4 Failure Mechanisms

Integrated circuits are incredibly complex devices, and few of them are perfect. Most contain subtle weaknesses and flaws, which predispose them toward eventual failure. Such components can fail catastrophically and without warning after operating perfectly for many years. Engineers have traditionally relied on quality assurance programs to uncover hidden design flaws. Operation under stressful conditions can accelerate many failure mechanisms, but not every design flaw can be found by testing. The designer must therefore find and eliminate as many of these flaws as possible.

The layout of an integrated circuit contributes to many types of failures. If the designer knows about potential weaknesses, then safeguards can be built into the integrated circuit to protect it against failure. This chapter discusses a number of failure mechanisms that can be partially or entirely prevented by layout precautions.

4.1 ELECTRICAL OVERSTRESS

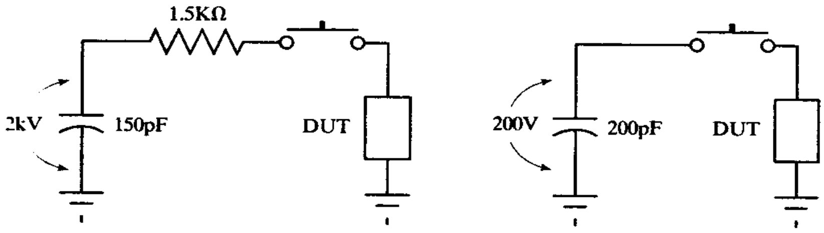
The term *electrical overstress* (EOS) refers to failures caused by the application of excessive voltages or currents to a component. Layout precautions can minimize the probability of three common types of EOS failures. *Electrostatic discharge* (ESD) is a form of electrical overstress caused by static electricity. The addition of special protective structures to vulnerable bondpads can minimize ESD failures. *Electromigration* is a slow wearout mechanism caused by excessive current densities; it can eventually cause open circuits or shorts between adjacent leads. Electromigration failures can be prevented by making leads wide enough to handle the maximum operating currents. The *antenna effect* is an unusual failure mechanism caused by charge accumulation on gate electrodes during etching or ion implantation. The problems posed by the antenna effect can be minimized by following a few simple design guidelines.

4.1.1. Electrostatic Discharge (ESD)

Almost any form of friction can generate static electricity. For example, if you shuffle across a carpet in dry weather and then touch a metal doorknob, a visible spark will leap from finger to doorknob. The human body acts as a capacitor, and the act of shuffling across a carpet charges this capacitance to a potential of 10,000V or more. When a finger is brought near the doorknob, the sudden discharge creates a visible spark and a perceptible electrical shock. A discharge of as little as 50V will destroy the gate dielectric of a typical integrated MOS transistor. Voltages this low produce neither visible sparks nor perceptible electrical shock. Almost any human or mechanical activity can produce such low-level electrostatic discharges.

Proper handling precautions will minimize the risks of electrostatic discharge. ESD-sensitive components (including integrated circuits) should always be stored in static-shielded packaging. Grounded wrist straps and soldering irons can reduce potential opportunities for ESD discharges. Humidifiers, ionizers, and antistatic mats can minimize the buildup of static charges around workstations and machinery. These precautions reduce but do not eliminate ESD damage, so manufacturers routinely include special ESD protection structures on-board integrated circuits. These structures are designed to absorb and dissipate moderate levels of ESD energy without damage.

Special tests can measure the vulnerability of an integrated circuit to ESD. The two most common test circuits are called the human body model and the machine model.[1] The ***human body model*** (HBM) employs the circuit shown in Figure 4.1A. When the switch is pressed, a 150pF capacitor charged to a specified voltage discharges through the integrated circuit to ground. A 1.5kΩ series resistor limits the peak current through the part. Ideally each pair of pins would be independently tested for ESD susceptibility, but most testing regimens only specify a limited number of pin combinations to reduce test time. Each pair of pins is subjected to a series of positive and negative pulses; for example, five positive and five negative. Modern integrated circuits are routinely expected to survive 2kV HBM. Specific pins on certain parts may be required to survive up to 25kV HBM.



(A) Human body model

(B) Machine model

FIGURE 4.1 Representative ESD tests: 2kV human body model (A) and 200V machine model (B). In both circuits, DUT stands for *device under test.*

Figure 4.1B shows the circuit employed for the *machine model* (MM). A 200pF capacitor charged to a specified voltage discharges through the integrated circuit to ground. The test circuit contains no intentional series resistance, but practical testers incorporate some 20 to 40Ω in wiring, switches, sockets, and so forth. This, together with the series inductance of the wiring, limits the peak currents generated during testing. The machine model forms a much harsher test than the human body model; few parts can survive more than 500V under these conditions.

A third ESD test called the *charged device model* (CDM) is gradually replacing the machine model. The charged device model charges the integrated circuit package to a specified static potential and then discharges one pin to a low-impedance ground. Researchers believe that this model provides more accurate modeling of

[1] R. Lahri, J. Shibley, D. Merrill, H. Wang, and B. Bastani, "Process Reliability," in A. R. Alvarez, ed., *BiCMOS Technology and Applications,* 2nd ed. (Boston: Kluwer Academic, 1993), p. 170ff.

factory handling conditions than either the human body or the machine model. A typical testing regimen will specify 1 to 1.5kV CDM testing.

Effects

Electrostatic discharge causes several different forms of electrical damage, including gate oxide rupture, gate oxide degradation, and avalanche-induced junction leakage.[2] In extreme cases, ESD discharges can even vaporize metallization or shatter the bulk silicon.

Less than 50V can rupture the gate dielectric of a typical MOS transistor. The rupture occurs in nanoseconds, requires little or no sustained current flow, and is for all intents and purposes irreversible. The rupture typically shorts the gate and the backgate of the damaged transistor.[3] Capacitors that use oxide or nitride dielectrics are also vulnerable to this failure mechanism. An ESD discharge that strikes a pin connecting only to gates or to capacitors may cause oxide rupture in these devices. If the pin connects to any diffusions, then these usually avalanche before the gate oxide ruptures.

The integrity of a gate oxide can be compromised by an ESD event that does not actually rupture it. The weakened oxide can then fail at any time, perhaps after hundreds or thousands of hours of flawless operation. Sometimes the failure does not occur until the product has been delivered to the customer. Testing cannot screen out this type of delayed ESD failure; instead, vulnerable dielectrics must be protected against excessive voltages.

Although junctions are considerably more robust than dielectrics, they can still suffer ESD damage. An avalanching junction dumps a large amount of energy into a small volume of silicon. Extreme current densities can sweep metallization through contacts to short out underlying junctions. Excessive heating can also physically damage junctions by melting or shattering the silicon. These catastrophic forms of junction damage most often manifest themselves as short circuits. Avalanched junctions that do not fail outright usually exhibit increased leakage that may or may not cause the overstressed unit to fail parametric testing.

Preventative Measures

All vulnerable pins must have ESD protection structures connected to their bondpads. Some pins can resist ESD and therefore do not require additional protection. Examples include pins connected to substrate and to large diffusions, such as the collectors of power NPN transistors. These large junctions disperse and absorb the ESD energy before it can damage other circuitry. The power supply pins of most integrated circuits connect to a multitude of diffusions, and thus are also quite robust.

Pins connecting to relatively small diffusions, particularly those connected to the bases or emitters of small NPN transistors, are vulnerable to ESD-induced junction damage. Avalanching the base-emitter junction of an NPN transistor permanently degrades its beta. A circuit designer can sometimes eliminate the vulnerable junctions by rearranging the circuit. ESD protection should be added to any pin connecting to a base-emitter junction, or more generally, to any pin connecting to a relatively small diffusion. Because ESD vulnerabilities are difficult to predict, cautious designers add protection devices to all pins that might be even remotely vulnerable.

[2] Some of these are discussed in A. Amerasekera, W. van den Abeelen, L. van Roozendaal, M. Hannemann, and P. Schofield, "ESD Failure Modes: Characteristics, Mechanisms, and Process Influences," *IEEE Trans. Electron Devices*, Vol. 39, #2, 1992, pp. 430–436.

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

Pins that connect only to gates of MOS transistors or to deposited capacitor electrodes are extremely vulnerable to ESD-induced dielectric rupture. Special gate protection structures should be placed on all such pins. These structures usually include a significant amount of series resistance (typically 500Ω to 5kΩ) and therefore cannot be used on pins that must conduct more than a fraction of a milliamp.

Processes that employ thin emitter oxides are also susceptible to ESD-induced rupture. This vulnerability can be eliminated by ensuring that leads that connect to external bondpads do not cross any emitter region to which they do not connect. Alternatively, ESD structures similar to those used for protecting gates can protect the vulnerable bondpads. Most modern versions of the standard bipolar process employ thick emitter oxides, which eliminate the need for these precautions.[4]

Considerable ingenuity is often required to formulate successful ESD structures for analog integrated circuits. A dozen or more protection circuits are often required to satisfy the large range of voltages and the several types of vulnerable devices found in analog circuits. Section 13.4.3 discusses several commonly employed ESD structures and shows how these can be modified to meet a variety of special requirements.

4.1.2. Electromigration

Electromigration is a slow wearout phenomenon caused by extremely high current densities. The impact of moving carriers with stationary metal atoms causes a gradual displacement of the metal. In aluminum, electromigration only becomes a concern when current densities approach $5 \cdot 10^5$A/cm^2. Although this may seem a tremendous current density, a minimum-width lead in a submicron process can experience electromigration at currents of only a few milliamps.[5]

Effects

Despite its homogenous appearance, aluminum metal is a polycrystalline material. The individual crystals, or *grains*, normally abut one another. Electromigration causes metal atoms to gradually move away from the grain boundaries, forming voids between adjacent grains. This causes a decrease in the lead's effective cross-sectional area and raises the current density seen by the remainder of the lead. Additional voids form and gradually coalesce until they ultimately sever the lead.

The addition of refractory barrier metal changes the observed modes of failure. Since the refractory metal is relatively resistive, most of the current initially flows through the aluminum. Once voiding finally severs the aluminum, the underlying refractory metal bridges the gap and continues to conduct current. Refractory metals strongly resist the effects of electromigration, so the lead will not completely fail. Instead, the formation of voids in the aluminum causes the lead's resistance to gradually and somewhat erratically increase. More ominously, aluminum metal displaced by voiding sometimes shorts adjacent leads together. The cross-sectional area of the aluminum portion of a lead therefore determines how much current it can safely conduct, regardless of the presence or absence of refractory barrier metal.

Refractory barrier metal is often used to prevent electromigration failures in contacts and vias because it ensures electrical continuity across steep sidewalls after the thin aluminum metallization at these points succumbs to electromigration-induced voiding. Lateral extrusion does not normally occur in contacts or vias since

[4] Even thick emitter oxides can rupture under certain conditions; see "Dielectric Breakdown of Emitter Oxide," *Semiconductor Reliability News*, Vol. IV, #1, 1992, p. 1.

[5] Assuming a lead width of one micron and a thickness of 5000Å, a current of 2.5mA will produce a current density of $5 \cdot 10^5$A/cm^2.

a contiguous sheet of metal covers the entire structure. Likewise, resistance changes caused by voiding are usually small compared to the inherent resistance of the contact or via structure. Where these resistance changes cannot be tolerated, the designer can insert additional contacts or vias to help reduce the current density.

Preventative Measures

The first line of defense against electromigration consists of process improvements. Aluminum metallization is now routinely doped with 0.5 to 4% copper to enhance electromigration resistance.[6] Copper accumulates at the grain boundaries, where it inhibits voiding by increasing the activation energy required to dislodge metal atoms from the lattice. Copper-doped aluminum exhibits five to ten times the current handling capability of pure aluminum.[7] The electromigration resistance of leads can be further improved by using compressively stressed protective overcoats that confine the metal under pressure and inhibit void formation. Refractory barrier metal can also help prevent electromigration failures in contacts and vias. Most manufacturers do not rely upon refractory metal to protect other oxide steps because of the risk of lateral extrusion. Instead, the leads are designed so that the aluminum portion of the metallization can handle the full current over all portions of the metal pattern except at contact and via openings.

Processing techniques can minimize electromigration, but there remains some maximum current density that cannot be exceeded without risking eventual metallization failure. The design rules for each process thus define a maximum allowed current per unit width. Typical values are 2mA/μm (50mA/mil) for leads that do not cross oxide steps and 1mA/μm (25mA/mil) for those that do. These values depend on the thickness of the metallization and its composition, and on the anticipated operating temperature (Section 14.3.3). Consider a lead that must conduct 50mA following the electromigration limits specified above. If this lead routes across field oxide in order to avoid oxide steps, then it need be only 25μm wide; otherwise its width must increase to 50μm. The lead cannot widen abruptly at the oxide steps because the current only gradually flows out from a narrow lead into a wider one. The wider lead should extend beyond the oxide step in either direction for a distance at least twice its greatest width.

Excessive current can also cause bondwires to overheat and fail. In practice, a typical one-mil gold bondwire that is 50mil (1.25mm) in length can safely conduct about an amp, while a similar aluminum wire can conduct about 750mA. If the anticipated currents exceed these limits, then the design will require larger-diameter bondwires or multiple bondwires placed in parallel (Section 14.3.3).

4.1.3. The Antenna Effect

Dry etching uses intense electrical fields to generate an ionizing plasma. During the etching of the gate poly and the oxide sidewall spacers, electrostatic charges may accumulate on the gate poly. The resulting voltages may become so large that current flows through the gate oxide. Although the amount of energy involved is usually inadequate to rupture the gate oxide, it still degrades its dielectric strength. The amount of degradation is proportional to the total charge that passes through the gate oxide divided by the total gate oxide area (Section 11.1.2). Each poly region collects an electrostatic charge proportional to its own area. A small gate oxide

[6] I. Ames, F. M. d'Heurle, and R. E. Horstmann, "Reduction of Electromigration in Aluminum Films by Copper Doping," *IBM J. of Research and Development*, Vol. 14, #4, 1970, pp. 461–463.

[7] S. S. Iyer and C. Y. Ting, "Electromigration Study of Al-Cu/Ti/Al-Cu System," *Proc. International Reliability Physics Symp.*, 1984, pp. 273–278. See also Lahri, *et al.*, p. 166.

region connected to a large poly geometry can suffer disproportionate damage. This mechanism is sometimes called the *antenna effect* because the large poly area acts as an antenna to collect the charge that flows through the vulnerable gate oxide. Gate oxide damage due to the antenna effect has also been observed during ion implantation of the source/drain regions.[8,9]

The magnitude of the antenna effect is proportional to the ratio between exposed conductor area and gate oxide area. During the patterning of the polysilicon, the poly is the exposed conductor. Similarly, during the patterning of the first level of metal, the metal is the exposed conductor. Separate area ratios must be computed for each conductor layer. One also computes separate ratios for PMOS and NMOS gate oxides because the two may not break down at exactly the same voltage. Conductor/gate area ratios of several hundred are usually required to produce significant damage. Most layouts contain few geometries with ratios this large, so antenna-effect damage is usually limited to a few locations on the die. Figure 4.2A shows an example of a layout that can produce a conductor-gate area ratio large enough to trigger this type of failure. The gate lead of NMOS transistor M_1 has been elongated to facilitate connection to transistor M_2. The elongated lead has sufficient area to endanger the small gate oxide region of transistor M_1. This vulnerability can be eliminated by inserting a metal jumper in the poly lead next to transistor M_1 (Figure 4.2B). This jumper drastically reduces the area of the poly geometry connected to M_1's gate oxide, which in turn reduces the conductor/gate area ratio to a safe value.

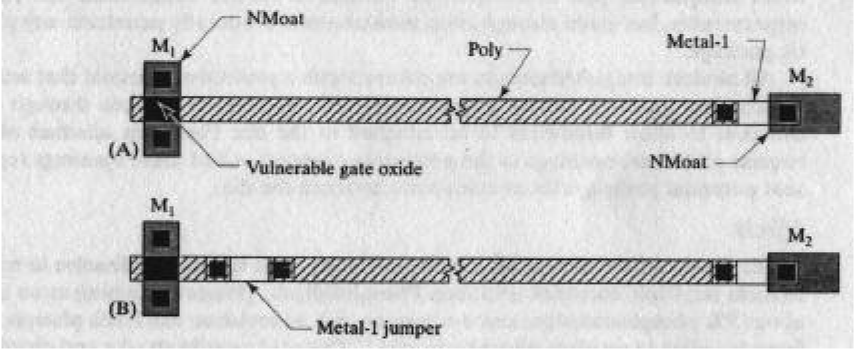


FIGURE 4.2 A layout susceptible to the antenna effect (A) can be made immune by the addition of a metal jumper (B).

Electrostatic damage can also occur during the etching of metal layers. Metal regions connecting to diffusions rarely pose any problem because these diffusions provide paths through which the electrostatic charges can leak away. The topmost layer of metal is almost immune to antenna effects because every geometry on this layer connects to a diffusion somewhere on the die, but the lower metal layers do not necessarily connect to diffusions until the top metal layer is in place. During the etching of a lower metal layer, any geometry that does not connect to a diffusion through layers already present can potentially collect a damaging electrostatic charge.

Antenna effects on lower metal layers can be eliminated by inserting short jumpers on the top metal layer to minimize the conductor area of the lower metal

[8] T. Watanabe and Y. Yoshida, "Dielectric Breakdown of Gate Insulation Due to Reactive Ion Etching," *Solid State Technology*, Vol. 27, #4, 1984, pp. 263–266.

[9] K. Markus, C. M. Osburn, P. Magill, and S. M. Bobbio, "Thin-oxide Degradation Along Feature Edges During Reactive Ion Etching of Polysilicon Gates," *J. Vac. Sci. Technol. A*, Vol. 12, #4, 1994, pp. 1339–1345.

layer attached to small gate oxide regions. This solution resembles that shown in Figure 4.2B. In cases where top-metal jumpers are not feasible, damage can still be avoided by ensuring that the lower-metal lead connects to a diffusion. If the layout does not include any conveniently located diffusion, consider adding a minimum-size NMoat/P-epi or PMoat/N-well diode (Section 10.2). These diodes may affect circuit operation, so they must not be added without consulting the circuit designer.

4.2 CONTAMINATION

Integrated circuits are vulnerable to certain types of contaminants. Assuming that the device has been properly manufactured, very low levels of contaminants will initially exist inside the plastic encapsulation. Plastic mold compounds have been carefully formulated to provide the highest possible degree of resistance to penetration by external contaminants, but no plastic is entirely impregnable. Contaminants seep in along the interface between the metal pins and the plastic, or they directly penetrate the plastic itself. Two major contamination issues faced by modern plastic-encapsulated dice are *dry corrosion* and *mobile ion contamination.*

4.2.1. Dry Corrosion

The aluminum metal system will corrode if exposed to ionic contaminants in the presence of moisture. Only trace amounts of water are necessary to initiate this so-called *dry corrosion.* Since moisture and ionic contaminants are both ubiquitous, integrated circuits must depend on their encapsulation to protect them. Early mold compounds had little moisture resistance. Newer compounds are more impermeable, but given enough time, moisture will eventually penetrate any plastic package.[10]

All modern integrated circuits are covered with a protective overcoat that acts as a secondary moisture barrier. Unfortunately, openings must be made through this overcoat to allow bondwires to be attached to the die. Fuse trim schemes often require additional openings in the protective overcoat. All of these openings represent potential pathways for contaminants to reach the die.

Effects

Water alone cannot corrode aluminum, but many ionic substances dissolve in water to form relatively corrosive solutions. Phosphosilicate glasses containing more than about 5% phosphorus represent a corrosion risk, as moisture can leach phosphorus from the glass to produce phosphoric acid.[11] This acid rapidly attacks and dissolves aluminum, causing open circuit failures. Many modern processes use nitride or oxynitride protective overcoats to ensure that moisture cannot reach the phosphosilicate glass that lies beneath. Alternatively, the phosphorus content of the glass can be reduced by using a combination of boron and phosphorus as dopants. Both of these elements reduce the softening point of a glass, so a *borophosphosilicate glass* (BPSG) will require less phosphorus to achieve the same softening point as a phosphosilicate glass.

Halogen ions in water solution can also corrode aluminum.[12] Common salt, or sodium chloride, provides an abundant source of chloride ions. Moisture seeping

[10] J. E. Gunn and S. K. Malik, "Highly Accelerated Temperature and Humidity Stress Technique (HAST)," *Reliability Physics, 19th Annual Proc.*, 1981, pp. 48–51.

[11] W. M. Paulson and R. W. Kirk, "The Effects of Phosphorus-Doped Passivation Glass on the Corrosion of Aluminum," *Proc. Reliability Physics Symposium*, 1974, pp. 172–179.

[12] M. M. Ianuzzi, "Reliability and Failure Mechanisms of Non-hermetic Aluminum SIC's in an Environment Contaminated with Cl_2," *(sic), IEEE Trans. Comp. Hyb. Man. Tech.*, 6, 1983, pp. 191–201.

into the package of an integrated circuit can transport chloride ions to the surface of the die where they can begin to corrode the aluminum metal system. Significant levels of bromides do not normally occur in the environment, but plastic encapsulants often contain organobromine flame retardants. These flame retardants begin to decompose and release bromide ions at temperatures in excess of about 250°C, which limits the storage and soldering temperatures these packages can safely withstand.[13]

Preventative Measures

Although contamination may seem completely beyond the control of the layout designer, several measures can be taken to minimize vulnerabilities in the protective overcoat. The designer should minimize the number and size of all PO openings. A production die should not include any openings that are not absolutely necessary for its manufacture. If the designer wishes to include additional testpads for evaluation, then these should occupy a special test mask. When the part is released to production, the test mask should be replaced by a production PO mask that seals the test pads under protective overcoat.

Metal should overlap bondpad openings on all sides by an amount sufficient to account for misalignment. The metal bondpads will then protect the underlying oxide from the entry of moisture and other contaminants. Openings made for polysilicon or metal fuses should be made as small as possible, and no circuitry of any sort except the fuse element itself should appear within the opening.

4.2.2. Mobile Ion Contamination

Many potential contaminants dissolve in silicon dioxide at elevated temperatures, but most lose their mobility at normal operating temperatures because they become bound into the oxide macromolecule. The alkali metals are exceptions to this rule and remain mobile in silicon dioxide even at room temperature.[14] Of these so-called *mobile ions*, sodium is by far the most common and the most troublesome.

Effects

Mobile ion contamination induces parametric shifts, most noticeably in MOS transistor threshold voltages. Figure 4.3A shows the gate oxide of an NMOS transistor contaminated by sodium during manufacture. The positively charged sodium ions

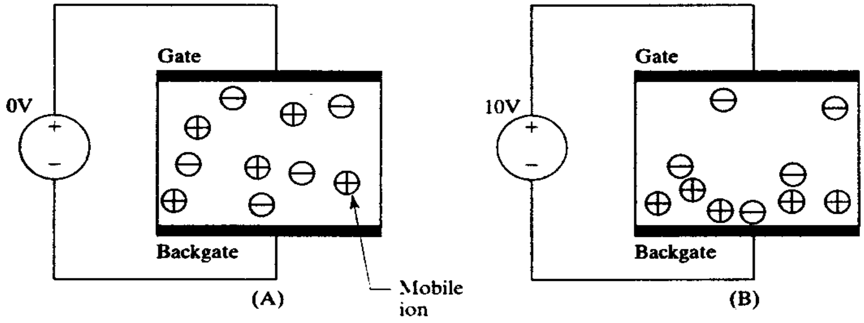


FIGURE 4.3 Behavior of mobile ions under bias: ions that were randomly distributed through the oxide (A) shift in unison in response to a positive gate bias (B).

[13] T. Raymond. "Avoiding Bond Pad Failure Mechanisms in Au-Al Systems," *Semiconductor International*, Sept. 1989, pp. 152–158.

[14] Actually, only lithium, sodium, and potassium (and perhaps hydrogen) qualify as mobile ions in silicon. The heavier alkali metals rubidium and cesium are far less mobile: B. E. Deal, "The Current Understanding of Charges in the Thermally Oxidized Silicon Structure," *J. Electrochem. Soc.*, Vol. 121, #6, 1974, pp. 198C–205C.

are initially distributed throughout the oxide. An equal number of negatively charged ions (anions) are also introduced. Unlike the sodium ions, these anions remain rigidly locked into the oxide macromolecule.[15]

Figure 4.3B shows the same gate dielectric after an extended period of operation under a positive gate bias. The positively charged gate electrode has repelled the mobile sodium ions down toward the oxide-silicon interface. Since the negative ions do not move, the redistribution of the sodium ions results in a net separation of charges within the oxide.[16] The presence of positive charges near the channel of the NMOS transistor decreases its threshold voltage. The magnitude of the threshold voltage shift depends on sodium ion concentration, gate bias, temperature, and time. Many analog circuits require that threshold voltages match within a few millivolts. Even low concentrations of mobile ions can produce shifts of this magnitude.

Mobile ion contamination can produce long-term failures when the slow drift of threshold voltages eventually causes a circuit to exceed its parametric limits. If the faulty devices are removed from operation and are baked at 200°C for a few hours, the mobile ions redistribute and the threshold shifts vanish. This treatment is only temporary; the threshold drift returns as soon as electrical bias is restored. Although analog circuits are particularly susceptible to parametric shifts caused by mobile ions, even digital circuits will eventually fail if the threshold voltages shift too far. Early metal-gate CMOS logic was plagued by threshold voltage shifts caused by severe sodium contamination.

Preventative Measures

Some mobile ions inevitably become incorporated in an integrated circuit during manufacture. This source of contamination can be minimized by using purer chemicals and improved processing techniques. MOS processes typically take extraordinary steps to ensure process cleanliness, but these alone cannot entirely eliminate threshold voltage variations.

Manufacturers of metal gate CMOS attempted to stabilize threshold voltages by adding phosphorus to the gate oxide.[17,18] Phosphorus stabilization had the desired effect of immobilizing alkali metal contaminants, but it also introduced a new problem. The electrically charged phosphate groups shift slightly under strong electrical fields even though they are bound to the oxide macromolecule. Phosphosilicate glasses therefore exhibit the same problem that they were intended to cure! All is not lost, though, because this *dielectric polarization* is not as severe a problem as mobile ion contamination. The threshold shift caused by a given voltage bias remains relatively small—a few tens of millivolts. The threshold shifts caused by dielectric polarization are also much more predictable than those produced by mobile ions, so circuit designers can predict whether a given circuit configuration will be adversely affected or not.[19] The threshold voltage shifts were finally eliminated altogether by using phosphorus-doped polysilicon gates rather than phosphorus-doped gate oxides. Phosphorus-doped polysilicon immobilizes

[15] Negative countercharges may not always exist, particularly if the contaminants enter the oxide after manufacture. Threshold shifts still occur because of image effects produced by the presence of electrical charges within the insulating oxide.

[16] N. E. Lycoudes and C. C. Childers, "Semiconductor Instability Failure Mechanisms Review," *IEEE Trans. of Reliability*, Vol. R-29, #3, 1980, pp. 237–249.

[17] M. Kuhn and D. J. Silversmith, "Ionic Contamination and Transport of Mobile Ions in MOS Structures," *J. Electrochem. Soc.*, Vol. 118, 1971, pp. 966–970.

[18] S. R. Hofstein, "Stabilization of MOS Devices," *Solid-State Electronics*, Vol. 10, 1967, pp. 657–670.

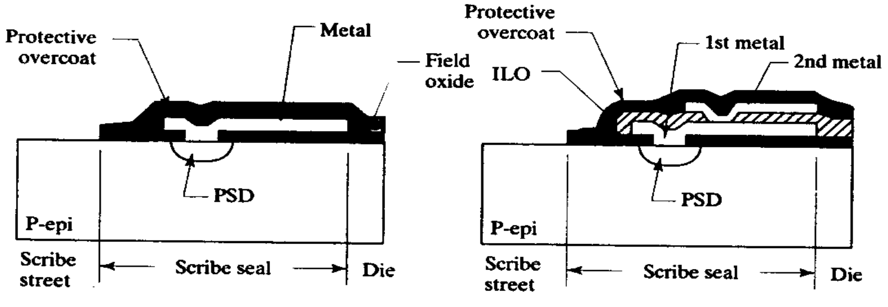
[19] E. H. Snow and B. E. Deal, "Polarization Phenomena and Other Properties of Phosphosilicate Glass Films on Silicon," *J. Electrochem. Soc.*, Vol. 113, #3, 1966, pp. 263–269.

alkali metals in much the same way as phosphorus stabilization without the added complication of dielectric polarization.

Moisture seeping into the integrated circuit's package can transport sodium in from the outside environment. Improved packaging materials can slow, but not stop, the ingress of sodium ions. The protective overcoat serves as a further barrier to mobile ions and can prevent them from reaching the vulnerable oxide layers in contact with the silicon. Protective overcoats typically consist of either silicon nitride, which is relatively impermeable to mobile ions, or phosphorus-doped glasses, which can immobilize them. The protective overcoat therefore serves as a final line of defense against impurities entering the die from outside.

Any opening through the protective overcoat represents a potential route for mobile ion contamination to enter the die. The metallization normally seals the bondpad openings, but scars left by probe needles can puncture the metal and expose the interlevel oxide (ILO) beneath. A minimum number of probe pads should be used, and these should be placed around the periphery of the die as far from sensitive analog circuitry as possible. Fuse openings through the protective overcoat also represent vulnerable points that should be kept far away from analog circuitry.

The scribe street surrounding the die typically consists of bare silicon because other materials either fracture or clog the saw blade. Contaminants can seep laterally into exposed oxide layers abutting the scribe street. Special structures, called *scribe seals*, placed around the periphery of the die can slow the ingress of contaminants. Figure 4.4A illustrates a typical scribe seal for a single-level-metal CMOS process. The first component of this scribe seal consists of a narrow contact strip surrounding the active area of the die. This contact must be a continuous ring uninterrupted by any gaps in order for it to block the lateral movement of mobile ions through the field oxide. A P-type diffusion placed underneath this contact allows it to double as a substrate contact. This arrangement is very convenient, as the metal plate forming part of the scribe seal also carries the substrate lead around the periphery of the die. The scribe seal also provides a guaranteed minimum area of substrate contacts.



(A) Single-level Metal

(B) Double-level Metal

FIGURE 4.4 Scribe seals for single- and double-level-metal variants of a CMOS or BiCMOS process. Depending on the manufacturer, various diffusions may also be placed over the scribe street.

The scribe seal also contains a second contamination barrier formed by flapping the protective overcoat into the scribe street directly on top of the exposed silicon. Any mobile ions attempting to penetrate the scribe seal must first surmount this flap-down and next pass the continuous contact ring before reaching the active regions of the die. Most processes prohibit direct contact between nitride and silicon because compressive stresses in the nitride spawn defects in the silicon lattice. The flap-down of protective overcoat over the scribe street is permitted because the scribe street does not contain any active circuitry that could be damaged by defects. Nitride should still not touch exposed silicon inside the active area of the die because defects spawned by the damaged silicon can propagate for some distance and may affect adjacent components.

Figure 4.4B shows a scribe seal for a double-level-metal CMOS process. This seal includes a third barrier consisting of a continuous via ring placed just inside the contact ring. This via ring helps prevent contaminants from entering the ILO between the two metal layers. In a triple-level-metal process, a second via ring would be added to protect the second ILO layer between metal-2 and metal-3.

The scribe seals shown in Figure 4.4 can protect almost any die, but the substrate contacts may require different diffusions depending on the process flow. For example, standard bipolar would substitute a combination of P-isolation and base for the PSD rings of Figure 4.4. The P-isolation would probably extend to the edge of the active die because most designers surround the die with a ring of substrate contacts placed underneath the grounded metallization. The functionality of the scribe seals remains the same regardless of the exact diffusions used.

4.3 SURFACE EFFECTS

Surface regions of high electric field intensity can inject hot carriers into the overlying oxide. Surface electric fields can also induce the formation of parasitic channels. Both of these mechanisms are referred to as *surface effects* because they occur at the interface between the silicon and the overlying oxide.

4.3.1. Hot Carrier Injection

Carriers are always in constant motion due to the random thermal vibrations responsible for their diffusion. Electric fields may produce a slow drift of carriers in one direction, but the resulting drift velocity is usually much smaller than the instantaneous velocities produced by thermal agitation. Consequently, electric fields rarely increase the instantaneous velocities of carriers by any perceptible amount (Figure 4.5A). Only at extremely high electric field intensities does the drift of carriers become so rapid that the instantaneous velocities actually increase (Figure 4.5B). The resulting carriers are called *hot carriers* because they move at speeds normally achieved only at elevated temperatures.

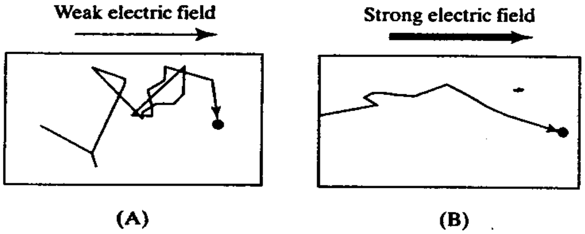


FIGURE 4.5 A weak electric field causes an overall drift of carriers but does not materially affect their instantaneous velocity (A), while a strong electric field actually increases the instantaneous velocity of the carriers (B).

Effects

MOS devices can generate hot carriers when operated in saturation at high drain-to-source voltages. As the drain-to-source voltage increases, the pinched-off portion of the channel slowly grows wider. The increase in width cannot keep pace with the applied voltage, so the electric field intensifies as the voltage increases. At high voltages, the electric field becomes large enough to generate hot carriers near the drain end of the transistor (Figure 4.6). NMOS transistors generate hot electrons, while

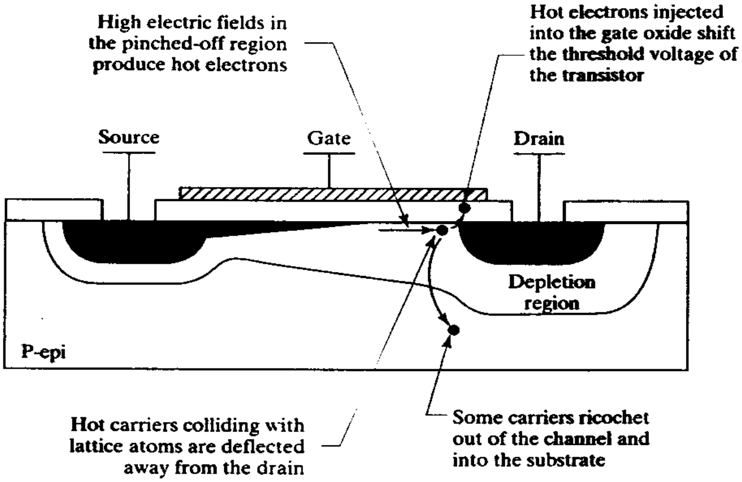

High electric fields in the pinched-off region produce hot electrons
Hot electrons injected into the gate oxide shift the threshold voltage of the transistor
Source
Gate
Drain
Depletion region
P-epi
Hot carriers colliding with lattice atoms are deflected away from the drain
Some carriers ricochet out of the channel and into the substrate

FIGURE 4.6 Simplified diagram showing the mechanism responsible for hot electron injection in an NMOS transistor.

PMOS transistors generate hot holes. Because of differences in effective mobilities, hot carrier production begins at substantially lower voltages in NMOS transistors than in PMOS transistors of similar dimensions. For example, if a 3μm NMOS experiences hot electron injection at 10V, then the equivalent PMOS would probably not experience significant hot hole injection below 20V.

Hot carriers produced at the drain end of the transistor collide with lattice atoms, and some few of the recoiling carriers travel upward into the overlying oxide. Most of these carriers pass through the oxide and return to the silicon, but a few become trapped at defect sites within the oxide. These trapped carriers represent a fixed oxide charge that gradually increases in magnitude as more carriers become trapped. This charge shifts the threshold voltage of the MOS transistor and can in turn affect the performance of the overall circuit.

Parametric shifts caused by hot carriers can be partially or completely reversed by baking the unbiased units at temperatures of 200 to 250°C for several hours. These temperatures impart sufficient thermal energy to the trapped carriers to free them and allow them to return to the silicon. The parametric shifts vanish as the fixed oxide charge dissipates. As in the case of mobile ions, the apparent cure is only temporary. As soon as bias is restored. hot carrier generation resumes and the threshold voltages begin to drift again.

Avalanching junctions also produce large numbers of hot carriers. Avalanche occurs near the surface in most diffused junctions because the dopant concentrations are highest there (Figure 4.7A). Some of the hot carriers produced by the avalanche process travel into the overlying oxide. In the case of a base-emitter junction, these carriers predominantly consist of hot holes that add to the positive fixed oxide charge.[20] As this charge increases, it induces a gradual widening of the depletion region at the surface (Figure 4.7B). The avalanche voltage slowly increases during

[20] G. Blasquez. G. Barbottin, and V. Boisson, "A Review of Passivation-Related Instabilities in Modern Silicon Devices," in G. Barbottin and A. Vapaille, eds., *Instabilities in Silicon Devices, Volume 2: Silicon Passivation and Related Instabilities* (Amsterdam: North-Holland, 1989). pp. 459–460. Blasquez, *et al.* state that Zener walkout in P+/N–junctions spontaneously reverses because some or all of the hot electron charge is not permanently trapped in the oxide macromolecule and can consequently dissipate over time even at room temperature. This effect is usually not observed in N+/P–junctions, presumably because the charge consists largely of holes.

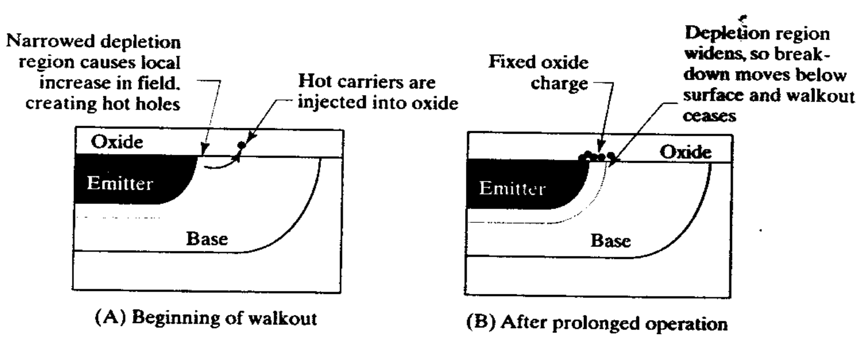

Narrowed depletion region causes local increase in field, creating hot holes
Hot carriers are injected into oxide
Oxide
Emitter
Base
(A) Beginning of walkout
Fixed oxide charge
Depletion region widens, so breakdown moves below surface and walkout ceases
Oxide
Emitter
Base
(B) After prolonged operation

FIGURE 4.7 Simplified diagrams showing Zener walkout mechanism: (A) initial condition of junction, in which hot carrier production occurs near the surface; (B) condition of junction after extended period of operation.

operation, a phenomenon called *Zener walk-out.*[21,22] If a junction diode's reverse breakdown is observed using a curve tracer, the knee of the breakdown curve will gradually "walk out" to higher and higher voltages due to the gradual widening of the depletion region at the surface in response to the accumulation of a fixed oxide charge. Since trapped oxide charges cause walk out, an unbiased high-temperature bake will at least partially reverse it. Depending on processing conditions, emitter-base Zeners can exhibit up to 200mV of walkout.[23] Experimental evidence suggests that the magnitude of Zener walkout diminishes when the process incorporates refractory barrier metal and silicided contacts,[24] although the mechanism responsible for this improvement is not apparent.

Preventative Measures

A *lightly doped drain* (LDD) structure can reduce or even eliminate hot carrier generation in a MOS transistor. Section 3.2.4 discusses the implementation of lightly doped drain structures in a typical polysilicon-gate CMOS process. If no lightly doped drain structure exists, or if the operating voltage exceeds the capabilities of the available structure, then the circuit must be redesigned to reduce the electrical stress imposed on the MOS transistors.

Transistors used as switches generate relatively few hot carriers. Such devices operate either fully on, in which case they are in the linear region, or fully off, in which case they are in cutoff. In neither case does current flow across a large drain-to-source voltage differential. Hot carriers are only generated during brief switching transitions between the two operating states. The average rate of hot carrier generation drops to a minute fraction of the value associated with continuous conduction, and the operating lifetime of the part increases by orders of magnitude. Transistors can withstand voltages far beyond the onset of hot carrier generation as long as switching transitions remain infrequent.

Long channel devices also gain some measure of protection against hot carrier effects. Hot carriers are still produced, but only in the vicinity of the drain. The rest of the channel remains unaffected, minimizing the overall impact of hot carriers on transistor parameters. A few extra volts of operating margin can often be obtained by increasing the channel length a few microns.

[21] J. F. Verwey, J. H. Aalberts, and B. J. de Maagt, "Drift of the Breakdown Voltage in Highly Doped Planar Junctions," *Microelectronics and Reliability*, Vol. 12, 1973, pp. 51–56.

[22] R. W. Gurtler, "Avalanche Drift Instability in Planar Passivated p-n Junctions," *IEEE Trans. on Electron Devices*, Vol. ED-15, #12, 1968, pp. 980–986.

[23] W. Bucksch, "Quality and Reliability in Linear Bipolar Design," *TI Technical Journal*, Nov. 1987, pp. 61–69.

[24] W. Bucksch, unpublished manuscript, 1988.

Ordinary base-emitter Zener diodes are surface devices and can therefore exhibit as much as several hundred millivolts of Zener walkout. Attempts have been made to minimize walkout in surface Zeners, but none have been notably successful. The Zener voltage can only be stabilized if the avalanche breakdown is confined to a subsurface region in order to keep hot carriers away from the vulnerable oxide-silicon interface. Such structures are usually called *buried Zeners* (Section 10.1.2).

4.3.2. Parasitic Channels and Charge Spreading

Any conductor placed above the silicon surface can potentially induce a *parasitic channel.* If the conductor bridges two diffusions, then a leakage current can flow through the channel from one diffusion to the other. Most parasitic channels are relatively long and cannot conduct much current, but even small currents can cause parametric shifts in low-power analog circuitry. Channels can sometimes form even in the absence of a conductor due to a mechanism called *charge spreading.* The addition of channel stops or field plates can suppress parasitic channel formation and so protect vulnerable circuitry.

Effects

Both PMOS and NMOS parasitic channels exist. A PMOS parasitic channel can form across any lightly doped N-type region. such as an N-tank in a standard bipolar process or an N-well in a CMOS or BiCMOS process. An NMOS parasitic channel can form across any lightly doped P-type region, such as the P-epi of a CMOS or BiCMOS process. or the lightly doped P-type isolation of standard bipolar processes. Both of these types of parasitic channels can cause a great deal of trouble.

PMOS parasitic channels can form underneath leads crossing lightly doped N-type regions. Consider a metal lead that crosses an N-tank containing a base diffusion (Figure 4.8A). The lead acts as the gate of a PMOS transistor and the N-tank as its backgate. The base region forms the source of the transistor and the isolation serves as its drain. A channel will form if the voltage difference between the lead and the base region exceeds the threshold voltage of the parasitic MOS transistor.[25] Since the thick-field oxide serves as the gate dielectric of this transistor, its threshold voltage is called the *PMOS thick-field threshold.* If the process has a 40V PMOS thick-field threshold, then the base must be biased at least 40V above the lead in order for a channel to form beneath the lead. A similar condition applies to any other potential parasitic PMOS: the P-type region serving as the source must rise

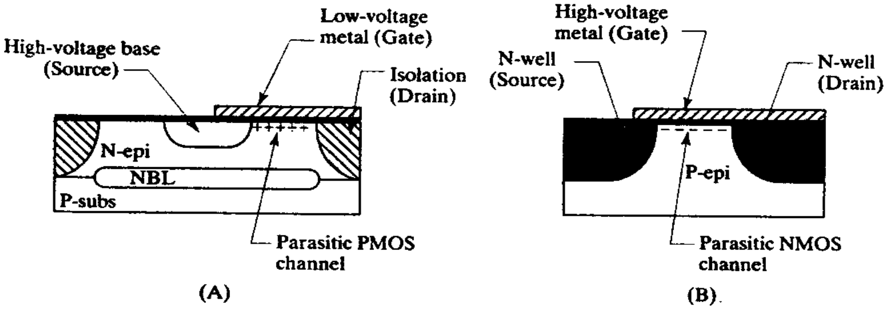


FIGURE 4.8 Parasitic PMOS in a standard bipolar process (A) and parasitic NMOS in an N-well CMOS process (B).

[25] "Bipolar Field Inversion," *Semiconductor Reliability News,* Vol. 3, #1, 1991, p. 7.

above the conductor acting as the gate by a voltage in excess of the PMOS thick-field threshold.

NMOS parasitic channels can form underneath leads crossing lightly doped P-type regions. Figure 4.8B shows a parasitic NMOS channel forming on an N-well CMOS die. This channel forms beneath the lead crossing the lightly doped P-epi. The lead acts as the gate and the P-epi as the backgate. Two adjacent wells serve as the source and the drain. A channel will form if the voltage difference between the gate and the source exceeds the NMOS thick-field threshold. In this case, the voltage on the lead must exceed the voltage on the N-well acting as the source by an amount equal to or greater than the NMOS thick-field threshold. Similar conditions apply to any other potential parasitic NMOS: the conductor serving as the gate must rise above the N-type region serving as the source by a voltage equal to or greater than the NMOS thick-field threshold.

The thick-field threshold voltages of a process depend on a number of factors, including conductor material, oxide thickness, substrate crystal orientation, doping levels, and processing conditions. Most processes quote only one value for the thick-field threshold, this being a minimum value obtained from a worst-case combination of conductors and diffusions. Other processes have undergone more extensive characterization to determine separate thick-field voltages for each combination of conductor and diffusion.

Designers sometimes invoke the body effect (Section 1.4.2) as justification for approaching or even exceeding the thick-field threshold. The body effect increases the apparent threshold voltage of the transistor when the backgate-source junction is reverse-biased. For example, the backgate of the parasitic PMOS in Figure 4.8A is probably biased to a higher voltage than the base. Unfortunately, backgate biasing cannot be relied on for any significant aid. The body effect is most significant in heavily doped backgates, whereas the backgate of a parasitic MOS is usually rather lightly doped. Furthermore, the threshold shift produced by the body effect varies as the square root of the backgate-to-source bias, so even a large backgate bias may not buy more than a few volts of margin.

Engineers once believed that channels could only form beneath conductors, but experience has shown otherwise. Channels can form whenever a suitable source and drain exist, even if no conductor exists to act as a gate. The mechanism underlying the formation of such channels is called *charge spreading*, and although some details still remain unclear, the basic principles are well understood.[26,27] The oxide and nitride films covering an integrated circuit are nearly perfect insulators. Electric current cannot flow through an insulator, but static electrical charges can accumulate on the surface of an insulator or along the interface between two dissimilar insulators. These static charges are not entirely immobile and can slowly shift or spread under the influence of electrical fields. In integrated circuits, the interface between the protective overcoat and the plastic encapsulation is susceptible to this phenomenon. If a nitride protective overcoat is used, then the oxide-nitride interface is also vulnerable. The rate of movement of such charges depends on temperature and on the presence of contaminants. Higher temperatures greatly accelerate charge spreading, as does the presence of even trace amounts of moisture.[28]

Charge spreading requires the presence of static electrical charges at the insulating interface. Experience has shown that these charges do exist and that they consist primarily of electrons, but the mechanisms that generate them are not fully understood.

[26] D. G. Edwards, "Testing for MOS IC Failure Modes," *IEEE Trans. Rel.*, R-31, 1982, pp. 9–17.

[27] Lycoudes, *et al.*, p. 240ff.

[28] E. S. Schlegel, G. L. Schnable, R. F. Schwarz, and J. P. Spratt, "Behavior of Surface Ions on Semiconductor Devices," *IEEE Trans. on Electron Devices*, Vol. ED-15, #12, 1968, pp. 973–980.

Hot carrier injection certainly contributes to charge spreading, but integrated circuits that do not produce hot carriers still exhibit charge spreading. Various hypothetical mechanisms have been postulated to account for these experimental observations. In practice, the source of the static charge is less important than its consequences.

Figure 4.9A shows a cross section of a standard bipolar die susceptible to charge spreading. The base region inside the tank is biased above the PMOS thick-field threshold. and therefore acts as the source of a parasitic PMOS transistor. The tank containing this base region is also, of necessity, biased above the PMOS thick-field threshold. Electrons present in the overlying insulating layers will tend to migrate toward the positively charged tank. Eventually, enough electrons may accumulate over the tank to induce a channel (Figure 4.9B). In effect, the static charge generated by charge spreading behaves as the gate electrode of an MOS transistor.

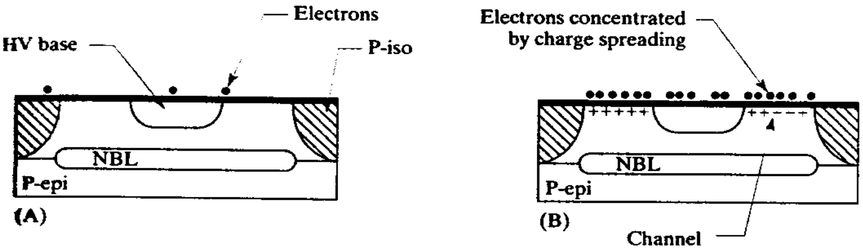


FIGURE 4.9 Cross section of a standard bipolar structure susceptible to charge spreading: (A) before and (B) after an extended period of operation under bias.

Standard bipolar appears to be more susceptible to charge spreading than does CMOS, probably because of less stringent process cleanliness. CMOS processes must minimize ionic contamination to maintain threshold voltage control; no such requirement exists for standard bipolar. The absence of excessive numbers of mobile ions gives CMOS and BiCMOS processes a certain degree of immunity to charge spreading.

Charge spreading produces parasitic PMOS transistors because it involves the accumulation of negative charges. The sources of these parasitic transistors consist of any P-regions that operate at voltages exceeding the PMOS thick-field threshold. The most vulnerable devices contain large, high-voltage P-regions operating at low currents—for example matched high-voltage HSR resistors. Failures tend to occur after long periods of high-temperature operation under bias. Moisture increases the mobility of surface charges, so environmental tests designed to detect moisture sensitivity often uncover charge spreading problems. The resulting parametric shifts resemble those produced by hot carrier injection in that they can be partially or completely reversed by baking the unbiased units at 200 to 250°C for several hours. The high temperature causes the accumulated static charges to disperse and restores an equilibrium between mobile ions and their fixed countercharges. This treatment does not constitute a permanent cure because the parametric drifts resume as soon as bias is restored.

Preventative Measures (Standard Bipolar)

NMOS channel formation can be suppressed in standard bipolar by coding base over all isolation regions. This *base-over-isolation* (BOI) requires no additional die area because the spacings required by the isolation are much larger than those required by the base. The BOI can therefore coincide with the isolation, or even slightly overlap it. Not all standard bipolar processes employ base-over-isolation; some already have a sufficiently heavily doped isolation diffusion to suppress channel formation.

Standard bipolar devices are susceptible to the formation of PMOS channels through charge spreading. Any tank that contains a P-type diffusion biased above the PMOS thick-field threshold requires protection in the form of field plates, channel stops, or a combination of both. Conservative designers usually derate the thick-field threshold of standard bipolar to account for this process's known propensity for charge spreading. For example, a designer might field plate and channel stop high-voltage P-regions operating above 30V even though the process has a rated PMOS thick-field threshold of 40V.

Figure 4.10 shows an example of a high-voltage HSR resistor vulnerable to parasitic channel formation. The tank containing the resistor connects to the positive supply to ensure isolation. A lead must route across the tank to connect to some adjacent low-voltage circuitry. A PMOS channel will form beneath this lead as soon as the voltage difference between the resistor and the lead rises above the PMOS thick-field threshold.

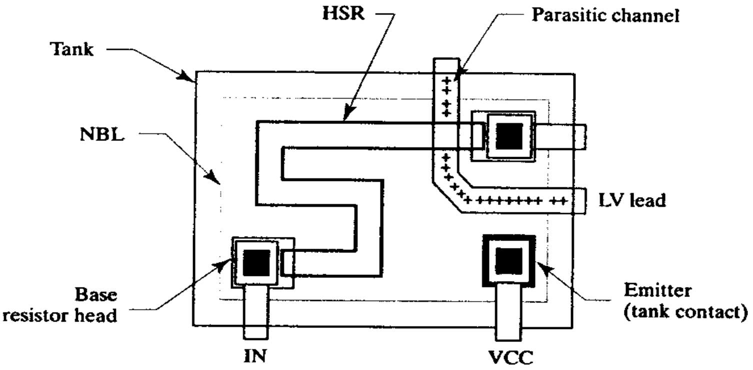


FIGURE 4.10 Example of a circuit susceptible to parasitic PMOS channel formation.

CMOS processes use channel stop implants to raise the thick-field thresholds. Standard bipolar does not include channel stop implants, but an emitter can be coded over selected regions of the N-tank to serve the same purpose. Figure 4.11A shows how emitter diffusions can disrupt the parasitic channels formed beneath a low-voltage lead. Each of the two minimum-width emitter strips disrupts a channel that would otherwise conduct current from the resistor to the isolation.

The emitter bars in Figure 4.11A extend slightly beyond the leads in either direction. These extensions will sever the channel even if the metal and the emitter misalign. The electric field also fringes out to either side of the lead. These fringing

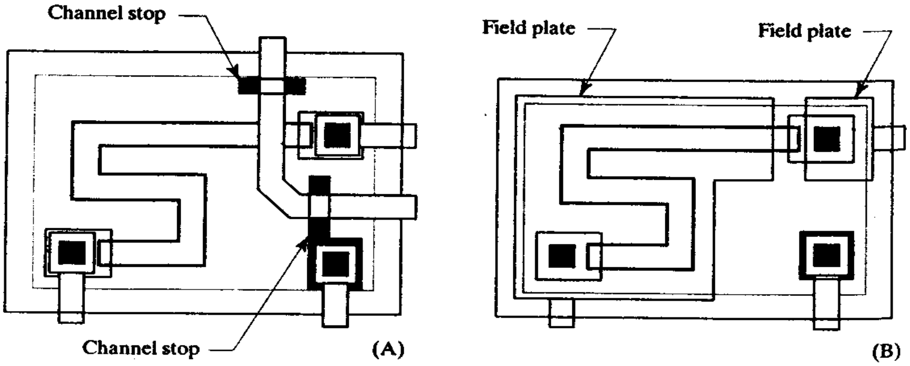


FIGURE 4.11 Two methods for preventing parasitic PMOS channels: (A) channel stops prevent channel formation beneath leads but do not stop charge spreading and (B) field plates provide relatively complete coverage, except possibly in the gap between the plates.

fields rarely extend laterally more than two or three times the oxide thickness, so the overlap of the emitter bar over the lead should equal the maximum photolithographic misalignment plus twice the oxide thickness. Assuming a two-level misalignment of 1μm and a 10kÅ thick-field oxide, the emitter bar should extend about 3 to 5μm beyond the lead on either side. These emitter bars are often called *channel stops*,[29] but they should not be confused with the blanket *channel stop implants* used in CMOS and BiCMOS processes. Channel stops are sometimes called *guard rings*, although this term is more properly applied to minority carrier guard rings (Section 4.4.2).

The channel stops in Figure 4.11A cannot, by themselves, prevent charge spreading. Even if a channel stop entirely encircles a serpentine resistor like that in Figure 4.11, parasitic channels can still form between its turns. If additional channel stops are placed between the turns, parasitic effects could still alter the effective width of the resistor by inverting the silicon along its edges. Some other technique must be used to supplement channel stops, especially for high-voltage diffused resistors.

Field plating can provide comprehensive protection against both parasitic channel formation and charge spreading. A field plate consists of a conductive electrode placed above a vulnerable diffusion and biased to inhibit channel formation.[30] Figure 4.11B shows an HSR resistor with field plates added. The low-voltage lead has been rerouted and a large plate of metal has been placed over the body of the resistor and connected to its positive terminal. The metal lead connecting to the negative end of the resistor has also been enlarged to protect the head of the resistor protruding beyond the main field plate. Both of these field plates must overlap the resistor enough to allow for outdiffusion, misalignment, and fringing fields. Assuming a two-level misalignment of 1μm, a maximum outdiffusion of 2μm, and a maximum fringing distance of 2μm, the total overlap must equal 5μm. Since the field plate consists only of metal, it can extend to fill the required area without enlarging either the resistor or its tank.

A field plate operates by providing an intentional gate for at least a portion of the MOS channel. This gate is biased to prevent the gate-to-source voltage of the parasitic transistor from exceeding the thick-field threshold. The presence of the conductive plate prevents the accumulation of static charges and thus suppresses charge spreading. Field plates also prevent modulation of carrier concentrations in the underlying silicon by acting as electrostatic shields. They therefore provide excellent protection against all types of electrostatic interactions, including conductivity modulation and noise coupling from overlying leads.

Most field plates contain gaps in which channels can still form. In the resistor of Figure 4.11B, a gap remains between the two field plates covering the resistor. Two methods exist for blocking these gaps. One method consists of flaring, or *flanging*, the ends of the field plate to elongate the channel as much as possible (Figure 4.12A). The close proximity of the parallel field plates induces a lateral electric field that sweeps static charges out of this region. The longer the potential channel, the greater the margin of safety provided by the flanges. The second method bridges the gaps between the field plates with short channel stops (Figure 4.12B). The emitter strips used for this purpose must overlap the field plates sufficiently to account for misalignment. This technique combines the strengths of a field plate with those of a channel stop to provide ironclad protection at all points.

[29] J. Trogolo and S. Sutton, "Surface Effects and MOS Parasitics," unpublished report, 1988, p. 13ff.

[30] Trogolo, *et al.*, p. 13ff.

The resistors in Figures 4.11 and 4.12 illustrate another important principle of field plating: the field plate biased to the highest potential should cover as much of the resistor as possible. If the low-voltage field plate were extended, it would provide less protection to the high-voltage end of the resistor. If the voltage difference between the tank and the field plate exceeds the thick-field threshold, then the field plate will actually induce channel formation. The field plate should extend from the high-voltage terminal of a vulnerable resistor to encompass as much of the resistor body as possible. The low-voltage terminal of the resistor should have just enough field plating to cover and protect the contact head. Matched resistors may require a slightly different field-plating strategy (Section 7.2.8).

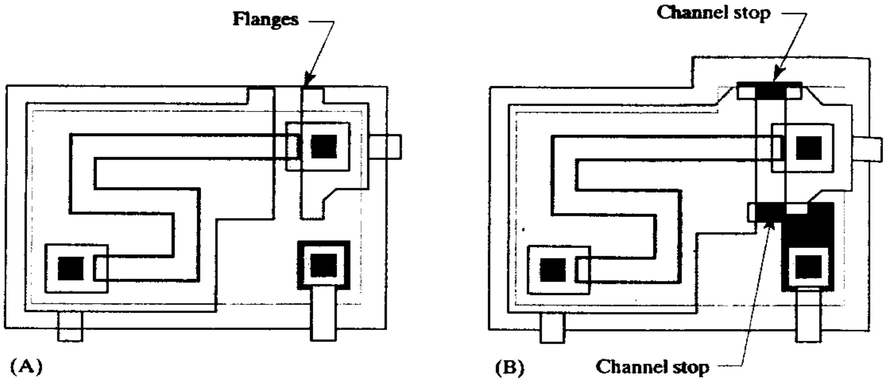


FIGURE 4.12 Improved field plating schemes: (A) flanged field plates and (B) combination of field plates and channel stops.

Figure 4.13 shows an interesting situation that sometimes occurs when laying out resistors. The two terminals of this device connect to a high potential and a low potential, respectively. The high-voltage end of the resistor needs protection against charge spreading, but the low-voltage end does not. Since the voltage drops linearly along the resistor, the field plate has been terminated partway down its length. Partial field plates should extend well beyond the point where the voltage drops below the thick-field threshold. In the case of Figure 4.13, as much of the resistor as possible has been field plated even though much of it apparently serves no useful function. The large safety margin obtained by this means costs nothing and helps ensure that the device will work even under worst-case conditions.

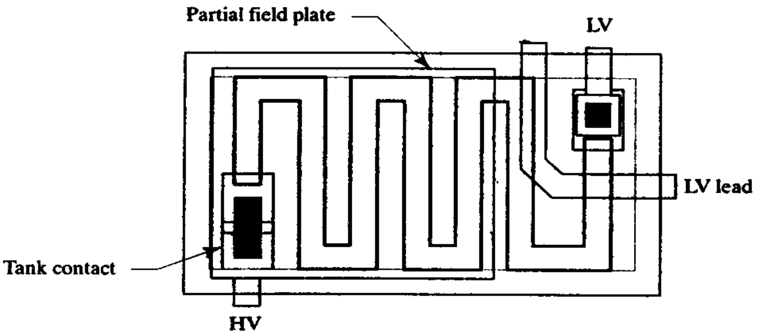


FIGURE 4.13 Example of partial field plating.

Figure 4.14 shows another example of selective field plating involving a multiple-collector lateral PNP transistor. The emitter, base, and one collector operate at voltages in excess of the thick-field threshold, while the remaining collector operates at

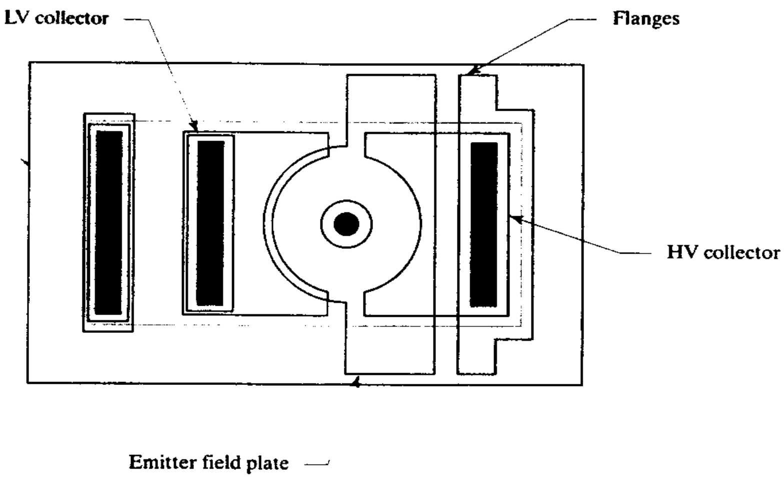


FIGURE 4.14 Field-plated, split-collector, lateral PNP with one low-voltage collector and one high-voltage collector. Flanging suppresses parasitic formation in the gaps between field plates.

a low voltage. The emitter field plate extends out from the emitter across the exposed surface of the base to a point just beyond the inner edge of the collector. The field plate need only overlap the collector by an amount equal to the maximum misalignment minus outdiffusion; certainly no more than 2 to 3μm. A second field plate extends outward from the high-voltage collector to block any parasitic channel that might form between the collectors. or from collector to isolation. No field plate surrounds the low-voltage collector since it does not require one. The field plates have been flanged to ensure that channels cannot form in the gaps. Channel stops could be added, but these would increase the size of the tank and are probably unnecessary.

To summarize, any P-type region biased in excess of the thick-field threshold acts as the source of a parasitic PMOS transistor. Field plates and channel stops ensure that no parasitic channels form from a high-voltage P-type region to any adjacent P-type diffusion. Field plating protects most of the device, while channel stops or flanges protect gaps left in the field plating. The official thick-field threshold of standard bipolar processes should be derated by 25% to provide an additional margin of safety against charge spreading. The following chapters describe additional examples of field plates and channel stops where appropriate.

Preventative Measures (CMOS and BiCMOS)

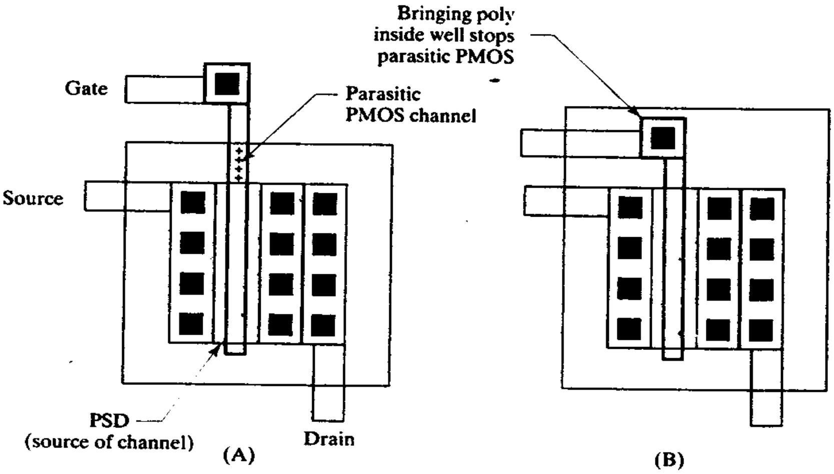
CMOS and BiCMOS processes usually incorporate channel stop implants to raise the thick-field threshold above the nominal operating voltage. The voltage rating of an N-well CMOS process is usually defined by NSD/P-epi breakdown, PSD/N-well breakdown, or gate oxide rupture, but some structures can withstand much higher voltages. The high N-well/P-substrate breakdown allows PMOS transistors to operate at elevated backgate voltages. These transistors will function normally as long as the drain-to-source voltage does not exceed the PSD/N-well breakdown voltage or the N-well punchthrough voltage. Similarly. an extended-drain NMOS using N-well as a lightly doped drain can withstand the full N-well/P-substrate breakdown voltage applied to its drain terminal.

The lightly doped N-well inverts in much the same manner as the lightly doped N-epi tanks of standard bipolar. Any N-well region containing a P-type diffusion biased above the thick-field threshold becomes vulnerable. As before, PMOS parasitic channels can be suppressed by field plates, channel stops, or a combination of both. The flanged field plate (Figure 4.12A) is especially attractive because the tighter CMOS layout rules allow a narrower gap between the flanges. The stronger lateral electric field makes close-spaced flanges particularly effective at preventing the accumulation of static charges. Flanged fieldplates are the method of choice for suppressing parasitic PMOS channels in high-voltage CMOS and BiCMOS structures.

Charge spreading is less prevalent in CMOS processes than bipolar ones, probably because of improved process cleanliness. Many CMOS designers consequently take a rather cavalier approach to field plates and channel stops. Such indiscretion is unwise considering the greatly reduced operating currents characteristic of modern CMOS designs. However, one special case does exist where charge spreading can be safely ignored. Most CMOS processes list a lower thick-field threshold for poly than for metal because the oxide layer beneath the poly consists only of thick-field oxide and MLO, while that beneath the metal contains an added layer of deposited ILO. Static charges can only accumulate at the interface between two dissimilar materials, so the lower thick-field thresholds associated with thinner oxides do not have any significance for charge spreading. Charge spreading becomes significant only at voltages beyond the highest thick-field threshold listed for the process.

Poly leads can induce parasitic channels if they run across an N-well containing a P-diffusion biased above the poly thick-field threshold. Figure 4.15A shows a typical example of a vulnerable structure consisting of the poly gate lead from a high-voltage PMOS transistor extending across the well and into the surrounding isolation. The well forms the backgate of the parasitic PMOS, the poly acts as the gate, the sources are the PSD regions of the PMOS transistor, and the drain is the P-epi isolation. As long as the backgate potential does not exceed the metal thick-field threshold, the channel can be interrupted by stopping the polysilicon lead short of the drawn edge of the well (Figure 4.15B). The channel can form only beneath the polysilicon lead, so a complete channel cannot form if the lead does not bridge the gap between source and drain. Charge spreading is unlikely to occur so long as the voltages involved do not exceed the highest thick-field threshold of the process. The minimum spacing between the poly and the drawn edge of the N-well should equal the photolithographic misalignment allowance, plus an extra 2 to 3μm to account for fringing fields. The outdiffusion of the well does not

FIGURE 4.15 The parasitic PMOS channel beneath a poly lead (A) can be eliminated by pulling poly inside the well (B).



provide any margin of safety because it becomes very lightly doped beyond its drawn boundaries.

The lightly doped P-type epi can also invert if a high-voltage lead runs across it (Figure 4.8B). The source and drain of this parasitic NMOS consist of two adjacent N-wells; the high-voltage lead acts as the gate, and the P-epi acts as the backgate. A parasitic channel will form if the voltage differential between the lead and an adjacent well exceeds the NMOS thick-field threshold. A channel stop can be inserted by running a thin bar or ring of PSD material down the center of the P-epi beneath the high-voltage lead. The PSD should extend beyond either edge of the lead by an amount sufficient to account for misalignment, plus an additional 2 to 3μm to allow for fringing effects. In many cases, the N-well to N-well spacing can accommodate a minimum-width PSD channel stop with little or no increase in the spacing between adjacent wells. A thin ring of PSD material can then encircle each well (Figure 4.16). This ring not only stops any possible leakage caused by charge spreading but also allows complete freedom to route the leads in any pattern desired. If the PSD rings are drawn when the wells are placed, or if they are automatically produced during mask generation, then the designer can subsequently ignore NMOS channel formation. These PSD rings correspond to the base-over-isolation (BOI) scheme used for the same purpose in standard bipolar designs.

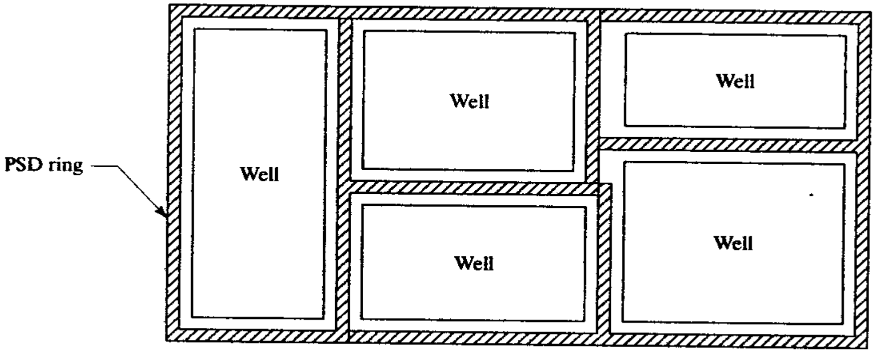


FIGURE 4.16 Sample layout showing the use of PSD rings to prevent NMOS channels.

4.4 PARASITICS

All integrated circuits contain electrical elements not required for their operation. These include reverse-biased isolation junctions, and resistances and capacitances between various diffusions and depositions. The circuit does not benefit from the presence of these *parasitic components*, but they can sometimes adversely affect its operation.

Parasitics are responsible for a number of different types of electrical failures. For example, capacitive coupling can inject noise into sensitive circuitry. The type of parasitics that will be discussed in this section concern the forward biasing of junctions that normally remain reverse-biased. When these junctions forward bias, current begins to flow between circuit nodes that normally remain isolated from one another. If these currents are small and the circuit is relatively insensitive to their presence, then these leakages may produce only subtle parametric shifts. Larger currents can catastrophically disrupt the operation of the circuit. The malfunctioning circuit may actually *latch up*, causing it to continue malfunctioning even after the removal of the triggering event. Latchup can cause physical destruction of an integrated circuit due to excessive power dissipation and consequent overheating. Even if the circuit does not self-destruct, normal operation can only be restored by interrupting the power.

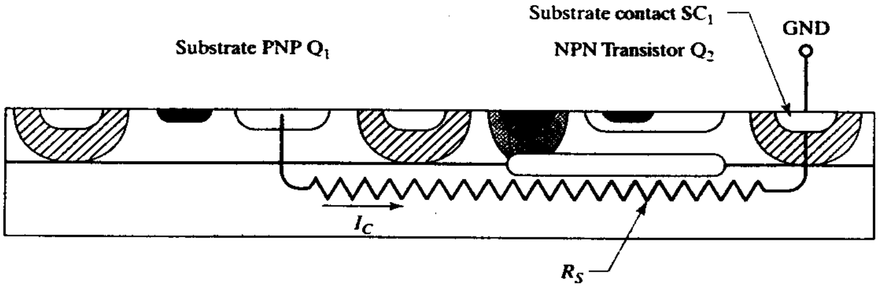
Two important parasitic mechanisms involve currents flowing through the substrate. ***Substrate debiasing*** occurs when parasitic currents induce voltage drops in a resistive substrate. If these voltage drops become large enough, they can forward-bias one of the isolation junctions. The forward-biased junction then injects current into other circuit nodes, causing potentially catastrophic malfunctions. ***Minority carrier injection*** occurs when a forward-biased junction injects minority carriers into the isolation, a tank, or a well. Some of these carriers diffuse several hundred microns before recombining and can easily cross reverse-biased junctions that block majority carrier flow.

4.4.1. Substrate Debiasing

Substrate debiasing becomes a problem when currents flowing through the substrate generate voltage drops of a few tenths of a volt or more. This substrate current consists of majority carriers that cannot surmount reverse-biased isolation junctions, but sufficient debiasing may cause one or more isolation junctions to forward-bias and inject minority carriers into active circuitry.

Figure 4.17 shows a typical example of substrate debiasing in a standard bipolar process. Substrate PNP transistor Q_1 injects its collector current I_C directly into the substrate. This current then flows laterally to substrate contact SC_1. Because of the presence of substrate resistance R_s, the substrate voltage immediately under NPN transistor Q_2 rises. Only a few hundred millivolts of substrate debiasing are necessary to forward-bias the collector-substrate junction of a saturated common-emitter NPN.

FIGURE 4.17 Cross section of a standard bipolar die showing potential substrate debiasing caused by substrate resistance R_s.



Effects

The voltage required to forward-bias a PN junction depends on both current density and temperature. Table 4.1 lists typical forward-bias voltages for the collector-substrate junction of a minimum-area NPN transistor constructed in a standard bipolar process. This table is useful for estimating susceptibility to substrate debiasing. For example, a circuit using 100μA minimum currents can probably tolerate 1μA of leakage. If it must operate at 125°C, then Table 4.1 indicates that substrate debias-

TABLE 4.1 Forward voltages for a typical collector-substrate junction of a minimum NPN transistor in standard bipolar, as a function of temperature and current.[31]

Current	25°C	85°C	125°C	150°C
10nA	0.43V	0.29V	0.19V	0.13V
100nA	0.49V	0.36V	0.27V	0.22V
1μA	0.55V	0.43V	0.35V	0.30V
10μA	0.61V	0.50V	0.43V	0.39V
100μA	0.67V	0.57V	0.51V	0.47V

[31] Based on V_{BE}(150°C, 1μA) = 0.3V.

ing must not exceed 0.35V. If the same circuit has to operate at 150°C, then it can tolerate no more than 0.30V of debiasing.

Figure 4.18 depicts the cross section of a standard bipolar wafer containing a single substrate current injector and a single substrate contact. R_1 models the lateral resistance through the substrate, while R_2 models the vertical resistance beneath the substrate contact. The total resistance of the substrate R_s equals the sum of the lateral and vertical components: $R_s = R_1 + R_2$.

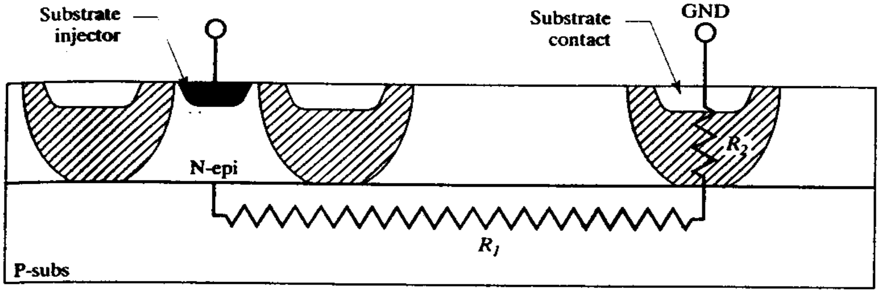


FIGURE 4.18 Simplified model of substrate debiasing in a standard bipolar process.

The relative magnitudes of R_1 and R_2 depend upon the process. Standard bipolar uses a lightly doped substrate and a heavily doped isolation diffusion, so $R_1 >> R_2$. The value of R_1 depends on various geometric factors, including the cross-sectional area of both the injector and the substrate contact as well as the distance between them. Typical values of R_1 range from hundreds to thousands of Ohms,[32] while R_2 rarely exceeds 10Ω.[33] The presence of a network of isolation diffusions criss-crossing the die complicates the computation of substrate resistance because the low sheet resistance of the isolation (usually about 10Ω/□) allows each substrate contact to extract current from a large area of isolation. This effect complicates the computations to such an extent that only empirical measurement or sophisticated computer simulation can yield accurate results.

Two points should be kept in mind when using lightly doped substrates. First, substrate resistance always increases with separation. A substrate contact placed adjacent to the injector will extract some of the current before it ever reaches the substrate. Contacts placed further away require the current either to flow through long stretches of isolation or to pass through the highly resistive substrate. Second, a substrate contact to a heavily doped isolation diffusion draws current not only from the substrate immediately beneath it but also from adjoining stretches of isolation. This effectively magnifies the area of substrate contacts, allowing even a minimum-size contact to have an effective area of many hundreds of square microns. Because of this effect, a scattering of minimum-size substrate contacts throughout the die will have a much lower effective resistance than a single large contact.

[32] An estimate of the lateral resistance R_1 can be obtained by examining the spreading resistance $R_{sp} = \rho/d$, where ρ is resistivity and d is the diameter of the points of contact. Spreading resistance assumes a semi-infinite slab of uniformly doped material and a probe separation much wider than the probe diameter: these conditions are only approximately met by typical substrate contacts. Assuming that the cross-sectional areas of injector and substrate contact are 1mil^2 each, this yields d = 28.7μm. A substrate with a resistivity of 10Ω·cm would have a spreading resistance of 3.48kΩ. More accurate results can be obtained by applying various correction factors: G. A. Gruber and R. F. Pfeifer, "The Evaluation of Thin Silicon Layers by Spreading Resistance Measurements," *National Bureau of Standards Special Publication 400–10*, Spreading Resistance Symposium, NBS, Gaithersburg, Maryland: June 1974.

[33] The vertical resistance through single-diffused isolation can be approximated by dividing the diffusion into multiple layers of constant doping. Computations for a diffusion with a surface doping of 10^{20}cm^{-3}, a minimum dopant concentration of 10^{17}cm^{-3}, and a depth of 5μm yield a resistance of about 4Ω/mil^2.

CMOS and BiCMOS processes usually employ a heavily doped substrate and a lightly doped epi, so $R_1 << R_2$. The value of R_1 is usually so small that it can safely be ignored. The value of R_2 depends on the thickness of the epi layer and its resistivity. A typical value is about 600kΩ/μm^2 (1kΩ/mil^2). This value can be used to compute the area of substrate contacts required for a CMOS or BiCMOS design, as explained in the following section.

Even on a heavily doped substrate, contacts placed immediately adjacent to a substrate injector will exhibit less resistance than ones placed far away. This *proximity effect* falls off rapidly with distance, and substrate contacts placed hundreds of microns away are no more effective than those placed on the opposite side of the die. The proximity effect occurs because carriers can flow directly to the adjacent contact, rather than having to flow down to the substrate, across, and up to a distant contact. Contacts placed immediately adjacent to a substrate injector can also help prevent localized debiasing of the highly resistive isolation, protecting adjacent tanks from injection from the isolation sidewalls.

Preventative Measures

Integrated circuits should inject as little current into the substrate as possible, as this not only minimizes substrate debiasing but also helps limit noise and cross-talk caused by modulation of the substrate potential. The collector current of substrate PNP transistors flows directly into the substrate, so these devices should be used sparingly, and no single device should conduct more than a milliamp or two. Lateral PNP and vertical NPN transistors can inject large substrate currents when they saturate, but techniques have been developed to minimize this problem (Sections 8.1.4–5). The exact requirements for substrate contacts depend on the nature of the substrate and isolation:

Heavily doped substrates. The contacts in the scribe seal can usually extract 5 to 10mA without undue debiasing. If higher substrate currents are anticipated, then the total area of contacts required can be computed using the following formula:[34]

$$A_c = 10\frac{\rho t_{epi} I_s}{V_d} \qquad [4.1]$$

This formula assumes a uniform lightly doped isolation, such as the P-epi of N-well CMOS and BiCMOS processes. A_c represents the required total area of substrate contacts in μm^2, ρ is the resistivity of the epi in Ω·cm, t_{epi} is the epi thickness in microns, I_s is the maximum substrate current in milliamps, and V_d is the maximum allowable debiasing in volts (from Table 4.1). The thickness of the epi is reduced by up-diffusion of dopants from the underlying substrate and from the presence of a heavily doped (if thin) contacting diffusion, such as PSD. Consider a die with a P-epi resistivity of 10Ω·cm and an effective epi thickness of 7μm. If the substrate must conduct 20mA without more than 0.3V of debiasing, then 47,000μm^2 (72mil^2) of substrate contacts are required. Subtracting the area of substrate contacts in the scribe seal yields the required area of additional contacts. These can be inserted wherever space exists in the layout. As a precaution against localized debiasing, substrate contacts should ring any device injecting more than 1mA.

[34] This formula is derived from the fundamental equation $R = \rho t/A$. It neglects fringing effects, which tend to reduce the effective resistance of small substrate contacts. The formula provides a first-order approximation of the worst-case substrate resistance.

Lightly doped substrates with heavily doped isolation. No simple formula exists for computing the area of substrate contacts required to protect a lightly doped substrate from debiasing. A scattering of ten or twenty substrate contacts across the die will, when combined with the scribe seal. handle at least 5 to 10mA. Any device that injects 100μA or more should have substrate contacts located nearby, and any device that injects 1mA or more should be ringed with as much substrate contact as possible. Sensitive low-current circuitry should reside at least 250μm (10mil) away from any substantial source of substrate injection, since debiasing on lightly doped substrates tends to localize around the point of injection. Once the layout has been completed, additional substrate contacts should be scattered throughout the layout wherever room exists. A large number of small substrate contacts scattered throughout the layout will prove more effective than a few large contacts. Even with all of these precautions, designs that inject more than 10mA into the substrate may experience debiasing. Apart from adding more substrate contacts or moving sensitive circuits away from substrate injectors, the only remedies for such problems are the addition of a heavily doped substrate or the use of backside contacting.

Lightly doped substrates with lightly doped isolation. A few processes use a lightly doped substrate in combination with a very resistive isolation. This situation can arise when a BiCMOS design is constructed on a lightly doped substrate to save costs. Such designs cannot rely on the scribe seal to extract more than a few milliamps of substrate current. Large numbers of substrate contacts scattered across the die will help extract substrate current. but some degree of localized substrate debiasing is almost inevitable. Sensitive circuits should be located far away from major sources of substrate injection. Since substrate modulation can inject substantial noise into high-impedance circuitry. consider placing wells under resistors and capacitors to isolate them from substrate noise coupling. Sensitive MOS circuitry may also employ NBL to isolate NMOS transistors from the substrate (Section 11.2.2). In some cases, it may be possible to add strips of heavily doped material to the isolation without increasing the well-to-well spacings (Figure 4.16). This strategy effectively converts the design into one that uses a lightly doped substrate in conjunction with a heavily doped isolation. This stratagem substantially reduces the number and area of substrate contacts required to extract large substrate currents. Backside contact can also provide a large reduction in substrate resistance, but it is difficult to obtain Ohmic contact to a lightly doped substrate unless a backside diffusion is performed to increase the surface doping concentration.

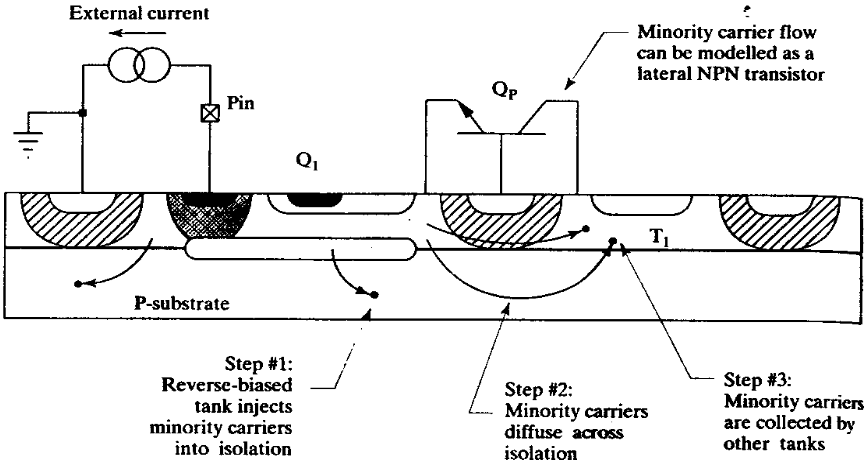
4.4.2. Minority-Carrier Injection

Junction isolation relies on reverse-biased junctions to block unwanted current flow. The electric fields set up by depletion regions repel majority carriers. but they cannot block the flow of minority carriers. If any isolation junction forward-biases, it will inject minority carriers into the isolation. Many of these carriers recombine, but some eventually find their way to the depletion regions isolating other devices.

Effects

Figure 4.19 shows a cross section of a standard bipolar circuit. Suppose that the collector of NPN transistor Q_1 connects to a pin of the integrated circuit, and that the external circuitry experiences occasional transient disturbances that pull current out of this pin. If transistor Q_1 is off, then these transients pull its tank below ground, forward-biasing the collector-substrate junction of Q_1 and injecting minority carriers (electrons) into the substrate. Most of these carriers recombine, but some diffuse across to other tanks, such as T_1.

FIGURE 4.19 Example of minority-carrier injection into the substrate of a standard bipolar process. Lateral NPN transistor Q_P models the transit of minority-carriers across the isolation.


External current
Pin
Q_1
Q_P
Minority carrier flow can be modelled as a lateral NPN transistor
T_1
P-substrate
Step #1: Reverse-biased tank injects minority carriers into isolation
Step #2: Minority carriers diffuse across isolation
Step #3: Minority carriers are collected by other tanks

The transit of minority carriers across the isolation is analogous to the flow of minority carriers through a bipolar transistor. The tank pulled below ground acts as the emitter of lateral NPN transistor Q_P. The isolation and substrate act as the base of this transistor, and any other reverse-biased tank acts as a collector. Each reverse-biased tank forms a separate parasitic transistor corresponding to Q_P. The betas of these parasitic lateral NPN transistors are very low because most of the minority carriers recombine in transit. The parasitic bipolar between two adjacent tanks might have a beta of 10, but the beta between two widely separated tanks might not even reach 0.001. Even such low gains can cause circuit malfunctions. Suppose that a forward-biased tank injects a minority current of 10mA into the substrate. If the parasitic associated with another tank has a beta of 0.01, then this tank will collect 100μA of current—easily enough to disrupt the operation of a typical analog circuit.

Substrate contacts cannot, by themselves, stop minority-carrier injection since minority carriers travel by diffusion and not by drift. Minority carriers are best collected by reverse-biased junctions. However, substrate contacts still provide majority carriers to feed recombination. Since most minority carriers recombine in the isolation, substrate contacts remain necessary to prevent substrate debiasing.

In some cases, minority-carrier injection can cause a circuit to latch up. Early CMOS processes suffered from a form of this malady that has since come to be called *CMOS latchup.*[35] Figure 4.20A shows the cross section of a portion of a CMOS die consisting of an NMOS transistor M_1 and a PMOS transistor M_2. In addition to these two desired MOS transistors, this layout contains two parasitic bipolar transistors. Lateral NPN transistor Q_N's emitter is the source of M_1, its base is the isolation, and its collector is the N-well of M_2. Lateral PNP transistor Q_P's emitter is the source of M_2, its base is the N-well, and its collector is the isolation. Figure 4.20B shows the two parasitic bipolar transistors drawn in a more familiar fashion. In this schematic, R_1 represents the well resistance of M_2, and R_2 represents the substrate resistance. These two resistors normally ensure that both bipolar transistors remain off. As long as this remains the case, neither parasitic conducts any current and the integrated circuit works as intended. When a transient disturbance turns on either transistor, the current flowing through this device will turn on the other parasitic as well. Each transistor then supplies the other's base current. Once both tran-

[35] R. R. Troutman, "Recent Developments in CMOS Latchup." *IEDM Tech. Dig.*, 1984, pp. 296–299.

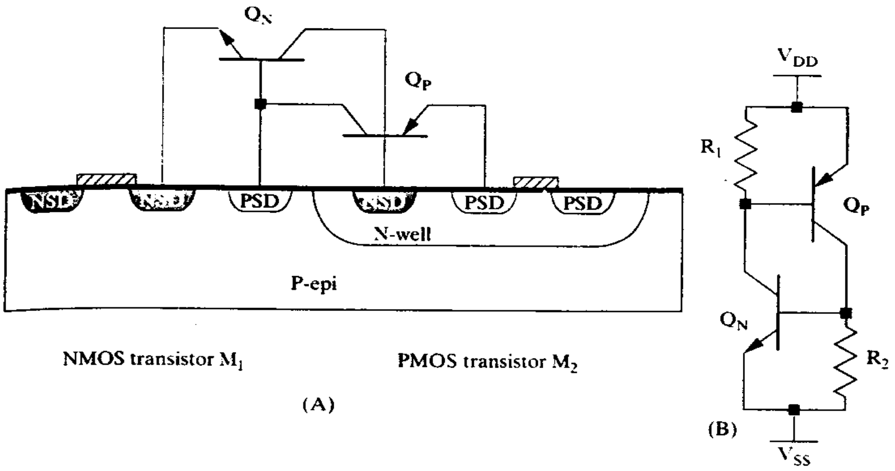


FIGURE 4.20 (A) Cross section of a CMOS die showing diffusions that form the two parasitic bipolar transistors Q_P and Q_N; (B) equivalent schematic showing these transistors along with well resistance R_1 and substrate resistance R_2.

sistors begin to conduct, they will continue to do so even if the transient disturbance that initiated conduction is removed. The circuit has *latched up* and it will remain in this state until power is removed. The integrated circuit can actually conduct so much current that it overheats and self-destructs. Even if this does not occur, latchup causes circuit malfunctions and excessive supply current consumption.

CMOS latchup can be triggered in one of two ways. If the source of NMOS transistor M_1 is pulled below ground, it will inject minority carriers (electrons) into the substrate, turning on parasitic transistor Q_N. This transistor will then turn on Q_P. Alternatively, the source of PMOS transistor M_2 may be pulled above the well. It will then inject minority carriers (holes) into the well and will turn on parasitic transistor Q_P. This transistor then turns on Q_N. Latchup can only occur if the product of the betas of transistors Q_N and Q_P exceeds unity. If the beta product is less than unity, then transient disturbances may occur, but the circuit cannot actually latch up (Section 11.2.7). CMOS latchup also requires the existence of four distinct semiconductor regions arranged in the sequence PNPN. A discrete device called a *silicon controlled rectifier* (SCR) has this same four-layer structure. CMOS latchup is sometimes described in terms of a parasitic SCR consisting of PSD, N-well, P-epi and NSD. This point of view is analogous to the transistor approach discussed above because the SCR operates in the same manner as the pair of coupled transistors.

The obvious way to stop CMOS latchup consists of reducing the beta of either or both parasitic transistors. If the product of these betas is less than unity, then latchup cannot occur. This is usually achieved by increasing layout spacings, which in turn increases the width of the neutral base regions of the parasitic lateral transistors. Alternatively, the amount of dopant present in the neutral base region of one (or both) parasitic transistors may be increased. Both of these approaches increase the Gummel number of one or both transistors and reduce the beta product.

Although many CMOS processes claim immunity to latchup, these claims are true only in a somewhat narrow sense. The PNPN structure inherent in the CMOS transistors of such a process lacks sufficient gain to establish regenerative feedback, but minority-carrier injection still occurs. The collected carriers can still cause circuit malfunctions, and if positive feedback exists in the circuit, these malfunctions can still cause a form of latchup. The significance of this observation is frequently underestimated. Any integrated circuit that experiences unanticipated minority-carrier injection can potentially latch up. Even if it does not actually do so, it is still likely to malfunction. Not only do electrons injected into the substrate pose a potential threat, but so do holes unintentionally injected into wells or tanks.

Preventative Measures (Substrate Injection)

Fundamentally, there are four ways to defeat minority-carrier injection: (1) eliminate the forward-biased junctions that cause the problem, (2) increase the spacing between components, (3) increase doping concentrations, and (4) provide alternate collectors to remove unwanted minority carriers. All of these techniques provide some benefit, and in combination they can correct almost any minority-carrier injection problem.

The simplest solution, at least in theory, consists of eliminating the forward-biased junctions that inject minority carriers. This goal is often very difficult to achieve. In a standard bipolar process, tanks must not go below substrate by more than about 0.3V or they will inject minority carriers into the substrate. In an N-well CMOS process, no well and no NSD region residing in the epi may go below substrate potential. If the voltage on a pin slews rapidly, parasitic inductance can cause transients that pull the pin above supply or below ground. The faster the node slews, the smaller the parasitic inductance required to cause such transients. Modern switching speeds have become so fast that the inductance of pin and bondwire alone often cause objectionable transients. Substrate injection has become difficult, if not impossible, to eliminate.

Minority-carrier injection into the substrate will cause fewer problems if potential injectors are separated from sensitive circuitry. In most designs, only a few devices connect to pins. With a little forethought, these devices can be placed far away from sensitive circuitry. In many cases, the layout will naturally favor this sort of separation. For example, power transistors inject minority carriers during transients. Since these transistors form part of the output circuitry, they will typically be placed far away from sensitive input circuitry to minimize electrical and thermal feedback. This same arrangement also minimizes the circuit's vulnerability to minority-carrier injection.

Additional dopant added to the isolation regions of the die will reduce the gain of the parasitic lateral bipolar. CMOS and BiCMOS processes often employ P+ substrates for just this reason. All other factors being equal, a process incorporating a heavily doped substrate will provide greater immunity to electrical upsets than one that uses a lightly doped substrate. However, a heavily doped substrate cannot, by itself, prevent minority carriers from moving laterally through isolation regions separating adjacent tanks or wells. In order to obtain the full benefits of the heavily doped substrate, the process must use a heavily doped isolation, or the designer must add suitable guard rings.

The isolation doping can also be increased by adding a deep-P+ diffusion. Most CMOS processes do not include any suitable diffusion. Some BiCMOS processes include one for constructing certain components (such as DMOS transistors)—usually as part of a process extension. Standard bipolar processes sometimes offer a deep-P+ process extension for constructing high-current lateral PNP transistors. If a suitable diffusion exists, it can be placed in the isolation regions of the die to help increase the isolation doping. This technique can help offset the very light doping of the PBL portion of an up-down isolation system, and can minimize lateral conduction of minority carriers between adjacent tanks or wells.

Minority carriers are collected in disproportionate numbers by reverse-biased junctions near the point of injection. Not only do carriers have less distance to travel to reach a nearby junction, but the nearer junctions also block the flow of carriers to more distant ones. Designers can take advantage of this *shadow effect* to erect deliberate barriers to the flow of minority carriers by placing reverse-biased junctions between the point of injection and vulnerable diffusions. A reverse-biased junction used in this manner is called a *minority-carrier guard ring*. Figure 4.21 shows a typical layout for such a guard ring in a standard bipolar process. Tanks T_1 and T_2 connect to pins that may experience voltage transients. These are surrounded by a third tank, T_3, that collects a significant fraction of the minority carriers injected by T_1 and T_2. Tanks T_1 and T_2 connect to pins, so it is quite natural to place them along one side of the die or even in a corner (as shown in Figure 4.21). This not only minimizes the length of interconnecting leads but also eliminates the need for guard rings along two edges.

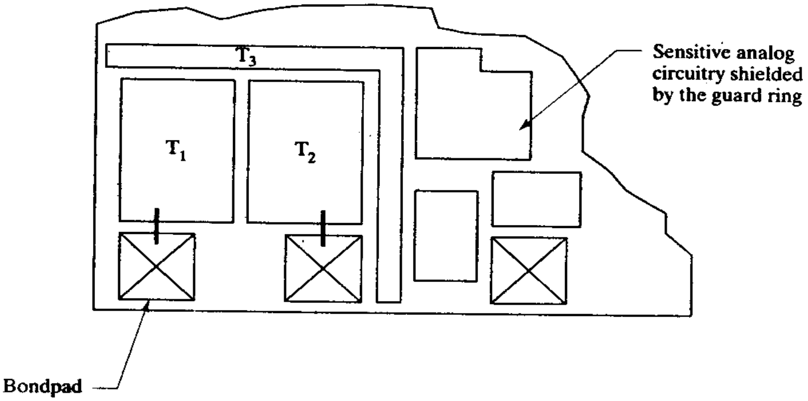


FIGURE 4.21 Sample electron-collecting minority-carrier guard ring (T_3) implemented in a standard bipolar process.

The key to designing efficient guard rings consists of making them deep, wide, and low-resistance. The deeper the guard ring, the larger the fraction of passing minority carriers it can collect. N-well makes a more effective electron-collecting guard ring than NSD, and an epi tank makes a better electron-collecting guard ring than an emitter diffusion. If the guard ring can be connected to produce a large reverse bias, then the depletion region surrounding it will widen and the collection surface will be forced even deeper into the silicon. Thus, electron-collecting guard rings connected to the positive supply become more effective than those connected to substrate potential. Since diffusing minority carriers move randomly, some will actually be collected by the bottom of the guard ring's depletion region. A wider guard ring will therefore collect more minority carriers than a narrow one will. Also, narrow diffusions do not penetrate as deeply as wide diffusions because dopants become diluted by lateral dispersion. A diffusion two or three times wider than minimum will contain enough dopant to obtain the maximum possible junction depth. Low resistance also helps improve the effectiveness of a guard ring, especially if it cannot be strongly reverse-biased. Collected carriers can forward-bias a high-resistance guard ring and cause it to re-inject minority carriers. The lower the vertical resistance of the guard ring, the larger the current it can collect before it saturates and re-injects.

Figure 4.22 shows cross sections of two minority-carrier guard rings designed to collect electrons injected into the substrate. Figure 4.22A shows a substrate guard ring for standard bipolar.[36] This guard ring includes all four available N-type materials: N-epi, deep-N+, NBL, and emitter. The NBL helps obtain the maximum possible junction depth, while deep-N+ and emitter reduce the vertical resistance. The wider this structure, the more effectively it will collect minority carriers. Most designers

[36] W. Davis, *Layout Considerations*, unpublished manuscript, 1981, p. 53.

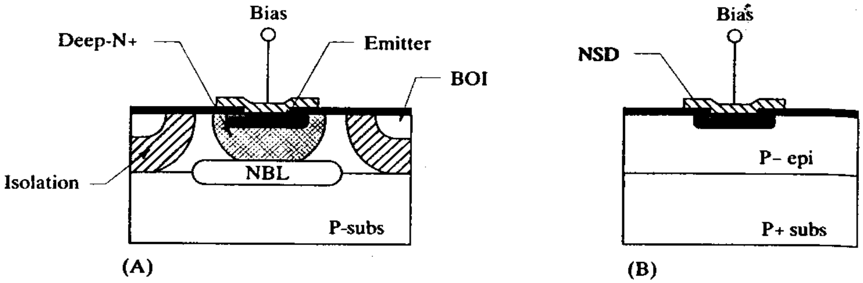


FIGURE 4.22 Cross sections of two representative electron-collecting guard rings: (A) standard bipolar[37] and (B) CMOS.

compromise between efficiency and area by making the deep-N+ strip no more than twice minimum width and by spacing the other layers accordingly. If possible, this guard ring should connect to the highest supply voltage available on the die. The guard ring will still function connected to substrate potential, but it may saturate unless all parts of the guard ring connect to the substrate terminal by a direct metal run. This is probably not practical in a single-level metal process since gaps must be left in the metallization to allow leads to pass through. Single-level-metal guard rings should connect to the positive supply and should contain as few gaps as possible.

BiCMOS layouts can produce electron-collecting guard rings similar to those in Figure 4.22A, although in this case N-well replaces the N-epi tank. Electron-collecting guard rings are considerably more difficult to construct in CMOS-only processes. The N-well becomes extremely resistive in the absence of deep-N+ and NBL, and most CMOS devices operate at relatively low voltages. Figure 4.22B shows an alternate CMOS minority-carrier guard ring. NSD has a relatively low resistance, but it is too shallow to capture more than a small percentage of the electrons in the substrate. The wider the NSD strip, the more effective the guard ring. A width of at least 8 to 10μm is recommended, although narrower guard rings do provide some benefit. The NSD guard ring should, if possible, connect to a supply pin. The reverse bias across the NSD-epi junction drives the depletion region deeper into the epi and increases the apparent depth of the guard ring. In low-voltage processes, a strongly reverse-biased NSD guard ring will often generate secondary carriers due to impact ionization. This problem can be minimized by connecting the low-voltage NSD guard ring to ground instead of to a power supply.

Guard rings of the type shown in Figure 4.22A can reduce substrate injection by a factor of 10 to 100 providing they are used in conjunction with a heavily doped substrate. The P–/P+ interface between the lightly doped epi and the heavily doped substrate repels minority carriers (Section 8.1.5), constraining them to remain within the relatively thin epi layer. This greatly improves the collection efficiency of the guard ring.[38] A simple modification to the electron-collecting guard ring of Figure 4.22A can further increase its attenuation. Instead of connecting the guard ring directly to a supply voltage, it is connected back to the substrate so that a majority-carrier current flows through the substrate beneath the guard ring (Figure 4.23). This type of guard ring is intended to protect against minority carriers originating on only one side of the ring—in this case, the right side. The deep-N+ sinker in the center of the guard ring collects most of these minority carriers. The resulting cur-

[37] C. Jones, "Bipolar Parasitics," unpublished report, 1988, p. 43.

[38] L. S. White, G. R. M. Rao, P. Linder, and M. Zivitz, "Improvement in MOS VLSI Device Characteristics Built on Epitaxial Silicon," in *Silicon Processing*, American Society for Testing and Materials STP 804, 1983, pp. 190–205.

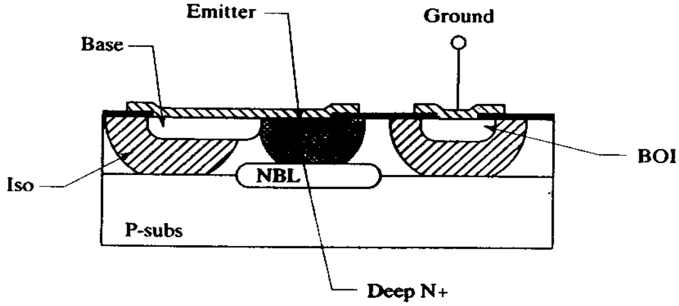


FIGURE 4.23 Cross section of an improved minority-carrier guard ring for collecting electrons injected into the substrate.

rent flows out of the sinker and into the attached metal plate. The base/iso diffusion at the left side of the structure re-injects this current into the substrate in the form of majority carriers. Since the nearest substrate contact lies on the other side of the structure, the majority carriers flow back underneath the guard ring. This current locally debiases the substrate and creates an electric field that opposes minority-carrier flow. The minority carriers are forced upward and toward the guard ring, where they are ultimately collected, or they are held in the substrate until they recombine. The inventor[39] claims an attenuation factor in excess of one million for this structure. While this degree of attenuation may not be achieved in every process, this guard ring will provide more attenuation than those shown in Figure 4.22, especially at higher currents.

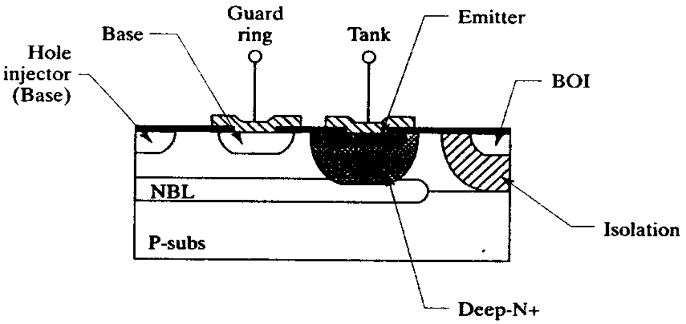
The modified guard ring in Figure 4.23 suffers from several drawbacks. It provides enhanced attenuation of carriers flowing in only one direction—in this case from right to left. Substrate contacts cannot be placed near the guard ring on the side facing the injected carriers. This structure also relies upon deliberate debiasing of the substrate, which could potentially forward-bias adjacent junctions. The principle behind this style of guard ring also applies to the design of ordinary guard rings of the sort shown in Figure 4.22. In all cases, it is better to place the electron-collecting guard ring adjacent to the injector, and to place substrate contacts inside of the guard ring. The majority-carrier substrate current that flows under the guard ring generates an electric field opposing the flow of minority carriers, thereby enhancing the performance of the guard ring.

Minority-carrier guard rings can also prevent holes injected into a tank from reaching the substrate and debiasing it. This situation can occur whenever a bipolar transistor saturates, regardless of whether this transistor is an NPN or a PNP (Section 8.1.4–5). Power transistors can easily inject tens or even hundreds of milliamps into the substrate. A heavily doped layer such as NBL can reduce minority-carrier injection from a P-type region through an N-well or N-epi tank to substrate. NBL also helps to minimize tank or well resistance and therefore makes it more difficult to develop the debiasing required to trigger CMOS latchup. CMOS processes generally do not incorporate NBL due to the cost of the extra masking step and to manufacturing difficulties associated with the fabrication of buried layers. Standard bipolar and analog BiCMOS processes frequently use NBL to reduce NPN collector resistance. If NBL exists, it should be added to all tanks or wells that can tolerate its presence, in order to minimize substrate injection and to improve latchup immunity.

[39] F. Van Zanten, U.S. Patent # 4,466,011, 1984.

Figure 4.24 shows a hole-collecting guard ring constructed in a standard bipolar process. As a first line of defense against hole injection into the substrate, the tank is floored with NBL and ringed with deep-N+. Any hole attempting to reach the isolation must pass through one or the other of these heavily doped regions. The large population of electrons in these regions enhances recombination and prevents most holes from successfully crossing. Far more importantly, the N+/N− boundary exhibits a built-in potential gradient caused by the outdiffusion of majority carriers that helps confine holes inside the tank until they recombine or are collected (Section 8.1.5).

FIGURE 4.24 Cross section of a minority-carrier guard ring for collecting holes injected into a tank.[40]



The guard ring in Figure 4.24 includes a base ring placed just inside the deep N+. This ring usually connects to ground, but it will remain reasonably effective even if it is tied to the tank terminal. Any hole impinging on the depletion region surrounding the base ring will be drawn across by the electric field. Holes become majority carriers inside the base diffusion and can be removed through the contact. The base ring collects almost all of the holes as long as the tank contains NBL. Even without the deep-N+ ring around the outside edge of the tank, the base ring will collect at least 90% of the holes. This type of guard ring is largely ineffectual without NBL.

CMOS devices may experience hole injection into wells if a PSD region rises above the well potential, as might occur if a pin connected to a PMOS source or drain rises above supply. Effective hole collection rings cannot be constructed in a pure CMOS process due to the absence of NBL. The doping gradient of the well causes a downward drift of holes toward the underlying substrate, rendering PSD guard rings placed around the edges of the well ineffectual. CMOS processes must therefore rely upon low-resistance substrate contacts to extract any hole current injected into the substrate.

BiCMOS processes can construct hole-collection rings similar to those in Figure 4.24. These rings are not quite as effective on BiCMOS processes as on standard bipolar because the graded profile of the BiCMOS well opposes the potential barrier raised by the NBL. This problem is exacerbated by the relatively light doping of BiCMOS NBL regions required to avoid autodoping the P-epi. Despite these problems, an overall efficiency in excess of 95% is usually achievable by using base hole-collection rings in combination with NBL and deep N+. Section 13.2 discusses several additional types of hole guard rings and some of the difficulties associated with constructing hole guard rings in a BiCMOS process.

[40] Jones, p. 18ff.

Preventative Measures (Cross-injection)

Circuit upsets caused by minority-carrier injection into a tank or well can often be eliminated by placing each potential emitter of minority carriers in its own tank or well. As a rule, any PMOS transistor whose source connects to an external pin should occupy its own well. Similarly, any base resistor, HSR resistor, or lateral PNP collector connecting to a pin is best placed in its own tank. The small amount of extra space required to construct separate tanks or wells will be amply repaid by the elimination of even one parasitic. If, on the other hand, several devices connect to a common pin, then these can all occupy a common tank or well.

The hole-collection rings discussed previously have been designed to minimize injection of holes into the substrate. Another type of minority-carrier guard ring can prevent holes injected by one device from interfering with the operation of other devices in the same tank or well, a problem called *cross-injection.* Consider the case where two lateral PNP transistors occupy a common tank. If either transistor saturates, some fraction of the carriers it emits will be collected by the adjacent transistor. The resulting increase in collector current may disturb the operation of the circuit, particularly if the devices were intended to match one another. Cross-injection can be prevented by placing each transistor in its own tank, but this wastes area because of the large spacings associated with the isolation diffusion. A more compact solution employs a type of minority carrier guard ring called a *P-bar* (Figure 4.25).[41]

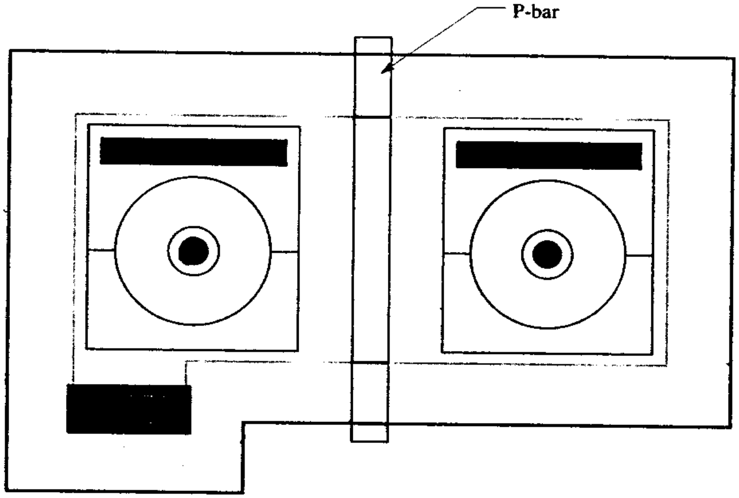


FIGURE 4.25 Example of a P-bar used to prevent cross-injection between two lateral PNPs.

A P-bar consists of a minimum-width strip of base diffusion placed between the two transistors.[42] Each end of the P-bar extends out into the isolation far enough to guarantee electrical contact. This arrangement ensures that the P-bar electrically connects to the isolation without requiring contacts. Now suppose that the lateral PNP on the left side of the P-bar saturates and begins to inject holes into the tank. In order for these holes to reach the lateral PNP transistor on the right, they must first pass underneath the P-bar. The base diffusion forming the bar reaches fairly deeply into the epi and leaves little room for carriers to pass underneath it. Most of

[41] Davis, p. 27; Jones, p. 10.

[42] Jones, p. 10.

the holes traveling from left to right will be collected by the P-bar and shunted to ground. This structure thus acts as a specialized type of minority-carrier guard ring. The presence of NBL beneath the P-bar ensures a low-impedance path for base current passing from the right-hand transistor to the tank contact at the lower left corner of the tank. The tank contact on the left side of this structure therefore suffices for both transistors.

Although the collection efficiency of P-bars can only be determined through empirical measurement, several observations are in order. As the tank bias increases, the depletion region surrounding the bar deepens and progressively pinches off the N-epi underneath it. Devices operating at high tank-to-substrate potentials therefore obtain a higher degree of isolation from a P-bar than devices operating near substrate potential. A wider P-bar also increases collection efficiency, in part because the pinched portion of the tank becomes wider and in part because the wider base region diffuses deeper into the epi. Even a minimum-width P-bar provides a high degree of isolation against minority-carrier cross-injection due to the up-diffusion of the underlying NBL and the formation of a depletion region beneath the P-bar.

The P-bar has many applications. Bipolar circuits often contain current mirrors composed of lateral PNP transistors with a common base connection. These transistors often occupy a common tank, but if one transistor saturates then the currents provided by the adjacent transistors increase. P-bars placed between the saturating transistor and the adjacent devices will prevent this effect without unduly enlarging the tank. Another common application consists of an NPN driving either a lateral or a substrate PNP transistor, in which the collector of the NPN connects to the base of the PNP. Minority-carrier conduction from the PNP to the NPN can initiate positive-feedback latchup by triggering the SCR inherent in this structure. A P-bar placed between the transistors may suppress the latchup, although this is not guaranteed unless the collection efficiency of the bar exceeds the reciprocal of the beta product of the two transistors.

P-bars also find use in CMOS processes, where they typically consist of PMoat. This type of P-bar exhibits a lower collection efficiency than its bipolar counterpart due to the shallowness of the PMoat diffusion and the absence of NBL. The lack of a buried layer greatly increases the well resistance beneath the P-bar, so prudence dictates the inclusion of well contacts on both sides of the bar. This structure can help increase a circuit's latchup immunity without requiring separate wells. If one PMOS transistor in the tank has a source or drain connecting to an outside terminal, then a transient can potentially forward-bias this PMoat into the well. The resulting minority-carrier injection can disturb adjacent transistors and can even lead to latchup. The strategic placement of a few minimum-width PMoat P-bars can provide considerable protection against this sort of cross-injection without consuming as much area as separate wells require.

Another type of minority-carrier guard ring called an *N-bar* can also protect against minority-carrier cross-injection. An N-bar consists of a strip of deep-N+ placed between two devices occupying a common tank (Figure 4.26).[43] The N-bar typically serves as a tank contact for the devices around it since the spacings surrounding the deep-N+ are large enough to allow room for both emitter diffusion and a contact. The doping gradient surrounding the N-bar repels minority carriers, and most of the carriers that overcome this gradient recombine inside the deep-N+ before they pass through it. Unfortunately, the N-bar generally stops short of the

[43] Davis, p. 31.

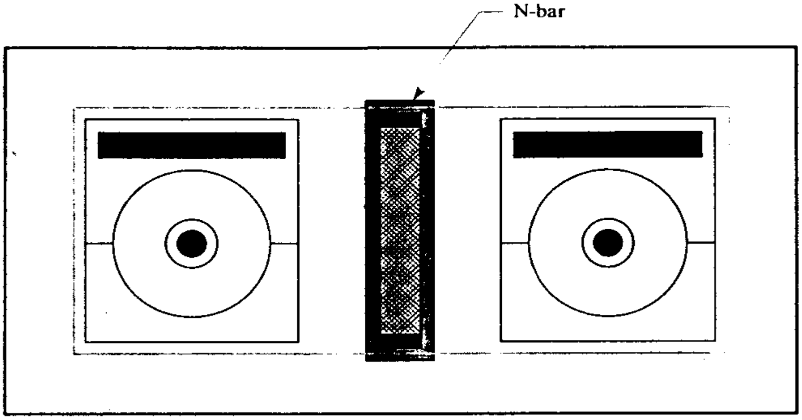


FIGURE 4.26 Example of an N-bar used to simultaneously provide tank contact and to minimize cross-injection between two lateral PNP transistors.

P-isolation on either side of the tank to avoid forming an N+/P+ junction that would break down at a relatively low voltage. These gaps allow minority carriers to bypass the N-bar, so an N-bar usually exhibits a lower collection efficiency than a P-bar. Still, the combination of a highly efficient collector contact with a moderately effective minority-carrier guard ring sometimes finds applications in high-current lateral PNP current mirrors and similar circuits.

4.5 SUMMARY

This chapter discusses a number of common failure mechanisms of integrated circuits. Table 4.2 summarizes these mechanisms, along with typical symptoms and suggested corrective actions. Even a cursory glance at the table reveals the interdisciplinary nature of the subject. Some mechanisms are primarily electrical, while others depend upon chemical or electrochemical processes. Some of these failure mechanisms require a knowledge of device physics to counteract them, while others require knowledge of processing and packaging technology. Only by amassing a working knowledge of many fields can one hope to design integrated circuits that will function reliably over a lifetime of use.

4.6 EXERCISES

Refer to Appendix C for layout rules and process specifications.

4.1. A certain copper-doped aluminum alloy can safely operate at current densities of $5{\cdot}10^5$A/cm^2. If the metallization thickness equals 8kÅ, but thins by 50% when passing over oxide steps, then how much current can a 10μm-wide lead carry across an oxide step?

4.2. Propose a scribe seal structure for a single-level-metal standard bipolar process. Draw a cross section of this structure and explain the purpose of each of its components.

4.3. Lay out a 15kΩ, 8μm-wide HSR resistor. Field plate the resistor as well as possible, including flanges where necessary. The field plate should overhang HSR by at least 6μm and base by at least 8μm.

4.4. Modify the layout from Exercise 4.3 to include channel stops constructed from emitter diffusion. Assume that the channel stops must overlap the field plates by 4μm.

4.5. Construct a minimum-size, standard-bipolar, lateral PNP using a circular emitter geometry. Fully field-plate both the emitter and the collector, leaving space for base metallization. Assume the emitter field plate must overlap the collector by 2μm and the collector field plate must overhang the collector by 8μm.

TABLE 4.2 Summary of failure mechanisms.

Failure Mechanism	Symptoms	Corrective Actions*
Electrostatic discharge (ESD)	Gate oxide ruptures either immediately or after delay, junctions shorted or leaky.	***Add ESD protection devices, do not route leads over thin emitter oxide.***
Electromigration	Open or short circuits after long-term operation, usually at high temperature.	Use copper-doped aluminum, use refractory barrier metal, ***use adquate lead widths, use adequate bondwires.***
Antenna effect	Small gate oxides connected to large conductors suffer delayed failure.	***Reduce ratio of conductor area to gate oxide area, add diodes.***
Dry corrosion	Open circuit failures, moisture accelerates failure.	Use nitride PO, ***minimize PO openings.***
Mobile ions	Threshold shifts under high-temp biased operation, relaxes after unbiased bake.	Use phosphosilicate glass, use polysilicon gate MOS, ***minimize PO openings, use adequate scribe seals.***
Hot carriers (in MOS)	Threshold shifts under high-temp biased operation, relaxes after unbiased bake.	***Limit drain-source voltages, use LDD structures, use long-channel devices.***
Zener walkout	Breakdown voltage drifts, relaxes after unbiased bake.	***Use buried Zener*** (if available).
Parasitic channels & Charge spreading	Leakage currents at high voltage. If they appear after high-temp biased operation and relax after high-temp bake, charge spreading is responsible.	Use (111) silicon, add channel stop implants, ***add base-over-iso, use channel stops, use field plates.***
Substrate debiasing	Latchup, parametric shifts that occur under specific biasing conditions.	***Maximize substrate contact, place contacts near injectors.***
Minority-carrier injection into the substrate	Latchup, parametric shifts that occur under specific biasing conditions.	Use P+ substrate, ***maximize substrate contact, separate sensitive circuitry, add NBL to shared wells, use deep-P+ in isolation, add guard rings.***
Minority carrier cross-injection	Latchup, mismatches between merged devices.	***Use P-bar or N-bar, place devices in separate tanks or wells.***

* Possible solutions listed in ***bold italics*** are under the control of circuit and layout designers; the remaining solutions can only be implemented by process engineers.

4.6. Compute the area of substrate contacts necessary to extract 25mA from a die that uses an 8μm-thick, 10Ω·cm, P-type epi layer on top of a 0.01Ω·cm P-type substrate. Assume a maximum allowed debiasing of 0.3V.

4.7. Lay out a standard-bipolar NPN transistor with a 20μm by 40μm emitter. Arrange the transistor to minimize the distance between the emitter and collector contacts. The transistor should include deep-N+ in the collector to reduce collector resistance. Place an electron-collecting guard ring around this transistor, following the cross section shown in Figure 4.22.

4.8. Lay out a 2000/5 PMOS transistor. Divide the transistor into a sufficient number of fingers to obtain a roughly square aspect ratio. Construct a hole-collecting guard ring similar to that in Figure 4.24 that encircles the PMOS transistor. Make sure that the NBL overlaps the deep-N+ diffusion by at least 4μm to provide an adequate seal at the point of intersection.

4.9. Lay out an example of a P-bar separating two minimum-size standard bipolar lateral PNP transistors. The P-bar should extend at least 4μm into the isolation to ensure electrical continuity.

4.10. Several failed devices have been de-encapsulated (*decapped*) for microscopic examination. Suggest at least one failure mechanism consistent with each of the following observations:

a. A metal trace from a bond pad has melted open.

b. A greenish deposit covers the bond pads.

c. The gate oxide of a minimum-size NMOS has ruptured at one point, shorting the poly gate to the underlying epi.

d. A thin, dark filament appears across the base of a large NPN transistor. The transistor's base-collector junction appears shorted.

4.11. A new high-voltage, low-current operational amplifier has just completed burn-in testing. Sample units were operated under bias at 150°C for 1000 hours. Parametric testing reveals that the input offset voltages of the amplifier have drifted several millivolts during testing, and the supply currents have increased by 20%. What failure mechanisms might be responsible for these symptoms, and how can the designer determine what to fix?