection, while M_2 and M_3 represent more distant portions. Resistors R_1 and R_2 model the resistance of the polysilicon gate. If the gate drive voltage V_G rises rapidly, then transistor M_1 turns on before transistors M_2 and M_3 due to the RC time delays caused by gate resistance and capacitance. The portion of the finger nearest its termination (M_1) begins to discharge load capacitance C_L before the other portions of the transistor can take up their share of the load. If the voltage across C_L is large, then the current density through M_1 may become large enough to damage the device. This type of failure only occurs if the rise time of the gate drive voltage V_G is smaller than the gate time delay, which is typically about a few nanoseconds. Rapid transient overloads of this sort often occur in ESD protection circuits and MOS gate drivers.

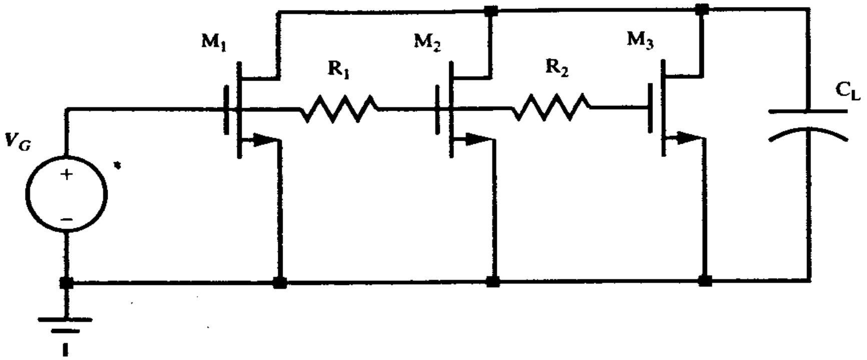

FIGURE 12.11 Model of a long MOS gate finger driving a capacitive load.

Bipolar transistors can usually withstand rapid transient overloading because they dissipate energy into the relatively large volume of their collector-base depletion regions. MOS transistors are more vulnerable because they dissipate energy within the small volume of their pinched-off channel. The bipolar transistor also has the advantage of dissipating energy deep within the silicon, far away from contacts that are prone to damage from overheating. Bipolar transistors are therefore favored for pulsed-power applications such as gate drivers and ESD protection circuitry.

MOS Switches versus Bipolar Switches

MOS transistors can conduct large currents at remarkably low drain-to-source voltages. The behavior of the transistor under these conditions can be derived from the Shichman-Hodges equation for the linear region. The equation is first rearranged in the form

$$I_D = k\,(V_{GS} - V_t)\,V_{DS} + \frac{kV_{DS}^2}{2} \qquad [12.2]$$

The quadratic term becomes negligible at low drain-to-source voltages ($V_{DS} << V_{GS} - V_t$). The above equation then simplifies to

$$I_D \cong k\,(V_{GS} - V_t)\,V_{DS} \qquad [12.3]$$

This equation reveals a linear relationship between the drain-to-source voltage V_{DS} and the drain current I_D. The transistor therefore behaves as if it were a resistor whose value $R_{DS(on)}$ equals

$$R_{DS(on)} \cong \frac{1}{k(V_{GS} - V_t)} \qquad [12.4]$$

The *on resistance* $R_{DS(on)}$ varies inversely with device transconductance and inversely with effective gate voltage V_{gst}. In theory, the on resistance can be reduced to arbitrarily small values by increasing *W/L*. In practice, the resistance of the metallization system and the bondwires places a lower limit on the achievable on resistance. Typical values of $R_{DS(on)}$ for large integrated power transistors are 50 to 500mΩ.

MOS transistors can achieve much lower forward voltage drops than bipolar transistors. The difference between the built-in potentials of the collector-base and the emitter-base junctions sets a lower limit on the achievable saturation voltage of a bipolar transistor. This quantity, called the *intrinsic saturation voltage,* usually equals at least 50mV for vertical bipolar transistors. Lateral bipolar transistors have zero intrinsic saturation voltages, but no junction-isolated transistor can approach zero saturation voltage without experiencing substrate injection (Section 8.1.4). Depending on junction temperature and current density, practical saturation voltages range from 0.1V to as much as 1V. This performance is much poorer than that of a 250mΩ MOS transistor with a forward voltage drop of only 25mV at 100mA. Clearly, MOS transistors make much better switches than do bipolar transistors.

MOS power switches also require much less sophisticated predrive circuitry than bipolar power switches. A bipolar switch must have adequate base drive to ensure that it remains in saturation at the maximum load current. The excess base drive is wasted when the transistor operates at lower currents. The most efficient types of base drive circuits monitor the collector current and provide just enough base drive to ensure that the transistor always remains saturated. These *proportional base drive* circuits are generally rather complicated. Most base drive circuits incorporate additional components to extract stored base charge from the saturated transistor, but even with these additional components the switching speed of power bipolar transistors remains less than 500kHz.[5] MOS power transistors require neither proportional drive nor special turnoff circuitry, and they can easily achieve switching speeds of several megahertz. MOS transistors are therefore overwhelmingly superior to bipolar transistors for high-power, high-speed switching applications such as switched-mode power supplies.

12.2.1. Conventional MOS Power Transistors

MOS transistors generally do not require ballasting because they do not experience thermal runaway or secondary breakdown. The same finger layouts used to construct small-signal transistors therefore serve equally well for power applications.

MOS power transistors are usually specified in terms of their on-resistance $R_{DS(on)}$ measured at a specified gate voltage V_{GS} and junction temperature. The $R_{DS(on)}$ of a power MOS transistor typically increases by 50% when the junction temperature rises from 25°C to 125°C, and it varies about ±30% over process. The metallization resistance becomes significant for on resistances of less than an Ohm, and the equation for $R_{DS(on)}$ then becomes

$$R_{DS(on)} \cong \frac{1}{k(V_{GS} - V_t)} + R_M \qquad [12.5]$$

where R_M is the sum of the resistance of the source and drain metallization. This metallization resistance is difficult to compute because it depends on transistor geometry. Many designers avoid the need for determining R_M by relying on measured $R_{DS(on)}$ data. This method requires that one measure the $R_{DS(on)}$ of a sample device whose layout resembles that of the proposed power device. The measured $R_{DS(on)}$ is then used to compute a figure of merit called the *specific on-resistance* R_{SP}

$$R_{SP} = A_d R_{DS(on)} \qquad [12.6]$$

where A_d represents the drawn area of the sample layout. The specific on-resistance is usually given units of $\Omega \cdot mm^2$. Smaller values of R_{SP} indicate increasingly area-efficient layouts.

Once the specific on-resistance has been determined, one can use equation 12.6 to compute the area required to obtain any desired on resistance. The biggest prob-

[5] This switching speed applies only to large power devices operated in deep saturation. Small-signal devices can operate at much faster speeds, especially if they do not enter saturation.

lem with this technique lies in obtaining an accurate estimate of the specific on-resistance. This is much more difficult than it might seem. The measured value of R_{SP} should not include the resistances of bondwires and leadframe because these do not scale with device area. This is best done by providing Kelvin connections (Section 14.3.2) to the sample device. Alternatively, one can measure the resistance of the leads and bondwires of a dummy unit that contains no die. $R_{DS(on)}$ computations based on R_{SP} do not include the bondwire and leadframe resistance, so these must be added to obtain the total $R_{DS(on)}$.

The specific on-resistance also varies with device area and aspect ratio. Any one value of R_{SP} only applies to a limited range of device sizes and aspect ratios. In practice, one cannot rely on an empirical value of R_{SP} to scale the $R_{DS(on)}$ of a device by more than a factor of two or three. If R_{SP} values are available for a range of device sizes, then one can interpolate between these measurements to find the R_{SP} value for a device of intermediate size. A similar process can be used to account for variations in aspect ratio.

Alternatively, one can attempt to compute the metallization resistance based on an analysis of the geometry of the transistor. Hand calculations yield only approximate results because of the large number of assumptions and simplifications required to render the problem tractable. A computerized finite-element analysis provides more accurate results because it takes into account a larger number of geometric factors. In either case, the computations require a detailed knowledge of the metallization pattern. The following sections describe two of the most popular metallization patterns for MOS power transistors. Many other patterns have been proposed for specific applications, but most of these are not sufficiently general to merit further discussion.

The Rectangular Device

Figure 12.12 shows a diagram of a simple double-level-metal pattern that produces a compact rectangular device. The top portion of the diagram shows only the interdigitated metal-2 patterns for the source and the drain fingers. The lower portion of the diagram shows these metal-2 patterns superimposed over the metal-1 fingers. Each of these fingers consists of a narrow strip of metal-1 spanning the entire width of the transistor and containing one row of contacts and vias. The metal-2 buses running up the left and right sides of the transistor collect the currents from all of the fingers and feed the source and drain terminations, which may either lie at the top or the bottom of the transistor.

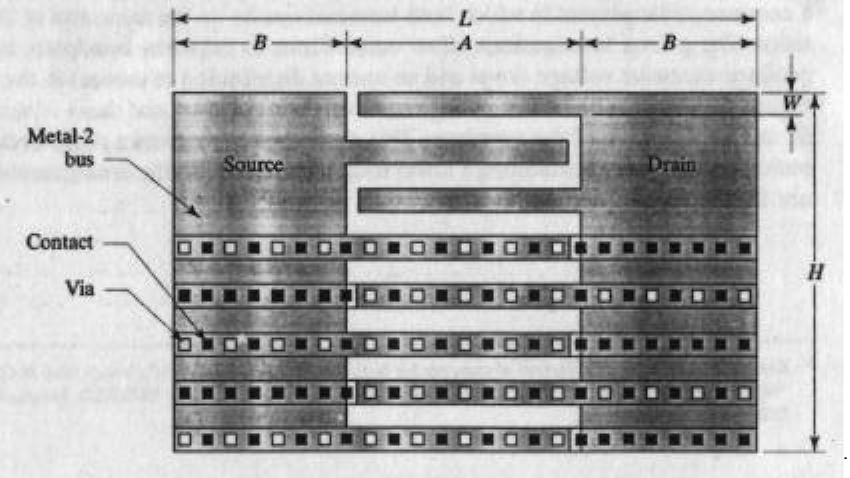

FIGURE 12.12 Metallization pattern for a rectangular MOS power device.

The metallization pattern of Figure 12.12 provides the maximum possible amount of metallization for both the source and the drain. To understand how this has been achieved, consider a single finger in isolation from the rest of the device—for example, the bottom finger. Current flows along this finger from right to left, finally exiting into the metal-2 bus connecting to the source. No vias exist on the rightmost third of the finger. The current flowing through this portion of the finger must pass entirely through metal-1. Each contact feeds a small amount of current into the metal, so the magnitude of the current increases as it flows leftward. Metal debiasing becomes an increasingly serious concern as the magnitude of the current increases. Once the current reaches the middle third of the finger, a portion of it flows upward through vias to reach a strip of metal-2. The current now flows through a sandwich of metal-1 and metal-2. The magnitude of the current continues to increase as it flows leftward. Once the current reaches the leftmost third of the finger, it flows up into the metal-2 bus and out to the termination of the transistor.

The magnitude of the voltage drop along a source finger, V_{SM}, increases from right to left, while the magnitude of the voltage drop along a drain finger, V_{DM}, increases from left to right (Figure 12.13). The sum of these voltage drops $V_{DM} + V_{SM}$, varies less than either of the terms that comprise it. This not only ensures approximately equal conduction through all parts of the transistor, but it also reduces the overall $R_{DS(on)}$. This principle can be applied to the metallization patterns for almost any type of power device, but it is particularly applicable to MOS power transistors where metallization resistance plays such an important role.[6]

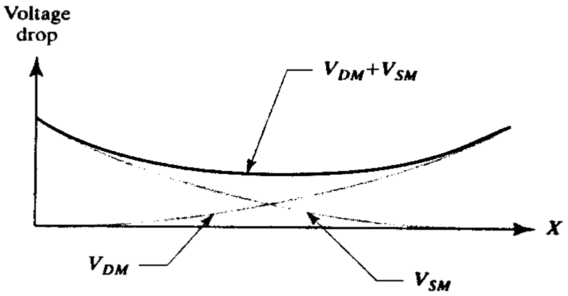

FIGURE 12.13 Graph of voltage drops across a lateral section of the power transistor of Figure 12.12.

The metal-2 buses along the left and right sides of the transistor collect the currents flowing from the individual source and drain fingers. These currents flow vertically to the terminations of the device. A variety of different termination arrangements are possible, some of which perform better than others. Figure 12.14A shows a common arrangement in which both terminations lie on the same end of the transistor. The paired terminations allow connections to adjacent bondpads, but they produce excessive voltage drops and an uneven distribution of current in the device. Figure 12.14B shows a better arrangement where the source and drain terminations lie at opposite ends of the transistor. This arrangement delivers a more even distribution of current and exhibiting a lower total resistance than the arrangement in Figure 12.14A.

[6] Krieger analyzes the distribution of currents for both parallel and antiparallel current flow in G. Krieger, "Nonuniform ESD Current Distribution Due to Improper Metal Routing," *EOS/ESD Symposium Proc.*, EOS-13, 1991, pp. 104–108.

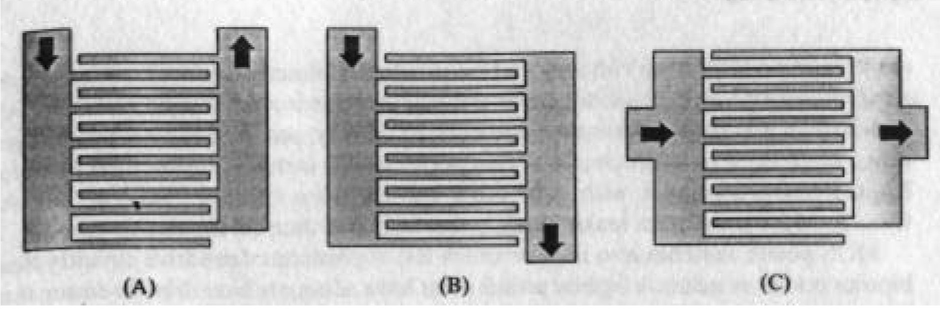

FIGURE 12.14 Three different metallization patterns for rectangular power transistors. The arrows indicate the flow of current to the termination points.

Figure 12.14C shows another method sometimes used to minimize metallization resistance. The termination points lie midway along either bus, so the current does not have to flow through the full length of the buses. This arrangement does minimize resistance, but it also produces an uneven current distribution. For most purposes, the layout in Figure 12.14B is superior to those in both Figure 12.14A and 12.14C.

The Diagonal Device

The rectangular layout in Figure 12.14 has one glaring flaw. The width of the metal-2 buses remains constant, yet the current flowing through them varies. Tapered buses can minimize debiasing and can provide a more uniform distribution of current among the fingers of the transistor. Figure 12.15 shows a layout employing tapered buses. The fingers of the transistor are arranged in a diagonal pattern that naturally produces trapezoidal metal-2 buses on either side of the device. The source and drain terminations must lie on opposite ends of the transistor. This device is more difficult to construct than the rectangular layout shown in Figure 12.12, and computer simulations are required to determine its optimum dimensions. The triangular areas beneath the metal-2 buses must be filled with circuitry in order to obtain the full packing density promised by this layout. Many designers prefer to use rectangular layouts, such as the one in Figure 12.14B, which are easier to construct and to optimize.

Computation of R_M

Accurate calculations of the metallization resistance become very complex and generally require computer modeling. The metallization pattern in Figure 12.14B

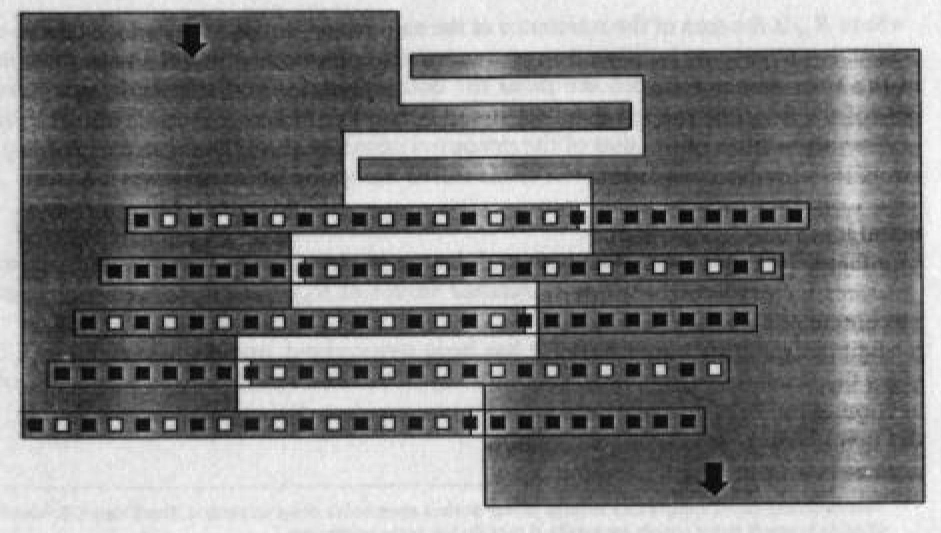

FIGURE 12.15 Metallization pattern for a diagonal MOS power device.

represents an exception, as it is possible to estimate its metallization resistance using the formula

$$R_M = \frac{B^2 R_{S1}}{2W\, N_D L} + \frac{A R_{S12}}{2W\, N_D} + \frac{H R_{S2}}{2B} \qquad [12.7]$$

where N_D equals the number of drain fingers (or half the total number of source/drain fingers), R_{S1} equals the sheet resistance of metal-1, R_{S2} equals the sheet resistance of metal-2, and R_{S12} equals the sheet resistance of a parallel combination of metal-1 and metal-2. Figure 12.12 shows the relationship between dimensions A, B, W, L, and H. This derivation assumes that each source/drain finger conducts an equal amount of current and that the current flowing through a finger increases linearly along its length. The formula also neglects the variation in voltage across the width of the metal-2 buses.

The above equation can be analyzed (Appendix D) to determine the optimum width, B, of the metal-2 buses, which equals

$$B = \left(\frac{R_{S12}}{R_{S1}}\right) L \qquad [12.8]$$

Assuming that both metal-1 and metal-2 have the same composition, equation 12.8 becomes

$$B = \left(\frac{t_1}{t_1 + t_2}\right) L \qquad [12.9]$$

where t_1 and t_2 are the thicknesses of metal-1 and metal-2, respectively. This equation provides a means of sizing the metal-2 buses. If the thickness of metal-1 equals or exceeds the thickness of metal-2, then the buses should each extend across half of the transistor. In this case, the interdigitated region described by dimension A vanishes entirely. If the thickness of metal-2 exceeds the thickness of metal-1, then the buses should not entirely cover the transistor, and an interdigitated region should exist between them. Although equations 12.8 and 12.9 were derived specifically for the structure in Figure 12.14B, they concern only a single finger and so apply also to the structures in Figures 12.14A and 12.14C.

Other Considerations

The metal leads and bondwires that connect a power transistor to its load can substantially increase $R_{DS(on)}$. These resistances depend on the general size and shape of the transistor and its placement relative to its bondpads. The calculations form part of the floorplanning process discussed in Section 14.2.

The connection of the gate lead also merits consideration. The resistance of long stretches of polysilicon substantially slows the switching of large power transistors. This resistance can be minimized by connecting the individual gate fingers with metal jumpers. Connecting both ends of the gate further reduces the gate resistance by a factor of about four (Figure 12.16).

Other factors worth considering when laying out power transistors include the placement of backgate contacts and guard rings. Power transistors may use either interdigitated or distributed backgate contacts, depending on which technique provides the smallest overall area (Section 11.2.7). Devices requiring independent connection to the backgate must use interdigitated backgate contacts spaced away from the neighboring source fingers. PMOS transistors constructed in analog BiCMOS processes may use NBL to obtain a solid backgate connection without the need for interdigitated or distributed backgate contacts.

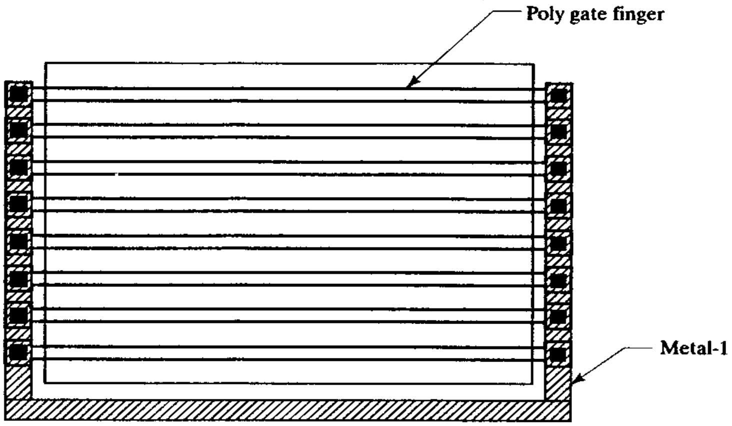

FIGURE 12.16 Gate metallization connecting both ends of the gate fingers reduces gate resistance.

Some applications momentarily forward-bias the backgate diode of a power MOS transistor. If this occurs, then the transistor not only requires an extensive network of backgate contacts to provide recombination current, but it also requires an efficient system of minority carrier guard rings.

Transistors constructed in the epi pose a particular problem because no way exists to block the flow of minority carriers to the substrate. This situation occurs when NMOS transistors are constructed in N-well CMOS processes. In BiCMOS processes, one can sometimes isolate these NMOS transistors using a combination of deep-N+ sinker and NBL (Section 11.2.2). An isolated NMOS cannot inject electrons into the substrate because the isolation structure acts as an electron-collecting guard ring. If an isolated structure is not feasible, then the designer must fall back on guard rings placed around the periphery of the transistor. These guard rings are often quite effective when combined with a heavily doped substrate. If the process uses a lightly doped substrate, then the guard rings should be made as wide as possible to prevent minority carriers from passing underneath them. Substrate contacts should be placed on the far side of the guard ring from the point of injection. Majority carrier current flowing underneath the guard ring through the lightly doped substrate generates an electric field that opposes the flow of minority carriers underneath the guard ring. One can sometimes arrange the wells of adjacent power transistors so they also act as guard rings.

Some circuits use a small transistor to sense the current passing through a much larger one. Ideally, the sense transistor should consist of a number of segments scattered throughout the transistor. so the average of these segments represents the average operating conditions of the power transistor. It is difficult to embed sense transistors within the interior of the power transistor. Instead, these devices usually lie at the ends of gate fingers along one or two sides of the device. If only one sense transistor segment can be used, then this should occupy the center of one side of the device. Two sense segments should occupy the center of opposite sides of the device. Four sense segments should occupy pairs of sites located symmetrically around an axis passing through the centroid of the power transistor. If the sense segments can lie within the confines of the power transistor. they should occupy locations approximately half way from the center of the transistor to its periphery, and should be located in symmetric locations as discussed above.

Figure 12.17 shows a typical example of a single embedded sense transistor located on the end of a gate finger. In practice. the power transistor would have a much larger number of fingers, and the one chosen for the sense device would lie as near to an axis of symmetry of the device as possible. The sense device shares common gate, source, and backgate connections, but it has an independent drain connection.

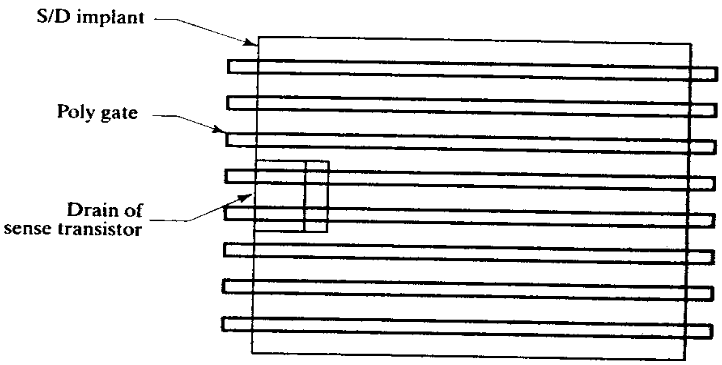

FIGURE 12.17 Construction of an embedded sense transistor (contacts and metallization not shown).

Nonconventional Structures

The conventional self-aligned poly-gate transistor consists of a series of interdigitated source and drain fingers. Although this arrangement possesses the virtue of simplicity, it does not produce the densest possible layout. Other designs can achieve lower specific on-resistances by tightly packing arrays of cleverly shaped source and drain elements. The *waffle transistor* in Figure 12.18A exemplifies this concept. It uses a mesh of horizontal and vertical poly strips to divide the source/drain implant into an array of squares. Each square contains a single contact. By alternately connecting these contacts to the source and drain metallization, one can arrange four drains around each source and four sources around each drain. The drain and source metallization consists of a series of diagonal strips of metal-1. which are usually combined with an interdigitated metal-2 pattern similar to those used for conventional transistors.

An analysis of the W/L ratios achieved for a given device area will show that the waffle transistor provides an increase in packing density equal to

$$\frac{(W/L)_w}{(W/L)_c} = \frac{2S_d}{L_d + S_d} \quad [12.10]$$

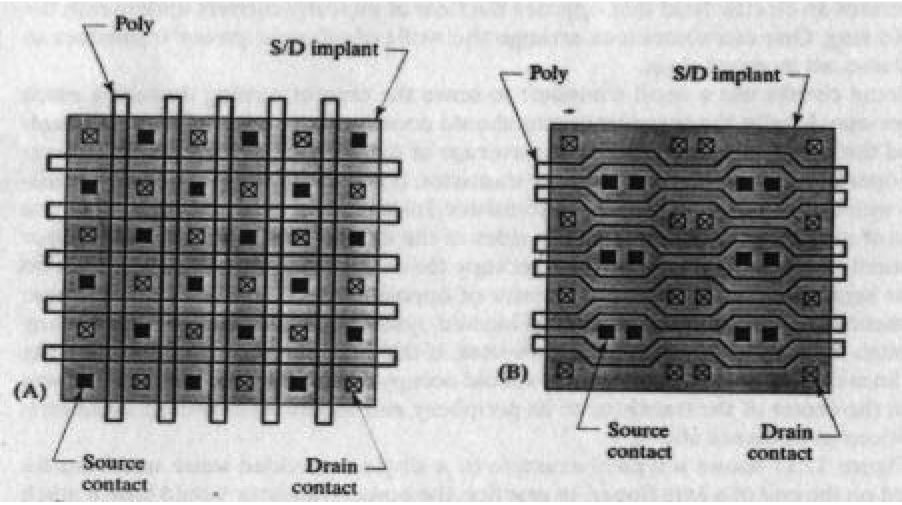

FIGURE 12.18 Nonconventional MOS transistor layouts: (A) waffle and (B) bent-gate. The drain contacts have been drawn differently than the source contacts to aid in their identification.

where $(W/L)_w$ of the waffle transistor and $(W/L)_c$ of the conventional interdigitated transistor are measured from two devices consuming equal die areas. The waffle transistor offers better packing than the conventional interdigitated transistor as long as the spacing between the gates S_d exceeds the gate length L_d. Almost all power transistors meet this requirement. Suppose the layout rules specify a minimum drawn gate length of 2μm, a minimum contact width of 1μm, and a minimum spacing poly-to-contact of 1.5μm. Using these rules, equation 12.10 indicates that the waffle transistor provides approximately 33% more transconductance than the interdigitated-finger transistor. A more precise estimate of the benefits of the waffle transistor would account for the differences between drawn gate length and effective gate length and would include corner conduction terms for the waffle transistor, neither of which are included in Equation 12.10.

The waffle transistor has three crucial defects. First, the above analysis does not consider the effects of metallization resistance. The metallization invariably contributes a significant portion of the $R_{DS(on)}$ of the transistor, and in thin CMOS metal systems it often becomes the dominant factor. If one assumes that the metallization contributes about half the total $R_{DS(on)}$, then the improvement gained by using the waffle layout drops by half, or from 33% to 16% for this example. The situation is actually even worse because the waffle layout is difficult to properly metallize. The metal-1 fingers must repeatedly cross the gate poly, which almost certainly introduces significant step-induced metal thinning. Second, the waffle transistor contains a large number of bends in its channels. These bends produce sharp corners in the source/drain regions that avalanche at lower voltages than the remainder of the transistor. Localized avalanche limits the amount of energy the waffle transistor can dissipate. This limitation becomes apparent in ESD testing, where the performance of waffle transistor may fall short of that of the conventional layout. Fillets or chamfers applied to the corners of the source/drain squares will largely eliminate this problem, and a transistor that includes them may provide even better ESD performance than a conventional layout.[7] Third, the waffle transistor makes no provision for backgate contacts. Unless the transistor is used in combination with a heavily doped substrate or a buried layer to provide backgate contact, it is quite susceptible to backgate debiasing and latchup. No simple way exists to add interdigitated or distributed backgate contacts to a waffle layout.

Figure 12.18B shows a *bent-gate transistor* that avoids most of the difficulties of the waffle transistor, while offering some unique advantages. The bent gates increase the gate width while simultaneously allowing the gate strips to pack more closely together. This layout readily accommodates distributed backgate contacts without sacrificing undue die area. It also avoids the use of 90° bends in favor of gentler 135° bends, which are less prone to localized avalanche. The diagonal arrangement of the source and drain contacts also provides additional source/drain ballasting that can improve robustness under extreme conditions, such as those encountered during ESD testing. These benefits, combined with the ease of inserting extensive networks of distributed backgate contacts, make this device ideal for applications that routinely experience transient overloads.

12.2.2. DMOS Transistors

High-voltage transistors require short, heavily doped backgates and wide, lightly doped drift regions. These are more-conveniently produced by diffusing the backgate into the drift region than *vice versa*. The *double-diffused MOS*, or DMOS, uses

[7] L. Baker, R. Currence, S. Law, M. Le. C. Lee, S. T. Lin. and M. Teene, "A 'Waffle' Layout Technique Strengthens the ESD Hardness of the NMOS Output Transistor," *EOS/ESD Symposium Proc.*, EOS-11. 1989, pp. 175–181.

this approach to produce short-channel, high-voltage transistors optimized for use as power devices.

Like the DDD transistor, the DMOS relies on the self-alignment of two diffusions driven through a common oxide opening. An N-channel DMOS is fabricated by diffusing boron and arsenic into lightly doped N-type silicon (Figure 12.19). Boron outdiffuses more rapidly than arsenic, producing a moderately doped P-type region enclosing a shallower and more-heavily doped N-type region. The heavily doped arsenic core forms the source of the DMOS transistor, while the surrounding, moderately doped boron diffusion forms the backgate. The channel length of the transistor equals the difference between the surface outdiffusion distances of the boron and arsenic implants, which depends solely on doping concentrations and drive times.

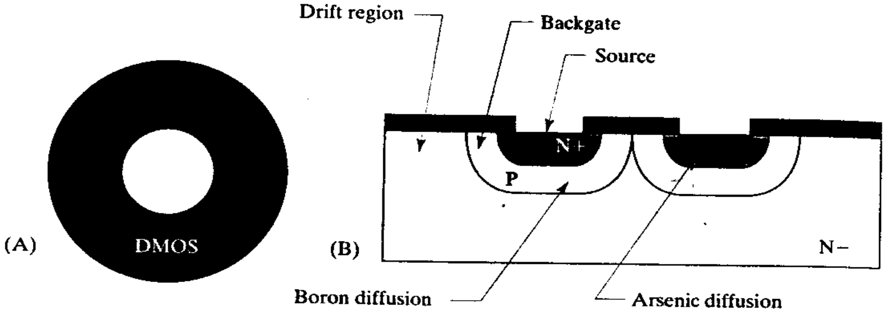

FIGURE 12.19 Layout of (A) DMOS mask geometry and (B) the resulting pattern of diffusions.

The lightly doped N-type region surrounding the backgate serves as the drift region of the DMOS transistor. This drift region must be contacted by means of a heavily doped N-type region. Discrete DMOS transistors often occupy a lightly doped N-type epi deposited on a heavily doped N-type substrate. The drain current flows down through the epi to the substrate and exits through the backside of the die. The epi thickness determines the width of the drift region and hence the maximum operating voltage of the transistor. In order to integrate this *vertical DMOS*, the drain region must be isolated from the substrate. This can be achieved by placing the transistor in an N-well furnished with NBL and deep-N+. This resolves the isolation issue, but the transistor still exhibits excessive drain resistance at low forward voltages because of incomplete depletion of the drift region. Drain resistance represents a major challenge for constructing low-voltage DMOS transistors. The epi thickness cannot be reduced too far or the tail of the NBL will intersect the DMOS diffusions, so another approach must be tried.

The Lateral DMOS Transistor

Most integrated DMOS transistors use a shallow, heavily doped N-type diffusion placed next to the DMOS backgate to extract the drain current. This type of device is called a *lateral* DMOS, or LDMOS.[8] The separation of the backgate and drain contact diffusions determines the width of the lateral drift region. This drift region is designed to fully deplete through at a relatively low voltage. This type of transistor does not require NBL or deep-N+, although these are often added to minimize substrate injection in the event that the backgate forward-biases into the drain.

[8] J. D. Plummer and J. D. Meindl, "A Monolithic 200-V CMOS Analog Switch," *IEEE J. Solid-State Circuits*, Vol. SC-11, #6, 1976, pp. 809–817.

The DMOS backgate contact represents something of a problem. Not only is the backgate relatively lightly doped, but it is also extremely narrow. A heavily doped P-type diffusion can contact the backgate, but only if it also contacts the N+ source. The resulting P+/N+ junction may leak so badly that the backgate cannot be isolated from the source. Most DMOS transistors use an annular geometry containing a central P+ plug that serves as a backgate contact. The P+ plug shorts to the source of the transistor through a single contact opening that covers both (Figure 12.20A). DMOS transistors are considered asymmetric devices because their backgate usually connects to their source and because the diffused backgate is more lightly doped near the drain than near the source.

Figure 12.20 shows a simple LDMOS transistor constructed in an N-well analog BiCMOS process. An annular DMOS implant consisting of arsenic and boron defines the source and backgate regions of the device, and an enclosing N-well serves as its drift region. The P+ plug contacting the backgate consists of a PSD implant whose outer edge coincides with the inner edge of the DMOS implant. The extrinsic drain consists of a ring of NSD implanted around the transistor. The spacing between the NMoat and the DMOS geometries determines the width of the drift region. The poly gate overlaps the thick-field oxide over the drift region to form a field-relief

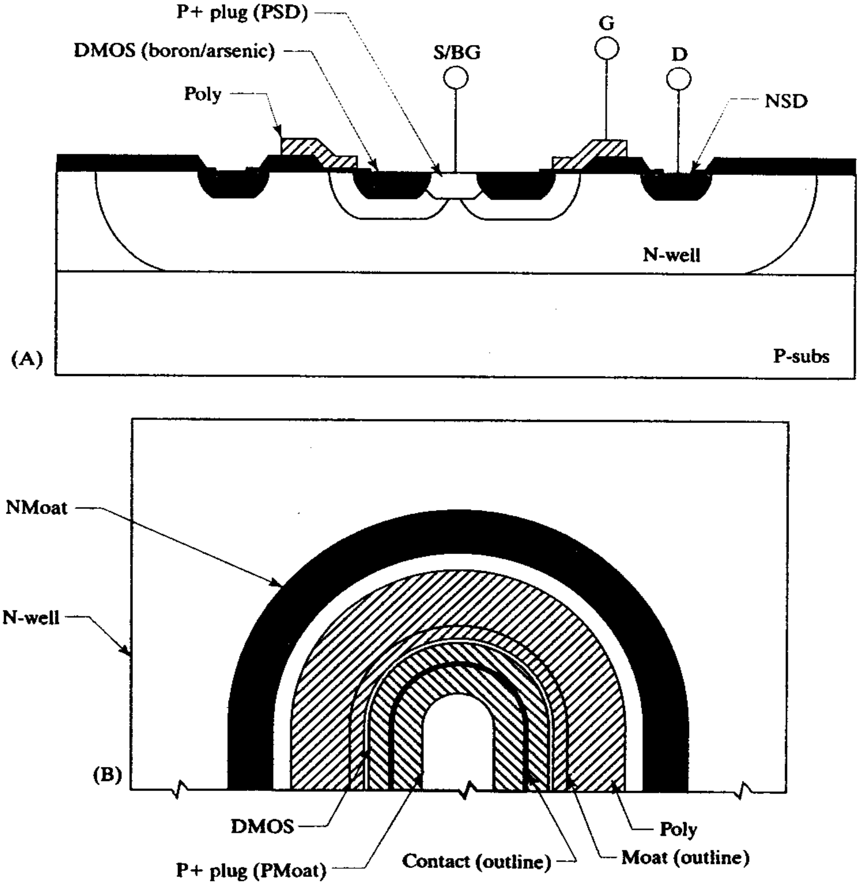

FIGURE 12.20 Layout and cross section of a lateral DMOS transistor constructed in analog BiCMOS.

structure, allowing the transistor to withstand large drain-to-gate voltage differentials without requiring a thick gate oxide. Many elaborations upon this structure exist.[9]

The smallest DMOS transistor uses a circular ring of DMOS implant, like that shown in Figure 12.19. Larger widths can be obtained by connecting many small annular devices in parallel, but such arrays of annular transistors pack rather loosely. The transistor in Figure 12.20 uses an elongated annular geometry to improve packing density. A large power device consists of an array of annular DMOS transistors interdigitated with drain contacts. This type of structure resembles a conventional interdigitated transistor and can use any of the metallization patterns discussed in Section 12.2.1.

NBL often forms part of a DMOS transistor, but its presence does not always prove beneficial. Consider the case of a DMOS transistor without NBL. The drain-backgate and drain-substrate depletion regions both widen as the drain voltage increases. If the well is sufficiently shallow, then it will punch through from backgate to substrate long before the well-backgate junction avalanches. This punchthrough only causes problems if the source of the transistor connects to a voltage other than substrate potential. NBL can prevent punchthrough by constraining the depletion region, but in so doing it intensifies the electric field and reduces the well-backgate avalanche voltage. The DMOS transistor therefore has two separate drain-to-source voltage ratings. A device whose source connects to substrate potential can omit NBL and can obtain a higher operating voltage. Devices whose sources connect to voltages other than substrate potential require NBL and are therefore constrained to lower operating voltages. Various improved versions of the lateral DMOS structure have been proposed.[10]

The DMOS NPN

The DMOS structure in Figure 12.20 contains a parasitic NPN transistor. The source of the DMOS acts as the emitter of this transistor, the backgate as its base, and the drain as its collector. This parasitic NPN has a heavily doped emitter that enhances its emitter injection efficiency, a thin, moderately doped base that reduces its Gummel number, and a wide, lightly doped collector that minimizes the Early effect. The performance of the DMOS NPN can approach that of a conventional CDI NPN, making it a useful alternative to the latter device.

Figure 12.21 shows a layout and cross section of a typical DMOS NPN. This structure uses a circular DMOS implant to form the base and emitter of the transistor. The drawn emitter area equals the drawn area of the DMOS implant. The emitter is contacted by a central plug of NSD, and the base is contacted by a ring of PSD surrounding the DMOS implant. The boron DMOS implant must overlap the PSD implant sufficiently to allow for misalignment. This usually requires that the two implants practically abut one another. The extrinsic collector consists of NBL and deep-N+, just as in a conventional NPN transistor.

[9] The RESURF (reduced surface field) structure eliminates avalanche breakdown due to electric field intensification at the corners of the drain: J. A. Appels and H. M. J. Vaes, "High Voltage Thin Layer Devices (RESURF Devices)," *IEEE 25th Int. Electron Devices Meeting*, 1979, pp. 238–241. For a typical integrated bipolar/CMOS/DMOS process, see A. Andreini, C. Contiero, and P. Galbiati, "A New Integrated Silicon Gate Technology Combining Bipolar Linear, CMOS Logic, and DMOS Power Parts," *IEEE. Trans. Electron Devices*, Vol. ED-33, #12, 1986, pp. 2025–2030.

[10] For example, the RESURF structure has been touted for reducing R_{sp}. See C.-Y. Tsai, J. Arch, T. Efland. J. Erdeljac, L. Hutter, J. Mitros, J.-Y. Yang, and H.-T. Yuan, "Optimized 25V, 0.34m$\Omega \cdot$ cm^2 Very-Thin-RESURF (VTR), Drain-Extended IGFETs in a Compressed BiCMOS Process," *IEDM*, 1996, pp. 469–472.

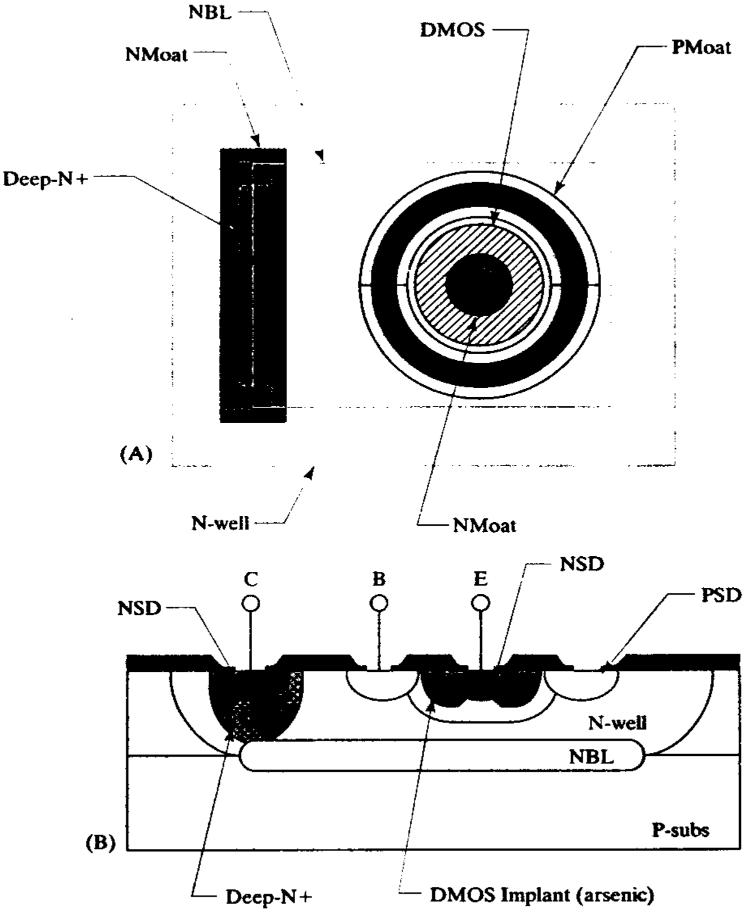

FIGURE 12.21 Layout and cross section of a DMOS NPN (omitting poly field plate).

A P+/N+ junction appears in the conventional DMOS structure between the P+ backgate contact and the N+ DMOS implant. The potential for leakage across this junction is of no concern in a DMOS transistor because the source always shorts to the backgate. The same is not true in the DMOS NPN because these diffusions form its base and emitter. Leakage can be avoided by reducing the dosage of the arsenic DMOS implant. NSD must now be added to the source regions to allow Ohmic contact to the lightly doped arsenic implant.

The structure in Figure 12.21 omits the moat geometry normally covering the DMOS implant, and allows thick-field oxide to grow over it. Dopant segregation and oxidation-enhanced diffusion drive the arsenic emitter deeper into the boron base, reducing the base width of the transistor. Conducting a field oxidation over the DMOS implant thus increases the beta of the DMOS NPN.

All DMOS NPN transistors contain a parasitic DMOS transistor connected between collector and emitter. The structure in Figure 12.21 does not show the poly gate electrode required to suppress this parasitic device. The poly electrode, or *field plate,* must cover the exposed boron DMOS implant with sufficient overlap to allow for misalignment. This field plate is usually connected to the emitter, since this connection shorts the gate and source of the parasitic DMOS.

12.3 THE JFET TRANSISTOR

Junction field-effect transistors (JFETs) were used throughout the 1970s and the early 1980s as substitutes for the less-reliable MOS devices of that era. JFETs were often used in the input stages of operational amplifiers to obtain input leakage currents several orders of magnitude smaller than those generated by the best bipolar circuits.[11] JFETs were also used as analog switches and as current sources.

Standard bipolar easily accommodated the steps required to construct simple JFET structures. The resulting *BiFET* processes merged bipolar and JFET transistors in much the same way that modern BiCMOS processes merge bipolar and CMOS. These BiFET processes were primarily used to construct low-input-current and low-noise operational amplifiers. The older BiFET processes have become largely obsolete because modern BiCMOS processes generally offer better performance (although low-noise BiMOS amplifiers can still outperform their BiCMOS counterparts).

JFET transistors remain of interest because they can be constructed on many existing processes without requiring any additional masking steps, and the resulting devices can replace high-value resistors in startup circuits. The following sections provide a brief overview of the operation and construction of JFETs, with an emphasis on structures compatible with standard bipolar and analog BiCMOS processes.

12.3.1. Modeling the JFET

Although the I-V characteristics of the JFET broadly resemble those of the depletion-mode MOS transistor, the underlying physics of the two devices are quite different. Most textbooks derive the JFET equations from fundamental principles, but so many assumptions are made along the way that the results have little practical value. This section discusses only those aspects of the theoretical model required to understand the sizing of JFET transistors and leaves the remaining details to other texts.[12]

The *pinchoff voltage* V_P of a JFET equals the minimum drain-to-source voltage V_{DS} required to pinch off the drain end of the channel when the gate-to-source voltage V_{GS} equals zero. In theory, the pinchoff voltage of an ideal JFET equals

$$V_P \cong 1.9 \cdot 10^{-16} N_C t^2 \qquad [12.11]$$

where N_C equals the doping concentration of the channel in atoms/cm^3 and t equals the channel thickness in microns.[13] In practice, the channel doping usually varies with depth and the pinchoff voltage must be determined empirically. This is done by examining the I-V characteristics of the JFET for $V_{GS} = 0$. The drain current I_D remains approximately constant at high drain-to-source voltages. As V_{DS} decreases, a point is eventually reached at which the drain current begins to diminish (Figure 12.22B). The pinchoff voltage equals the drain-to-source voltage at this inflection point.

The *saturation current* I_{DSS} of a JFET equals the drain current at $V_{GS} = 0$ and $V_{DS} = V_P$. If one assumes a uniformly doped channel with resistivity ρ, width W, length L, and thickness t, then the saturation current equals

$$I_{DSS} = \frac{V_P t}{3\rho}\left(\frac{W}{L}\right) \qquad [12.12]$$

[11] The advantages of JFET input stages vanish at higher temperatures because the input current of a JFET increases exponentially with temperature, while the input current of a base-current-compensated bipolar circuit increases somewhat more slowly.

[12] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed. (New York: John Wiley & Sons, 1986), p. 202ff.

[13] The full equation is $V_P = qN_C t^2/2\varepsilon$, where q is the charge on the electron and ε is the permittivity of silicon.

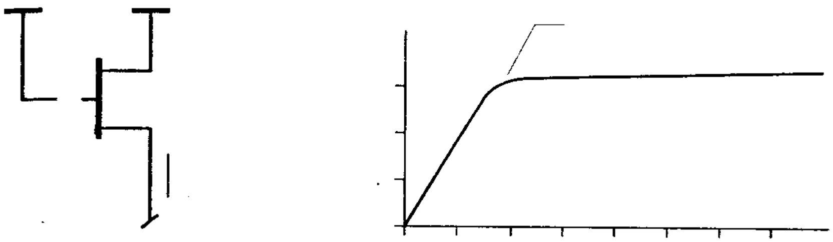

FIGURE 12.22 (A) A schematic of a P-JFET connected as a current source and (B) an I-V plot of the same device showing $V_P = 8V$ and $I_{DSS} = 16\mu A$.

This equation can also be used for nonuniformly doped channels providing that one empirically determines the effective channel resistivity ρ by measuring devices of different widths and lengths and fitting these measurements to the equation. Several factors complicate the extraction of I_{DSS}. The width and length used in the equation do not exactly correspond to the drawn dimensions of the device, any more than the effective width and length of MOS transistors exactly correspond to their drawn dimensions (Section 11.2.1). Correction factors δW and δL relate the effective width W_{eff} and effective length L_{eff} to the drawn width W_d and drawn length L_d

$$W_{eff} = W_d + \delta W \qquad [12.13A]$$

$$L_{eff} = L_d + \delta L \qquad [12.13B]$$

For devices with channel widths of less than 10μm, the value of I_{DSS} can be accurately determined only by measuring a device having the desired channel width. Devices that channel lengths of less than 10μm are better avoided because a variety of short-channel effects complicate the task of sizing the transistors.

12.3.2. JFET Layout

Practical JFET devices can be created using existing layers of a standard bipolar or an analog BiCMOS process. Figure 12.23 shows one type of N-channel JFET compatible with standard bipolar processing. This device is sometimes called an *epi-FET* because its channel consists of a portion of the N-type epitaxial layer. The epi-FET is also called an *epi pinch resistor*, particularly when it operates in its linear region (Section 5.5.5). The thickness of the channel has been greatly reduced by placing a base diffusion over the epi. The updiffusion of the underlying substrate causes substantial grading of the backgate-body junction and renders the constant-doping approximations that underlie equations 12.11 and 12.12 of questionable validity. These devices are usually sized by interpolating between the I_{DSS} currents measured on an array of test devices.

The tank geometry determines the drawn width W_d while the base pinch plate determines the drawn length L_d. The effective width of an epi-FET is substantially smaller than its drawn width because of isolation outdiffusion. The relationship between effective width and drawn width becomes nonlinear for small widths because of diffusion interactions between the opposing sidewalls of the channel. The width correction factor δW may therefore vary with width, especially for small widths. The length correction factor δL also varies with length because of the presence of the extrinsic source/drain regions on either end of the channel, but δL has little effect upon devices having channel lengths of at least 50μm.

The base pinch plate extends into the surrounding isolation and shorts the gate of the epi-FET to substrate. Most epi-FETs are used as startup devices in which the grounded-gate configuration is quite acceptable. If the drain-to-source voltage across the epi-FET is large enough, then its drain will draw a current equal to the

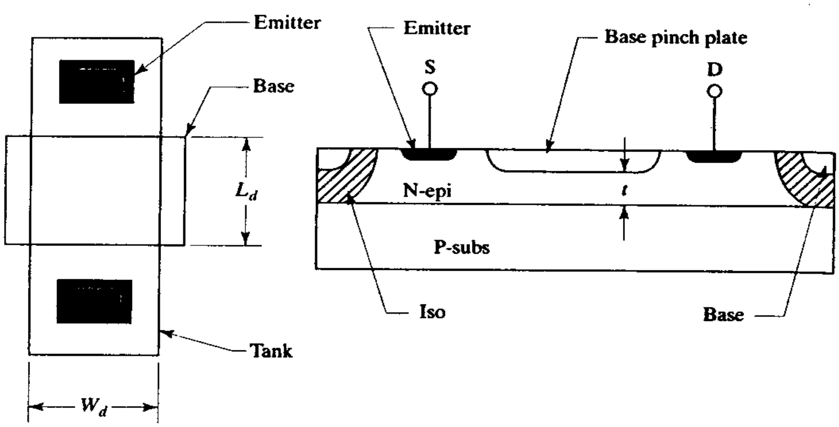

FIGURE 12.23 Layout and cross section of an N-channel JFET constructed in standard bipolar. The gate connects to the substrate and is accessed through an adjacent substrate contact.[14]

saturation current I_{DSS}. In practice, most epi-FETs have such large pinchoff voltages that they do not fully saturate under normal operating conditions, and they therefore resemble pinch resistors. The main advantages of the epi-FET include high breakdown voltage and low transconductance, which together allow it to replace a much larger pinch resistor. The operating voltage of an epi-FET is limited only by the breakdown of the epi-base junction. JFETs are immune to hot-carrier-induced threshold shifts because they do not contain a gate dielectric. They are also immune to the parasitic channel formation and conductivity modulation because the base pinch plate serves as a field plate covering the active region of the device.

Epi-FETs are designed for compactness rather than for precision. These transistors normally use the minimum channel width, even though wider devices exhibit less variability. The channel is frequently serpentined to fit into unused areas in the layout. Contacts are usually placed over the base pinch plate and connected to substrate potential. Although not strictly necessary, these contacts help minimize variations in epi-FET current caused by substrate debiasing. Any contact to the base pinch plate also serves as a substrate contact in its own right and helps extract stray substrate currents flowing near the epi-FET. Some designers use rounded bends in serpentine epi-FETs believing that these increase the breakdown voltage by preventing electric field intensification. Although this practice causes no harm, it provides little or no benefit because the exposed edge of the base pinch plate usually breaks down before the isolation sidewalls.

Analog BiCMOS processes can construct an *N-well JFET* analogous to the epi-FET in Figure 12.23 by substituting N-well for the tank and NMoat for emitter (Figure 12.24). The resulting device usually has a lower pinchoff voltage than its epi-FET counterpart due to the graded nature of the well. The pinchoff voltage can be reduced still further by growing field oxide over the base pinch plate, as the resulting oxidation-enhanced diffusion drives the base deeper into the N-well.

N-well JFETs and epi-FETs vary in several ways. One critical difference concerns the overlap of the base pinch plate over the channel. In the epi-FET, the isolation diffuses inward and the base pinch plate need overlap the tank only slightly, if at all. The base pinch plate of the N-well JFET must overlap the well by a much greater distance because the N-well diffuses outward rather than inward. The high resistance of the P-epi also makes it desirable to add contacts directly to the base pinch plate

[14] A similar device is discussed in D. J. Hamilton and W. G. Howard, *Basic Integrated Circuit Engineering* (New York: McGraw-Hill, 1975), p. 170.

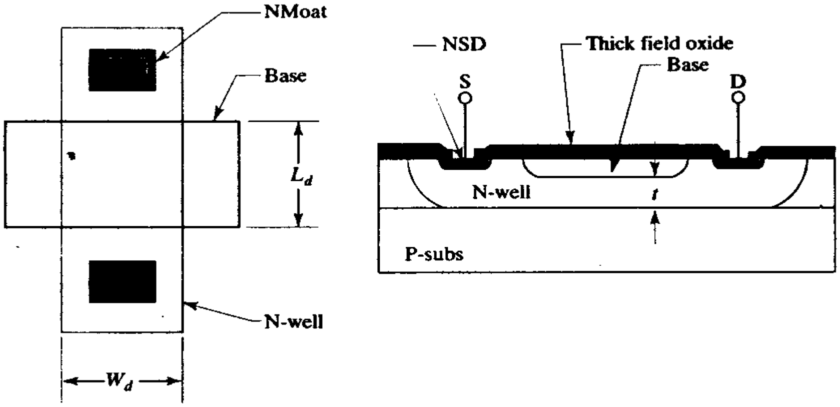

FIGURE 12.23 Layout and cross section of an N-well JFET constructed in an analog BiCMOS process (the contacts to the base plate have been omitted for clarity).

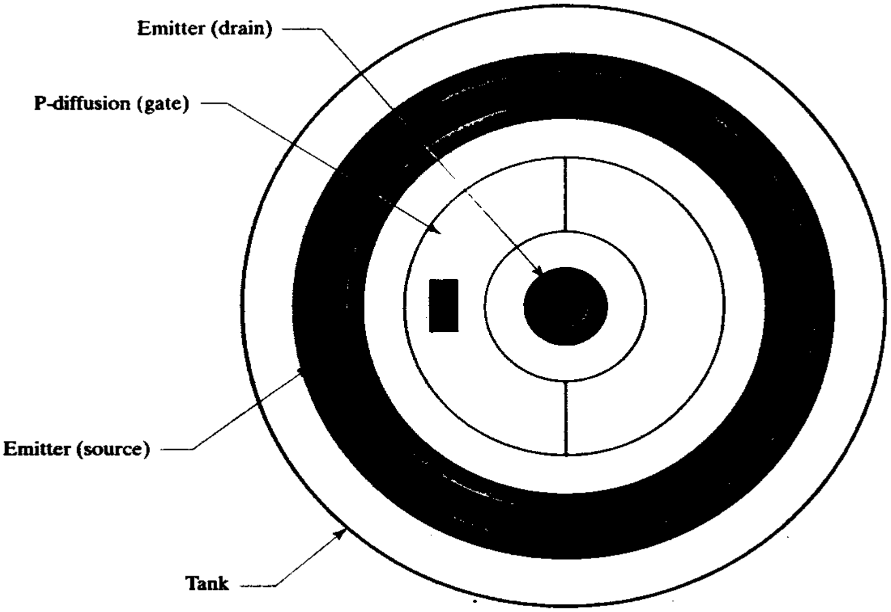

FIGURE 12.24 Annular circularly symmetric N-channel JFET.

rather than to rely on the presence of substrate contacts elsewhere on the die. These contacts should not reside over the channel of the N-well JFET because the moat region required for the contact alters the thickness of the channel. Instead, the contacts should be located next to the device and should connect to it by a strip of base or PSD diffusion.

An N-well JFET with a minimum-width channel has a much lower pinchoff voltage than one with a wider channel. If the channel is covered by thick-field oxide, the narrowest devices may even pinch off entirely and become unusable. These effects occur because the dopant in a narrow N-well diffuses laterally as well as vertically, leaving a lower overall doping concentration within the channel. The dopant in a wider well diffuses laterally near the edges, but the center of the well still retains a high dopant concentration. The wider device thus has a higher pinchoff voltage than the narrow one. If necessary, the pinchoff voltage of an N-well JFET can be increased by placing a moat region above the base or by substituting a PSD implant for the base implant. The PSD implant usually gives so high a pinchoff voltage that the device cannot saturate under normal operating conditions. Thus it behaves as a nonlinear pinch resistor rather than as a true FET (Section 5.5.9).

P-channel JFETs can be constructed in both standard bipolar and analog BiCMOS processes, but those constructed from existing diffusions leave much to be desired. The standard bipolar device has the same structure as a base pinch resistor (Figure 3.15). The analog BiCMOS device has a similar structure, consisting of base pinched by NSD rather than by emitter. The operating voltages of these devices are limited by the avalanche of the base-emitter and base-NSD junctions, respectively. The pinchoff voltages of both devices greatly exceed their respective breakdown voltages, so neither device ever saturates. Both of these devices are really nothing more than nonlinear pinch resistors (Section 5.5.3). A special N-type implant must be added to the process to construct a true P-JFET capable of operating in saturation. This implant must have a slightly shallower junction depth than the base diffusion, and a doping concentration just sufficient to invert the base diffusion. A shallower diffusion yields too large a pinchoff voltage, and a more heavily doped one produces too low a breakdown voltage. No suitable diffusion exists in either standard bipolar or analog BiCMOS, although one can be added as a process extension. Previous-generation BiCMOS processes were generally derived from standard bipolar by the addition of just such an extension. The P-channel transistors constructed in this way are called *double-diffused JFETs* because their gates are produced by the diffusion of the N-implant into the base. The layout and cross section of the double-diffused P-JFET are essentially the same as that of the base pinch resistor in Figure 3.15, with the substitution of the new N-implant for the emitter. New processes rarely support the P-JFET extension because CMOS transistors have largely supplanted JFETs.

All of the layouts previously discussed short the gate to the backgate, which in the N-channel device consists of the substrate. In order to use the N-channel JFET in any application other than as a grounded current source, one must first separate the gate and the backgate electrodes by using an annular structure similar to that in Figure 12.24. The gate of the annular N-JFET consists of a ring-shaped P-type diffusion placed inside an N-epi tank. Tank contacts placed inside and outside this ring serve as the drain and source, respectively. This arrangement minimizes the drain capacitance at the cost of increased source capacitance.

The schematic symbol used for the annular N-JFET is exactly the same as that used for the conventional N-JFET (Figure 1.30A). The two can be differentiated by examining the connection of the gate electrode. The conventional layout in Figure 12.23 should be used if the gate connects to substrate potential. If the gate connects to any other potential, the transistor must use the annular layout shown in Figure 12.24. The substrate forms the backgate of the annular device. The width and length of annular JFET devices are computed using the rules presented for annular MOS transistors (Section 11.2.6).

12.4 MOS TRANSISTOR MATCHING

A wide variety of analog circuits use matched MOS transistors. Some circuits, such as differential pairs, rely on matching of gate-to-source voltages, while others, such as current mirrors, rely on matching of drain currents. The biasing conditions required to optimize voltage matching differ from those required to optimize current matching. One can optimize MOS transistors either for voltage matching or for current matching, but not simultaneously for both.

The relationship between biasing and voltage matching is easily derived from the Shichman-Hodges equations (Section 11.1.1). Suppose two matched MOS transistors operate at the same drain current I_D. If the transistors were ideal devices, then they would develop exactly the same gate-to-source voltage V_{GS}. In practice, mismatches cause the gate-to-source voltages of the two transistors to differ by an amount $\Delta V_{GS} = V_{GS1} - V_{GS2}$. Assuming that the transistors operate in saturation, as is usually the case, then the offset voltage ΔV_{GS} equals

$$\Delta V_{GS} \cong \Delta V_t - V_{gst1}\left(\frac{\Delta k}{2k_2}\right) \qquad [12.14]$$

where ΔV_t equals the difference between the threshold voltages of the two transistors, Δk equals the difference between their device transconductances, V_{gst1} equals the effective gate voltage of the first transistor. and k_2 equals the device transconductance of the second (Appendix D). The offset voltage ΔV_{GS} depends on device dimensions due to the presence of the device transconductance k_2 in the denominator. Similarly, the offset voltage depends on biasing conditions because of the presence of the effective gate voltage V_{gst1} in the equation. These dependencies are unique to MOS transistors and are not shared by bipolar transistors (Section 9.2).

The MOS designer can minimize the offset voltage ΔV_{GS} by reducing the effective gate voltage V_{gst} of the matched transistors. MOS circuits that depend on voltage matching therefore benefit from the use of large W/L ratios and low operating currents. The improvements obtainable in this manner are limited by the onset of subthreshold conduction and by the presence of threshold mismatches. As a practical matter, reducing V_{gst} below about 0.1V produces little improvement in voltage matching.

MOS circuits relying on current matching behave quite differently. The mismatch between two drain currents, I_{D1} and I_{D2}, can be specified in terms of a ratio I_{D2}/I_{D1} equal to

$$\frac{I_{D2}}{I_{D1}} \cong \frac{k_2}{k_1}\left(1 + \frac{2\Delta V_t}{V_{gst1}}\right) \qquad [12.15]$$

The mismatch in drain currents actually increases at low effective gate voltages due to a larger contribution from the threshold mismatch ΔV_t (Appendix D). MOS circuits relying on current matching should operate at reasonably large effective gate voltages to avoid exacerbating threshold voltage variations. The optimal value of V_{gst} depends on many factors and is difficult to quantify. As a practical matter, one should endeavor to maintain a nominal V_{gst} of at least 0.3V (and preferably 0.5V) in MOS transistors generating matched currents. Larger effective gate voltages may provide some additional benefit, but most applications cannot spare the headroom to support a higher V_{gst}.

In summary, MOS circuits that generate matched voltages should operate at low effective gate voltages, while MOS circuits that generate matched currents should operate at high effective gate voltages. For most purposes, a nominal V_{gst} of 0.1V or less will provide optimal voltage matching, and a nominal V_{gst} of 0.3V or more will provide optimal current matching. Assuming that the circuit designer adjusts the biasing of the transistors to these values, the matching now depends almost entirely on the care taken in transistor layout. The next three sections discuss layout considerations that affect MOS matching.

12.4.1. Geometric Effects

The size, shape, and orientation of MOS transistors all affect their matching. Large transistors match more precisely than small ones because increased gate area helps minimize the impact of localized fluctuations. Long-channel transistors match more precisely than short-channel ones because longer channels reduce linewidth variations and channel-length modulation. Transistors oriented in the same direction match better than those oriented in different directions because of the anisotropic nature of monocrystalline silicon. This section discusses the impact of these and other geometric factors on MOS transistor matching.

Gate Area

MOS mismatches have been experimentally measured for a number of processes. These measurements reveal that the magnitude of the threshold voltage mismatch varies inversely with the square root of the active gate area. This relationship can be expressed in terms of the effective channel dimensions W_{eff} and L_{eff} as follows

$$s_{V_t} = \frac{C_{V_t}}{\sqrt{W_{eff}\,L_{eff}}} \qquad [12.16]$$

where s_{V_t} is the standard deviation of the threshold voltage mismatch and C_{V_t} is a constant.[15] The value of C_{V_t} is empirically determined by measuring the random mismatch between pairs of transistors of different sizes. The results only apply to transistors closely resembling the test devices used to derive C_{V_t}. The relationships between drawn dimensions and effective dimensions are not always known, and sometimes the drawn dimensions W_d and L_d must be substituted for the effective dimensions W_{eff} and L_{eff}. This substitution will have little effect on the accuracy of the predictions as long as both dimensions of the transistor are several times greater than minimum.[16]

Strictly speaking, equation 12.16 only applies to MOS transistors that have been carefully laid out to ensure optimal matching. Poorly matched transistors often exhibit gross defects that do not scale as predicted. Once these gross defects have been eliminated, the residual threshold mismatches usually follow equation 12.16 quite precisely. Theoretical studies suggest that residual threshold mismatches stem mostly from statistical fluctuations in the distribution of backgate dopants.[17] Statistical fluctuations in the distribution of fixed oxide charge may also play a minor role.

Random short-range variations also appear to determine the residual transconductance mismatches observed between well-matched devices. If the transconductance mismatch is described as a normalized ratio s_k/k, then it varies with effective dimensions, W_{eff} and L_{eff}, as follows

$$\frac{s_k}{k} = \frac{C_k}{\sqrt{W_{eff}\,L_{eff}}} \qquad [12.17]$$

where C_k is a constant. Possible causes for short-range variations in transconductance include linewidth variation, gate oxide roughness, and statistical variations in mobility. The relative importance of these causes is not known, although several authors have suggested that mobility variations predominate.

Gate Oxide Thickness

Many designers believe that MOS transistors with thin gate oxides match better than those with thick gate oxides. At first glance, the evidence seems to support this hy-

[15] K. R. Lakshmikumar, R. A. Hadaway, and M. A. Copeland, "Characterization and Modeling of Mismatch in MOS Transistors for Precision Analog Design," *IEEE J. Solid-State Circuits*, SC-21, #6, 1986, pp. 1057–1066.

[16] Substituting drawn for effective dimensions will have very grave effects if either the width or the length of the matched devices is small; see S. J. Lovett, M. Welten, A. Mathewson, and B. Mason, "Optimizing MOS Transistor Mismatch." *IEEE J. Solid-State Circuits*, Vol. 33, #1, 1998, pp. 147–150.

[17] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching Properties of MOS Transistors," *IEEE J. Solid-State Circuits*, Vol. SC-24, #5, 1989, pp. 1433–1439. Also see Lakshmikumar, *et al.*, p. 1059.

pothesis, but factors other than oxide thickness are probably at work. The low-voltage, thin-oxide transistors are usually produced by some form of constant-field scaling (Section 11.2.5) that affects not only gate oxide thickness but also backgate doping. If, as research seems to indicate, backgate doping is the dominant cause of threshold voltage mismatch, then constant-field scaling should decrease mismatch by a factor of *S*, where *S* is the scaling factor. Constant-field scaling also decreases oxide thickness by a factor of *S*. This coincidence may account for the empirically observed relationship between oxide thickness and threshold voltage mismatch. Regardless of the exact cause, scaling MOS transistors to smaller dimensions does seem to improve their threshold voltage matching. This effect does not extend to transconductance matching, which appears to remain largely independent of scaling.

Channel Length Modulation

Channel length modulation can cause severe mismatches between short-channel transistors operating at different drain-to-source voltages. The systematic mismatch between the transistors is proportional to the difference between their drain-to-source voltages, and inversely proportional to their channel length. Drawn lengths of 15 to 25μm are generally adequate for noncritical applications such as current distribution networks. Greater precision can be obtained by operating the matched transistors at similar drain-to-source voltages. for example, through the addition of cascodes. MOS designers rarely use source degeneration to combat channel length modulation because the low transconductance of MOS transistors makes it difficult to obtain adequate degeneration without using extremely large resistors.

Orientation

The transconductances of MOS transistors depend on carrier mobilities, and these in turn exhibit orientation-dependent stress sensitivities. MOS transistors oriented along different crystal axes will therefore exhibit different transconductances under stress. Since all packaged devices experience some stress, these mismatches can only be avoided by orienting matched transistors in the same direction. The devices in Figure 12.25A, which are oriented along the same crystal axis, match better than the devices in Figures 12.25B and 12.25C, which are not. Stress-induced mobility variations can induce current matching errors of several percent between rotated devices.[18] The use of tilted wafers may induce current matching errors of as much as 5%.[19]

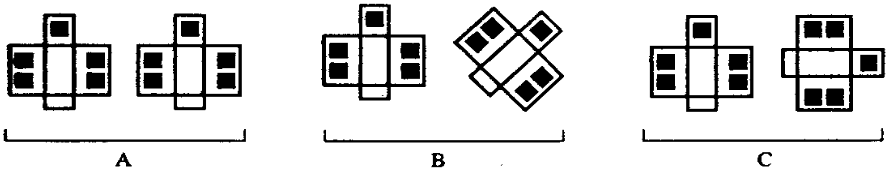

FIGURE 12.25 (A) Devices oriented in the same direction match more precisely than (B,C) those oriented in different directions.

Editing can easily introduce orientation errors if the design has not been properly partitioned. Consider a circuit that contains two matched transistors: M_1, located in cell X_1; and M_2, located in cell X_2. During top-level layout, the designer decides to rotate cell X_1 by 90°. Although this operation seems innocuous, it actually introduces a 90° difference between the orientations of M_1 and M_2. Errors of this sort can be prevented by grouping matched devices together in the same cells. This can some-

[18] Pelgrom, *et al.*, p. 1436.

[19] J. E. Chung, J. Chen, P.-K. Ko, C. Hu, and M. Levi, "The Effects of Low-Angle Off-Axis Substrate Orientation on MOSFET Performance and Reliability," *IEEE Trans. on Electron Devices*, Vol. 38, #3, 1991, pp. 627–633.

times make the schematic more difficult to comprehend, but it greatly reduces the risk of inadvertently introducing matching errors during editing.

MOS transistors that do not self-align must follow very strict orientation rules. Consider the asymmetric extended-drain NMOS transistors, M_1 and M_2, in Figure 12.26A. Each of these transistors is a mirror image of the other. The channel lengths of M_1 and M_2 are both defined by the overhang of their poly gates beyond their respective N-well regions. Suppose that photolithographic misalignment causes the poly gates to shift to the right. This misalignment increases the channel length of M_1 and decreases the channel length of M_2. These mismatches are easily eliminated by ensuring that the matched devices are superimposable, as are M_3 and M_4 in Figure 12.26B. Even fully self-aligned transistors may experience slight orientation-dependent mismatches due to diagonal shifting of the source/drain implants (Section 12.4.4).

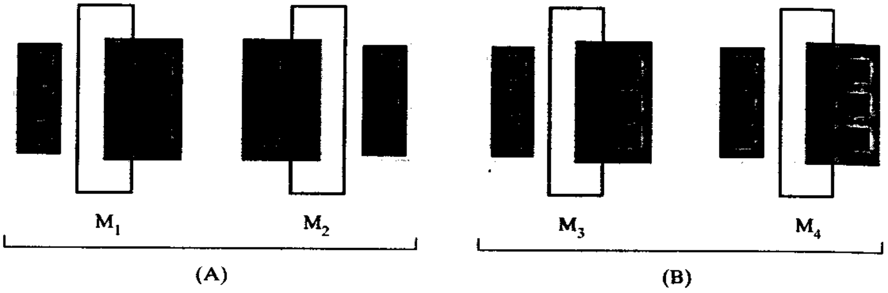

FIGURE 12.26 Extended-drain transistors that are (A) mirror images of one other experience mismatches that do not affect (B) superimposable transistors.

12.4.2. Diffusion and Etch Effects

The previous section examined sources of mismatch that depended solely upon geometry. Certain other types of mismatch are caused by the presence or absence of other structures near the matched transistors. For example, the presence of poly regions near the gate electrodes can cause slight variations in polysilicon etch rates. These variations produce mismatches in the effective widths and lengths of the matched transistors. Similarly, the placement of other diffusions near the channel may influence the backgate dopant concentration and may therefore cause variations in both threshold voltage and transconductance. This section discusses these and other sources of interaction-dependent mismatch.

Polysilicon Etch Rate Variations

Polysilicon does not always etch uniformly. Large poly openings clear more quickly than small ones because etchant ions have freer access to the sides and bottom of the large opening. The edges of the large opening therefore exhibit some degree of overetching by the time the smaller openings clear. This effect can cause variations in the gate lengths of poly-gate MOS transistors. Consider the layout in Figure 12.27A. The gate of transistor M_2 faces adjacent gates on both sides, but the gates of transistors M_1 and M_3 face an adjacent gate on only one side. The outside edges of the gates of M_1 and M_3 experience more erosion than the corresponding edges of the gate of M_2, so the gate lengths of M_1 and M_3 are slightly shorter than the gate length of M_2.

The etch rate variations experienced by MOS transistors are usually smaller than those experienced by poly resistors (Section 7.2.4), because poly gates do not lie as close together as poly resistor segments do. Many MOS transistors also use relatively long channel lengths. Even so, transistors that must achieve moderate or precise current matching should use dummy gates to ensure uniform etching. Failure to do so may produce current mismatches of 1% or more. Figure 12.27B shows an example

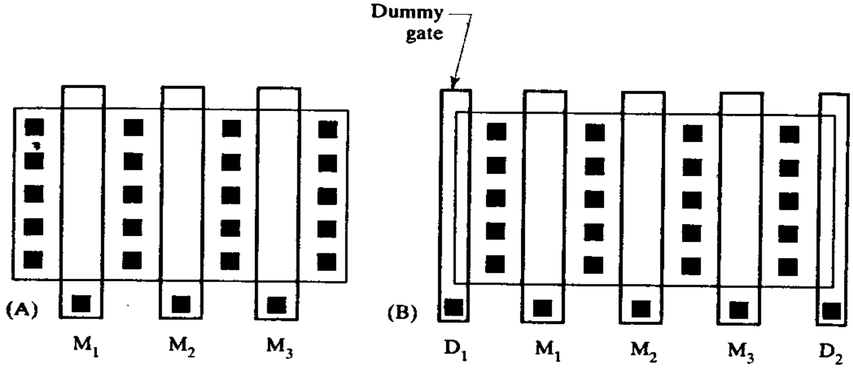

FIGURE 12.27 MOS transistor arrays (A) without dummy gates and (B) with dummy gates.

of an array of MOS transistors incorporating dummies. Most designers make the dummy gates the same width as the active ones, but this precaution is not strictly necessary because the width of the poly strips is far less significant than their spacing. Dummies D_1 and D_2 are therefore made as narrow as possible while still allowing space for a contact. The spacing between the dummies and the actual gates must exactly equal the spacing between the actual gates themselves.

Since the dummies are not actual transistors, they do not require the presence of source/drain regions along their outside edges. The source/drain implant can therefore terminate on top of the dummies, as it does in Figure 12.27B. This should not introduce significant mismatches as long as the moat geometry extends beyond the inner edge of the dummy gate electrodes by a few microns to ensure that the edge of the dummy rests on thin gate oxide.

The dummy gate electrodes should be electrically connected to prevent them from floating at unknown potentials. Although this precaution is not strictly necessary, it helps ensure that the electrical characteristics of the transistors are not affected by the formation of spurious channels or depletion regions beneath the dummies. Some designers connect the dummies to the adjacent gate electrodes, but this practice is not recommended because it increases terminal capacitances and leakage currents. A better practice consists of connecting the dummies to the backgate potential.

Many designers interconnect multiple gate electrodes with a strip of polysilicon. While this is undeniably convenient, it may introduce etch rate variations due to the presence of an adjacent polysilicon geometry. For the best possible matching, one should use simple rectangular strips of polysilicon connected by metal.

Contacts Over Active Gate

For reasons not well understood, the placement of contacts over the active gate regions of MOS transistors sometimes induces significant threshold voltage mismatches. One possible explanation for this effect is the presence of metal above the active gate (see Section 12.4.3). Another potential mechanism for contact-induced mismatches involves the localized silicidation of contacts. In processes where the gate poly is sufficiently thin, some silicide may actually penetrate entirely through the gate poly. The presence of silicide at the oxide interface drastically alters the work function of the gate electrode in the vicinity of the contact and can cause gross threshold voltage mismatches. Changes in grain size, dopant distributions, and stress patterns may also play a role in generating contact-induced mismatches. Figure 12.27 illustrates the proper placement of gate contacts in extensions of the poly gate electrodes. This precaution ensures that the contacts reside over thick-field oxide, where they cannot significantly alter transistor properties.

Annular transistors such as those in Figure 11.19 present a special problem because they require contacts to be placed over active gate regions. Matched annular transistors should only be used if absolutely necessary. If they are used, then they should incorporate identical arrangements of minimal numbers of small gate contacts. In annular extended-drain transistors, the gate contacts should reside over the field-relief regions so they rest on field oxide rather than on gate oxide. This precaution effectively locates the contacts outside the active gate region. In cases in which the field relief region is not wide enough to accommodate the contacts, they should still be located as far inside it as possible to take advantage of the zone of intermediate oxide thickness just inside the edges of the moat region (the bird's beak).

Diffusions Near the Channel

Deep diffusions can affect the matching of nearby MOS transistors. The tails of these diffusions extend a considerable distance beyond their junctions, and the excess dopants they introduce can shift the threshold voltages and alter the transconductances of nearby transistors. The deep-N+ sinker of the analog BiCMOS process represents one example of a deep diffusion. All sinkers and similar diffusions should be spaced away from matched channels by at least twice their junction depth.

Wells also qualify as deep diffusions. N-well geometries should not be placed near matched NMOS transistors to prevent the tail of the N-well dopant distribution from intersecting the channels of the matched transistors. PMOS transistors should be placed far inside the edges of their enclosing N-well regions to prevent outdiffusion from causing variations in backgate doping. In all cases, a spacing from the active gate regions equal to or greater than twice the junction depth of the deep diffusion should limit interactions to negligible levels (Figure 12.28).

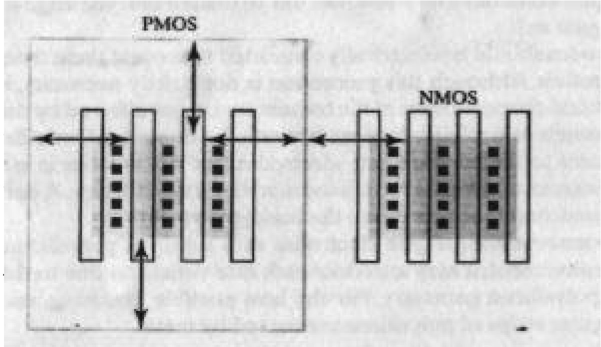

FIGURE 12.28 Spacings between drawn well boundaries and active gate regions.

Because MOS transistors are surface devices, they are vulnerable to the surface discontinuities produced by the NBL shadow. The channels of matched MOS transistors should be placed far enough away from NBL boundaries to allow for both misalignment and pattern shift. If the pattern shift has not been characterized, assume that it can displace the NBL shadow by up to 150% of the epi thickness. Thus the spacing from the active gate regions to the edge of the nearest NBL region should equal at least 150% of the epi thickness. Although this substantially increases the overlap of NBL over the matched transistor, much of this space is already required to satisfy the increased well spacings discussed previously.

PMOS versus NMOS Transistors

NMOS transistors usually match more precisely than PMOS transistors. This phenomenon has been observed on a number of different processes, including both

P-well and N-well variants. Several authors have reported that PMOS transistors exhibit 30 to 50% more transconductance mismatch than comparable NMOS transistors.[20] Some studies have also found increased threshold mismatches in PMOS transistors, although these do not appear to be as significant as the differences in transconductance matching.

The mechanisms responsible for the differences between PMOS and NMOS transistors are not well understood. Possible culprits include increased backgate doping variability, the presence of buried channels, and orientation-dependent stress effects. Several authors have suggested that the increased variability stems (at least in part) from differences in the threshold adjust implants, but this seems an unlikely explanation since so many different processes behave similarly.

12.4.3. Thermal and Stress Effects

Another important category of mismatches stems from long-range variations called *gradients*. The magnitude of gradient-induced mismatches depends on the separation between the effective centers, or *centroids*, of the matched devices. Providing that the devices are placed relatively close to one another, the variation ΔP in parameter P between two matched devices equals the product of the distance d between the centroids and the gradient ∇P along a line connecting the two centroids:

$$\Delta P \cong d\nabla P \qquad [12.18]$$

The impact of the gradient on matching depends on both the magnitude of the gradient and the distance between the centroids of the matched devices. Gradients that affect MOS matching include those of oxide thickness, stress, and temperature.

Oxide Thickness Gradients

The thickness of a grown oxide film depends on the temperature and composition of the oxidizing atmosphere used to create it. Although modern oxidation furnaces are very precisely controlled, slight variations of temperature and atmospheric composition still occur within the furnace tube. Thick oxide layers often exhibit a pattern of concentric rainbow-colored rings that betray the presence of a radial oxide thickness gradient. Gate oxides are too thin to exhibit interference colors, but they also tend to exhibit radial oxide thickness gradients. Devices placed close to one another have very similar oxide thicknesses, while devices placed further apart exhibit greater differences in oxide thickness. These differences directly affect threshold voltage matching.

Stress Gradients

Stress affects the device transconductance of MOS transistors by causing variations in carrier mobilities. As discussed in Section 7.2.6, the effects of stress on mobility depend on orientation. In bulk silicon, holes experience maximum stress dependence along the <110> axis and minimum stress dependence along the <100> axis. Similarly, electrons in bulk silicon experience maximum stress dependence along the <100> axis and minimum stress dependence along the <110> axis. Dice are oriented to the major wafer flat, which lies perpendicular to a <110> axis. Therefore electrons experience minimum stress-induced bulk mobility variation in directions

[20] Lakshmikumar, *et al.*, pp. 1060, 1062: Pelgrom, *et al.*, p. 1437.

aligned with the X- and Y-axes of a (100)-oriented die, while holes experience minimum stress-induced bulk mobility variations in directions oriented 45° to these axes.

The stress dependence of bulk mobilities drops to nearly zero along the preferred orientations, but unfortunately the same is not true of the *effective mobilities* of carriers confined to a channel. The stress dependence of the effective mobilities does decrease along the directions predicted by theory, but these minima are nowhere near as pronounced as in the case of bulk mobilities. A diagonal placement of a PMOS transistor may reduce the stress dependence of its device transconductance by only 50%, rather than by the 90% or more one would expect based on bulk mobility data.[21] The randomizing effects of carrier collisions with the oxide/silicon interface probably account for the reduced orientation dependence of effective mobilities, but not all researchers agree on the details of this mechanism. Given these uncertainties, there seems little reason to diagonally orient PMOS transistors. One should instead rely on proper design of common-centroid layouts to minimize stress sensitivity.

Stress has relatively little effect on voltage matching because the threshold voltages of MOS transistors are largely independent of stress. What small stress dependencies do exist are probably caused by stress-induced changes in the bandgap voltage of silicon. The threshold voltage generally does not exhibit more than a few millivolts of stress-induced variation, which can be reduced still further by using common-centroid layout techniques.

Metallization-induced Stresses

The routing of metal leads over the active gate region of MOS transistors has been shown to produce stress-induced mismatches of several percent.[22] Metallization may cause even larger mismatches if the wafers are not sufficiently annealed in a reducing atmosphere, as the deposition of metal above the gate oxide appears to introduce a surface state charge into the gate oxide (see Section 11.1.1).

Ideally, leads should never be routed across the active gate regions of matched MOS transistors. If leads must cross the MOS transistors, then consider adding dummy leads so that each MOS transistor is crossed by an identical segment of metallization at the same position along its channel. This precaution will minimize the impact of the metallization on matching but will not totally eliminate it, so for the highest precision one should entirely avoid routing leads across the active gate regions.

Thermal Gradients

The voltage matching of MOS transistors depends primarily on the matching of threshold voltages. Threshold voltages decrease with temperature at roughly the same rate as base-emitter voltages of bipolar transistors—about –2mV/°C. Most of the temperature coefficient stems from variations in the work functions of the gate and backgate materials with temperature, and it is therefore virtually independent of drain current.[23] Voltage-matched MOS transistors therefore exhibit about the same sensitivity to thermal gradients as bipolar transistors.

MOS and bipolar input differential pairs respond quite differently to offset trimming. In bipolar circuits, trimming the offset voltage to zero also trims its temperature dependence to zero. This happens because the equation for the offset voltage between two matched bipolar transistors contains only one significant source of tem-

[21] H. Mikoshiba, "Stress-sensitive Properties of Silicon-gate MOS Devices," *Solid-State Elect.*, Vol. 24, #3, pp. 221–232.

[22] H. Tuinhout, M. Pelgrom, R. P. de Vries, and M. Vertregt, "Effects of Metal Coverage on MOSFET Matching," *IEDM*, 1996, pp. 735–738.

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

perature variation: the thermal voltage V_T. The temperature dependence therefore scales directly with ΔV_{BE}, and when ΔV_{BE} has been trimmed to zero, it vanishes. The input offset voltages of MOS transistors are trimmed by adjusting drain current densities. This operation attempts to cancel the mismatch in threshold voltages by introducing a compensating offset in device transconductance. The temperature coefficient of threshold voltage is caused by different mechanisms than the temperature coefficient of transconductance is, so the two are not equal, and the trimming operation does not reduce the temperature coefficient to zero. Thus, while trimmed bipolar input differential pairs retain their very low offset voltages over temperature, trimmed MOS input differential pairs do not. Trimmed bipolar amplifiers and comparators therefore provide much better performance over temperature than do their MOS counterparts.

The current matching of MOS transistors depends primarily on the matching of device transconductances. These transconductances are directly proportional to effective carrier mobilities, which exhibit rather large temperature coefficients. At temperatures near 25°C, MOS device transconductances typically exhibit temperature coefficients of about +7000ppm/°C. Temperature variations in threshold voltage have little effect on current matching as long as the transistors operate at a relatively large effective gate voltage V_{gst}. The low transconductance of MOS transistors makes them much less sensitive to thermal gradients than bipolar transistors, but it also makes it difficult to improve matching by source degeneration. Instead of relying on degeneration resistors. one should use common-centroid layout techniques.

12.4.4. Common-centroid Layout of MOS Transistors

Gradient-induced mismatches can be minimized by reducing the distance between the centroids of the matched devices. Some types of layout can actually reduce the distance between the centroids to zero. These *common-centroid* layouts can entirely cancel the effects of long-range variations as long as these are linear functions of distance. Even if the variations contain a nonlinear component, they still remain approximately linear over short distances. The more compact the common-centroid layout can be made, the less susceptible it becomes to nonlinear gradients. The best layouts for MOS transistors combine exact alignment of the centroids with compactness.

The active gate region of an MOS transistor usually takes the form of a long, narrow rectangle. As in the case of resistors, MOS transistors are usually divided into segments, or *fingers*, to allow the construction of a compact array. The simplest types of arrays involve the placement of multiple device fingers in parallel. If these fingers are properly interdigitated, then the centroids of the matched devices will align at a point midway along the axis of symmetry bisecting the array. Figure 12.29 shows an example of a pair of matched MOS transistors laid out as an interdigitated array.

This layout uses the interdigitation pattern ABBA to ensure exact alignment of the centroids (Section 7.2.6). If source and drain fingers are denoted by subscripts, then the pattern becomes ${}_DA_SB_DB_SA_D$. Notice that the A-segment on the right has its drain on the right, while the A-segment on the left has its drain on the left. Similarly, the B-segment on the right has its source on the right, while the B-segment on the left has its source on the left. Each transistor thus contains one segment oriented in either direction. The reason for this precaution is rather subtle. Suppose one transistor consists entirely of segments with drains on the left, while a second transistor consists entirely of segments with drains on the right. If left-oriented and right-oriented segments differ in any way, then the two transistors will not match. If both transistors consist entirely of segments oriented in the same direction, then

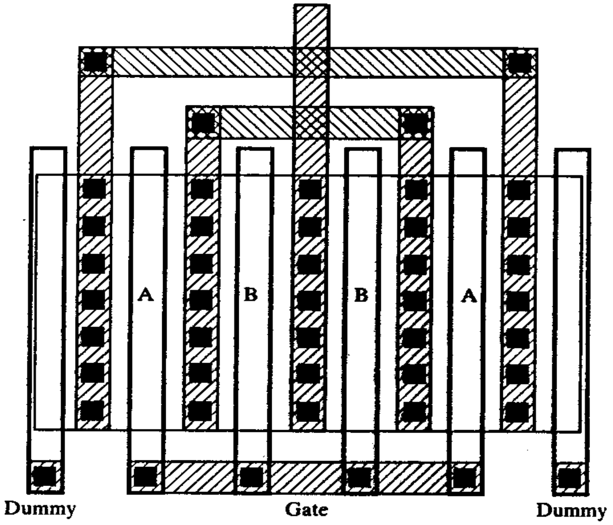

FIGURE 12.29 Interdigitated MOS transistors.

the effect of orientation on each transistor will be the same (Section 12.4.1). If each transistor consists of an equal number of left-oriented and right-oriented segments, then the effects of orientation will cancel and the transistors will again match.

More generally, if we define the *chirality* of a transistor as the fraction of right-oriented segments it contains minus the fraction of left-oriented segments it contains, then transistors having equal chirality will not experience orientation-dependent mismatches.[24] For example, a transistor having three right-oriented segments and one left-oriented segment has a chirality of 3/4 – 1/4 = 1/2. Similarly, a transistor having nine right-oriented and three left-oriented segments has a chirality of 9/12 – 3/12 = 1/2. Since these transistors have equal chirality, they do not exhibit any orientation-dependent mismatch. Most designers prefer to use transistors having chiralities of zero; in other words, transistors that consist of equal numbers of left- and right-oriented segments.

Orientation-dependent mismatches can develop in MOS transistors due to diagonal shifts in the source/drain implants. Such diagonal shifts occur when ion implantation is performed at an angle to prevent channeling.[25] Such *tilted implants* cause the source/drain regions on the left side of the gates to differ from the source/drain regions on the right side (Figure 12.30). If the matched devices are arranged in a pattern such as ${}_DA_SB_D$, then the drain of the left-hand device differs from the drain of the right-hand device. Similarly, the source of the left-hand device differs from the source of the right-hand device. Tilted implants have little effect upon the matching of transistors operated in the linear region, but saturated devices sometimes experience small transconductance differences. These mismatches become worse as the voltage drop across the device approaches maximum because tilted implants have an especially strong impact on hot-carrier generation.[26] These orientation dependencies cancel as long as the matched transistors have equal chirality.

[24] The term *chirality* refers to the asymmetry, or handedness, of an object. The term is most commonly encountered in stereochemistry.

[25] J. F. Gibbons, "Ion Implantation in Semiconductors—Part I: Range Distribution Theory and Experiment," *Proc. IEEE*, Vol. 56, #3, 1968, pp. 296–319.

[26] F. K. Baker and J. R. Pfiester, "The Influence of Tilted Source-Drain Implants on High-Field Effects in Submicrometer MOSFETs," *IEEE Trans. on Electron Devices*, Vol. 35, #12, 1988, pp. 2119–2124.

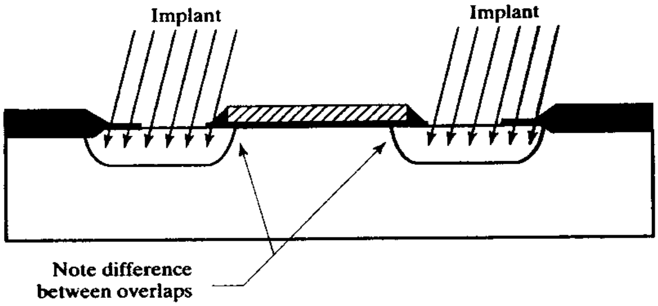

FIGURE 12.30 Diagonal shift in the source/drain regions of an implanted transistor due to the use of a tilted implant. The angle of implantation has been exaggerated for clarity.

Some newer ion implantation systems support on-axis implantation to help minimize the problems caused by tilted implants. The addition of an oxide layer helps minimize channeling, but some still occurs. BiCMOS processes sometimes use *tilted wafers* that have been cut off-axis to minimize pattern distortion. The tilted silicon lattice channels a portion of the ion beam, and this diagonal channeling may cause slight device asymmetries despite the use of on-axis implants.

The interdigitation patterns for common-centroid MOS transistor arrays are often difficult to construct, as it is not easy to satisfy all of the rules of common-centroid layout. MOS transistors must obey not only all four rules given in Section 7.2.6 but also a fifth rule—that of *orientation*. This additional rule ensures that tilted implants (and other device asymmetries) do not affect matching. The full set of rules for MOS devices are as follows:

1. **Coincidence:** The centroids of the matched devices should at least approximately coincide. Ideally, the centroids should exactly coincide.
2. **Symmetry:** The array should be symmetric around both the X- and Y-axes. Ideally, this symmetry should arise from the placement of segments in the array and not from the symmetry of the individual segments themselves.
3. **Dispersion:** The array should exhibit the highest possible degree of dispersion; in other words, the segments of each device should be distributed throughout the array as uniformly as possible.
4. **Compactness:** The array should be as compact as possible. Ideally, it should be nearly square.
5. **Orientation:** Each matched device should consist of an equal number of segments oriented in either direction; more generally, the matched devices should possess equal chirality.

Table 12.1 shows a few of the simpler interdigitation patterns used for MOS transistors. Source and drain fingers are denoted by subscripts, and sequences of segments that may be repeated are enclosed in parentheses: $({}_S A_D A)$. When a pattern includes more than one repeated sequence, each portion of the sequence in

1. $({}_S A_D A)({}_S B_D B_S B_D B)({}_S A_D A)_S$
2. $({}_D A_S B_D \text{-} {}_D B_S A_D)\text{-}({}_D A_S B_D \text{-} {}_D B_S A_D)$
3. $({}_D A_S B_D B_S A)_D$
4. $({}_S A_D A_S B_D B)_S(B_D B_S A_D A_S)$
5. $({}_S A_D A_S B_D B_S A_D A)_S$
6. $({}_S A_D A_S B_D \text{-} {}_S A_D A_S \text{-} {}_D B_S A_D A)_S$
7. $({}_S A_D A_S B_D B_S C_D C)_S(C_D C_S B_D B_S A_D A_S)$

TABLE 12.1 Sample interdigitation patterns for MOS transistor arrays.

parentheses must be replicated the same number of times. Certain patterns contain locations where the source/drain fingers cannot merge with one another; these are denoted by dashes. All of the entries in this table obey the rules of coincidence, symmetry, and orientation, but many of them are not as disperse nor as compact as possible. For example, consider patterns 1 to 4, all of which provide a 1:1 ratio between two matched devices. Pattern 1 lacks dispersion because it contains long runs of segments belonging to the same device. Pattern 2 contains gaps that make it less compact than the others. Patterns 3 and 4 both exhibit considerable dispersion because the segments appear in pairs throughout most parts of the array. However, the middle of pattern 4 contains a run of four segments belonging to the same device. The middle of pattern 3 contains a run of only two segments, so it provides better dispersion than pattern 4. In summary, pattern 3 should exhibit more precise matching than patterns 1, 2, and 4. The device of Figure 12.29 uses pattern 3.

Interdigitated MOS transistors do not provide the best possible cancellation of gradients because they rely on the symmetry of individual device segments to provide one of their two axes of symmetry. A two-dimensional common-centroid array provides a higher degree of symmetry because both axes of symmetry arise from the layout of the array, rather than from the segments comprising it. Two-dimensional common-centroid arrays are particularly useful for matching pairs of transistors of equal size, such as differential pairs. Layouts of this sort are called *cross-coupled pairs*. As with other common-centroid MOS layouts, care must be taken to ensure that orientation dependencies cancel.

Figure 12.31 shows the simplest possible cross-coupled pair. This layout follows the interdigitation pattern ${}_{D}A_{S}B_{D}/{}_{D}B_{S}A_{D}$, where the slash (/) separates the segments that occupy the upper two quadrants from those that occupy the lower two.[27] Not only does this produce a very compact layout, but it also satisfies the rule of orientation, because the two segments belonging to each matched device are oriented in opposite directions. This layout is especially suited for pairs of relatively small MOS transistors.

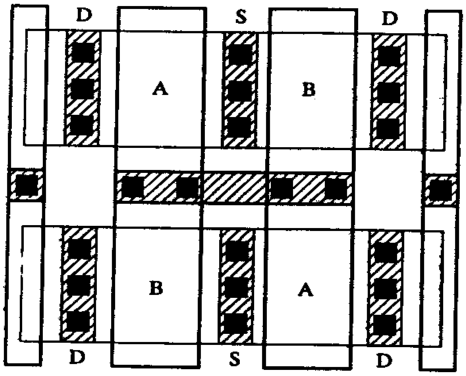

FIGURE 12.31 Cross-coupled MOS transistors.

Larger cross-coupled pairs are more difficult to construct. Most designers simply divide each transistor into two equal halves and place these halves in diametrically opposite corners of the array. A layout of this sort can be represented by the pattern XY/YX, where X and Y are subarrays composed entirely of segments of transistors A and B, respectively. A typical implementation of such an array is $({}_{S}A_{D}A)_{S}$

[27] Tsividis discusses a ${}_{D}A_{S}B_{D}/{}_{D}B_{S}A_{D}$ array, but without the dummies: Y. Tsividis, *Mixed Analog-Digital VLSI Devices and Technology* (New York: McGraw-Hill, 1997), p. 233.

$(B_DB_S)/({}_SB_DB)_S(A_DA_S)$. While this pattern satisfies most of the rules of interdigitation, it does not provide optimal dispersion. As the array grows larger, its lack of dispersion renders it increasingly susceptible to mismatches caused by nonlinear components of the variation. A much better pattern for large cross-coupled pairs is $({}_DA_SB_DB_SA)_D/{}_D(B_SA_DA_SB_D)$. If the array becomes very large, then additional dispersion can be introduced by elaborating this array in the vertical dimension, as shown in the following examples:

$$\begin{array}{lll}
{}_DA_SB_DB_SA_D & {}_DA_SB_DB_SA_D & {}_DA_SB_DB_SA_DA_SB_DB_SA_D \\
{}_DB_SA_DA_SB_D & {}_DB_SA_DA_SB_D & {}_DB_SA_DA_SB_DB_SA_DA_SB_D \\
 & {}_DA_SB_DB_SA_D & {}_DA_SB_DB_SA_DA_SB_DB_SA_D \\
 & {}_DB_SA_DA_SB_D & {}_DB_SA_DA_SB_DB_SA_DA_SB_D \\
 & & {}_DA_SB_DB_SA_DA_SB_DB_SA_D \\
 & & {}_DB_SA_DA_SB_DB_SA_DA_SB_D
\end{array}$$

The main drawback to the more elaborate patterns of this type lies in the difficulty of connecting the various segments together to form the full device. This becomes particularly difficult in cases where the gates of the two matched devices do not connect together. The simpler—and hence easier to connect—patterns generally serve for all except the most demanding applications.

12.5 RULES FOR MOS TRANSISTOR MATCHING

This section summarizes the previously given information in the form of a set of qualitative rules. These rules allow designers to construct matched MOS transistors, even if no quantitative matching data exists for the process in question. The rules use the terms *minimal, moderate,* and *precise* to denote increasingly precise degrees of matching, which may be interpreted as follows:

- **Minimal matching:** Typical three-sigma drain current mismatches of several percent. Minimal matching is often used for constructing bias current networks that do not require any particular degree of precision. This level of matching corresponds to typical offsets in excess of ±10mV and is therefore inadequate for voltage matching applications.
- **Moderate matching:** Typical three-sigma offset voltages of ±5mV or drain current mismatches of less than ±1%. Useful for constructing input stages of noncritical op-amps and comparators, where untrimmed offsets of ±10mV can be maintained.
- **Precise matching:** Typical three-sigma offset voltages of less than ±1mV or drain current mismatches of less than ±0.1%. This level of matching usually involves trimming, and the resulting circuit will probably meet specification within only a limited range of temperatures due to the presence of uncompensated temperature variations.

The following rules summarize the most important principles of MOS transistor matching:

1. Use identical finger geometries.
   Transistors of different widths and lengths match very poorly. Even minimally matched devices must have identical channel lengths. Most matched transistors require relatively large widths and are usually divided into sections, or *fingers*. Each of these fingers should have the same width and length as all others. Do not

attempt to match transistors of different widths and lengths, because the width and length correction factors, δW and δL, vary substantially from lot-to-lot.

2. Use large active areas.
   The active area of an MOS transistor equals the product of its channel width and length. Assuming that all other matching considerations have been addressed, the residual offset due to random fluctuations scales inversely with the square root of device area. Moderate matching usually requires active areas of several hundred square microns, while precise matching requires thousands of square microns.
3. For voltage matching, keep V_{gst} small.
   The offset voltage of a pair of matched MOS transistors contains a term dependent on device transconductance. This term scales with V_{gst}, so smaller values of V_{gst} provide better voltage matching. Reducing the V_{gst} below 0.1V provides little additional benefit because threshold voltage variations begin to dominate the offset equation. Most designers decrease V_{gst} by using larger W/L ratios because these simultaneously increase the active area of the transistors.
4. For current matching, keep V_{gst} large.
   The current mismatch equation contains a term dependent upon threshold voltage. This term scales inversely with V_{gst}, so large values of V_{gst} minimize its impact upon current matching. Circuits relying upon current matching should maintain a nominal V_{gst} of at least 0.3V. Moderately matched transistors should maintain a nominal V_{gst} of at least 0.5V whenever headroom allows. Precisely matched transistors should use the largest value of V_{gst} allowed by the configuration of the circuit, but in any event should equal at least 0.5V.
5. Orient transistors in the same direction.
   Transistors that do not lie parallel to one another become vulnerable to stress- and tilt-induced mobility variations that can cause several percent variation in their transconductance. This effect is so severe that even minimally matched transistors should lie parallel to one another. Matched transistors, especially those that are not fully self-aligned, should have equal chirality. This condition can be met by ensuring that each transistor contains an equal number of segments oriented in each direction.
6. Place transistors in close proximity.
   MOS transistors are vulnerable to gradients in temperature, stress, and oxide thickness. Even minimally matched devices should reside as close as possible to one another. Moderately or precisely matched transistors should be kept next to one another to facilitate common-centroid layout.
7. Keep the layout of the matched transistors as compact as possible.
   MOS transistors naturally lend themselves to long, spindly layouts that are extremely vulnerable to gradients. Common-centroid layout cannot entirely eliminate this vulnerability, so the designer should strive to create as compact an arrangement of matched devices as possible. This usually requires that each device be divided into a number of fingers.
8. Where practical, use common-centroid layouts.
   Moderately and precisely matched MOS transistors require some form of common-centroid layout. This can often be achieved by dividing each transistor into an even number of fingers and by then arranging these fingers in an

interdigitated array. Pairs of matched transistors should be laid out as cross-coupled pairs to take advantage of the superior symmetry of this arrangement.

9. Place dummy segments on the ends of arrayed transistors.
Arrayed transistors should include dummy gates at either end. These dummies need not have the same length as the actual gates, but the spacing between the dummy gates and the actual gates must equal the spacing between actual gates. The moat diffusion should extend at least several microns into the dummies to prevent the edge of the dummies from resting on the bird's beak. The dummy gates should be connected, preferably to a potential that prevents channel formation beneath them. This is most easily achieved by connecting the dummies to the backgate potential.

10. Place transistors in areas of low stress gradients.
The stress gradients reach a broad minimum in the center of the die. Any location ranging from the center of the die out halfway to the edges will fall within this broad minimum. Whenever possible, precisely matched transistors should reside within this low-stress area. Moderately and precisely matched transistors should reside at least 10mils (250μm) away from any side of the die. The stress distribution reaches a maximum in the die corners, so avoid placing any matched transistors near corners. PMOS transistors may experience slightly less stress dependence when oriented along [100] directions. This effect is not sufficiently pronounced to justify placing minimally or moderately matched transistors diagonally, but precisely matched transistors might benefit from this unconventional orientation. NMOS transistors should always be oriented horizontally and vertically.

11. Place transistors well away from power devices.
For purposes of discussion, any device dissipating more than 50mW should be considered a power device, and any device dissipating more than 250mW should be considered a major power device. Precisely matched transistors should reside on an axis of symmetry of the major power devices using an optimal symmetry arrangement (see Section 7.2.7). Moderately and precisely matched transistors should reside no less than 10 to 20mils (250 to 500μm) away from the closest power device. Minimally matched devices may be placed next to power devices, but only if they use some form of common-centroid layout.

12. Do not place contacts on top of active gate area.
Whenever possible, extend the gate poly beyond the moat and place the gate contacts over thick-field oxide. When this is not possible, minimize the number and size of the gate contacts and place them in the same location on each transistor. Consider placing the gate contacts of high-voltage annular transistors over the field-relief structure because this is not part of the active gate.

13. Do not route metal across the active gate region.
Whenever possible, avoid routing metal across the active gate region of precisely matched MOS transistors. Leads may route across moderately matched MOS transistors, but additional dummy leads should be added so that every section of the array of matched devices is crossed at the same location along its channel by an identical length of lead.

14. Keep all junctions of deep diffusions far away from active gate area.
The minimum spacing between a drawn well boundary and a precisely matched MOS transistor should equal at least twice the well junction depth.

Moderately and minimally matched transistors need only obey the applicable layout rules. Similar considerations apply to deep-N+ sinkers and other deep diffusions.

15. Place precisely matched transistors on axes of symmetry of the die.
Arrays of precisely matched transistors should be placed so that the axis of symmetry of the array aligns with one of the two axes of symmetry of the die. If the design contains large numbers of matched transistors, then reserve the optimal locations for the most critical devices.

16. Do not allow the NBL shadow to intersect the active gate area.
The NBL shadow should not fall across the active gate region of any precisely matched transistor. If the direction of the NBL shift is unknown, allow adequate overlap of NBL over the transistor on all sides. If the magnitude of the NBL shift is also unknown, then overlap NBL over the active gate region by at least 150% of the maximum epi thickness.

17. Connect gate fingers using metal straps.
Connect the gate fingers of moderately and precisely matched transistors using metal rather than poly. Minimally matched transistors can use a poly comb structure to simplify the connection of the gate electrodes.

18. Use thin-oxide devices in preference to thick-oxide devices.
Some processes offer multiple thicknesses of gate oxides. The transistors with thinner gate oxides generally exhibit better matching characteristics than those with thick gate oxides. Whenever circuit considerations allow, consider using thin-oxide transistors in preference to thick-oxide transistors.

19. Consider using NMOS transistors rather than PMOS transistors.
NMOS transistors generally match better than PMOS transistors. Whenever circuit considerations allow, consider using NMOS transistors rather than PMOS transistors.

12.6 SUMMARY

Many circuit designers think of MOS transistors primarily as building blocks for digital logic. While they are indispensable in this role, they also have many other important applications. Modern mixed-mode integrated circuits rely heavily upon MOS transistors for power switching and low-current analog functions.

MOS power transistors have revolutionized power switching. The new generation of high-efficiency switch-mode power supplies relies almost exclusively upon MOS power transistors. Similarly, virtually all low-voltage power distribution circuits use MOS transistors. Now that BiCMOS processes offer integrated power MOS transistors with specific on-resistances approaching those of discrete devices, it has become economically feasible to integrate many power switches into a single integrated circuit. The low on-resistances of the power switches also minimize power dissipation, allowing the use of compact surface-mount packages.

MOS transistors have also found many applications in analog signal processing circuitry. Although bipolar transistors remain entrenched in a few applications, the vast majority of analog circuitry can be implemented using MOS transistors. MOS circuits are usually smaller and consume less power than their bipolar counterparts. Even though most modern analog circuits are implemented in BiCMOS processes, the vast majority of the circuitry consists of MOS transistors. Circuit designers continue to develop new applications for MOS transistors, so future integrated circuits will probably incorporate an even larger percentage of MOS circuitry.

12.7 EXERCISES

Refer to Appendix C for layout rules and process specifications.

12.1. Suppose an extended-voltage transistor with a drain depletion width x_d equal to 10% of the pinched-off region width x_p can withstand a drain-to-source voltage of 10V. What drain-to-source voltage could a similar device withstand if x_d were increased to 50% of x_p?

12.2. Suggest a structure for a self-aligned extended-voltage PMOS transistor. Draw a cross section of a representative transistor using this structure.

12.3. If the thin gate oxide of an extended-drain NMOS having no field-relief structure can withstand 10V, then which of the following biasing conditions are allowable, and why?

a. Asymmetric NMOS, $V_{GS} = 6V, V_{DS} = 10V.$

b. Asymmetric NMOS, $V_{GS} = 7V, V_{DS} = 16V.$

c. Asymmetric NMOS, $V_{GS} = 3V, V_{DS} = 16V.$

d. Symmetric NMOS, $V_{GS} = -13V, V_{DS} = -16V.$

e. Symmetric NMOS, $V_{GS} = 20V, V_{DS} = 0V.$

12.4. Lay out asymmetric and symmetric extended-drain NMOS transistors, each having drawn dimensions of 2(15/10). The N-well drain geometry should abut the N-moat source geometry beneath the gate as shown in Figure 12.5. The overlap of the poly gate over the N-well should equal exactly 3μm. Include abutting backgate contacts for the asymmetric transistor. Why can't abutting backgate contacts be used for the symmetric transistor?

12.5. Compute the maximum theoretical $R_{DS(on)}$ for a 50000/2 NMOS power transistor operating at $5 < V_{GS} < 15V$ and $-40 < T_j < 150°C$. Assume that the device's threshold voltage equals 0.7 ± 0.2V with a temperature coefficient of −2mV/°C, and that its process transconductance equals $35uA/V^2$ ± 20% at 150°C.

12.6. Determine the specific on-resistance (in $\Omega \cdot mm^2$) for a power device having an $R_{DS(on)}$ of 165mΩ and an area of $2.26mm^2$. Use this information to determine the area required for a 100mΩ power transistor. Assume both $R_{DS(on)}$ values do not include bondwire or leadframe resistance.

12.7. Compute the ideal ratio *B/L* for a metal system consisting of a first layer of metal that is 7500Å thick and a second layer of metal that is 14000Å thick. Lay out a 20,000/2 PMOS transistor using this ratio and the analog BiCMOS layout rules in Appendix C. Divide the transistor into sufficient fingers to produce a roughly square aspect ratio. Fill the well with NBL and ring the outer edge of the well with deep-N+ to provide a backgate contact for the transistor. Include all necessary metallization.

12.8. Compute the approximate metallization resistance of the transistor in Exercise 12.7. Do not include the resistance of the metal-2 buses extending beyond the interdigitated region of the transistor.

12.9. Assume the source and drain leads of the transistor in Exercise 12.7 run 25μm from the drawn edge of the well to the edge of their respective bondpads, and that each bondpad connects to a 600μm-long 1mil-diameter gold bondwire. Calculate the total metallization resistance of the transistor, including bondwires. Assume that the resistance between the edge of the bondpad and the bondwire is negligible.

12.10. Lay out a minimum-size, circular-annular, lateral DMOS transistor using the analog BiCMOS rules in Appendix C, supplemented by the following rules for the DMOS layer:

1. **DMOS** width	5μm
2. **DMOS** spacing to **DMOS**	4μm
3. **DMOS** spacing to **PMOAT**	0μm
4. **POLY** extends into **DMOS**	2μm
5. **POLY** overhang of **DMOS**	4μm
6. **MOAT** overlap of **DMOS**	2μm
7. **CONT** extends into **DMOS**	2μm

A DMOS to PMOAT spacing of 0μm implies that the outer edge of the PSD plug should coincide with the inner edge of the annular DMOS ring. Include all necessary metallization.

12.11. If the length of the DMOS channel equals 1μm and the inner edge of the channel coincides with the outer edge of the drawn DMOS geometry, then what is the drawn width of the transistor constructed in Exercise 12.9?

12.12. Lay out a standard bipolar epi-FET having drawn dimensions of 30/8. Assume that the base pinch plate must extend at least 2μm into the isolation.

12.13. Construct a minimum-size circularly symmetric epi-FET. Include all necessary metallization. What are the drawn width and length of this device?

12.14. A cross-coupled NMOS differential pair of transistors, each having dimensions 100/10, has a three-sigma random mismatch of ±2.85mV. Estimate the three-sigma random mismatch of a similar differential pair where the transistors each have dimensions of 1000/5.

12.15. Lay out a pair of differential NMOS transistors, each having dimensions of 1000/5, to obtain the best possible matching. The transistors may be divided into as many or as few segments as desired. Assume that backgate contacts are required only along the edges of the array. Include all necessary metallization, including the links connecting individual source/drain fingers and the links connecting the gate fingers.

12.16. Lay out the MOS operational amplifier shown in Figure 12.32 following the recommendations for optimal matching. Use the poly-gate CMOS rules in Appendix C, and include all necessary backgate and substrate contacts. Assume that all PMOS transistors have backgates connecting to VDD.

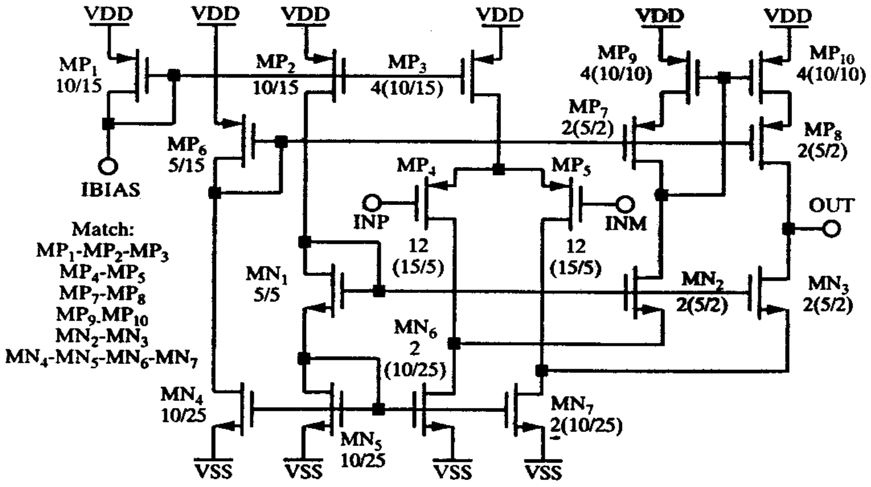

FIGURE 12.32 Folded-cascode MOS operational amplifier for Exercise 12.16.

12.17. Compare the matching of the following interdigitation patterns. Which pattern provides the best matching, and why?

a. ${}_SA_DA_SB_D-{}_DA_SA_D-{}_DB_SA_DA_S$

b. ${}_DA_SA_DA_SB_DB_SA_DA_SA_D$

c. ${}_DB_SA_DA_SA_DA_SA_DA_SB_D$

12.18. What are the chiralities of each of the following interdigitation patterns? Which patterns exhibit orientation-dependent mismatches?

a. ${}_DA_SB_DB_SA_D$

b. ${}_SA_D-{}_DB_SB_DB_S-{}_DA_S$

c. ${}_SA_DA_S-{}_SB_D-{}_DA_SA_D$