# 3 Representative Processes

Semiconductor processing has evolved rapidly over the past fifty years. The earliest processes produced only discrete components, mainly switching diodes and bipolar transistors. The first practical integrated circuits appeared in 1960.[1] These consisted of a few dozen bipolar transistors and diffused resistors connected to form simple logic gates. By modern standards, these early integrated circuits were terribly slow and inefficient. Refinements were soon made, and by the mid-1960s bipolar integrated logic offered clear advantages over discrete logic. The first analog integrated circuits appeared at about the same time; these consisted of matched transistor arrays, operational amplifiers, and voltage references. The standard bipolar process that was created to support these products remains in use today.

Integrated bipolar logic was fast but power-hungry. MOS integrated circuits held out the promise of a low-power alternative, but the metal-gate MOS processes of the 1960s suffered from unpredictable threshold voltage shifts. This problem was eventually conquered through the development of polysilicon-gate MOS processes in the early 1970s. MOS logic soon replaced bipolar logic and created vast new markets for microprocessors and dynamic RAM chips. Analog CMOS circuits of this era touted greatly reduced operating currents but provided only mediocre performance, so standard bipolar remained the process of choice for high-performance analog integrated circuits.

By the mid-1980s, customers were demanding the integration of both digital and analog functions onto a single mixed-signal integrated circuit. A new generation of merged bipolar-CMOS (BiCMOS) processes were soon developed specifically for mixed-signal design. Although these processes are complex and costly, they offer a level of performance unachievable by other means. The world of analog integrated circuits is dominated by these three processes: standard bipolar, polysilicon-gate CMOS, and analog BiCMOS. This chapter will analyze the implementation of a representative process of each type.

[1] J. S. Kilby, "Invention of the Integrated Circuit," *IEEE Trans. on Electron Devices*, Vol. ED-23, #7, 1976, pp. 648–654.

## 3.1 STANDARD BIPOLAR

Standard bipolar was the first analog integrated circuit process. Over the years, it has produced many classic devices, including the 741 operational amplifier, the 555 timer, and the 431 voltage reference. Even though these parts represent thirty-year-old technology, they are still manufactured in vast quantities today.

Standard bipolar is seldom used for new designs. CMOS offers lower supply currents, BiCMOS provides superior analog performance, and various advanced bipolar processes yield faster switching speeds. But the knowledge gained through first developing and then refining standard bipolar will never become obsolete. The same devices reappear in every new process, along with many of the same parasitic mechanisms, design tradeoffs, and layout principles. This chapter will therefore begin with an overview of standard bipolar.

### 3.1.1. Essential Features

Standard bipolar was shaped by a conscious decision to optimize the NPN transistor at the expense of the PNP. This decision rested on the observation that the NPN transistor employs electron conduction while the PNP transistor relies on hole conduction. The lower mobility of holes reduces both the beta and the switching speed of PNP transistors. Given equivalent geometries and doping profiles, an NPN will outperform a PNP by more than 2:1. Several additional processing steps are required to optimize both types of transistors simultaneously, so early processes optimized NPN transistors and avoided PNP transistors altogether. This decision met the requirements of bipolar logic, consisting as it does of NPN transistors, resistors, and diodes. When analog circuits were first constructed using standard bipolar, several types of PNP transistors were cobbled together from existing process steps. Although these transistors performed relatively poorly, they sufficed to design many useful circuits.

The standard bipolar process employs *junction isolation* (JI) to prevent unwanted currents from flowing between devices that are formed on the same substrate.[2] The components reside in a lightly doped N-type epitaxial layer deposited on top of a lightly doped P-type substrate (Figure 3.1). A deep-P+ *isolation diffusion* driven down to contact the underlying substrate provides isolation between components. Regions of N-epi separated from one another by isolation are called *tanks* or *tubs*. If the isolation is biased to a potential equal to or below the lowest-voltage tank, then reverse-biased junctions surround every tank. The substrate forms the floor of these tanks, and isolation diffusions form their sidewalls.

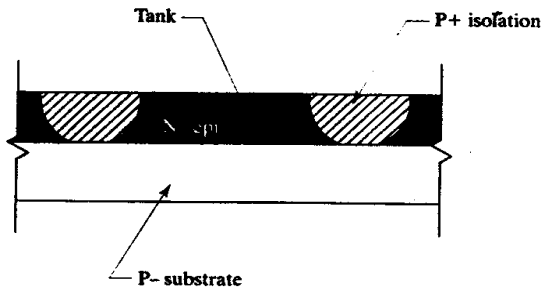


**FIGURE 3.1** Cross section of the junction isolation system employed for standard bipolar.

[2] R. N. Noyce, U.S. Patent #2,981,877, 1961.

Junction isolation has several significant drawbacks. The reverse-biased isolation junctions exhibit enough capacitance to slow the operation of many circuits. High temperatures can cause significant leakage currents, as can exposure to light or ionizing radiation. Unusual operating conditions can also forward bias the isolation junctions and inject minority carriers into the substrate. Despite these difficulties, junction-isolated processes can successfully fabricate most circuits. Junction isolation is also considerably cheaper than any of its alternatives.

## 3.1.2. Fabrication Sequence

The baseline standard bipolar fabrication sequence consists of eight masking operations. The significance of each step can be illustrated best by presenting the entire flow from starting material to finished wafer. Representative cross sections will be used to illustrate each step. When examining these cross sections (and all others in this text), keep in mind that the vertical scale of the drawings has been exaggerated by a factor of two to five for clarity. The lateral dimensions of a typical integrated device are so much greater than its vertical dimensions that a true-scale diagram would be virtually illegible. The cross sections therefore exaggerate the vertical scale by a factor of two to five. The substrate is also much thicker than depicted; the additional silicon serves to strengthen the wafer against warping and breakage.

### *Starting Material*

Standard bipolar integrated circuits are fabricated on a lightly doped (111)-oriented P-type substrate. The wafers are usually cut several degrees off-axis to minimize distortion of the NBL shadow (*pattern distortion*).[3] The use of (111) silicon helps suppress a parasitic PMOS transistor that is inherent in the standard bipolar process. The N-epi forms the backgate of this parasitic, while a lead crossing the field oxide above the tank acts as its gate electrode. A base region within the tank forms the source, and the drain consists of the P+ isolation (Figure 3.2). When the base diffusion is biased to a high voltage relative to the metal lead, a channel forms and allows current to flow from the base to the isolation. The threshold voltage of an MOS transistor formed under thick field oxide is called a *thick-field threshold*. The use of (111) silicon artificially elevates the PMOS thick-field threshold by introducing positive surface-state charges along the oxide-silicon interface.

### *N-Buried Layer*

The first processing step consists of growing a thin layer of oxide across the wafer. Photoresist spun onto this oxide is patterned using the N-buried layer (NBL) mask.

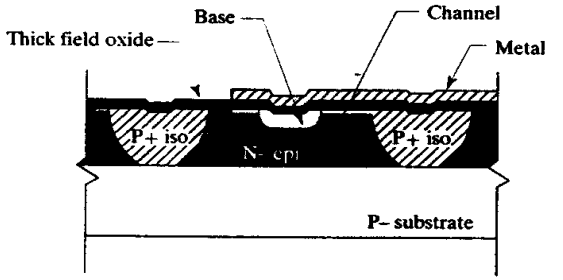


**FIGURE 3.2** Parasitic PMOS formation in standard bipolar.

[3] W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology* (Reading, MA: Addison-Wesley, 1994), p. 331.

After an oxide etch opens windows to the silicon surface, ion implantation or thermal deposition transfers an N-type dopant into the wafer. The N-buried layer customarily consists of either arsenic or antimony because the low diffusivities of these elements limit up-diffusion during subsequent processing. The brief drive following the deposition serves two purposes: first, it anneals out lattice damage; and second, it grows a small amount of oxide that forms a slight discontinuity at the silicon surface (Figure 3.3). This discontinuity will later produce an NBL shadow to which other masks can align.

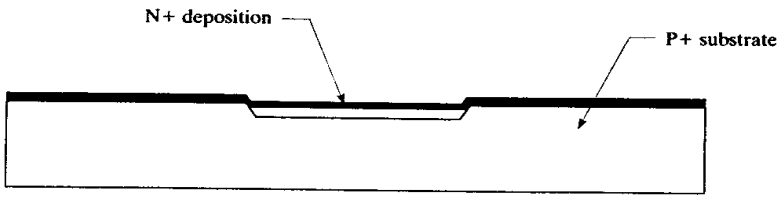


**FIGURE 3.3** Wafer after anneal of NBL implant.

*Epitaxial Growth*

The oxide layer that remains on the wafer is stripped prior to the growth of some 10 to 25μm of lightly doped N-type epi. Surface discontinuities propagate upward during epitaxial deposition at approximately a 45° angle along the axis. Upon completion of epitaxial growth, the NBL shadow will have shifted laterally a distance approximately equal to the thickness of the epi (Figure 3.4).

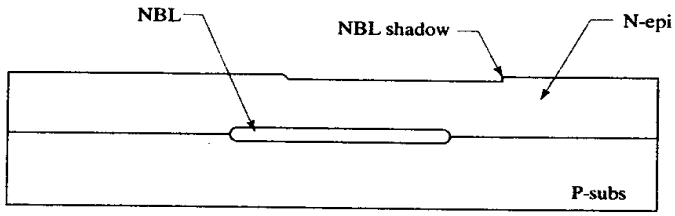


**FIGURE 3.4** Wafer after epitaxial deposition. Note the pattern shift exhibited by the NBL shadow.

*Isolation Diffusion*

The wafer is again oxidized, coated with photoresist, and patterned using the *isolation* mask. This mask must be aligned to the NBL shadow using a deliberate offset to correct for pattern shift. A heavy boron deposition followed by a high-temperature drive forces the isolation diffusion partway down through the epi layer. Oxidation also occurs during this drive, covering the isolation windows with a thin layer of thermal oxide. The drive stops before the isolation junction reaches the substrate, since later high-temperature processing steps (primarily the deep-N+ drive) will force the diffusion the rest of the way down. Figure 3.5 shows the wafer after this partial drive.

*Deep-N+*

A deep-N+ diffusion (sometimes called a *sinker*) allows low-resistance connection to the NBL. First a photoresist is applied and patterned using the deep-N+ mask. A heavy phosphorus deposition followed by a high-temperature drive forms the deep-N+ sinkers. The drive not only causes the deep-N+ to diffuse down to meet the upward-diffusing NBL but also completes the isolation drive. Sufficient time is

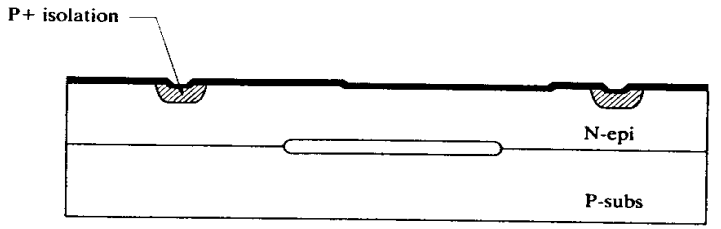


**FIGURE 3.5** Wafer after isolation deposition and partial drive.

allowed to overdrive the junctions by about 25%. Without this overdrive, the bottom of the isolation and deep-N+ diffusions would be very lightly doped. The overdrive simultaneously reduces the vertical resistance through both the isolation and the deep-N+ sinkers. The deep-N+ drive also forms the thick field oxide.

Both deep-N+ and isolation diffusions approach their final junction depths during the deep-N+ drive (Figure 3.6). These junctions will diffuse slightly deeper during subsequent processing, but all of the later diffusions are fairly shallow compared to deep-N+ and isolation, and therefore the tank regions appear fully formed in Figure 3.6. The NBL regions are normally spaced some distance inside the isolation diffusion to increase the tank breakdown voltage. Otherwise the N+/P+ junction formed by the intersection of NBL and isolation would avalanche at about 30V.

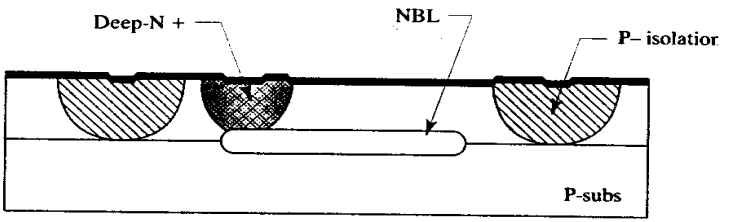


**FIGURE 3.6** Wafer after isolation drive.

### *Base Implant*

Next, photoresist spun onto the wafer is patterned using the base mask. An oxide etch clears windows through the field oxide to the silicon surface. A light boron implant conducted through these openings counterdopes the N-epi to form the base regions of the NPN transistors. Ion implantation allows precise control of base doping and thus minimizes process-derived beta variation. The subsequent drive anneals implant damage and sets the base junction depth. Oxide grown during this drive serves as a mask for the subsequent emitter deposition. Base is also implanted across the isolation regions to increase surface doping. This practice, called *base-over-isolation* (BOI), substantially increases the NMOS thick-field threshold without requiring the use of a separate channel stop. Figure 3.7 shows a cross section of the wafer following the base drive.

### *Emitter Diffusion*

The wafer is again coated with photoresist and patterned using the emitter mask. A subsequent oxide etch exposes the silicon surface in regions where NPN emitters will form and in regions where Ohmic contact must be made to the N-epi or the deep-N+ diffusion. A very concentrated phosphorus deposition forms the emitter.

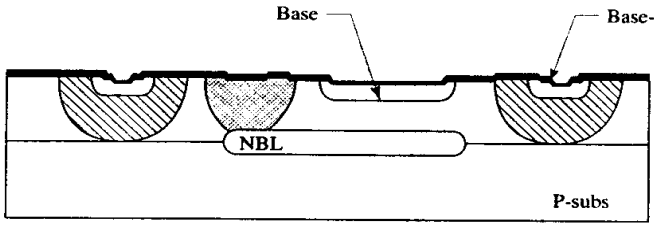


**FIGURE 3.7** Wafer after base drive.

A $POCl_3$ source is often used since precise control of emitter doping is unnecessary. A brief drive sets the final emitter junction depth and thereby determines the width of the active base region of the NPN transistors.

An oxide film grown over the emitter diffusion insulates it from subsequent metallization. Some processes employ dry oxidation for this step, but the short oxidation time results in a *thin emitter oxide* vulnerable to electrostatic discharges (Section 4.1.1). Alternatively, a wet oxidation can grow a *thick emitter oxide* that possesses a higher rupture voltage. Figure 3.8 shows a cross section of the wafer after the emitter drive.

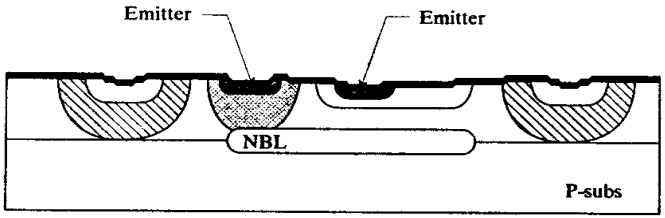


**FIGURE 3.8** Wafer after emitter drive.

Many older processes also incorporate an *emitter pilot* step to provide a means of adjusting NPN beta. A dummy wafer that is inserted into the wafer lot before base implant and removed after emitter deposition is used to conduct an experimental emitter drive. By monitoring the performance of the base-emitter junction formed on the pilot wafer, the actual drive can be adjusted on a lot-by-lot basis to target the desired NPN beta.

### *Contact*

All diffusions are now complete. The remaining steps form the metallization and apply the protective overcoat. The first step in this sequence forms contacts to selected diffusions. The wafer is again coated with photoresist, patterned using the contact mask, and etched to expose bare silicon. This process is sometimes called *contact OR*, in which OR stands for *oxide removal*.

### *Metallization*

A layer of aluminum-copper-silicon alloy is evaporated or sputtered across the entire wafer. This metal system typically incorporates 2% silicon to suppress emitter punchthrough and 0.5% copper to improve electromigration resistance. Standard bipolar employs relatively thick metallization, typically at least 10kÅ (1.0μm) thick, to reduce interconnection resistance and decrease vulnerability to electromigration. The metallized wafer is patterned using the metal mask and etched to form the interconnection system.

### *Protective Overcoat*

Next, a thick layer of *protective overcoat* (PO) is deposited across the entire wafer. Compressive nitride protective overcoats provide excellent mechanical and chemical protection. Some processes use a *phosphosilicate-doped glass* (PSG) layer either beneath a compressive nitride or as a replacement for it. Since the deposition of the protective overcoat occurs at moderate temperatures, it also sinters the aluminum metallization.

Finally, a layer of photoresist is applied and patterned using the PO mask. A special etch opens windows through the protective overcoat to expose areas of metallization for bonding. This composes the final fabrication step; the wafer is now complete. Figure 3.9 shows a fully processed wafer (the illustrated cross section does not include a bondpad opening).

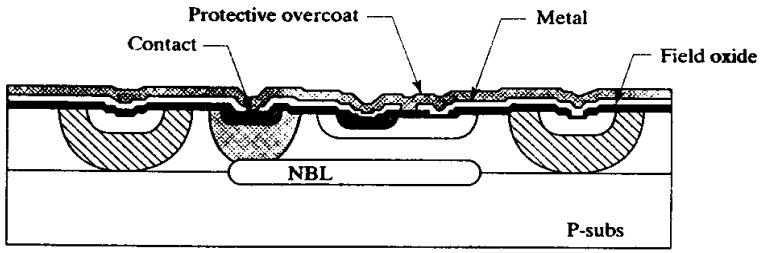


**FIGURE 3.9** Completed standard bipolar wafer.

## 3.1.3. Available Devices

Standard bipolar was originally developed to provide bipolar NPN transistors and diffused resistors. A number of other devices can also be fabricated using the same process steps, including two types of PNP transistors, several types of resistors, and a capacitor.[4] These devices form a basic component set suitable for fabricating a wide variety of analog circuits. Section 3.1.4 will examine several additional devices that require extensions to the baseline process.

*NOTE:* The dimensions of standard bipolar devices are often specified in mils. A *mil* equals 0.001 inches. The relevant conversion factors are 1mil = 25.4μm, and $1\text{mil}^2 \cong 645\mu\text{m}^2$. Another obscure unit of measurement sometimes used to specify junction depths is the *sodium line*, which equals one-half of the wavelength of the sodium spectrum D-line (1 line = 0.295μm).[5]

### *NPN Transistors*

Figure 3.10 shows a representative layout and cross section of a minimum-area NPN transistor. The collector of the NPN consists of an N-epi tank, and the base and emitter are fabricated by successive counterdopings. The carriers flow vertically from emitter to collector through the thin base region underneath the emitter diffusion. The difference between the base and emitter junction depths determines the effective base width. Since these dimensions are controlled entirely by diffusion processing, they are not subject to photolithographic misalignment, allowing a base width substantially smaller than the alignment tolerance. For example, a process with a 5μm minimum feature size can easily fabricate a 2μm base width.

---

[4] For a general overview of the standard bipolar process, see N. Doyle, "LIC Technology," *Microelectronics and Reliability*, Vol. 13, 1974, pp. 315–324.

[5] G. E. Anner, *Planar Processing Primer* (New York: Van Nostrand Reinhold, 1990), pp. 107–108.

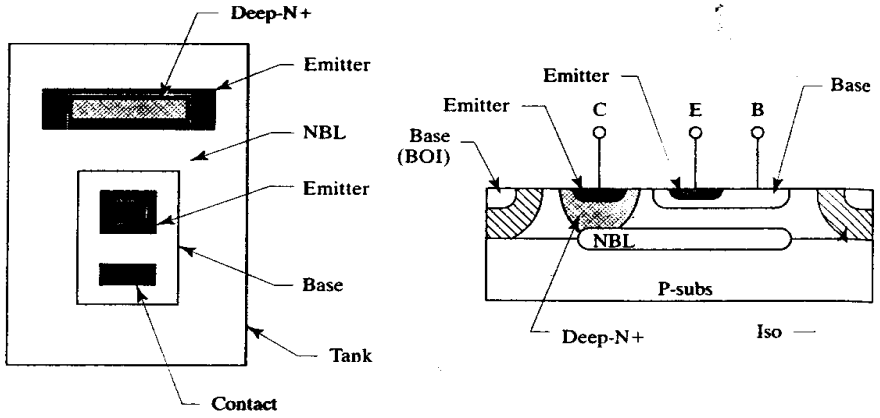


**FIGURE 3.10** Layout and cross section of an NPN transistor with deep-N+ and NBL.[6]

The collector consists of lightly doped N-type epi lying on top of heavily doped NBL. The lightly doped epi allows the formation of a wide collector-base depletion region without excessive intrusion into the neutral base. This enables the transistor to support high operating voltages while simultaneously minimizing the Early effect (Section 8.2). NBL and deep-N+ create a low-resistance pathway to the portion of the epi layer beneath the transistor's active base. By this means, the collector resistance of a minimum NPN can be reduced to less than 100Ω and the collector resistance of a power NPN can be reduced to less than 1Ω.

The high concentration of donors in the NBL effectively halts the downward growth of the collector-base depletion region. The distance between the bottom of the base diffusion and the top of the NBL determines the maximum operating voltage of the NPN transistor. Thicker epi layers allow higher operating voltages, up to a limit set by the breakdown of the base diffusion sidewall (typically 50 to 80V). The maximum operating voltage of a bipolar process is usually specified in terms of the avalanche voltage of the NPN collector to emitter with base open ($V_{CEO}$). Depending on epi thickness and doping, this voltage can range from less than 10V to more than 100V.

The vertical NPN is the best device fabricated in the standard bipolar process. It consumes relatively little area and offers reasonably good performance. Circuit designers try to use as many of these transistors as possible. Table 3.1 lists typical device parameters for a minimum-emitter NPN transistor in a 40V standard bipolar process.

**TABLE 3.1** Typical vertical NPN device parameters.

| Parameter | Nominal Value |
|---|---|
| Drawn emitter area | $100\mu m^2$ |
| Peak current gain (beta) | 150 |
| Early voltage | 120V |
| Collector resistance, in saturation | 100Ω |
| Collector current range for maximum beta | 5μA–2mA |
| $V_{EBO}$ (Emitter-base breakdown, collector open) | 7V |
| $V_{CBO}$ (Collector-base breakdown, emitter open) | 60V |
| $V_{CEO}$ (Collector-emitter breakdown, base open) | 45V |

[6] In many standard bipolar processes, the isolation spacings are much wider than this illustration suggests; see Appendix C.1 for a brief discussion and some typical spacing rules.

The NPN transistor can also act as a diode, the characteristics of which depend upon the terminals chosen to form the anode and cathode. The least series resistance and fastest switching speeds occur when the base and collector form the anode and the emitter forms the cathode. This configuration is sometimes called a *CB-shorted diode,* or a *diode-connected transistor.* Its only serious drawback consists of a low breakdown voltage, equal to the $V_{EBO}$ of the transistor, or about 7V. On the other hand, the relatively low $V_{EBO}$ allows a suitably-connected transistor to serve as a useful Zener diode. The breakdown voltage of this structure varies somewhat due to doping variations and surface effects, so a tolerance of at least ± 0.3V should be allowed.

### *PNP Transistors*

The standard bipolar process cannot fabricate an isolated vertical PNP because it lacks a P-type tank. A non-isolated vertical PNP transistor, called a *substrate PNP,* can be constructed using the substrate as a collector. The collector of this device always connects to the substrate potential of the die, which usually consists of either ground or the negative supply rail. Figure 3.11 shows a representative layout and cross section of this device.

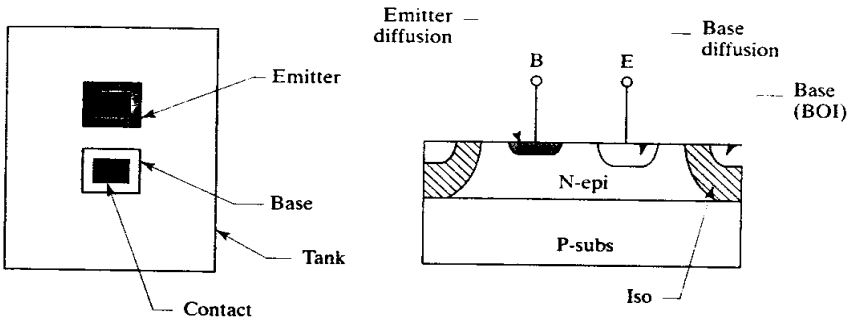


**FIGURE 3.11** Layout and cross section of a substrate PNP transistor. The substrate forms the collector and is contacted through substrate contacts (not shown).

The base of the substrate PNP consists of an N-tank, and the emitter is fabricated from base diffusion. The collector current must exit through the substrate and the isolation. The collector contact does not have to reside next to the substrate PNP since all isolation regions interconnect electrically through the substrate. The resistance of the isolation and substrate are, however, substantial. Substrate contacts placed adjacent to the transistor help extract the collector current and thus minimize voltage drops in the substrate (*substrate debiasing*) that might otherwise impair circuit performance (Section 4.4.1).

The difference between the final epi thickness and the base junction depth determines the base width of a substrate PNP. As in the case of the vertical NPN, the base width is unaffected by photolithographic tolerances. NBL must be left out of the substrate transistor because its presence severely reduces beta. Deep-N+ therefore serves no useful function in a substrate PNP. An emitter diffusion placed beneath the collector contact ensures the surface doping concentration necessary to achieve Ohmic contact, while simultaneously thinning the oxide. The epi thicknesses and doping concentrations of the standard bipolar process are calculated to optimize the vertical NPN transistor, but the substrate PNP performs respectably (Table 3.2). The choice of names for the emitter and base diffusions is somewhat unfortunate, as the *emitter* of a substrate PNP consists of *base* diffusion.

**TABLE 3.2** Typical PNP device parameters.

| Parameter | Lateral PNP | Substrate PNP |
|---|---|---|
| Drawn emitter area | 100μm² | 100μm² |
| Drawn base width | 10μm | N/A |
| Peak current gain (beta) | 50 | 100 |
| Early voltage | 100V | 120V |
| Typical operating current for maximum beta | 5–100μA | 5–200μA |
| $V_{EBO}$ (Emitter-base breakdown, collector open) | 60V | 60V |
| $V_{CBO}$ (Collector-base breakdown, emitter open) | 60V | 60V |
| $V_{CEO}$ (Collector-emitter breakdown, base open) | 45V | 45V |

The lack of an isolated collector limits the versatility of the substrate PNP. Another transistor, called a *lateral PNP*, trades off device performance for isolation. Figure 3.12 shows a representative layout and cross section of a minimum-geometry lateral PNP transistor. Both the collector and the emitter regions of the lateral PNP consist of base diffusions formed into an N-tank. As in the case of the substrate PNP, this tank serves as the base of the transistor. Transistor action in the lateral PNP occurs laterally outward from the central emitter to the surrounding collector. The separation of the two base diffusions sets the base width of the transistor. The emitter and collector of the lateral PNP are said to *self-align* because a single masking operation forms both regions. The base width of the lateral PNP can be precisely controlled because photolithographic misalignment does not occur between self-aligned diffusions. The effective base width of the transistor is considerably less than the drawn base width because of outdiffusion. This consideration limits the drawn base width to a minimum of about twice the base junction

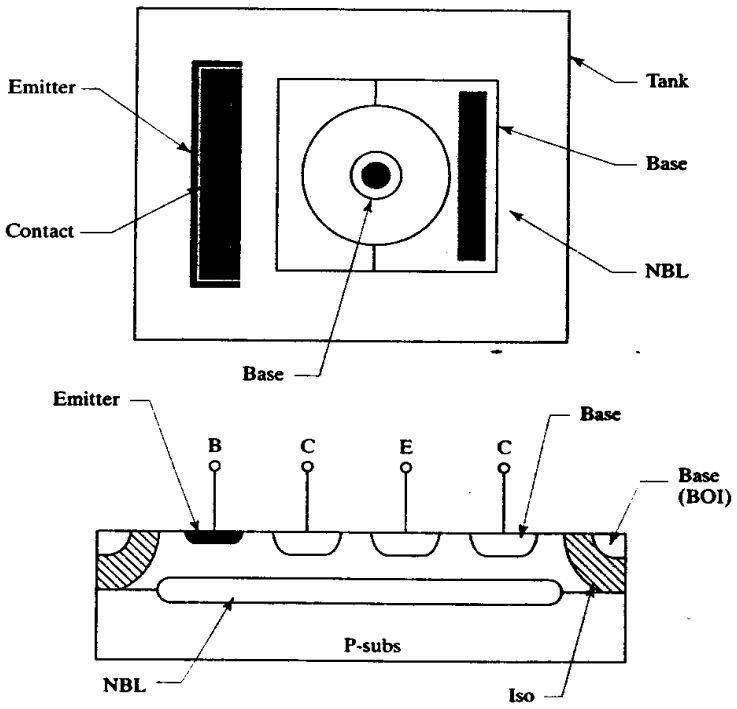


**FIGURE 3.12** Layout and cross section of a lateral PNP transistor. The collector of the transistor appears twice on the cross section because it encircles the emitter.

depth. Narrow-base lateral PNP transistors exhibit low Early voltages and low-voltage punchthrough breakdown, so wider base widths are often employed.

Some percentage of the carriers injected by the lateral PNP's emitter will actually flow to the substrate rather than to the intended collector. This undesired conduction path forms a parasitic substrate PNP transistor. Unless this parasitic is somehow suppressed, most of the current injected by the emitter will find its way to the substrate, and the lateral PNP will exhibit very low apparent beta. For reasons that will be explained in Section 8.2.3, NBL largely blocks substrate injection and therefore boosts the lateral PNP beta.

Lateral PNP transistors have lower effective betas than their Gummel numbers would indicate. A large number of recombination centers reside at the oxide-silicon interface, especially in (111) silicon. The surface recombination rate thus far exceeds that in the bulk. Much of the current flow in the lateral PNP occurs near the surface and is therefore subject to these elevated recombination rates.[7] Certain layout techniques can minimize surface effects and. with care, betas of fifty or more can be obtained. Lateral PNP transistors are also quite slow, due mainly to large parasitic junction capacitances associated with the base terminal.

Neither the lateral nor the substrate PNP transistor forms a true complement to the vertical NPN. Both are useful devices. but each has its drawbacks and limitations. Circuit designers tend to avoid routing active signal paths through PNP devices (especially laterals) because of their poor frequency response, but most analog circuits still contain PNP transistors in supporting roles. Table 3.2 lists typical device parameters for PNP transistors formed on a 40V standard bipolar process.

### *Resistors*

Standard bipolar does not include any diffusions intended specifically for fabricating resistors, but several types of resistors can be made using layers intended for other purposes. Typical examples include base, emitter, and pinch resistors, all three of which employ the relatively shallow base and emitter diffusions to achieve tighter spacings.

Each of the materials used to construct resistors possesses a characteristic *sheet resistance,* defined as the resistance measured across a square of the material contacted on opposite sides. Sheet resistance is customarily given in units of *Ohms per square* (Ω/□). It can be calculated from the thickness of the material and its doping concentration, but in the case of diffusions, nonuniform doping complicates these calculations (Section 5.2). In practice, sheet resistances are best determined by measuring sample resistors with known geometries constructed from the desired materials. Typical values for silicon diffusions range from 5 to 5000Ω/□.

A *base resistor* consists of a strip of base diffusion isolated by an N-tank and connected so that it will reverse-bias the base-epi junction (Figure 3.13). Connecting the tank to the more positive end of the resistor will ensure isolation. Alternatively, the tank can be connected to any point in the circuit biased to a higher voltage than the resistor. If a base resistor should forward-bias into its tank, a parasitic substrate PNP will inject current from the resistor into the substrate. NBL can help suppress this parasitic PNP should the base-epi junction momentarily forward-bias. Deep-N+ need not be added because the tank terminal does not draw significant current. Most standard bipolar processes produce base resistors with sheet resistances of 150 to 250Ω/□.

---

[7] R. S. Muller and T. I. Kamins. *Device Electronics for Integrated Circuits,* 2nd ed. (New York: John Wiley and Sons. 1986). pp. 366–368.

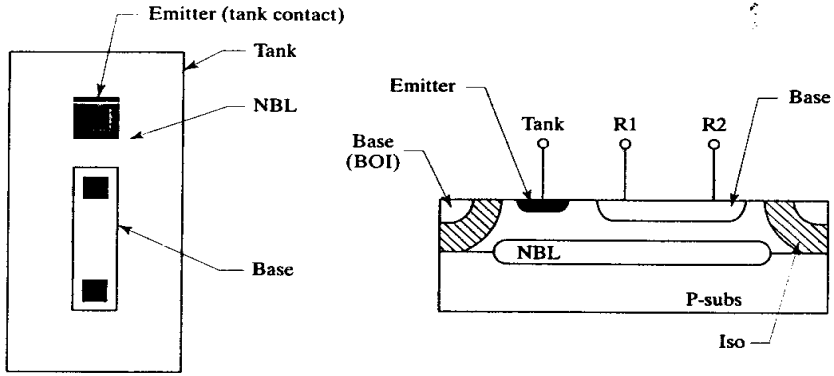


**FIGURE 3.13** Layout and cross section of a base resistor.

An *emitter resistor* consists of a strip of emitter diffusion isolated by a base diffusion enclosed within an N-tank (Figure 3.14). The base region is connected so that it will reverse-bias the emitter-base junction, while the tank is biased to reverse-bias the base-epi junction. The simplest way to achieve this goal consists of tying the base to the low-voltage end of the resistor and the tank to the high-voltage end. Various other connections are also feasible, so long as neither junction forward-biases. NBL is usually added to help suppress parasitic substrate PNP action. The emitter sheet resistance is relatively low (typically less than 10Ω/□), and the breakdown of the emitter-base junction limits the differential voltage across the resistor to about 6V.

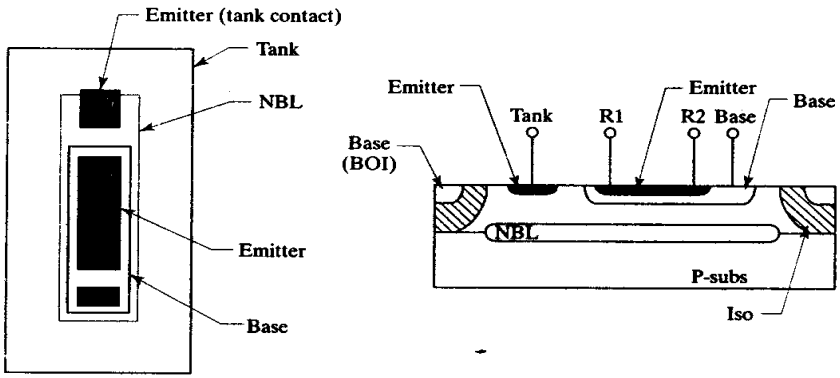


**FIGURE 3.14** Layout and cross section of an emitter resistor (note the presence of bias terminals for both the tank and the base region).

A *pinch resistor* consists of a combination of base and emitter diffusions (Figure 3.15). The emitter forms a plate overlapping the middle of a thin strip of base.[8] Contacts occupy the ends of the base strip, which project from under the emitter plate. The tank and emitter plate are both N-type and are therefore electrically united. A tank contact biases both to a voltage slightly more positive than the resistor in order to ensure isolation. The body of the resistor consists of the portion of the base diffusion beneath the emitter plate. This *pinched base* is thin

[8] R. P. O'Grady, "The 'Pinch' Resistor in Integrated Circuits," *Microelectronics and Reliability*, Vol. 7, 1968. pp. 233–236.

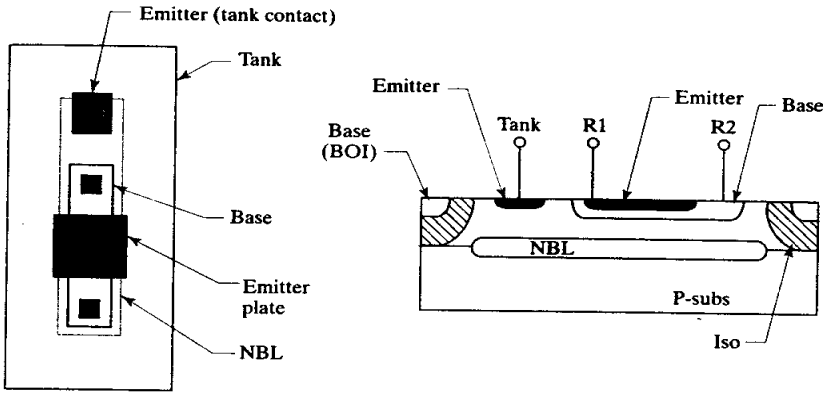


**FIGURE 3.15** Layout and cross section of a base pinch resistor.

and lightly doped and therefore its resistance may exceed 5000Ω/□. Emitter-base breakdown limits the differential voltage across the resistor to about 7V. Pinch resistors are notoriously variable—much more so than either emitter or base resistors. Worst of all, these resistors exhibit severe voltage modulation. The intrusion of depletion regions into the neutral base tends to further pinch the resistor, causing it to act much like a JFET (Section 12.3). Pinch resistors find application in startup circuits and other noncritical roles, but their many drawbacks prohibit more widespread use. Table 3.3 compares the performance of emitter, base, and pinch resistors.

| Parameter | Emitter | Base | Pinch |
|---|---|---|---|
| Sheet resistance | 5Ω/□ | 150Ω/□ | 3000Ω/□ |
| Minimum drawn width | 8μm | 8μm | 8μm |
| Breakdown voltage | 7V | 50V | 7V |
| Variability (15μm width) | ±20% | ±20% | ±50% or more |

**TABLE 3.3** Typical resistor device parameters.

### *Capacitors*

Standard bipolar was not intended to support capacitors. All of its oxide layers are so thick that they cannot be used to fabricate any but the smallest capacitors. However, the depletion region of a base-emitter junction exhibits a capacitance on the order of 0.8fF/μm$^2$ (0.5pF/mil$^2$), which can be used to construct a so-called *junction capacitor* (Figure 3.16). This capacitor consists of a base diffusion overlapping an emitter diffusion, both placed in a common tank. The emitter diffusion shorts to the tank, and the base-tank capacitance therefore adds to the base-emitter capacitance. The emitter plate must be biased positively with respect to the base plate to maintain a reverse bias across the base-emitter junction, and the differential voltage across the capacitor must not exceed the emitter-base breakdown voltage (about 7V). The resulting capacitance depends on bias and varies substantially (±50% or more). Junction capacitors are frequently used for compensating feedback loops, where their high capacitance per unit area makes up for their excessive variability.

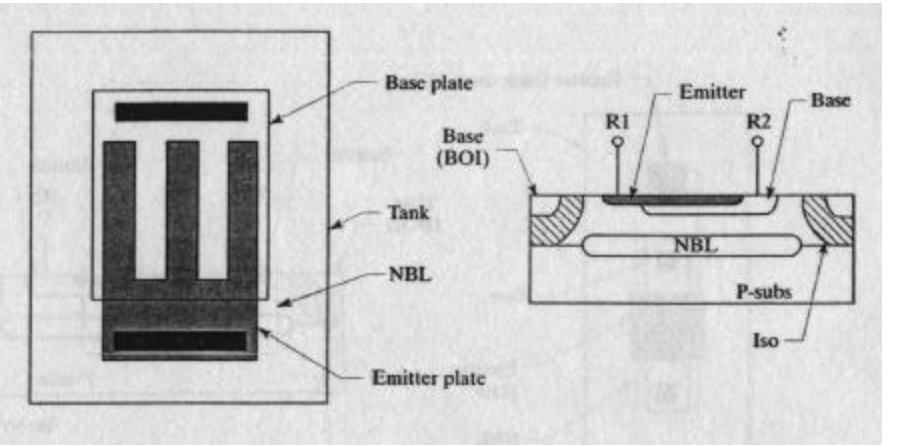


**FIGURE 3.16** Layout and cross section of a junction capacitor.

### 3.1.4. Process Extensions

Standard bipolar has spawned a large number of process extensions, five of which are discussed in this section. They include up-down isolation, double-level metal, Schottky diodes, high-sheet resistors, and super-beta transistors. The benefits of each extension must be weighed against the increased cost and process complexity.

*Up-down Isolation*

Standard bipolar employs deep-P+ junction isolation driven down through the epitaxial layer to the underlying substrate. Outdiffusion increases the width of the isolation by 20μm or more, limiting how closely components can be packed together. One means of reducing outdiffusion uses a *P-buried layer* (PBL) to supplement the P+ isolation. The resulting *up-down isolation* consists of an isolation diffusion drawn coincident with the PBL. The isolation diffuses down from the surface, while the PBL diffuses up from the epi-substrate interface (Figure 3.17). Each only has to cross half the distance of a regular isolation diffusion, and outdiffusion is therefore cut approximately in half.

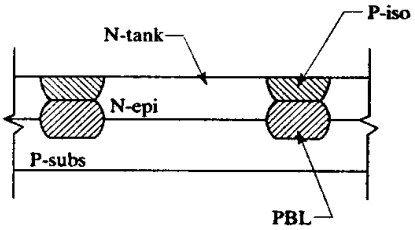


**FIGURE 3.17** Cross section of a typical up-down isolation system.

Up-down isolation does have one significant drawback. Process considerations limit the PBL implant dose, so the final PBL becomes very lightly doped, and vertical resistance through the up-down isolation greatly exceeds that of conventional top-down isolation. The PBL also requires an additional masking step and diffusion. Up-down isolation saves 15 to 20% die area, so a case-by-case analysis should be performed to determine if it is worthwhile.

*Double-level Metal*

Standard bipolar originated as a *single-level metal* (SLM) process. The lack of a second metal layer greatly complicated lead routing. Instead of crossing wires by

means of jumpers, diffusions were employed to form low-value resistors, called *crossunders* or *tunnels* (Section 13.3.2). Many devices can be custom-tailored to incorporate crossunders at the cost of compromising device performance and increasing die area. Single-level routing requires a deep understanding of device and circuit operation and an intuitive sense of topological connectivity. Most layout designers require years to master these skills.

*Double-level metal* (DLM) can be added to a standard bipolar process at the cost of two extra masks: vias and metal-2. The thickness of the first metal layer is often reduced to simplify planarization. Double-level metal is a useful, if somewhat costly, option. Lead routing no longer requires the use of customized devices, allowing component standardization and a considerable reduction of layout time and effort. Since metallization consumes a great deal of area, double-level metal can also reduce die area by up to 30%. These benefits are so attractive that manufacturers now routinely employ double-level metal for all new designs.

### *Schottky Diodes*

Standard bipolar originally used silicon-doped aluminum metallization. Modern processes usually employ a combination of silicidation and refractory barrier metallization to ensure adequate step coverage while maintaining low contact resistance. Along with its more obvious benefits, silicidation also offers the opportunity to fabricate reliable Schottky diodes. Although aluminum forms a rectifying Schottky barrier to lightly doped N-type silicon, the forward voltage of the resulting diodes varies unpredictably depending on annealing conditions. Certain silicides, most notably those of platinum and palladium, produce Schottky barriers with very stable and repeatable properties. The forward voltages of these Schottky diodes lie slightly below that of a moderately doped PN-junction diode, so they can serve as antisaturation clamps (Section 8.1.4).

Schottky diodes require contacts formed through thick-field oxide. A contact etch that just penetrates the field oxide will severely overetch the base and emitter contacts. Overetching can be prevented by performing two consecutive oxide removals, the first of which thins the field oxide over the Schottky contacts and the second of which creates the actual contact openings. The fabrication of Schottky diodes thus requires an additional masking step.

Figure 3.18 shows a typical Schottky diode layout. The anode consists of a rectifying contact to an N-epi tank while the cathode Ohmically contacts the same tank with the assistance of emitter diffusion. The addition of NBL and deep-N+ to the cathode

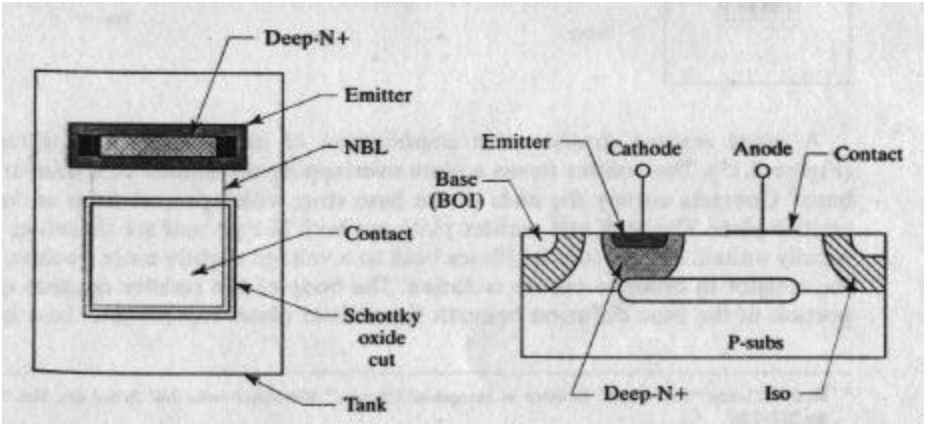


**FIGURE 3.18** Layout and cross section of a Schottky diode.

greatly reduces the series resistance of the diode. The anode employs two concentric oxide removals, the larger of which thins the field oxide and the smaller of which forms the actual contact. This two-stage process not only eliminates overetching of base and emitter contacts but also improves the step coverage of metallization to the Schottky contact. This structure readily scales to provide diodes of any size. Electric field intensification at the exposed edges of the Schottky contact opening limits this simple structure's breakdown voltage and causes excessive reverse-bias leakage. Section 10.1.3 discusses methods of circumventing these problems.

### *High Sheet Resistors*

Accurate resistors in standard bipolar are usually fabricated from base diffusion. Because the sheet resistance of this material rarely exceeds 200Ω/□, a typical die can incorporate a total of only 200 to 500kΩ of base resistance. Low-current circuits require more resistance than base diffusion can provide and more precision than pinch resistors can achieve. The only solution to this dilemma consists of adding a process extension to fabricate a precisely controlled high-sheet resistance (HSR) material.

The *high-sheet implant* consists of a shallow, lightly doped P-type implant. Depending on dose and junction depth, the sheet resistance of this implant can range from 1 to 10kΩ/□. The larger sheet values suffer from surface depletion effects that cause resistance variations, so most processes employ HSR implants of 1 to 3kΩ/□.

Figure 3.19 shows the layout and cross section of a typical HSR resistor. The body of the resistor consists of high-sheet implant, but small regions of base diffusion at either end of the resistor ensure Ohmic contact. The resistor occupies an N-tank and is isolated by the reverse-biased HSR-tank junction. The tank is often connected to the more positive end of the resistor, just as in the case of the base resistor. High-sheet resistors require one additional mask step and one dedicated implant. The cost of this process extension can usually be justified if the circuit includes more than 100 to 200kΩ of resistance.

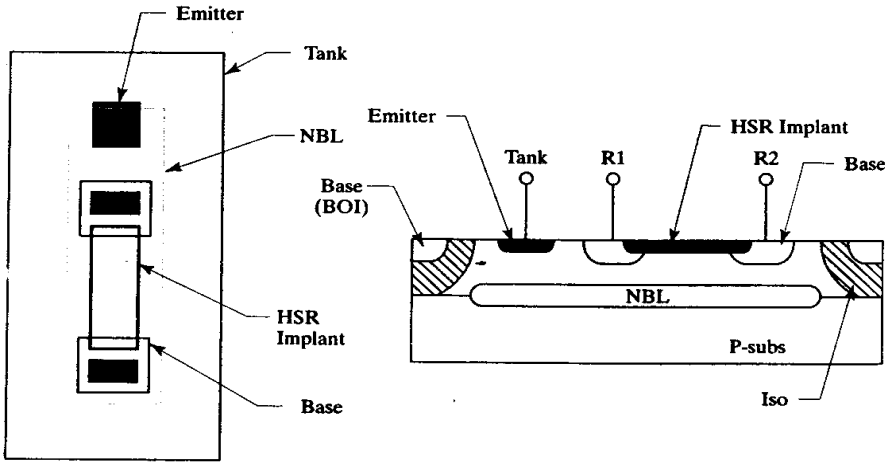


**FIGURE 3.19** Layout and cross section of a high-sheet resistor.

### *Super-beta Transistors*

The standard bipolar NPN provides a reasonable compromise between high beta and adequate operating voltage. Beta can be greatly increased by narrowing the base width. If the process incorporates two separate emitter implants, then one can

be optimized for beta and the other for operating voltage. The resulting *super-beta transistors* can achieve betas of 1000 to 3000 (Section 8.3.1). The use of an extremely thin and lightly doped base causes considerable depletion region intrusion into the neutral base and hence compromises not only operating voltage but also Early voltage. These highly specialized devices find application only in a limited range of circuits. For example, they have been used to fabricate bipolar operational amplifiers with extremely low input bias currents.

## 3.2 POLYSILICON-GATE CMOS

With the addition of two masking steps, standard bipolar can fabricate metal-gate PMOS transistors similar to those fabricated by early MOS processes (Figure 3.20). An N-tank serves as a backgate for the PMOS transistor; the backgate contact incorporates emitter diffusion to ensure Ohmic contact. None of the oxide layers in standard bipolar is sufficiently thin to serve as a gate oxide, necessitating the addition of a special masking step. Aluminum metal forms the gate electrode, while the source and drain consist of shallow P+ implants. Since standard bipolar does not include any suitable implant, another masking step patterns a special P-type source/drain (PSD) implant.

Practical MOS transistors require threshold voltages that lie within relatively narrow limits. Threshold voltages of less than 0.5V cause excessive leakage, while those of more than 1.5V unnecessarily reduce the available transconductance. The threshold voltage of an unadjusted (or *natural*) PMOS usually lies between 2 and 4V, necessitating a threshold adjust implant to shift it to the desired target of about 1V. The threshold voltage must lie within approximately ±0.5V of target. The metal gate PMOS transistor has difficulty maintaining even this minimal degree of control. The excess surface state charges introduced by using (111) silicon constitute one source of threshold variation; mobile ion contamination (Section 4.2.2) represents another.

Metal-gate MOS transistors also suffer from excessive overlap capacitance. The gate electrode is patterned by a different mask than the source and drain diffusions are. The gate must therefore overlap the source and drain sufficiently to form a continuous channel, even in the presence of photolithographic misalignments. The overlap between the gate and source causes a gate-to-source capacitance $C_{gs}$, while the

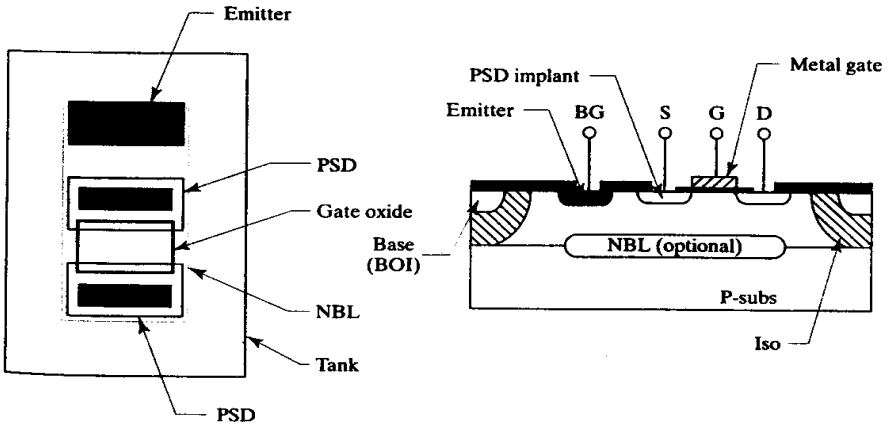


**FIGURE 3.20** Layout and cross section of a metal-gate PMOS transistor constructed in standard bipolar.

overlap between gate and drain causes a gate-to-drain capacitance $C_{gd}$. These parasitic capacitances slow the transistor because they must be charged and discharged during switching. The gate-to-drain capacitance is particularly deleterious because the voltage gain of the transistor multiplies its apparent value (a phenomenon called the *Miller effect*). These parasitic overlap capacitances must be minimized in order to allow the construction of high-speed logic circuitry.

A complementary NMOS transistor would greatly enhance the utility of the metal-gate PMOS process. Taken together, NMOS and PMOS transistors would allow the construction of versatile *complementary MOS* (CMOS) circuits. Unfortunately, standard bipolar cannot easily fabricate NMOS transistors because they require a lightly doped P-type backgate that does not exist in this process. The NMOS threshold voltage on suitably doped (111) silicon is slightly negative, so a threshold-adjust implant is required to form an enhancement device. Yet another masking step is required to raise the thick-field threshold in the lightly doped P-type backgate around the NMOS transistors to prevent parasitic channel formation. The relatively poor performance of metal-gate CMOS cannot justify the cost of five additional masking steps, especially when a nine-mask polysilicon-gate CMOS process can fabricate vastly superior transistors. The following section examines the construction and performance of this alternate process.

### 3.2.1. Essential Features

The polysilicon-gate CMOS process is optimized to form complementary PMOS and NMOS transistors on a common substrate. It does not support the construction of bipolar transistors and it offers only a limited range of passive components. Originally intended solely for manufacturing CMOS logic gates, with slight modifications this process can also fabricate a limited variety of analog circuits.

A key difference between polysilicon-gate CMOS and standard bipolar lies in the choice of substrate material. Standard bipolar employs (111) silicon to enhance the thick-field threshold by increasing the surface state density, while polysilicon-gate CMOS uses (100) silicon to reduce the surface state density in order to improve threshold voltage control. A second major innovation lies in the use of polysilicon rather than aluminum as the gate material. Polysilicon can safely withstand the high temperatures required to anneal the source/drain implants, so it can act as a mask to form self-aligned sources and drains. The effects of mobile ion contamination can also be minimized by doping the polysilicon gate with phosphorus. Poly gates thus offer not only faster switching speeds but also better control of threshold voltages.

The choice of threshold voltages forms one of the few differences between analog and digital CMOS processes. Most digital CMOS processes[9] target threshold voltages between 0.8V and 0.9V with a variation of about ±0.2V. Analog CMOS designers favor maximizing headroom by targeting threshold voltages around 0.7V ± 0.2V. Since (100) silicon dictates the use of threshold-adjust implants in either case, the threshold voltages can often be retargeted by simply changing the implant dosage. Analog CMOS also rules out the use of blanket silicidation (Section 3.2.4). Neither of these requirements fundamentally modify the polysilicon-gate CMOS process.

[9] The information provided in the text applies to processes with operating voltages of 5V or more. Lower-voltage processes require smaller threshold voltages and tighter control. For example, a 3V process will typically target a threshold voltage of 0.6 ± 0.15V.

## 3.2.2. Fabrication Sequence

The baseline polysilicon-gate CMOS fabrication sequence consists of nine masking operations. The processing steps required to fabricate a finished wafer will be presented in the order in which they are performed. The cross sections used to illustrate this process employ a vertical exaggeration of between two and five, just as did the cross sections previously presented for standard bipolar.

### *Starting Material*

CMOS integrated circuits are normally fabricated on a P-type (100) substrate doped with as much boron as possible in order to minimize substrate resistivity. This precaution helps provide a degree of immunity to *CMOS latchup* by minimizing substrate debiasing (Section 4.4). CMOS processes do not require NBL, so substrate doping is limited only by solid solubility.

### *Epitaxial Growth*

The first step of the CMOS process consists of growing a lightly doped P-type epitaxial layer on the substrate. This epitaxial layer, typically some 5 to 10μm thick, is considerably thinner than the one used for standard bipolar. NMOS transistors are formed directly into the epi layer, which serves as their backgate. Since this process needs no buried layers, epi-coated wafers can be stockpiled to serve as starting material for all types of products. Standard bipolar does not allow this economy of scale, since each product requires a uniquely patterned NBL.

In theory, CMOS processes do not require epitaxy since the MOS transistors can be grown directly into a P– substrate. Epitaxial processing increases costs, but it also improves latchup immunity by allowing the use of a P+ substrate. In addition, the electrical properties of the epitaxial layer are more precisely controllable than those of Czochralski silicon, resulting in better control of MOS transistor parameters.

### *N-well Diffusion*

After the wafer has been thermally oxidized, a layer of photoresist that has been spun onto it is patterned using the N-well mask. An oxide etch opens windows through which ion implantation deposits a controlled dose of phosphorus. A prolonged high-temperature drive creates a deep lightly doped N-type region called an *N-well* (Figure 3.21). The N-well for a typical 20V CMOS process has a junction depth of about 5μm. Thermal oxidation during the well drive covers the exposed silicon with a thin layer of *pad oxide.*

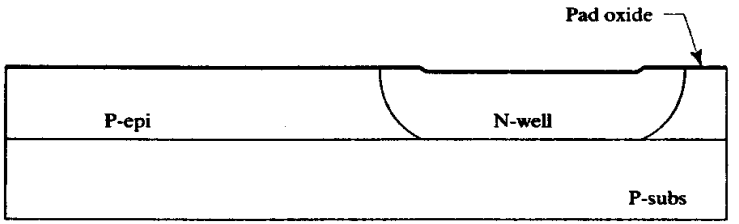


**FIGURE 3.21** Wafer after N-well drive.

In an *N-well CMOS process,* such as that illustrated in Figure 3.21, NMOS transistors occupy the epi, and PMOS transistors reside in the well. The increased total dopant concentration caused by counterdoping the well slightly degrades the mobility of majority carriers within it. The N-well process therefore optimizes the performance of the NMOS transistor at the expense of the PMOS transistor. As a side

effect, the N-well process also produces the grounded substrate favored by most circuit designers.

A *P-well CMOS process* uses an N+ substrate, an N– epitaxial layer, and a P-well. NMOS transistors are formed in the P-well and PMOS transistors in the epi. This process optimizes the PMOS transistor at the expense of the NMOS transistor, but the NMOS still outperforms its counterpart because electrons are more mobile than holes. A P-well process requires that the substrate connect to the highest-voltage supply instead of ground. Designs that employ multiple power supplies often have difficulty biasing an N-type substrate because of ambiguities in the sequencing of the supplies.

Both P-well and N-well CMOS processes exist. The N-well process offers a slightly better NMOS transistor, and it allows the use of a grounded substrate. N-well CMOS is also upwardly compatible with BiCMOS technology, as will become apparent later in this chapter. The N-well process has therefore been chosen to illustrate CMOS technology.

### *Inverse Moat*

The CMOS process employs a thick-field oxide for much the same reasons as standard bipolar: it increases the thick-field threshold voltages, and it reduces parasitic capacitance between the metallization and the underlying silicon. Unlike standard bipolar, CMOS processes employ LOCOS technology to selectively grow the field oxide, leaving only a thin pad oxide over the regions where active devices will be formed. The locally oxidized regions of the die are called *field regions*, while the areas protected from oxidation are called *moat regions*.

The LOCOS process uses a patterned nitride layer formed by first depositing nitride across the entire wafer, then patterning this nitride using the inverse moat mask, and finally employing a selective etch to remove the nitride over the field regions (Figure 3.22). The photomask used for this step is called the *inverse moat* mask because it consists of a color reverse of the moat regions. In other words, the mask codes for areas where moat is absent, not where it is present.

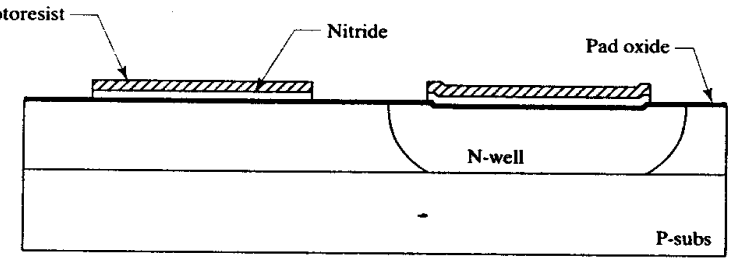


**FIGURE 3.22** Wafer after nitride deposition and inverse moat pattern.

The nitride layer used for LOCOS must lie on top of a thin oxide layer (the *pad oxide*) because the conditions of nitride growth induce mechanical stresses that can cause dislocations in the silicon lattice. The pad oxide provides mechanical compliance and prevents strains produced by the nitride growth from damaging the underlying silicon.

### *Channel Stop Implants*

The CMOS process deliberately minimizes threshold voltages in order to produce practical MOS transistors. The LOCOS field oxide will raise the thick-field thresholds, but not by enough to support supply voltages of more than a few volts. Practical CMOS processes nearly always require additional measures to ensure that

the thick field thresholds exceed the operating voltages. Dopants are usually selectively implanted beneath the field regions to raise the threshold voltages of the thick-field transistors. P-epi field regions receive a P-type *channel stop implant,* while N-well field regions receive an N-type channel stop implant. The formation of channel stops thus requires two successive ion implantations.

Various techniques have been developed to produce channel stops. The method presented here involves the use of a blanket boron implant followed by a patterned phosphorus implant. The boron implant uses the photoresist left from patterning the LOCOS nitride. This mask exposes the field regions where channel stops will be deposited, so all of these regions receive the blanket boron implant (Figure 3.23A). This step sets the thick-field threshold in the epi regions.

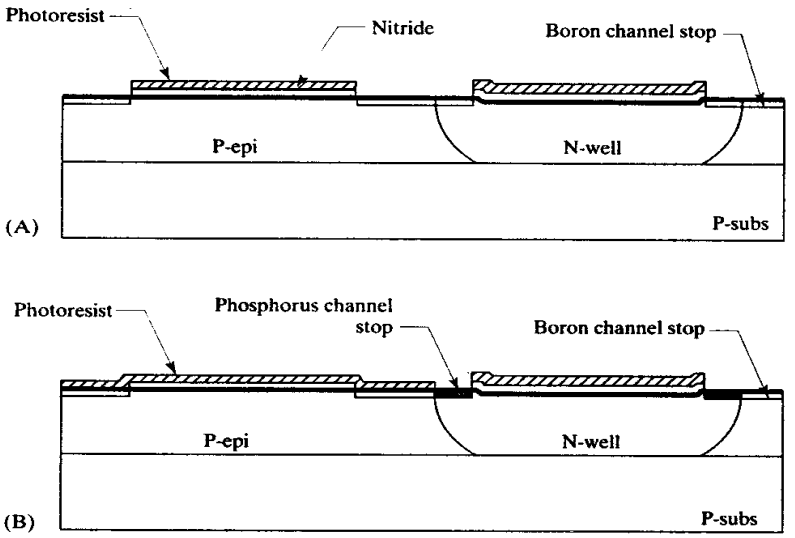


**FIGURE 3.23** Wafer after blanket boron channel stop implant (A) and after selective phosphorus channel stop implant (B).

The wafer is again coated with photoresist immediately after the boron implant. The previous photoresist can remain in place since the channel stop implant will not affect the moat regions that lie beneath it. The recoated wafer is patterned using the channel stop mask, exposing only N-well field regions. The subsequent phosphorus implant counterdopes the previous blanket boron implant and raises the NMOS thick-field threshold above the maximum operating voltage (Figure 3.23B). Following the phosphorus implant, all photoresist is stripped from the wafer in preparation for LOCOS oxidation.

### *LOCOS Processing and Dummy Gate Oxidation*

Steam is often used to increase the rate of LOCOS oxidation; alternately the furnace pressure can be raised to five or ten times atmospheric. After LOCOS oxidation, a suitable etchant strips away the remnants of the nitride block mask. Figure 3.24 shows a cross section of the resulting wafer. The curved transition region, called a *bird's-beak,* at the edges of the moat results from oxidants diffusing under the edges of the nitride film.

The *Kooi effect* (Section 2.3.4) causes nitride deposits to form underneath the pad oxide around the edges of the moat. These deposits can potentially cause gate oxide integrity failures, but they can be eliminated by a dummy gate oxidation. A

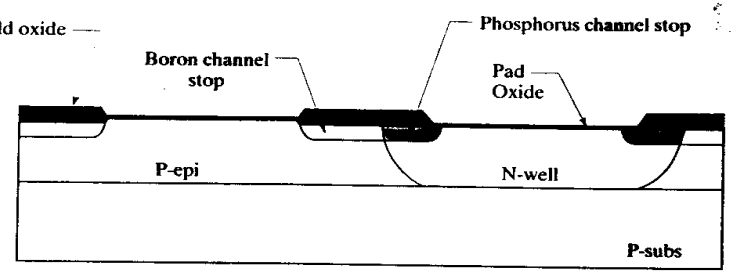


**FIGURE 3.24** Wafer after LOCOS oxidation and nitride strip.

brief etch strips away the thin pad oxide without substantially eroding the thick-field oxide. Next, a brief dry oxidation grows a thin layer of oxide called a *dummy gate oxide* (or *sacrificial gate oxide*) in the moat regions. Any nitride deposits that remain will gradually oxidize. All of the nitride will be consumed if the dummy gate oxidation continues for a sufficient length of time.

### *Threshold Adjust*

The use of (100) silicon helps stabilize the threshold voltages of the MOS transistors, but the backgate dopings and gate electrode materials preclude the achievement of usable threshold voltages without threshold adjust implants. For example, the unadjusted PMOS threshold might range from –1.5 to –1.9V, while the NMOS threshold might range from –0.2 to 0.2V. One or two threshold adjust implants (also known as $V_t$ *adjusts*) retarget the threshold voltages to the desired targets, usually 0.7V for NMOS and –0.7V for PMOS transistors.

Two methods exist for adjusting threshold voltages. The first method employs two separate implants, one to set the PMOS $V_t$ and the other the NMOS $V_t$. The use of two implants allows independent optimization of both thresholds. Many processes do not require this degree of flexibility. These processes can use a single $V_t$ adjust to simultaneously reduce the PMOS threshold and increase the NMOS threshold. If this implant is properly performed, a nominal threshold voltage of 0.7 to 0.9V can be obtained for both types of MOS transistors. Figure 3.25 illustrates this approach.

After the wafer has been coated with photoresist, the $V_t$ adjust mask is used to open windows over areas where MOS transistors will form. The boron $V_t$ adjust implant penetrates the dummy gate oxide to dope the underlying silicon. After the $V_t$ adjust implant, the dummy gate oxide is stripped away to reveal bare silicon in the moat regions.

The true gate oxidation employs dry oxygen to minimize excess charge incorporation due to surface states and fixed oxide charges. This oxidation must be very brief, because gate oxides are exceedingly thin. A 10V MOS transistor typically requires a 300Å (0.03μm) gate oxide, while a 3V transistor may employ an oxide less than 100Å (0.01μm) thick. This gate oxide will form the dielectric of the MOS

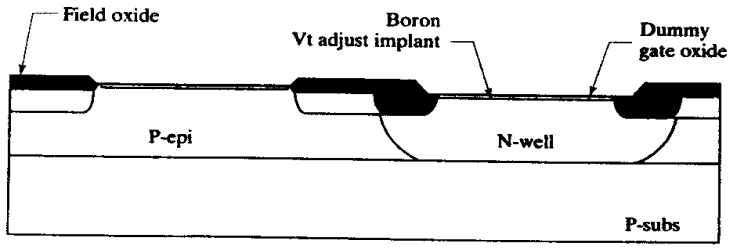


**FIGURE 3.25** Wafer after $V_t$ adjust implant.

transistors; it also covers the regions where source and drain implants will later occur.

### *Polysilicon Deposition and Patterning*

The polysilicon layer used to form gate electrodes is heavily doped with phosphorus to reduce its resistivity to about 20 to 40Ω/□. Although gate leads do not conduct DC current, switching transients do draw substantial AC current, and low-resistance gate polysilicon greatly improves the switching speeds of MOS circuitry. Phosphorus doping simultaneously adjusts the work function of the polysilicon to produce threshold voltages compatible with a single-step $V_t$ adjust. Phosphorus-doped gate polysilicon also minimizes threshold voltage variation due to mobile ions, allowing threshold voltage control of ±0.1 to 0.2V. While it is possible to dope polysilicon during deposition, most processes first deposit intrinsic polysilicon and subsequently dope it using conventional deposition or implantation techniques.

The deposited polysilicon layer must now be patterned using the poly mask (Figure 3.26). Modern submicron processes can fabricate polysilicon gates less than 0.5μm long, and any variation in gate length directly affects the transconductance of the resulting transistors. Thus, the patterning and etching of poly form the most critical photolithographic steps of a CMOS process. The simple process discussed here produces a minimum channel length of about 2μm and therefore does not require as high a degree of precision as submicron processes, but polysilicon patterning still remains its most challenging photolithographic step.

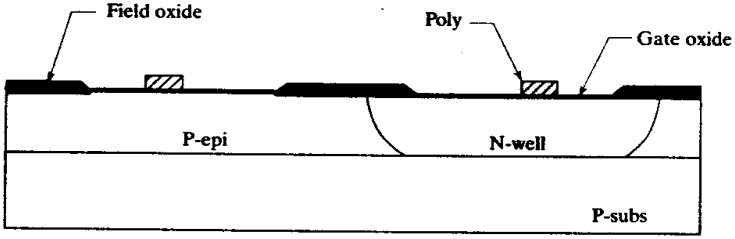


**FIGURE 3.26** Wafer after polysilicon deposition and pattern. For simplicity, the channel stop and threshold adjust implants do not appear in this or subsequent cross sections.

### *Source/Drain Implants*

The completed polysilicon gates now act as masks that self-align the source/drain implants for both the PMOS and NMOS transistors. These implants can be performed in either order. In the illustrated process, the N-type source/drain (NSD) implant occurs first, followed by the P-type source/drain (PSD) implant.

The NSD implant begins with the application of photoresist to the wafer, followed by patterning using the NSD mask. Shallow, heavily doped N-type regions are then formed by implanting arsenic through the exposed gate oxide. The polysilicon gate blocks this implant from the regions directly underneath the gate and therefore minimizes the gate/source and gate/drain overlap capacitances. Once the NSD implant has been completed, the photoresist residue is stripped from the wafer. The PSD implant begins with the application of a second photoresist layer patterned using the PSD mask. A shallow, heavily doped P-type region is formed by implanting boron through the exposed gate oxide. As with the NSD implant, the PSD implant self-aligns to the polysilicon and the PMOS transistors also exhibit minimal overlap capacitance. Following the PSD implant, photoresist is again stripped from the wafer.

A brief anneal activates the implanted dopants and slightly thickens the oxide over the source and drain regions. This anneal is the final high-temperature step of

the process, corresponding to the emitter drive of standard bipolar. Figure 3.27 shows a cross section of the wafer following the source/drain anneal.

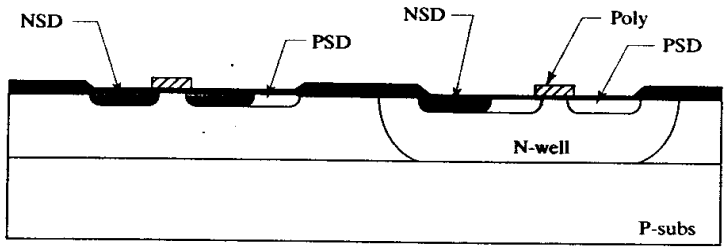


**FIGURE 3.27** Cross section of the wafer after NSD and PSD implants and anneals. The backside contact implants abut the source implants to save space.

### *Contacts*

Despite further oxidation during the source/drain anneal, the oxide covering the moat regions remains thin and therefore vulnerable to oxide rupture. Most processes deposit a *multilevel oxide* (MLO) before contact patterning. The MLO thickens the oxide over the moat regions and at the same time coats and insulates the exposed polysilicon pattern. Metal leads can now run over moat regions and polysilicon gates without risk of oxide rupture.

After the wafer is again coated with photoresist, the contact regions are patterned using the contact mask. Ohmic contacts form to the heavily doped source and drain without difficulty, but the backgate regions are far too lightly doped to allow direct Ohmic contact. The addition of NSD and PSD implants in the vicinity of the backgate contacts overcomes this difficulty. Contacts opened over polysilicon allow contact to the gate electrodes.

### *Metallization*

The shallow NSD and PSD diffusions are vulnerable to junction spiking. Most CMOS processes employ a combination of contact silicidation and refractory barrier metallization to ensure reliable contact to the source/drain regions. After formation of silicide in the contact openings, a thin film of refractory metal sputtered over the wafer precedes a much thicker layer of copper-doped aluminum. The metallized wafer is coated with photoresist and patterned, using the metal mask. A suitable etchant then removes unwanted metal to form the interconnection pattern. Most processes also include a second layer of metallization. In such a process, another layer of oxide deposited over the first metal pattern insulates it from the second metal pattern. This second deposited oxide is usually called an *interlevel oxide* (ILO). Some form of planarization minimizes the nonplanarities caused by the first metal pattern to ensure adequate second metal step coverage. Vias etched through the ILO connect to a second metal layer deposited and patterned in much the same way as the first.

### *Protective Overcoat*

A protective overcoat is now deposited over the final layer of metallization, both to provide mechanical protection and to prevent contamination of the die. The protective overcoat must resist penetration by mobile ions, so it normally consists of either a thick phosphosilicate glass (PSG), a compressive nitride layer, or both.

After coating with photoresist, the wafer is patterned using the *protective overcoat* (PO) mask. A suitable etchant removes the overcoat over selected areas of

metallization and allows attachment of bondwires to the integrated circuit. This composes the final fabrication step; the wafer is now complete. Figure 3.28 shows a cross section of the resulting wafer, with only a single level of metal for simplicity. No bondpad openings exist in the illustrated portion of the die. This cross section includes an NMOS transistor on the left and a PMOS transistor on the right.

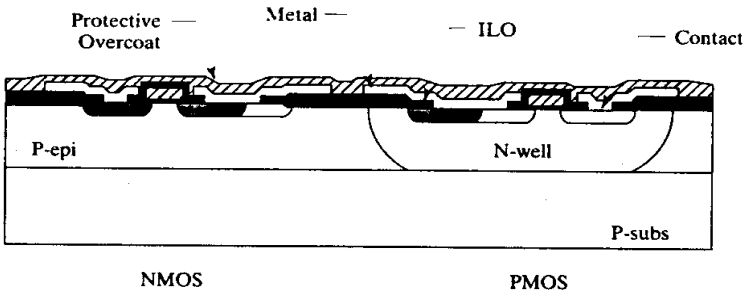


**FIGURE 3.28** Cross section of the completed polysilicon-gate CMOS wafer.

## 3.2.3. Available Devices

Polysilicon-gate CMOS was originally developed to provide relatively low-voltage NMOS and PMOS transistors. The same process steps can fabricate several other components, including natural MOS transistors, a substrate PNP, several types of resistors, and a capacitor. Together these components allow the construction of a considerable variety of analog circuits. Section 3.2.4 examines process extensions that allow higher operating voltages and denser circuit packing.

### *NMOS Transistors*

Figure 3.29 shows a representative layout and cross section of an NMOS transistor. The source and drain regions consist of NSD implants that self-align to the polysilicon gate. Since the backgate of the NMOS consists of the P-epi (and by extension the substrate), any substrate contact on the die will serve as a backgate terminal for the transistor. Many layouts actually include separate backgate contacts immediately adjacent to each NMOS transistor even though these are not strictly necessary. The close proximity of these backgate contacts improves CMOS latchup immunity, and this arrangement ensures that an adequate number of substrate contacts are distributed throughout the layout. In cases where the source of the NMOS transistor connects to the substrate potential, a very compact layout can be achieved by butting the PSD substrate contact against the NSD source. PSD and NSD cannot abut one another if they connect to different potentials because the resulting P+/N+ junction leaks excessively.

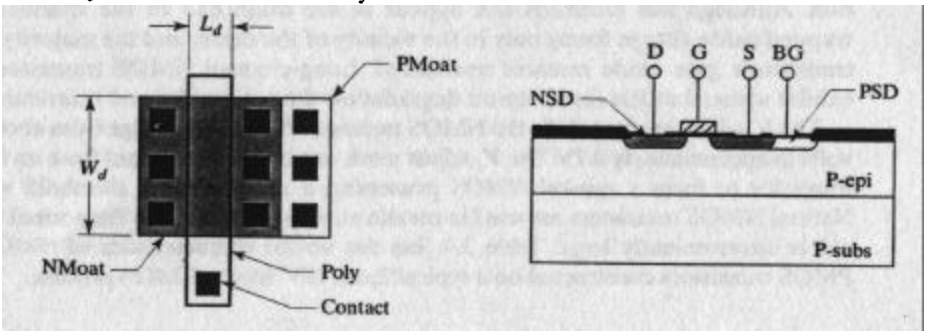


**FIGURE 3.29** Layout and cross section of an NMOS transistor. The source and backgate of this transistor are shorted together using metal (not shown).

Figure 3.29 illustrates the common practice of coding CMOS transistors using drawing layers called *NMoat* and *PMoat*. These layers do not correspond to individual masks but rather to combinations of several masks. A figure drawn on the NMoat layer simultaneously produces figures on both the NSD and moat masks. Similarly, a figure drawn on PMoat simultaneously generates figures on both the PSD and moat masks. The use of PMoat and NMoat drawing layers simplifies the layout by reducing the number of figures required to draw a transistor.

The arrays of small square contacts shown in Figure 3.29 are characteristic of CMOS processes. The etch rate of oxide windows varies somewhat depending upon their size and shape, and these variations become particularly severe for very small openings. Many processes therefore allow only a single size of contact opening, usually consisting of a minimum-dimension square. Larger contacts must consist of arrays of minimum contacts rather than larger oxide openings.

In Figure 3.29, $W_d$ and $L_d$ denote the *drawn width* and *drawn length* of the NMOS transistor, respectively. The names given to these two dimensions may seem counterintuitive since the length of the gate is actually the width of the drawn polysilicon strip, but the channel length is customarily defined as the separation between its source and drain regions. The transconductance of an MOS transistor is approximately proportional to the ratio of the channel width divided by the channel length ($W_d/L_d$). Short channel lengths generate more transconductance per unit area, but analog circuits often employ longer channels to reduce channel length modulation.

Hot electron degradation limits the simple NMOS transistor of Figure 3.29 to relatively low operating voltages. The depletion region across the pinched-off portion of the channel accelerates electrons to high velocities. Some of these *hot electrons* collide with lattice atoms and ricochet out of the channel into the overlying gate oxide, where they become trapped. Hot electron injection causes a gradual shift in threshold voltage due to the slow accumulation of a fixed oxide charge. Eventually the threshold voltage shifts so far that the circuit ceases to meet parametric specifications.

Hot electron injection only occurs when the NMOS transistor operates in saturation with a relatively large drain-to-source bias. In the linear region the drain-to-source voltage is too small to produce hot electrons, and in cut-off no conduction occurs. NMOS transistors used as switches experience hot electron injection only during switching transients. If the switching frequency remains fairly low, then the total quantity of hot electrons produced over the operating lifetime of the integrated circuit remains acceptably small. Two different operating voltages are often specified for NMOS transistors. Junction breakdown and punchthrough limit the *blocking voltage rating*, which applies to transistors used as switches and as components of low-frequency digital logic. The somewhat lower *operating voltage rating* determined by the onset of hot electron degradation applies to transistors that operate for an appreciable length of time in saturation (as do the majority of analog transistors).

Increasing the length of the transistor reduces the impact of hot electron injection. Although hot electrons still appear at the drain end of the transistor, the trapped oxide charge forms only in the vicinity of the drain, and the majority of the transistor's gate oxide remains unaffected. Long-channel NMOS transistors thus exhibit somewhat less hot electron degradation than short-channel transistors.

The $V_t$ adjust implant shifts the NMOS transistor threshold voltage from about zero volts to approximately 0.7V. The $V_t$ adjust mask can block the implant from an NMOS transistor to form a *natural NMOS* possessing a relatively low threshold voltage. Natural NMOS transistors are used in certain analog circuits where the normal threshold is inconveniently large. Table 3.4 lists the device characteristics of NMOS and PMOS transistors constructed on a typical 2μm, 10V analog CMOS process.

| Parameter | NMOS | PMOS |
|---|---|---|
| Minimum channel length | 2μm | 2μm |
| Gate oxide thickness | 400Å | 400Å |
| Threshold voltage (adjusted) | 0.7V | –0.7V |
| Threshold voltage (natural) | 0V | –1.4V |
| Transconductance ($W_d/L_d$ = 10/10) | 50μA/V$^2$ | 20μA/V$^2$ |
| Operating voltage | 7V | 15V |
| Blocking voltage | 15V | 15V |

**TABLE 3.4** Typical polysilicon-gate CMOS device parameters.[10]

## *PMOS Transistors*

Figure 3.30 shows a representative layout and cross section of a PMOS transistor. This device resides in an N-well that acts as its backgate. Any number of PMOS transistors can occupy the same well as long as their backgates all tie to the same potential. The relatively deep N-well outdiffuses substantially and the layout dimensions associated with it become quite large. Merging PMOS transistors in the same tank therefore saves substantial area. While the backgate of an NMOS transistor inherently connects to substrate, the backgate of a PMOS transistor can connect to any potential greater than or equal to its source. The N-well backgate thus provides an extra degree of freedom that analog designers frequently employ to enhance circuit performance.

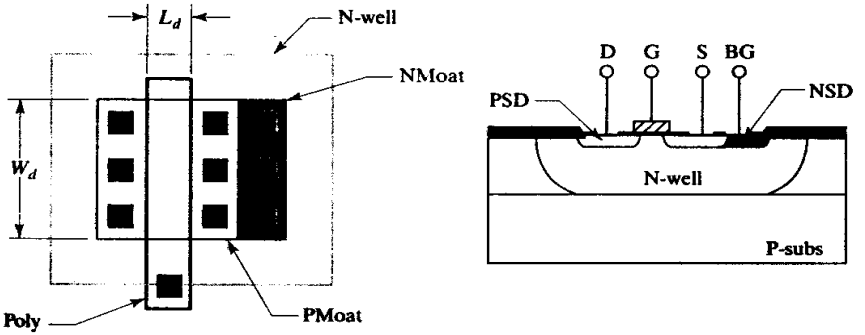


**FIGURE 3.30** Layout and cross section of a PMOS transistor. The backgate and source of this transistor are shorted together using metal (not shown).

PMOS transistors are subject to *hot hole degradation,* but this causes fewer problems than hot electron degradation because holes are less mobile than electrons. A higher electric field and therefore a larger drain-to-source voltage is required to accelerate holes to velocities sufficient to inject charge into the oxide. Junction avalanche and punchthrough often limit PMOS transistors to voltages where hot hole degradation remains unimportant. Higher-voltage PMOS transistors will encounter hot-carrier problems similar to those of NMOS transistors.

A *natural PMOS* transistor can be fabricated by blocking the $V_t$ adjust implant from the channel region of the device. Natural PMOS transistors possess inconveniently high threshold voltages, usually in excess of a volt. These transistors see only

[10] These parameters are roughly analogous to those of the 3μm Advanced LinCMOS™ process described in R. K. Hester, L. Hutter, L. Le Toumelin, J. Lin, and Y. Tung. "Linear CMOS Technology," *TI Technical Journal,* Vol. 8, #1, 1991, pp. 29–41. (The trademark belongs to Texas Instruments.) The transconductance figures given in this text are those used in the Schichman-Hodges equation $I_d = 1/2\ k\ (W/L)\ (V_{gs} - V_t)^2$.

occasional use in analog circuit design. Table 3.4 lists typical device parameters for a 2μm, 10V PMOS transistor.

### *Substrate PNP Transistors*

The only bipolar transistor available in an N-well process is a substrate PNP. Figure 3.31 shows a typical layout and cross section for this device. The emitter consists of a PSD implant formed in an N-well that acts as the base of the transistor. A slug of NSD provides Ohmic contact to the N-well. The collector of this device consists of the P+ substrate and P-epi surrounding the well.

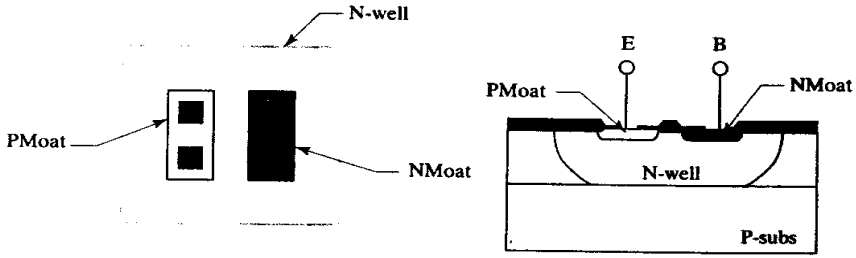


**FIGURE 3.31** Layout and cross section of a substrate PNP transistor. The collector is connected by means of substrate contacts (not illustrated).

Although CMOS processes do not deliberately optimize bipolar components, the substrate PNP can still provide reasonably good performance. Its beta may approach that of a standard bipolar device (50 to 100), but it often drops to much lower values if the emitter diffusion is silicided, as is the case in a clad-moat process. Since this transistor injects current into the substrate, care must be taken to provide adequate substrate contact. The main component of collector resistance consists of the lightly doped P-epi layer interposed between the P+ substrate and the PSD diffusion beneath the substrate contacts. A large substrate contact area is necessary to prevent substrate debiasing. A typical CMOS integrated circuit incorporates enough substrate contact in the scribe street to accommodate 10 to 20mA of substrate current. If higher substrate currents will occur, then unused areas of the die should be filled with substrate contacts.

Although a lateral PNP transistor can theoretically be constructed in an N-well process, the absence of NBL encourages substrate injection, and only a small fraction of the total emitter current reaches the intended collector. These transistors exhibit extremely low apparent gains, rendering them of limited use to the analog circuit designer.

### *Resistors*

The most useful of the four resistors available in polysilicon-gate CMOS consists of doped polysilicon (Figure 3.32). Although gate polysilicon exhibits a sheet resistance of only 20 to 30Ω/□, very narrow widths and spacings allow substantial resistance per unit area. A 2μm process can produce polysilicon resistors as area-efficient

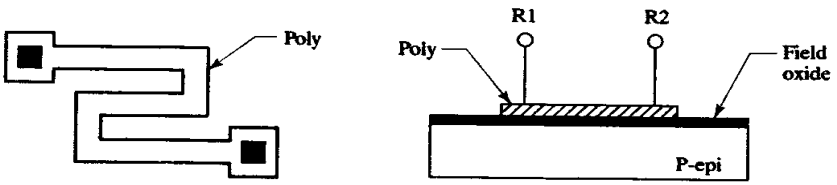


**FIGURE 3.32** Layout and cross section of a poly resistor.

as standard bipolar base resistors. Submicron processes can provide remarkable amounts of resistance from narrow polysilicon traces, but the tolerances and matching of such resistors leave much to be desired. In a clad-poly process, a silicide block mask becomes necessary to obtain sufficient sheet resistance to construct practical poly resistors.

The polysilicon resistor of Figure 3.32 consists of a strip of poly deposited on top of field oxide. Contacts at either end allow it to be connected into the circuit. Oxide completely isolates this resistor, enabling it to be biased in any manner desired. Poly resistors can withstand large voltages relative to the substrate (100V or more) and can operate below substrate potential or above the positive supply voltage. The thick-field oxide also reduces parasitic capacitance between the resistor and the underlying substrate. Oxide isolation has one drawback: it isolates the resistor thermally as well as electrically. A poly resistor that dissipates sufficient power will experience permanent resistance variations due to self-induced annealing. Extreme power dissipation will melt or crack polysilicon long before diffused resistors of similar dimensions suffer damage. This behavior allows the construction of polysilicon fuses for wafer-level trimming, but it makes poly resistors undesirable for certain specialized applications, such as ESD protection.

Figure 3.33A shows the layout and cross section of an NSD resistor formed by contacting either end of a strip of NSD diffusion. NSD typically has a sheet resistance of 30 to 50Ω/□. Avalanche breakdown of the relatively shallow NSD diffusion also limits the operating voltage of this resistor, typically to no more than 10 to 15V. A similar resistor can also be constructed from PSD (Figure 3.33B). This resistor consists of a strip of PSD diffusion contained in an N-well region. The well must be biased above the resistor to maintain isolation. The well is therefore connected either to the more positive end of the resistor or to a high-voltage node (for example, the positive supply). PSD resistors also suffer from limited sheet resistance and a relatively low breakdown voltage.

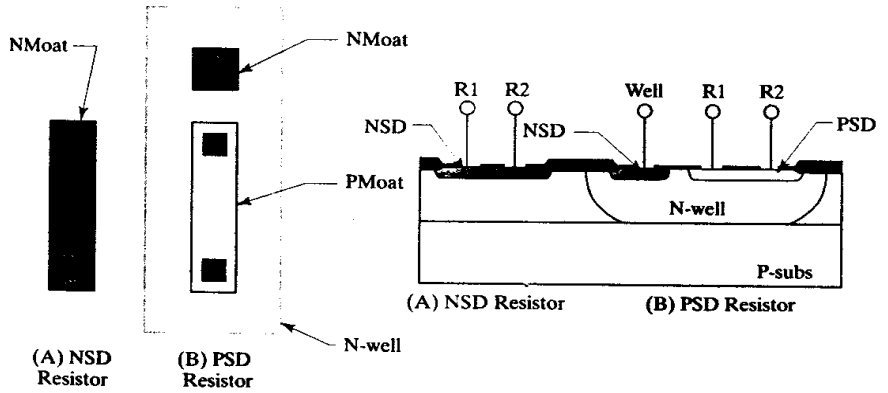


**FIGURE 3.33** Layout and cross section of an NSD resistor (A) and a PSD resistor (B).

Another type of resistor consists of a strip of N-well with contacts at either end. NSD placed beneath these contacts prevents the formation of rectifying Schottky barriers. The large spacings required to allow for outdiffusion partially offset the high sheet resistance of the well (typically some 2 to 5kΩ/□). Well resistors are notoriously variable. Slight differences in doping and outdiffusion, voltage modulation of the depletion regions, and surface effects can all cause significant variations in well resistance. Proper layout precautions can help minimize these sources of variability, but most designers prefer to employ narrow polysilicon resistors. Table 3.5 summarizes the properties of the four types of resistors available in a typical CMOS process.

**TABLE 3.5** Typical resistor device parameters.

| Parameter | Poly | PSD | NSD | Well |
|---|---|---|---|---|
| Sheet resistance | 20Ω/□ | 50Ω/□ | 50Ω/□ | 2000Ω/□ |
| Minimum drawn width | 2μm | 3μm | 3μm | 5μm |
| Breakdown voltage | >100V | 15V | 15V | 50V |
| Variability (5μm width) | ±30% | ±20% | ±20% | ±50% (10μm) |

*Capacitors*

The gate oxide used to fabricate MOS transistors can also be employed to construct capacitors. One plate of the capacitor consists of doped polysilicon and the other of a diffusion, typically N-well. Figure 3.34 shows the layout and cross section of one type of MOS capacitor. The capacitance of this device typically ranges from 0.5 to 1.6fF/μm$^2$ (0.3 to 1.0pF/mil$^2$), depending on oxide thickness. The tight control of gate oxide thickness characteristic of modern MOS processes results in typical tolerances of ± 20% as long as the well electrode remains at least 1V above the poly electrode. Failure to maintain adequate bias across the capacitor will cause a dramatic drop in capacitance (Section 6.2.2). The main drawbacks of these capacitors consist of excessive bottom-plate parasitic junction capacitance and series resistance, and of nonlinearity effects at certain voltages.

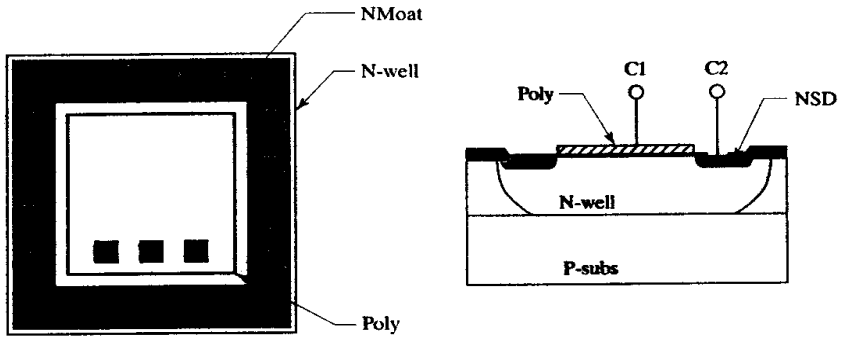


**FIGURE 3.34** Layout and cross section of a PMOS capacitor.

## 3.2.4. Process Extensions

CMOS process extensions tend to focus on improving the PMOS and NMOS transistors rather than on constructing additional devices. One set of extensions seeks to provide higher operating voltages by suppressing hot carrier degradation. Another set of extensions focuses on reducing the size of the transistors to improve packing. Unlike standard bipolar, little emphasis is placed on providing a large number of specialized components because most CMOS fabs build primarily digital products. The large expenditure of time and money required to construct additional devices cannot be justified by the relatively small volume of analog CMOS products. Analog BiCMOS presents an entirely different picture because this process caters specifically to analog design. Many of the features and process extensions of BiCMOS can be retrofitted into a CMOS process if sufficient economic incentive exists.

*Double-level Metal*

The early CMOS processes used one metal and one poly layer (a combination sometimes called *1½-level metal*). Polysilicon can be used to jumper most signals,

so routing is much easier than for single-level metal. Autorouting software still cannot efficiently route 1½-level metallization, so most modern CMOS processes add a second layer of metal. Analog CMOS layouts can benefit from using this second metal layer to increase packing density. Since planarization becomes more difficult as device sizes shrink. CMOS metallization tends to be substantially thinner than that used for standard bipolar. Higher-power CMOS circuits such as output drivers often benefit from combining multiple metal layers to form a thicker conductor.

Double-level metal adds two mask steps to the process: one for vias and one for metal-2. An *interlevel oxide* (ILO) deposited between the two metal layers provides insulation, and some form of planarization improves planarity for the second-level metal. Although these extra processing steps increase the cost of the wafer, most CMOS fabs routinely employ double-level metal for the majority of their products. Some processes use three, four, or even five metal layers to further reduce the area required for interconnection in high-density autorouted logic.

### *Silicidation*

CMOS processes make extensive use of silicides. In addition to contacts, gate poly is often silicided to reduce its sheet resistance. Some processes even silicide source and drain diffusions to reduce their parasitic resistance. Processes with submicron feature sizes may use all of these forms of silicidation. Older processes with larger feature sizes cannot construct sufficiently fast transistors to justify siliciding their gates and source/drain regions, but they usually employ silicided contacts to prevent contact spiking.

A Schottky diode can be formed on an N-well CMOS process that employs platinum or palladium silicides. The Schottky anode consists of a silicided contact while the cathode consists of N-well contacted by means of an NSD diffusion. The lack of a buried layer increases the internal resistance of this diode and prevents it from being used for high-current applications. The exposed edges of the silicide limit the operating voltage, but techniques exist for suppressing edge breakdown without creating any additional process steps (Section 10.1.3). CMOS processes do not require a Schottky contact mask since a moat region can serve the same function. Note that some silicides do not form usable Schottky barriers. For example, titanium silicide produces such a low forward voltage that the resulting Schottky diodes leak excessively.

Silicidation of the gate poly reduces its sheet resistance to about 2Ω/□, which is too low for constructing most resistors. Poly resistors can still be formed in a silicided-gate process by adding a *silicide block mask*. This mask patterns the silicide layer, either by preventing metal deposition over resistors or by allowing its removal in these areas prior to sintering. The use of a silicide block mask complicates the silicidation process, but is necessary for most analog designs.

Some processes also silicide NSD and PSD diffusions to form so-called *clad moats*. These cannot be used as resistors since their sheet resistances approximate that of their silicide layer (typically about 2Ω/□). A silicide block mask can prevent the silicidation of selected PSD and NSD regions to allow their use as resistors. The silicidation of emitter regions also reduces the beta of substrate PNP transistors (Section 8.3.3). A silicide block mask can sometimes improve substrate PNP beta to the point where this device becomes a practical component.

### *Lightly Doped Drain (LDD) Transistors*

If MOS transistors must operate at high drain-to-source voltages, then precautions become necessary to prevent hot carrier degradation. Typical NMOS transistors

with 3μm channel lengths can indefinitely withstand 5 to 10V, while PMOS transistors of the same dimensions can withstand 15 to 20V. Voltages beyond these limits require alternative structures.

The strong electric field across the pinched-off channel of a saturated MOS transistor causes hot carrier degradation. The electric field intensity diminishes if the depletion region can somehow be widened. In a conventional transistor, the depletion region cannot intrude to any significant extent into the heavily doped drain. If the drain diffusion is more lightly doped, then the depletion region can extend into the drain as well as into the channel and the electric field intensity will decrease. Such *lightly doped drain* (LDD) transistors can withstand substantially higher drain-to-source voltages than can conventional *singly doped drain* (SDD) devices.

LDD transistors actually use two drain diffusions, one forming a lightly doped *drift region* near the edge of the gate, and the other forming a more heavily doped region beneath the contact. The heavily doped diffusion reduces the drain resistance of the structure and allows the transistor to retain most of the performance of a conventional SDD device. One process for forming LDD transistors uses an *oxide sidewall spacer* to self-align the two drain diffusions, enabling precise control of the width of the drift region.[11] Figure 3.35 illustrates the steps required to fabricate this structure. Immediately after patterning the polysilicon gate, a shallow implant self-aligned to the edges of the gate polysilicon deposits the lightly doped drain. The wafer is coated with a thick layer of isotropically deposited oxide. The oxide at the edges of the polysilicon gate is deeper than the oxide over adjacent regions of the wafer. An anisotropic dry etch removes most of the deposited oxide, but some remains along the edges of the gates even after the planar surfaces have entirely cleared. The etch is halted so that filaments of oxide remain along either side of the gate; these form the desired oxide sidewall spacers. These spacers are approximately as wide as the polysilicon gate is thick. A second drain implant self-aligned to the edges of the oxide sidewall spacers forms the heavily doped portions of the LDD structure. The width of the lightly doped drift region approximately equals the width of the sidewall spacer, typically about 0.5μm.

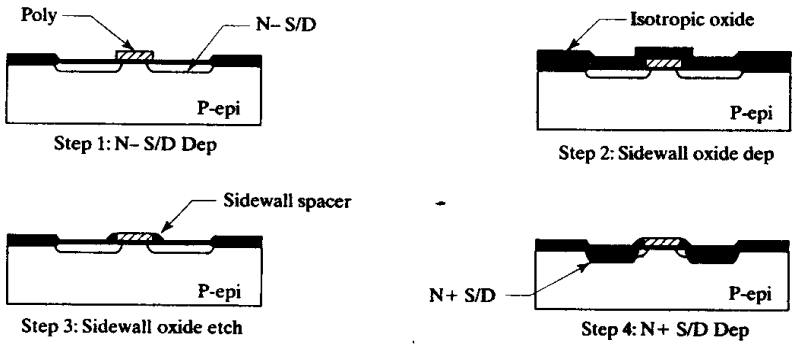


**FIGURE 3.35** Steps required to fabricate an LDD NMOS transistor.

Only the drain terminal of the NMOS transistor requires an LDD structure, but no simple way exists to block the formation of the sidewall spacer, so the source terminal of the NMOS also receives an LDD structure. The resulting transistor is *symmetric* in the sense that source and drain can be interchanged without affecting device performance. The PMOS transistor also receives the oxide sidewall spacers.

[11] R. H. Eklund, R. A. Haken, R. H. Havemann, and L. N. Hutter, "BiCMOS Process Technology," in A. R. Alvarez, ed., *BiCMOS Technology and Applications*, 2nd ed. (Boston: Kluwer Academic, 1993), pp. 90–95.

but no lightly doped diffusion. For reasons discussed in section 12.1.1, a channel forms underneath these sidewall spacers. The transistor, therefore, appears to have a slightly longer channel than its drawn dimensions suggest.

The LDD process described previously forms transistors that are suitable for drain-to-source voltages of 10 to 20V. No additional masks are required if all transistors receive both source/drain implants (N– S/D and N+ S/D). Purchasing an additional mask to selectively block the N– S/D implant may provide some additional benefits. Short-channel transistors do not require LDD processing because they break down by punchthrough before hot carrier generation becomes significant. The presence of N– S/D in short-channel NMOS transistors therefore serves no useful purpose. The transistors can pack more densely if the drawn channel dimensions can be reduced without causing premature punchthrough. One way to achieve this goal consists of selectively blocking the N– S/D implant from the short-channel devices. By leaving out the N– S/D, the drawn channel length can be decreased by 0.5 to 1.0μm without affecting the effective dimensions of the device. The purchase of separate N– S/D and N+ S/D masks can make a significant impact on high-density, low-voltage logic circuitry that does not require LDD transistors.

### *Extended-Drain, High-Voltage Transistors*

Oxide sidewall spacers allow the construction of fairly conventional MOS transistors that can withstand drain-to-source voltages of 10 to 20V. Higher voltages require a different approach to constructing the lightly doped drain region to protect against avalanche breakdown as well as hot carriers. Practical high-voltage MOS transistors can be constructed using only the existing masks of the standard N-well polysilicon-gate CMOS process. These *extended-drain* devices do not self-align, so they are inherently long-channel devices that exhibit substantial overlap capacitance. Even so, they suffice for constructing many high-voltage circuits.

Figure 3.36 shows a sample high-voltage, extended-drain NMOS. This transistor uses an N-well region as an extremely lightly doped drain. Since the well is both relatively deep and very lightly doped, it possesses a breakdown voltage in excess of 30V. A plug of NSD provides Ohmic contact to the drain. The source of the transistor consists of an NSD region without the addition of N-well. Since source and drain employ different diffusions, this device is an *asymmetric* MOS transistor.

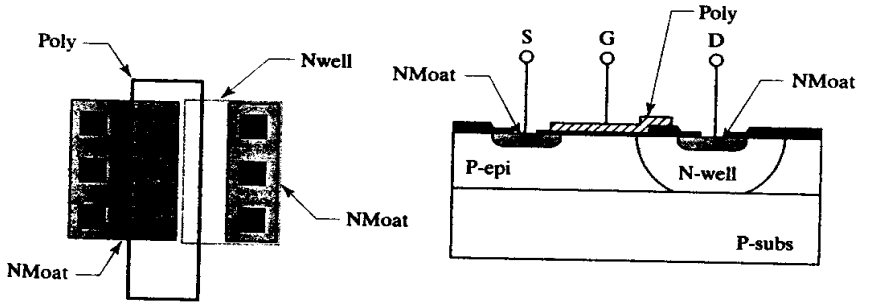


**FIGURE 3.36** Layout and cross section of an extended-drain NMOS. The backgate is common to the substrate.

The gate oxide of a high-voltage MOS transistor presents something of a dilemma. The 15V CMOS gate oxide is 300 to 400Å thick and can safely withstand 20 to 25V. Higher voltages will rupture the oxide and destroy the device. A separate gate oxidation can form a thicker oxide for the high-voltage devices, but increasing the oxide thickness decreases the device transconductance. The best solution consists of thickening the gate oxide only over the lightly doped drain where the highest field

stresses occur (Figure 3.36). The thin gate oxide remains in place over the channel and ensures a high device transconductance.

High-voltage, extended-drain PMOS transistors use the P-type channel-stop as a lightly doped drain. These devices are discussed in Section 12.1.2.

## 3.3 ANALOG BiCMOS

The incessant demand for higher levels of integration has driven the evolution of ever more complex and costly processes. Not only must more devices fit into the same die area, but the performance of these devices must steadily improve to satisfy new applications. By the early 1980s, customers demanded *mixed-signal* integrated circuits incorporating both analog and digital subsystems on a common substrate. A typical mixed-signal integrated circuit contains 90 to 95% digital circuitry and 5 to 10% analog circuitry.[12] CMOS logic overwhelmingly outperforms bipolar logic in packing density and power requirements, so the first attempts to fashion mixed-signal circuits employed unmodified CMOS processes. Analog CMOS circuitry had been designed for decades, so few manufacturers envisioned any difficulty in building the last percentages of the mixed-signal system. But these manufacturers soon discovered that difficulty does not correspond to component count. Although the analog components compose only a few percent of the total devices, they often consume the majority of the design effort. The inferior performance of analog CMOS requires even more design resources to compensate for process inadequacies.

After a few years of failures and qualified successes, most manufacturers began to realize that the analog portions of a mixed-signal system require tailor-made components. The true benefits of using such components lie not in improved performance, but in faster cycle times and higher probabilities of success. The late 1980s saw the development of a new generation of processes specifically aimed at the construction of mixed-signal integrated circuits. These *analog BiCMOS* processes are usually based on CMOS process flows, but are augmented by the addition of bipolar transistors, high-sheet poly resistors, and other special devices.

### 3.3.1. Essential Features

Analog BiCMOS processes are characterized by their complexity. Most require at least fifteen masks, and the more exotic variants use upward of thirty masks. The penalties of complexity include higher wafer costs, longer manufacturing times, and lower process yields. Set against these disadvantages are the benefits of higher-performance analog circuitry, reduced design effort, and faster design cycles. By the mid-1990s, the majority of new analog designs used some form of analog BiCMOS processing.

Because analog BiCMOS is still an evolving technology, the different manufacturers have not yet settled upon a common definition of the process. Most agree that practical BiCMOS must consist of a standard CMOS flow with the addition of a minimum number of steps to construct adequate bipolar transistors. Most also agree that deep P+ isolation is undesirable because it requires a prolonged high-temperature drive. One alternative form of isolation uses the N-well to form the collector of the NPN transistor. The base and emitter then consist of successive

[12] Reckoned by device count, not area. The relatively few analog devices on a mixed-mode integrated circuit tend to take up an inordinate amount of die area because analog circuitry cannot take full advantage of the reduced dimensions of modern transistors while retaining adequate performance.

counterdopings of the well. The collector-epi junction serves to isolate this style of bipolar transistor, thus the name *collector-diffused-isolation* (CDI).[13] Figure 3.37 shows an NPN transistor constructed using CDI.

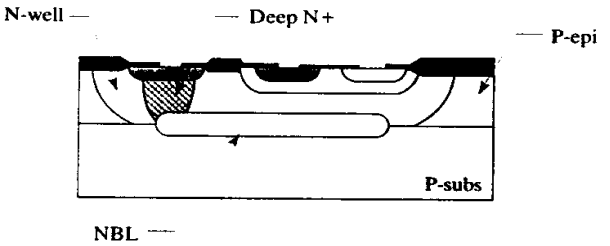


**FIGURE 3.37** Details of collector-diffused-isolation (for an NPN transistor). Only the portions of the CDI system related to the collector are labeled.

The specific steps required to construct bipolar transistors on a CMOS process have been hotly debated for some years. While no consensus has yet emerged, analysis shows that three masks are probably needed: NBL, deep-N+, and base.

NBL is the most reluctantly conceded mask, but also the most vital. Some CMOS fabs do not possess epi reactors and lack the experience necessary to implement buried layers. Predictably, attempts have been made to eliminate NBL, and equally predictably, the results have been unsatisfactory. NBL greatly reduces NPN collector resistance—one of the major parasitics of this device. NBL also provides higher NPN operating voltages on the thin epi favored by modern CMOS processing because it blocks vertical punchthrough breakdown. Furthermore, NBL suppresses parasitic substrate PNP action. Without it, lateral PNP transistors become almost impossible to construct. Any practical analog BiCMOS process will almost certainly include NBL.

Standard bipolar uses a deep-N+ sinker to further reduce the collector resistance of power NPN transistors. Although a power MOS transistor can substitute for a power NPN in many applications, deep-N+ remains necessary for creating fully effective minority carrier injection guard rings. Furthermore, due to the graded nature of the well, CDI NPN transistors often exhibit excessive vertical resistance between the collector contact and the NBL. These transistors saturate prematurely, limiting low-voltage operation, complicating device modeling, and causing undesired substrate injection. Most BiCMOS processes include deep-N+, if only as a process extension.

The base diffusion sets the NPN transistor gain, breakdown voltage, and Early voltage. Some processes have attempted to construct NPN transistors using layers scavenged from other devices, with mixed results. Attempts to construct NPN transistors using the diffusions that normally form DMOS transistors have succeeded in some processes but not in others. Questions still remain as to the suitability of the DMOS NPN for applications requiring matching or conducting large currents (Section 12.2.2). Extended-base transistors that use the P-epi or a shallow P-well as a base region have also been successfully constructed, but these devices have several drawbacks. They require more area than conventional CDI devices, especially for small devices, and they often have low Early voltages because they lack a drift region (see Section 8.3.2).

[13] B. T. Murphy, S. M. Neville, and R. A. Pedersen, "Simplified Bipolar Technology and Its Application to Systems," *IEEE J. Solid-State Circuits*, Vol. SC-5, #1, 1970, pp. 7–14.

### 3.3.2. Fabrication Sequence

This section discusses an analog BiCMOS process based on N-well polysilicon-gate CMOS.[14] The N-well provides collector-diffused isolation; NBL, deep-N+, and base are added to create bipolar transistors. Double-level metal has been added to simplify interconnection. This process requires fifteen masks, one of which is used twice for a total of sixteen masking operations.

#### *Starting Material*

The substrate material chosen for analog BiCMOS consists of a P+ (100) substrate cut several degrees off the crystal axis to minimize pattern distortion. The use of NBL in conjunction with a P+ substrate dictates the insertion of an additional epitaxial deposition into the process. Without this step, the NBL would directly contact the substrate to form an N+/P+ junction with a very low breakdown voltage. A lightly doped P-epi layer some 20μm thick therefore resides between substrate and NBL. Three factors determine the thickness of this first epi layer: the up-diffusion of the underlying substrate, the down-diffusion of the NBL, and the width of the depletion region required to withstand the maximum anticipated operating voltage (typically 30 to 50V). The first epi deposits on an unpatterned wafer, so epi-coated wafers may be stockpiled for use as starting material. One could alternatively sacrifice the P+ substrate to eliminate the need for a first epi deposition, but the use of a lightly doped substrate makes the process very susceptible to latchup and substrate debiasing (Section 4.4.1).

#### *N-buried Layer*

A brief thermal oxidation grows a thin layer of oxide across the wafer. This oxide is patterned using the N-buried layer (NBL) mask, and an oxide etch opens windows to the silicon surface. Ion implantation deposits an N-type dopant, either arsenic or antimony, in these windows. A brief drive conducted in an oxidizing ambient anneals out lattice damage and causes the formation of the surface discontinuity required for subsequent mask alignment.

#### *Epitaxial Growth*

After the NBL anneal, the oxide is stripped and the wafers are returned to the epitaxial reactor for deposition of a second P-epi layer. Surface discontinuities propagate upward through the epi at approximately a 45° angle along a [100] axis determined by the tilt of the wafer. After epitaxial growth, the NBL shadow will have shifted laterally a distance approximately equal to the thickness of the second epi, typically about 10μm. Figure 3.38 shows the wafer after the second epi deposition.

During the growth of the second epi, the reactant gases leach some of the NBL dopant from the surface of the wafer and redeposit it elsewhere, a process called *lateral autodoping*.[15] This mechanism can cause the formation of a thin layer of N-type silicon at the interface between the first and second epi layers, shorting adjacent wells. Autodoping can be limited by using antimony as the dopant or by conducting the epitaxy at reduced pressure. In either case, the BiCMOS NBL is liable to be somewhat more lightly doped than that used for standard bipolar.

---

[14] Eklund, *et al.*, pp. 120ff. Also see J. Erdeljac, B. Todd, L. Hutter, K. Wagensohner, and W. Bucksch, "A 2.0 micron BiCMOS Process Including DMOS Transistors for Merged Linear ASIC Analog/Digital/Power Applications," *Proceedings 1992 Applied Power Elect. Conf.*, 1992. pp. 517–522.

[15] M. W. M. Graef, B. J. H. Leunissen, and H. H. C. de Moor, "Antimony, Arsenic, Phosphorus, and Boron Autodoping in Silicon Epitaxy," *J. Electrochem. Soc.*, Vol. 132, #8, 1985, pp. 1942–1954.

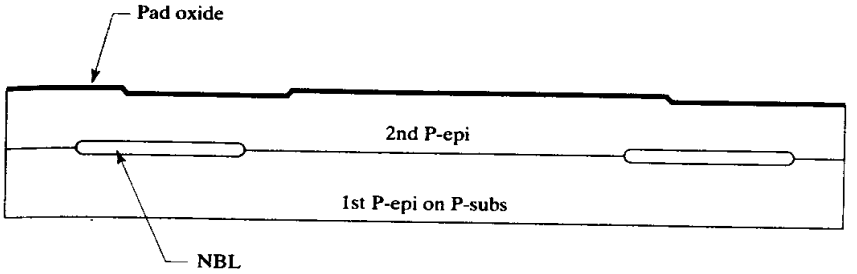


**FIGURE 3.38** Wafer after second epitaxial deposition. The P+ substrate is not explicitly shown, but it lies beneath the first P-epi layer.

### *N-well Diffusion and Deep-N+*

A thin layer of oxide is now grown and patterned using the N-well mask. Ion implantation deposits phosphorus, which is subsequently driven down to form the well diffusion. This drive stops before the well and NBL collide to permit the timely insertion of the deep-N+ deposition into the process flow. Additional oxide grown during the well drive allows patterning of the subsequent deep-N+ diffusion. After a heavy dose of phosphorus is implanted, the drive continues until the well and the deep-N+ both overlap the NBL by about 25% of their respective junction depths in order to minimize vertical resistance. Figure 3.39 shows a cross section of the wafer after the drive.

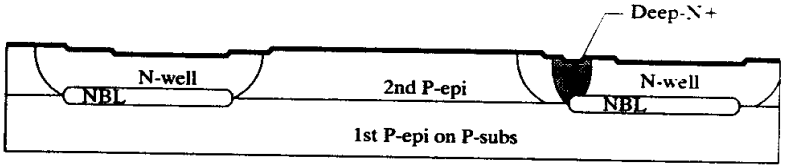


**FIGURE 3.39** Wafer after N-well and deep-N+ deposition and drive.

The N-well diffusion influences a number of device parameters for both PMOS and bipolar transistors. Compromises must be made between the two types of devices to the detriment of either or both. For example, short-channel PMOS transistors require a moderately doped well to suppress punchthrough while bipolar transistors need a lightly doped well to form their collector drift region. The doping of the well therefore targets a compromise value that dictates a minimum channel length of 2 to 3μm. If one desires to fabricate shorter channel lengths, then additional wells must be added to allow independent optimization of the MOS and bipolar components of the process.

### *Base Implant*

A uniform, thin layer of pad oxide is grown after the remnants of the previous oxide patterns have been stripped from the wafer. The wafer is patterned using the base mask, and boron is implanted through the pad oxide to form P-type regions, which are subsequently annealed under an inert ambient. Later high-temperature processing steps complete the base drive and set the final base-junction depth. Figure 3.40 shows the wafer after the base anneal. Triple counterdoping degrades the beta of the CDI NPN by raising the total base dopant concentration and hence the recombination rate in the neutral base. This can be partially offset by using a relatively lightly doped base implant. The final base-sheet resistance therefore equals several times that of standard bipolar.

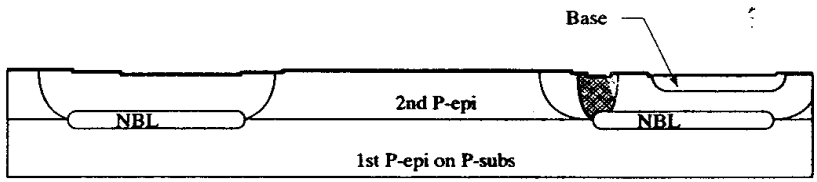


**FIGURE 3.40** Wafer after base implant and anneal.

### *Inverse Moat*

Analog BiCMOS uses the same LOCOS technology employed by polysilicon-gate CMOS. A thick layer of LPCVD nitride, patterned using the inverse moat mask, is etched to expose the regions where field oxide will eventually form (Figure 3.41). As in the case of polysilicon-gate CMOS, the MOS transistors occupy moat regions that are not covered by thick-field oxide. Moat regions also serve two additional purposes:

- Base regions are enclosed by moats to prevent oxidation-enhanced diffusion from overdriving the base.
- Schottky contacts are enclosed by moat to allow their etching to proceed simultaneously with base and emitter contacts.

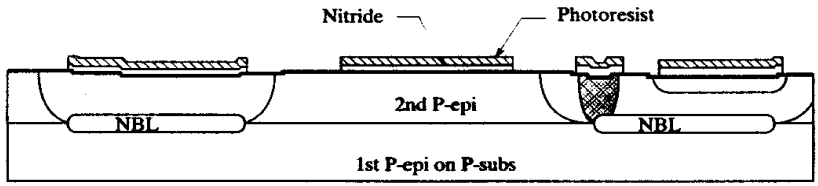


**FIGURE 3.41** Wafer after nitride deposition and inverse moat etch.

The inverse moat mask consists of a color reverse of a combination of NMoat, PMoat, base, and Schottky contact drawing layers. Some processes generate the inverse moat mask automatically during pattern generation; other processes require the layout designer to code moat geometries over some or all of the aforementioned drawing layers.

### *Channel Stop Implants*

Since analog BiCMOS uses (100) silicon, channel stop implants are required to raise the thick-field threshold above the operating voltage. The channel stop strategy for analog BiCMOS parallels that for polysilicon-gate CMOS. A blanket boron channel stop adjusts the thick-field threshold over the P-epi, while a patterned phosphorus channel stop sets the thick-field threshold over well regions. The boron channel stop is implanted using the patterned photoresist left from the inverse moat masking operation. A second coat of photoresist is applied and patterned, using the channel stop mask. This mask exposes only the regions of N-well that will ultimately lie beneath the thick-field oxide. A phosphorus implant offsets the previously deposited boron and increases the surface concentration in these well regions. Figure 3.42 illustrates the wafer cross section after the completion of both channel stop implants and the subsequent photoresist removal.

### *LOCOS Processing and Dummy Gate Oxidation*

The LOCOS oxidation employs either steam or elevated pressures to increase the rate of oxide growth. Afterward, the nitride layer and the underlying pad oxide are stripped away. A dummy gate oxidation removes any lingering nitride residue (Figure 3.43).

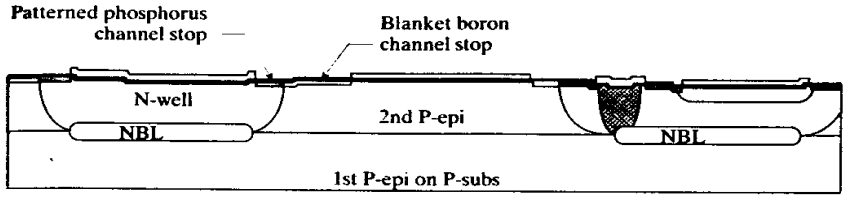


**FIGURE 3.42** Wafer after channel stop implants.

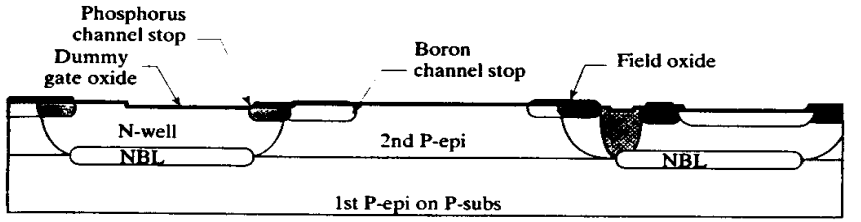


**FIGURE 3.43** Wafer after LOCOS oxidation, nitride strip, and dummy gate oxidation.

### *Threshold Adjust*

A single boron $V_t$ adjust implant simultaneously raises the NMOS threshold and lowers the PMOS threshold. With suitable well and epi doping concentrations, both of these thresholds can be simultaneously adjusted to the desired target of 0.7 ± 0.2V. These MOS threshold voltages approximately coincide with the base-emitter voltages of bipolar transistors, although differences in the underlying physics preclude any true matching between the two types of devices.

The threshold implant process consists of coating the wafer with photoresist, patterning the resist with the $V_t$ adjust mask, and implanting the necessary dose of boron through the dummy gate oxide. The final gate dielectric consists of some 300Å of high-quality dry oxide grown after the removal of the dummy gate oxide (Figure 3.43). Figure 3.44 shows a cross section of the resulting wafer.

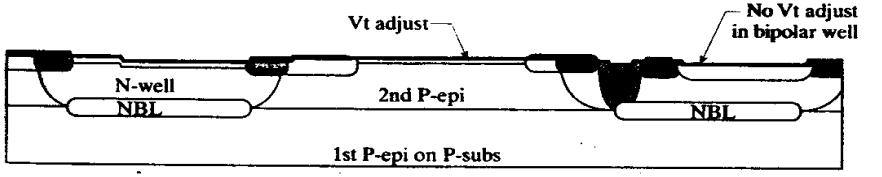


**FIGURE 3.44** Wafer after $V_t$ adjust implant.

### *Polysilicon Deposition and Pattern*

The gates of the MOS transistors consist of heavily doped N-type polysilicon formed by depositing intrinsic polysilicon and subsequently doping it with a blanket phosphorus deposition. The patterning step uses the poly-1 mask, so named because a second poly layer may be added as a process option. Figure 3.45 shows the wafer after poly-1 pattern.

### *Source/Drain Implants*

Analog BiCMOS typically produces bipolar transistors with 10 to 20V breakdown voltages. The MOS transistors should ideally be capable of withstanding similar voltages. The breakdown voltages of PSD and NSD can be raised to 15 to 20V without difficulty. Punchthrough can be averted by increasing the minimum channel

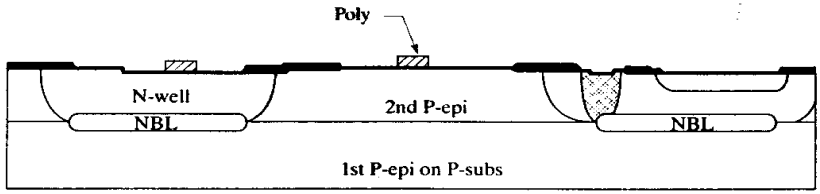


**FIGURE 3.45** Wafer after polysilicon deposition and pattern. The channel stop and threshold adjust implants do not appear in this or subsequent cross sections.

length to 2 to 3μm. The addition of a lightly doped drain will suppress hot electron generation in the NMOS transistors.

Photoresist is spun onto the wafer and patterned using the N– S/D mask. A phosphorus implant forms lightly doped source and drain regions that self-align to the polysilicon gate. Isotropic deposition of oxide and subsequent anisotropic etching form sidewall spacers on either side of the gates. A second photoresist layer, patterned using the N+ S/D mask, defines a heavier and somewhat deeper N+ S/D implant that aligns to the edges of the oxide sidewall spacers. The lightly doped drain consists of that portion of the N– S/D implant that lies beneath the sidewall spacers. If all NMOS transistors receive LDD structures, then the N– S/D mask can be reused for the N+ S/D implant.

A 10 to 20V PMOS transistor does not require a lightly doped drain, eliminating the need for a P– S/D implant. The P+ S/D implant occurs after the formation of the sidewall spacers, so the PMOS transistor channel length increases by twice the width of the sidewall spacer (Figure 3.46). This increase in width can be offset by reducing the drawn length of the PMOS gate proportionately.

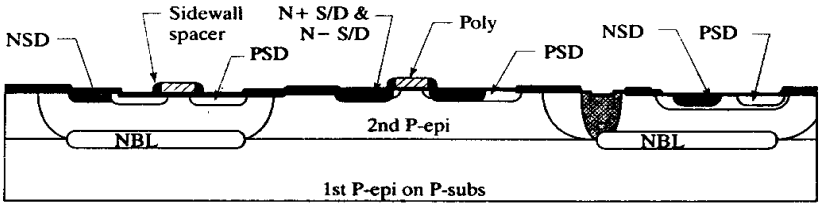


**FIGURE 3.46** Wafer after N– S/D, N+ S/D, and P+ S/D implants.

### *Metallization and Protective Overcoat*

The double-level metal flow requires five masks: contact, metal-1, via, metal-2, and protective overcoat. The contacts are silicided to control resistance and, coincidentally, to allow the formation of Schottky diodes. Refractory barrier metal ensures adequate step coverage across the nearly vertical sidewalls of contacts and vias. Copper-doped aluminum minimizes electromigration susceptibility, and a thick layer of compressive nitride protective overcoat provides a mechanical and chemical barrier between the metallization and the encapsulation. Figure 3.47 shows the cross section of a completed wafer, which includes NPN, NMOS, and PMOS transistors.

### *Process Comparison*

Analog BiCMOS uses all of the same steps as polysilicon-gate CMOS. Three additional masks (NBL, deep-N+, and base) are inserted at opportune points in the process as shown by the shaded entries in Table 3.6. NBL deposition must occur before the growth of the second epi. Deep-N+ and base must occur early in the process because these deep diffusions require high temperatures and long drive times.

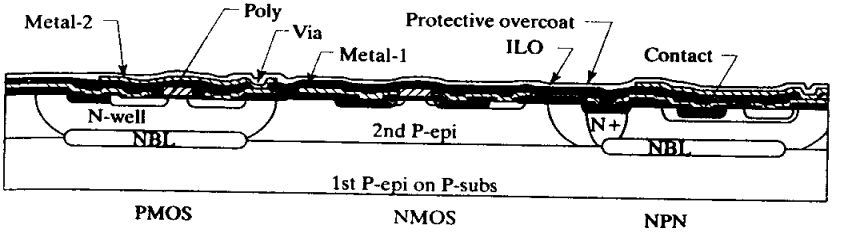


**FIGURE 3.47** Completed wafer, showing metal-1 and metal-2 layers.

| Mask | Polysilicon-gate CMOS | Analog BiCMOS |
|---|---|---|
| 1. NBL | | First epi growth<br>NBL deposition/anneal |
| | Epi growth | Second epi growth |
| 2. N-well | N-well deposition/drive | N-well deposition/drive |
| 3. Deep-N+ | | Deep-N+ deposition/drive |
| | Pad oxidation | Pad oxidation |
| 4. Base | | Base implant/anneal |
| 5. Inverse moat | Nitride deposition/pattern | Nitride deposition/pattern |
| | Blanket boron channel stop | Blanket boron channel stop |
| 6. Channel stop | Patterned phosphorus channel stop | Patterned phosphorus channel stop |
| | LOCOS oxidation | LOCOS oxidation |
| | Nitride/pad oxide strip | Nitride/pad oxide strip |
| | Dummy gate oxidation | Dummy gate oxidation |
| 7. $V_t$ Adjust | Threshold adjust implant | Threshold adjust implant |
| | True gate oxidation | True gate oxidation |
| 8. Poly-1 | Polysilicon deposition | Polysilicon deposition |
| | Poly implant/anneal | Poly implant/anneal |
| 9. NSD | N– S/D implant | N– S/D implant |
| | Sidewall spacer formation | Sidewall spacer formation |
| 10. NSD (again) | N+ S/D implant | N+ S/D implant |
| 11. PSD | P+ S/D implant | P+ S/D implant |
| | MLO deposition | MLO deposition |
| 12. Contact | Contact OR | Contact OR |
| | Platinum sputter/sinter/etch | Platinum sputter/sinter/etch |
| 13. Metal-1 | 1st metal deposition/etch | 1st metal deposition/etch |
| | ILO deposition/planarization | ILO deposition/planarization |
| 14. Via | Via etch | Via etch |
| 15. Metal-2 | 2nd metal deposition/etch | 2nd metal deposition/etch |
| 16. PO | PO deposition/etch | PO deposition/etch |

**TABLE 3.6** Comparison of analog BiCMOS and polysilicon-gate CMOS processes.

### 3.3.3. Available Devices

All of the devices available in polysilicon-gate CMOS also exist in analog BiCMOS. These devices include LDD NMOS and SDD PMOS transistors, four types of resistors (poly-1, PSD, NSD, and N-well), gate oxide capacitors, and Schottky diodes. Extended drain transistors can also be created without using any additional masks. Sections 3.2.3 and 3.2.4 discuss the preceding devices in detail. Additional analog

BiCMOS components include a CDI NPN transistor, lateral and substrate PNP transistors, and base resistors.

### *NPN Transistors*

Figure 3.48 shows the layout and cross section of a minimum-geometry NPN transistor. The collector of the NPN consists of an N-well region into which the base and emitter (consisting of NSD) are successively diffused. The inclusion of NBL beneath the active region of the transistor and the addition of a deep-N+ sinker help minimize collector resistance. NSD implanted on top of the sinker ensures Ohmic contact. In a similar manner, PSD allows contact to the lightly doped base. The general appearance of this transistor closely resembles that of the standard bipolar transistor in Figure 3.10. There are, however, several subtle differences.

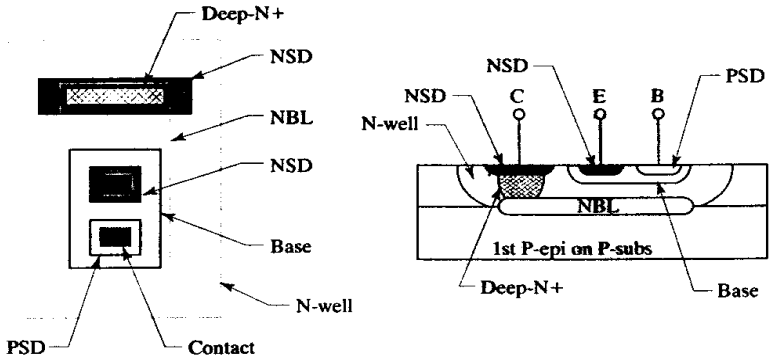


**FIGURE 3.48** Layout and cross section of an NPN transistor with deep-N+ and NBL.

The use of three successive counterdopings increases recombination in the neutral base and therefore decreases the gain of the BiCMOS NPN transistor. Lighter base doping partially compensates for this effect. Conventional circuit design techniques require a minimum beta of about fifty. Allowing for a tolerance of ±50%, the nominal beta target becomes seventy-five, which is only about half that offered by a standard bipolar NPN transistor. Higher betas could be achieved, but only at the expense of degrading other device characteristics.

The graded well exhibits an extremely high collector resistance if deep-N+ is not placed beneath the collector contact. This resistance causes a soft transition from the saturation to the forward active region, which resembles the effects of quasisaturation (Section 8.3.2). The transistor may also intrinsically saturate even when the terminal voltages seem to indicate that saturation cannot occur (Section 8.1.4). These problems can be prevented by adding a deep-N+ sinker. Transistors used as Zeners do not require deep-N+, even in analog BiCMOS, because the collectors of these devices do not conduct significant current.

The analog BiCMOS NPN transistor does not perform as well as the standard bipolar NPN transistor, but it still serves for most applications (Table 3.7). Analog BiCMOS also allows much smaller emitter areas than standard bipolar. This benefit does not translate into a proportional reduction in transistor area since many other spacings contribute to the size of the final device. Still, the minimum-geometry analog BiCMOS transistor requires only about 30% of the room of its standard bipolar counterpart.

### *PNP Transistors*

Analog BiCMOS supports both substrate and lateral PNP transistors. Figure 3.49 shows the layout and cross section of a representative substrate PNP transistor. This device is

| Parameter | Standard Bipolar | Analog BiCMOS |
|---|---|---|
| Drawn emitter area | $100\mu m^2$ | $16\mu m^2$ |
| Peak current gain (beta) | 150 | 80 |
| Early voltage | 120V | 120V |
| Collector resistance, in saturation | 100Ω | 200Ω |
| Typical operating current range for maximum beta | 5μA–2mA | 1–200μA |
| $V_{EBO}$ (Emitter-base breakdown, collector open) | 7V | 8V |
| $V_{CBO}$ (Collector-base breakdown, emitter open) | 60V | 50V |
| $V_{CEO}$ (Collector-emitter breakdown, base open) | 45V | 25V |

**TABLE 3.7** NPN device parameters for standard bipolar and analog BiCMOS.

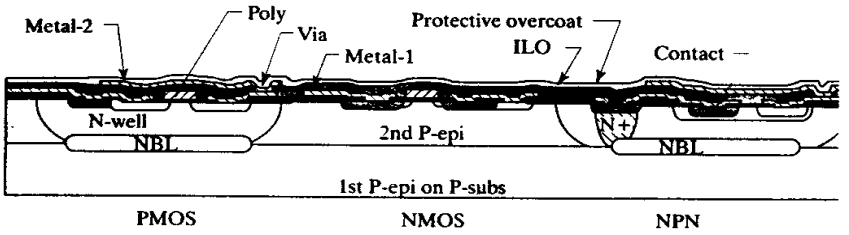


**FIGURE 3.49** Layout and cross section of a typical substrate PNP transistor. The substrate acts as the collector.

constructed by implanting PSD into an N-well region. The PSD implant forms the emitter of the transistor while N-well forms the base. A small NSD plug ensures Ohmic contact to the lightly doped N-well. The emitter of the substrate PNP cannot consist of base diffusion because this reaches so deeply into the well that it compromises the breakdown voltage of the transistor. The characteristics of the analog BiCMOS substrate PNP broadly resemble those of a substrate PNP formed in standard bipolar.

Figure 3.50 shows a layout and cross section of a lateral PNP transistor constructed in analog BiCMOS. This device employs the base diffusion to form the emitter and collector, which both reside in an N-well region that forms the base. The addition of NBL to the transistor serves several purposes. First, it acts as a depletion stop, which allows the lateral PNP to withstand higher operating voltages without suffering punchthrough. Second, NBL blocks punchthrough breakdown and allows the base implant to be used in place of the shallower PSD implant. The deeper base implant enhances emitter sidewall injection and therefore raises the beta of the device. NBL also helps to minimize substrate injection. The graded nature of the well reduces the effectiveness of the NBL as a minority carrier barrier, but substantial benefits remain. Without NBL, the lateral transistor would exhibit an apparent beta of less than 10; with NBL the beta can easily exceed 100.

Additional implants must surround the contacts since both the N-well and the base diffusion are too lightly doped to allow direct Ohmic contact. A rectangle of PSD encloses both the emitter and the collector, but this implant only penetrates the moat regions that form the actual collector and emitter regions of the device. The thick LOCOS oxide blocks the PSD implant from reaching the base of the transistor and prevents the P-type implant from shorting the emitter and collector. Similarly, an NSD plug allows Ohmic contact to the N-well. Deep-N+ is not normally

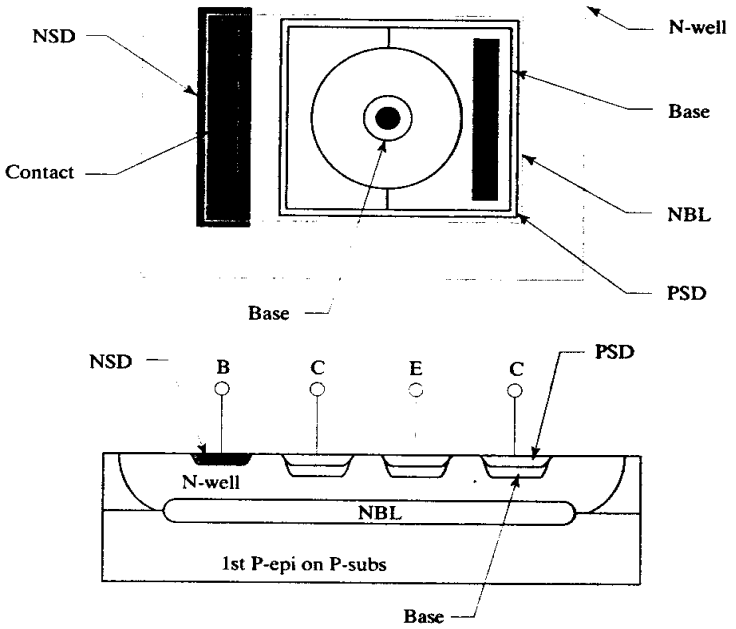


**FIGURE 3.50** Layout and cross section of a lateral PNP transistor.

required, although it may become necessary in larger transistors that conduct significant base current.

The minimum-geometry lateral PNP transistor can attain a peak beta well in excess of 100. Fine-line photolithography allows a narrower base width than standard bipolar and greatly reduces the area of a minimum emitter. As the emitter area decreases, the ratio of periphery to area increases. This enhances the desired lateral injection of carriers at the expense of undesired vertical injection. The graded nature of the well and the presence of the channel stop implant generate a doping gradient that forces carriers away from the surface, diminishing surface recombination losses. The use of (100) silicon instead of (111) silicon also reduces surface recombination by minimizing the surface state charge. All of these effects together produce a transistor having a beta of up to 500.

The use of a lightly doped base implant for the emitter reduces emitter injection efficiency and causes a corresponding reduction in beta. Large PNP transistors consist of arrays of many minimum emitters to achieve a high area-to-periphery ratio. Table 3.8 lists several important parameters for both types of PNP transistors available in analog BiCMOS.

**TABLE 3.8** Typical PNP device parameters for analog BiCMOS.

| Parameter | Lateral PNP | Substrate PNP |
|---|---|---|
| Drawn emitter area | 16μm² | 16μm² |
| Drawn base width | 5μm | — |
| Peak current gain (beta) | 120 | 100 |
| Early voltage | 80V | 100V |
| Typical operating current for maximum beta | 1–20μA | 1–50μA |
| $V_{EBO}$ (Emitter-base breakdown, collector open) | 45V | 45V |
| $V_{CBO}$ (Collector-base breakdown, emitter open) | 45V | 45V |
| $V_{CEO}$ (Collector-emitter breakdown, base open) | 30V | 45V |

### *Resistors*

Analog BiCMOS base resistors consist of rectangles of base material occupying an N-well. Contact is made to either end of the resistor through a PSD plug. The well contact contains an NSD implant to increase the surface doping of the N-well. As with base resistors in standard bipolar, the addition of NBL serves not only to block possible minority carrier injection into the well during transients, but also to raise the operating voltage by preventing punchthrough between the resistor and the underlying substrate. Figure 3.51 shows the cross section and layout of a typical base resistor. The analog BiCMOS base diffusion is relatively lightly doped and typically exhibits a sheet resistance of 500Ω/□. While the high sheet resistance provides compactness, it also complicates resistor layout because the diffusion becomes vulnerable to surface depletion effects (Section 5.3.3). Few designers actually employ base resistors; most prefer to pay the small additional cost for a process extension for high-sheet polysilicon resistors. Pinch resistors are also available, but they offer little or no advantage over minimum-width poly resistors.

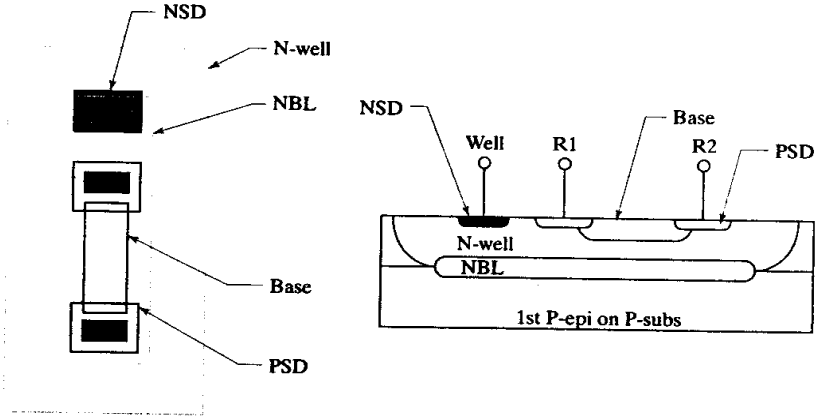


**FIGURE 3.51** Layout and cross section of a base resistor.

## 3.4 SUMMARY

This chapter has examined three representative processes in detail: the standard bipolar process used to construct early analog integrated circuits, a polysilicon-gate CMOS process favored for digital logic, and an analog BiCMOS process, which merges the best of these two technologies onto a common substrate. Variations of these three processes fabricate the bulk of the low-cost, high-volume analog integrated circuits available today.

Standard bipolar is the oldest of the many available bipolar processes, most of which offer much faster switching speeds. The newer processes often replace junction isolation with partial or complete dielectric isolation to simultaneously reduce component area and minimize parasitic capacitance. Diffusions consist of much thinner layers, often created through innovative use of polysilicon as a doping source. Self-alignment and improved dimensional control allow the construction of far smaller geometries. These improvements have increased switching speeds by two orders of magnitude, allowing modern bipolar transistors to operate at speeds approaching fifty gigahertz.[16] The highly specialized processes used to obtain such performance are fully as complex and costly as the most sophisticated of CMOS processes.

[16] These speeds are only obtained using specialized processing techniques, such as the formation of silicon-germanium layers within the base of the transistors.

CMOS processes have undergone their share of evolutionary advances. Gate lengths have shrunk relentlessly as digital designers have sought to pack ever larger numbers of transistors onto limited amounts of silicon real estate. Once gate lengths fall below 1μm, a variety of short-channel effects dictate the use of much more elaborate structures. The backgate doping levels must increase to combat premature punchthrough and elaborate implantation strategies become necessary to confine the channel to the desired region immediately beneath the gate oxide. Cost and complexity have soared as processes push steadily deeper into the submicron regime seeking the elusive ultimate limits of device performance.

BiCMOS technologies seek to merge the best of both worlds. They are therefore heir to both the complexities of submicron CMOS and the elaborate bipolar structures required to obtain high switching speeds. Almost any bipolar or CMOS innovation can be merged into a BiCMOS process, and most have been.

## 3.5 EXERCISES

Refer to Appendix C for layout rules and process specifications.

**3.1.** Lay out the NPN transistor shown in Figure 3.10. Use a minimum-size square emitter and a minimum contact overlap of the deep-N+ sinker. Allow room for all necessary metallization.

**3.2.** Draw a cross section of the NPN transistor shown in Exercise 3.1 to scale, assuming an epi thickness of 10μm, NBL up-diffusion from the epi-substrate junction of 3μm. NBL down-diffusion from the epi-substrate junction of 4μm, an isolation junction depth of 12μm, a deep-N+ junction depth of 9μm, a base junction depth of 2μm, and an emitter junction depth of 1μm. Assume 80% outdiffusion where necessary. Oxide nonplanarities and deposited layers need not be shown.

**3.3.** Lay out the lateral PNP transistor shown in Figure 3.12. Use the minimum possible basewidth. Allow room for all necessary metallization, including a circular metal field plate connecting to the emitter contact and overhanging the inner edge of the collector region by 2μm.

**3.4.** Lay out a 500Ω base resistor following the example in Figure 3.13. Make the base resistor 8μm wide, and widen the contacts as much as the width of the resistor allows.

**3.5.** Lay out a 25kΩ base pinch resistor following the example in Figure 3.15. Assume that all of the resistance comes from the portion of the base beneath the emitter plate. Make the base strip 8μm wide and overlap the emitter plate over the base strip by at least 6μm. NBL should overlap the base strip in the pinched region by at least 2μm.

**3.6.** Lay out a fingered junction capacitor similar to the one shown in Figure 3.16. Make each of the three emitter fingers 50μm long. The emitter plate should overlap the base by at least 6μm; minimize all other dimensions.

**3.7.** Lay out the Schottky diode shown in Figure 3.18, assuming a contact opening that is 25 by 25μm. Overlap metal over the Schottky contact layer (SCONT) by no less than 4μm.

**3.8.** Lay out a 20kΩ HSR resistor. Make the width of the HSR resistor 8μm. The contacts should have the same width as the HSR resistor body. Assume that the base heads contribute negligible resistance and compute the value of the resistor based only on the drawn length of the HSR segment between the base heads.

**3.9.** Lay out an NMOS transistor with a drawn width of 10μm and a drawn length of 4μm following the example in Figure 3.29. Allow room for all necessary metallization.

**3.10.** Draw a cross section of the NMOS shown in Exercise 3.9 to scale, assuming a well junction depth of 6μm, PSD and NSD junction depths of 1μm, a gate oxide thickness of 350Å (0.035μm), and a polysilicon thickness of 3kÅ. Ignore the $V_t$ adjust and the channel stop implants. Assume 80% outdiffusion where necessary. Assume the silicon surface is planar, and ignore details of the metallization system.

**3.11.** Lay out a PMOS transistor with a drawn width of 7μm and a drawn length of 15μm following the example in Figure 3.30. Assume NBL is not used. Include all necessary metallization.

**3.12.** Lay out a PSD resistor with a value of 200Ω following the example in Figure 3.33B. Make the resistor minimum width, and abut the well contact with one end of the resistor to save space.

**3.13.** Lay out a 3pF poly capacitor following the example in Figure 3.34. Include all necessary metallization. Assume that contacts and vias can reside on top of the poly plate.

**3.14.** Lay out the BiCMOS NPN transistor shown in Figure 3.48. The NBL should overlap the base region by at least 2μm. Use minimum emitter dimensions and include all necessary metallization.

**3.15.** Draw a cross section of the NPN from Exercise 3.14 to proper scale, assuming the dimensions given in Exercise 3.10. Assume NBL diffuses upward by 3μm and downward by 2μm. In addition, assume a first epi thickness of 7μm, a deep-N+ junction depth of 5μm, and a base junction depth of 1.5μm. Ignore channel-stop implants and the effects of LOCOS field oxidation on surface planarity. Assume that the silicon surface is planar, and ignore the details of the metallization system.

**3.16.** Lay out the BiCMOS lateral PNP transistor shown in Figure 3.50. Assume a minimum basewidth. Unlike the device from Exercise 3.3, this transistor does not require a metal field plate over the base region. NBL should overlap the outer edge of the collector by at least 1.0μm. Include all necessary metallization.

**3.17.** Lay out the resistor-transistor logic NOR gate shown in Figure 3.52A using standard bipolar layout rules. Place $Q_1$ and $Q_2$ in the same tank, and use as small a plug of deep-N+ as possible to contact the collector of this tank. Assume the emitters of $Q_1$ and $Q_2$ are both minimum-size. Place $R_1$ in its own tank. Provide at least one substrate contact. Surround this contact with base: this base region may touch, but not extend into adjacent tanks. Label all inputs and outputs.

**3.18.** Lay out the CMOS NOR gate shown in Figure 3.52B using poly-gate CMOS layout rules. The *W* and *L* values for each transistor are shown on the schematic in the form of a fraction; 5/3 indicates a drawn width of 5μm and a drawn length of 3μm. Place all PMOS transistors in the same well, and connect this well to $V_{DD}$. Provide at least one substrate contact. Bring all inputs and outputs up to second-level metal, and label them appropriately.

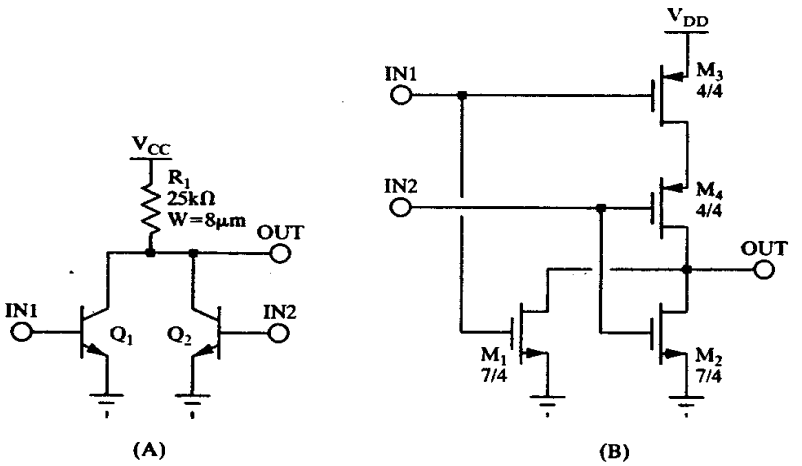


**FIGURE 3.52** Circuits for Exercises 3.17 and 3.18.