13 *Special Topics*

The previous chapters have presented the details of constructing and matching resistors, capacitors, diodes, and transistors. Integrated circuits also contain a number of more specialized components, including merged devices, guard rings, tunnels, bondpads, and ESD protection devices.

Merged devices appear separate from one another in schematics, but they are combined in the layout. Mergers not only save space but in some cases also improve performance. The designer must weigh the benefits of mergers against the possibility of introducing unexpected interactions between merged devices.

Guard rings prevent minority carriers injected by one device from interfering with the operation of another device. Not only do guard rings prevent latchup, but they also block noise coupling that might otherwise interfere with the operation of low-power circuitry.

Tunnels are low-value resistors used as signal crossing points. Single-level-metal layouts usually include at least a few tunnels. Multiple-level-metal layouts do not require them, but they occasionally offer a convenient way to route leads to otherwise-inaccessible areas of the die. Tunnels introduce parasitics that can degrade circuit performance, so each proposed tunnel must be carefully analyzed for potential problems.

Bondpads allow the connection of the integrated circuit to the external world. Most bondpads require electrostatic discharge (ESD) protection circuitry. *Trimpads* and *testpads* are accessible only during wafer-level probing, so they do not require ESD protection.

13.1 MERGED DEVICES

The largest spacings in standard bipolar are those associated with the isolation diffusion. Most circuits contain components whose tanks are connected to the same potential. Considerable space can be saved by placing these devices in common tanks. Figure 13.1A shows three minimum-geometry NPN transistors laid out side-by-side, while Figure 13.1B shows the same three transistors merged into a common tank.

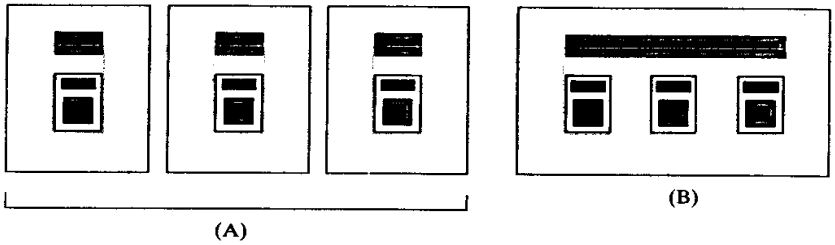


FIGURE 13.1 (A) Three separate NPN transistors and (B) the same transistors merged into one tank.

The merged devices require only about 70% of the area of the separate devices. Additional area can be saved by reducing the size of the collector contact.

The largest spacings in CMOS and BiCMOS processes are those associated with the isolated well (or wells). The merger of components into common wells can again save considerable die area. This is particularly true for designs containing large numbers of small components, such as minimum-size MOS transistors.

The injudicious merger of devices that should have remained separate has caused many circuit malfunctions, some of which have proved remarkably difficult to diagnose and fix. Consequently, designers have become somewhat reluctant to merge devices, even when mergers would obviously result in significant area savings.

Virtually every failure caused by merged devices can be traced to one of three sources: minority carrier injection, Ohmic debiasing, or capacitive coupling. If a designer understands these three mechanisms, then he or she can discern which mergers are safe and which should be avoided. Minority carrier injection is by far the most serious problem encountered in designing merged devices. Trouble can occur whenever a device injects minority carriers into a shared region such as a tank or a well. Some of the injected minority carriers transit to other devices, where they are collected by reverse-biased junctions. The flow of minority carriers between devices that should remain isolated causes leakage currents to appear at unexpected points in the circuit. These currents can cause circuit malfunctions ranging from subtle parametric shifts to catastrophic latchup. The following section discusses several device mergers that are known to have minority carrier cross-injection problems.

13.1.1. Flawed Device Mergers

Figure 13.2A shows a split-collector lateral PNP transistor. This structure essentially consists of a pair of merged lateral PNP transistors sharing common emitter and base regions. Under normal operating conditions, collectors C_1 and C_2 both remain

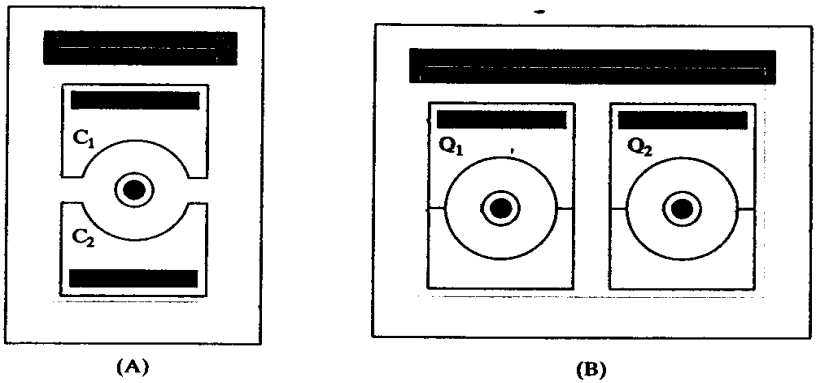


FIGURE 13.2 Merged lateral PNP structures susceptible to cross-injection: (A) split-collector lateral PNP and (B) two lateral PNP transistors merged in a common tank.

reverse-biased. Holes flow radially from the shared emitter to the two collectors, and each collector intercepts about half of the total emitter current. Now suppose that collector C_1 saturates. The minority carriers that should have been absorbed by C_1 are now re-injected from its surfaces. Most of these re-injected carriers flow to the isolation sidewalls and thence to the substrate, but some also flow from collector C_1 to collector C_2. This *cross-injection* increases the current flowing out of C_2 and imbalances the apparent ratio between the two collectors.

Most designers know that the saturation of a split-collector PNP causes cross-injection, but many overlook the possibility of cross-injection between separate lateral PNP transistors occupying a common tank. Figure 13.2B shows a tank containing two lateral PNP transistors, Q_1 and Q_2. Normally the collectors intercept virtually all of the minority carriers injected by their respective emitters, so the two transistors remain effectively isolated from each other, even though both occupy the same tank. Now suppose that Q_1 saturates while Q_2 continues to operate in the normal active region. The holes injected by the emitter of Q_1 transit to its collector, but when this collector forward-biases, they are re-injected back into the tank. Since the collectors of Q_1 and Q_2 face each other across a relatively narrow gap, most of the carriers launched from Q_1 toward Q_2 reach its collector. Hence when Q_1 saturates, about a quarter of its emitter current flows to Q_2.

The simplest way to prevent cross-injection between lateral PNP transistors is to place each transistor in its own tank. However, this wastes so much space that designers have devised other methods of preventing (or at least minimizing) cross-injection between merged devices. For example, two lateral PNP transistors can be separated from one another by a P-bar or an N-bar (Section 4.4.2). These bars block most, but not all, of the cross-injected minority carriers. The designer should consider what would happen to the circuit if, for example, 5% of the current injected by the saturating device were to reach adjacent devices. If this amount of current could cause a malfunction, then the devices would require separate tanks.

Another example of minority carrier injection involves the merger of an NPN transistor, Q_2, driving a lateral PNP, Q_1 (Figure 13.3).[1] The collector of Q_2 is electrically common to the base of Q_1. This merged device will probably operate satisfactorily as long as Q_1 does not saturate. If it does, then holes re-injected by its collector will flow to the base of NPN Q_2. This additional base current drives Q_2 harder than before and provides additional collector current. The increased collector current feeds Q_1 and increases its emitter current. This situation is a classic example of

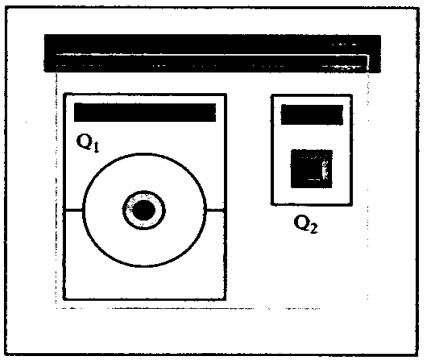


FIGURE 13.3 Example of a device merger prone to latchup due to minority-carrier injection.

[1] C. Jones, "Bipolar Parasitics," unpublished manuscript, 1987, pp. 27–29.

SCR latchup. Once latchup has been triggered, it will continue until the power is interrupted. The die may overheat and self-destruct at high supply voltages; otherwise it simply malfunctions and consumes excessive supply current. The potential for latchup makes this structure risky, even if the lateral PNP never saturates during normal operation. If some transient condition saturates the lateral PNP, then the structure will latch up.

The use of P-bars or N-bars to suppress device latchup is potentially risky. In order for latchup to occur, the product of the beta of the PNP, β_P, and the beta of the NPN, β_N, must exceed unity. The addition of a bar between the two transistors adds a third term, η_c, representing the fraction of minority carriers intercepted by the bar. The condition for latchup (Section 11.2.7) then becomes

$$\beta_N \beta_P (1 - \eta_c) > 1 \qquad [13.1]$$

The beta of a standard bipolar NPN transistor may equal 300 or more. The beta of the lateral PNP is lower than that of a conventional transistor because the actual basewidth equals the separation between the transistors. Even so, the beta of the lateral PNP could easily equal ten. In order to prevent latchup with these betas, the efficiency of the bar must exceed 0.997, or 99.7%. N-bars will almost certainly fall short of this efficiency, and P-bars may not always achieve it. If there is the slightest chance that a lateral PNP may saturate, then it should not occupy the same tank as an NPN transistor.

Figure 13.4 illustrates another pair of merged devices prone to latchup.[2] This example is particularly noteworthy because the latchup stems from the interaction of two separate mechanisms, namely Ohmic debiasing and minority carrier injection. This structure places an NPN transistor Q_1 in the same tank as a base resistor R_1. The NPN is configured as an emitter follower and one end of the base resistor connects to the supply. The tank contact serves as the collector contact of Q_1 and the tank contact of R_1. Both the base-collector junction of Q_1 and the base-tank junction of R_1 remain reverse-biased under normal conditions. This situation may change if Q_1 draws enough collector current through the shared tank contact. If the voltage drop between the tank contact and the intrinsic collector of Q_1 becomes large enough, then the positive end of R_1 will forward-bias into the tank. This is an example of *Ohmic debiasing*. Some of the minority carriers injected by R_1 reach the base

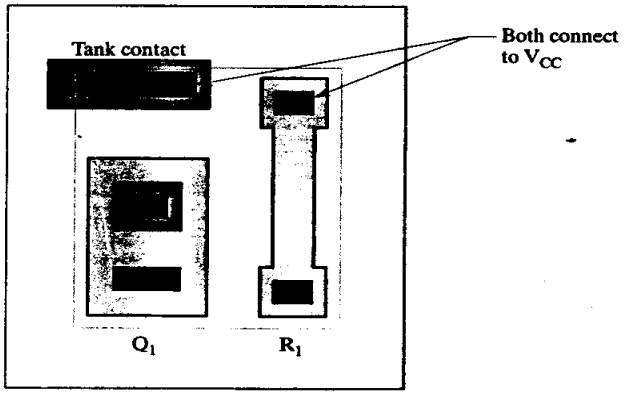


FIGURE 13.4 Another example of a device merger prone to latchup due to minority-carrier injection.

[2] W. F. Davis, *Layout Considerations*, unpublished manuscript, 1981, pp. 32–33.

of Q_1, where they provide additional base drive. Transistor Q_1 now pulls even more collector current. The additional collector current increases the Ohmic debiasing experienced by R_1, causing R_1 to inject additional holes into the tank. The resulting positive feedback causes the circuit to latch up. As with the previous example, latchup occurs because of the presence of an SCR, which in this case consists of base resistor R_1, the shared tank, and the base and emitter of Q_1.

Although the structure in Figure 13.4 contains an SCR, it will not latch up unless triggered by a voltage drop produced by Ohmic debiasing within the tank. The voltage drop required to trigger the SCR equals about 0.3V at 150°C (see Section 4.4.1). If the NPN transistor conducts an average of 100μA, then the tank resistance must equal 3kΩ to produce 0.3V of debiasing. The vertical resistance between the tank contact and the NBL can equal hundreds, if not thousands, of Ohms if the transistor lacks a deep-N+ sinker. The inclusion of even a minimum plug of deep-N+ reduces the tank resistance to no more than a few hundred Ohms. The structure in Figure 13.4 is therefore likely to latch up without deep-N+, but it is very unlikely to do so if a sinker is present.

Ohmic debiasing can also cause capacitive coupling of noise into sensitive nodes. Using the merger of Figure 13.4 as an example, suppose that the debiasing is insufficient to actually trigger latchup. Even so, the current flowing through Q_1 still causes some voltage drop within the tank. If the operation of Q_1 causes a rapid fluctuation of the collector current, then this will in turn generate a high-frequency ripple on the tank voltage. This signal can couple through the capacitance of the reverse-biased junction surrounding R_1. If R_1 is part of a sensitive circuit, then the noise injected by tank voltage fluctuations can cause problems. Designers should avoid merging noisy circuitry and sensitive circuitry in the same tank. Although some such mergers function satisfactorily, many do not.

Figure 13.5 shows another pair of problematic merged devices. This example merges an NPN transistor, Q_1, with a Schottky diode, D_1. The collector of Q_1 connects to the cathode of D_1 through the tank. Since Schottky diodes are majority-carrier devices, this might appear a safe merger. Unfortunately, this is not the case. Most Schottky diodes incorporate a field-relief guard ring consisting of a P-type diffusion. This guard ring begins to inject minority carriers into the tank as soon as the voltage across the Schottky rises above the forward voltage of the guard-ring junction. The series resistance of small Schottky diodes can equal hundreds or thousands of Ohms, so their guard rings are easily debiased into conduction.

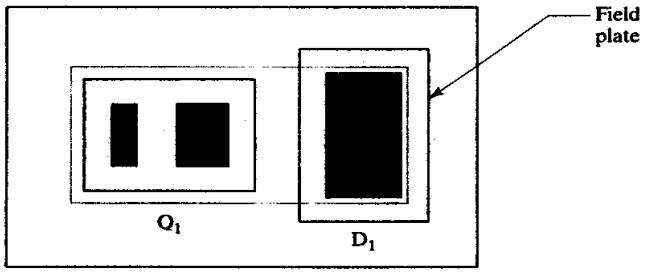


FIGURE 13.5 Another structure prone to minority-carrier cross-injection, consisting of an NPN transistor, Q_1, and a Schottky diode, D_1.

The structure in Figure 13.5 uses a field-plated Schottky diode instead of a guard-ringed Schottky. This eliminates the possibility of the guard ring forward-biasing into the tank, but low levels of cross-injection can still occur due to the presence of the Schottky barrier itself. Although a rectifying Schottky junction conducts primarily by means of majority carriers, small numbers of minority carriers are also injected into

the semiconductor side of the junction. The minority carriers injected by Schottky diode D_1 can travel to the base of Q_1, where they appear as additional base drive. This mechanism causes parametric shifts and could potentially trigger latchup.

The type of malfunction that occurs in the circuit in Figure 13.5 can potentially occur in an ordinary NPN transistor. If a heavily doped diffusion does not entirely enclose the collector contact, the portion of the contact touching the lightly doped epi forms a Schottky barrier. This barrier can inject minority carriers in much the same way as the Schottky diode in Figure 13.5. This problem normally occurs in structures that have been incorrectly laid out, but misalignments caused by poor photolithography has been known to produce the same result in structures that pass all applicable design rules.

13.1.2. Successful Device Mergers

This section presents two device mergers of the sort often encountered in standard bipolar layouts. There are countless possible mergers, and the examples given here only provide a general impression of what a skilled designer can achieve. Additional mergers can be discovered in the layout of many standard bipolar integrated circuits.

The *Darlington pair* in Figure 13.6A consists of a power NPN transistor, Q_1, and a smaller predrive transistor, Q_2, both sharing a common collector connection. Each transistor also has an associated base turnoff resistor. The layout of Figure 13.6B shows how all four of these components can occupy the same tank. The tank contact consists of a bar of deep-N+ placed along the left edge of the tank. The deep-N+ sinker does not encircle the power device, Q_2, because such an arrangement would consume additional die area. In general, only saturating NPN transistors and large power devices require full deep-N+ rings. Q_2 is unlikely to saturate because the extrinsic collector-to-emitter voltage V_{CE} cannot drop below the sum of the extrinsic V_{BE} of Q_2 and the extrinsic V_{SAT} of Q_1. The extrinsic V_{BE} of Q_2 probably approaches 1V at high current levels, so the tank contact can debias by the better part of a volt before Q_2 begins to saturate. The deep-N+ bar in Figure 13.6B probably exhibits no more than 5 to 10Ω of vertical resistance, so it can handle several hun-

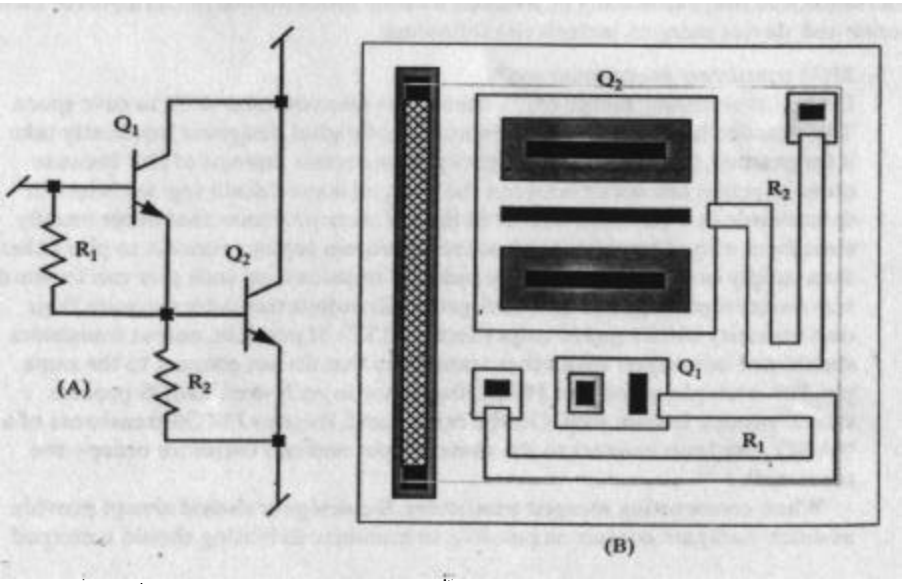


FIGURE 13.6 (A) Schematic and (B) layout of a merged Darlington pair (metal omitted for clarity).

dred milliamps of current without debiasing enough to allow Q_2 to saturate. Q_1 can saturate, but the current flow through this device is small enough to limit substrate injection to manageable levels.

Even if Q_1 saturates, it will not interfere with the operation of the Darlington. When Q_1 saturates, it will be delivering as much current as possible into Q_2. Holes injected by Q_1 into the common tank will be collected by R_1, R_2, or Q_2. The holes collected by R_1 will either return to the base of Q_1 or will flow to the base of Q_2. The influx of additional base drive into either transistor causes no harm because both transistors are already conducting all the current they can. Holes collected by Q_2's base simply represent additional base drive for Q_2. Holes collected by R_2 will probably flow to the emitter of Q_2, there to combine with the much larger current flowing through this transistor. Again, the influx of extra current causes no harm. In conclusion, it makes little difference whether or not Q_1 saturates.

The base contact of each NPN transistor also serves as a contact head for its respective base turnoff resistor. This merger can save considerable area, but the HSR implant must be spaced far enough away from the emitter to prevent this implant from raising the doping concentration in the intrinsic base region of the NPN. This layout achieves the necessary separation without enlarging the transistor by running the HSR implant into the transistor base behind the base contact.

The devices in Figure 13.6 have been arranged to enable interconnection with one level of metal. The reader may wish to mentally trace the connections between the contacts. The collector lead enters the tank from the left and the emitter lead exits from the right. These leads can be as wide as desired. The lead connecting the emitter of Q_1 to the base of Q_2 passes between the tank contact and the body of Q_2.

Figure 13.7A shows another example of a circuit that can benefit from device mergers. Transistors Q_1 and Q_2 act as a differential pair. Three-fourths of the emitter

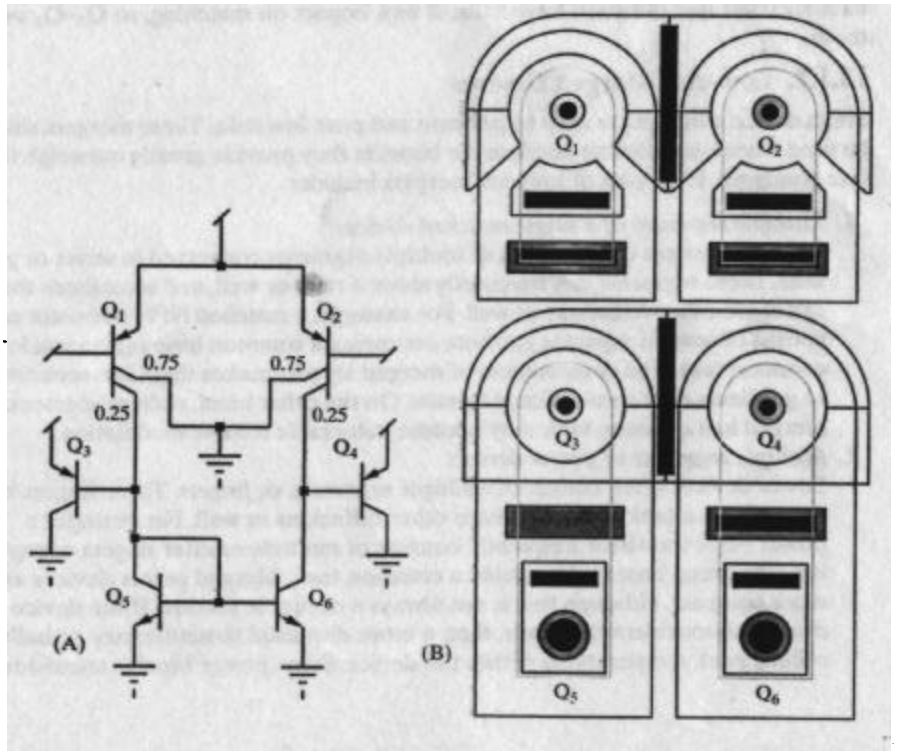


FIGURE 13.7 (A) Schematic and (B) layout of a merged operational amplifier input stage. The metallization has been omitted for clarity.

current of Q_1 and Q_2 is shunted to ground, while one-fourth feeds a current mirror consisting of Q_5 and Q_6. The output of this circuit is taken through a substrate PNP emitter follower, Q_4. An identical substrate PNP, Q_3, balances the load on the circuit and eliminates a systematic offset otherwise created by the base current of Q_4.

The layout of this circuit requires no fewer than four tanks. Q_1 and Q_2 have separate base connections and therefore require individual tanks. Q_3 and Q_5 can reside in a common tank, as can Q_4 and Q_6. The larger collectors of Q_1 and Q_2 connect to ground through an extension of the collector into the surrounding isolation. This saves considerable area by eliminating the spacing between the collector and the isolation. The size of these transistors can be further reduced by aligning the smaller collector segments to allow them to fit within a narrower tank (Figure 13.7B). The narrower tank cannot contain enough NBL to fully floor the active region of the transistor, but NBL outdiffusion prevents any significant leakage of minority carriers to the substrate.

The mergers of Q_3–Q_5 and Q_4–Q_6 present more significant problems. Each of these structures places a substrate PNP in the same tank as an NPN. The substrate PNP injects holes that could trigger latchup if they reached the merged NPN transistors. The illustrated layout circumvents this problem by substituting lateral PNP's for the substrate PNP's normally used in this role. The lateral PNP collector functions as a P-bar. This collector also extends out into the isolation to save space. The P-bar may not entirely block minority carrier flow, but several standard bipolar designs have successfully used this layout. Latchup has not been observed, and any low level of hole collection experienced by Q_5 is also experienced by Q_6. Since Q_5 and Q_6 balance one another, the circuit inherently tolerates low levels of cross-injection between Q_3–Q_5 and Q_4–Q_6.

The mergers of Q_3–Q_5 and Q_4–Q_6 are not quite identical because Q_4–Q_6 requires a tank contact while Q_3–Q_5 does not. Both mergers incorporate identical strips of emitter to ensure matching, but only Q_4–Q_6 includes a tank contact. The contact and its associated metallization have little, if any, impact on matching, so Q_3–Q_5 omits these.

13.1.3. Low-risk Merged Devices

Some device mergers are easy to perform and pose few risks. These mergers should be used whenever possible because the benefits they provide greatly outweigh their disadvantages. Examples of low-risk mergers include:

1. *Multiple segments of a single matched device.*
   Matched devices often consist of multiple segments connected in series or parallel. These segments can frequently share a tank or well, and sometimes they can share other diffusions as well. For example, a matched NPN transistor can consist of several separate emitters occupying a common base region inside a common tank. The compactness of merged layouts makes them less sensitive to gradients than conventional layouts. On the other hand, resistor segments merged in a common tank may become vulnerable to tank modulation.
2. *Multiple segments of power devices.*
   Power devices often consist of multiple segments, or *fingers*. These fingers normally share a tank, and may share other diffusions as well. For example, a power NPN transistor frequently consists of multiple emitter fingers occupying a common base region inside a common tank. Merged power devices are more compact, although this is not always a desirable feature. If the device dissipates considerable power, then a more dispersed structure may actually reduce peak temperatures within the device. Some power bipolar transistors

interdigitate base regions with strips of deep-N+ to spread power dissipation over a larger area while simultaneously reducing collector resistance.

3. *Sense transistors and their associated power transistors.*
   A power transistor sometimes has an associated sense transistor. The current passing through the sense transistor is smaller than, but proportional to, the current passing through the power transistor. The sense and power transistor must match fairly precisely despite the presence of large thermal gradients. Ideally, the sense device should consist of two equal segments located on an axis of symmetry passing through the power device. Each segment should reside approximately halfway between the center of the power device and its periphery so that its temperature roughly matches the average temperature of the whole power device. If the sense device cannot be divided into segments, then it should reside on the periphery of the power device on one of its axes of symmetry.
4. *Schottky clamps and their associated NPN transistors.*
   Schottky-clamped NPN transistors are usually constructed as merged devices. The merged Schottky clamp can use the same deep-N+ sinker as the transistor, and if necessary it can also use an extension of the NPN base region as a guard ring.
5. *NPN Darlington transistors.*
   Darlington transistors can use a layout similar to that in Figure 13.6. Most Darlington transistors are power devices that require custom layouts. Only a little additional time and effort are required to incorporate the predrive transistor and turnoff resistors in the same tank.
6. *Base turnoff resistors for NPN transistors.*
   NPN transistors often require base turnoff resistors. If these resistors are diffused devices, then they can occupy the same tanks as their associated NPN transistors.

13.1.4. Medium-risk Merged Devices

Another class of device mergers is easy to perform and offers substantial area savings, but these mergers are not without an element of risk. The risks involved are understood, and they can usually be avoided without much difficulty. Examples of moderate-risk device mergers include the following:

1. *MOS transistors in common wells.*
   Designers routinely merge MOS transistors into common wells to save space. This practice has become so widespread that digital designers frequently take it for granted. Actually, such mergers pose a certain amount of risk because cross-injection can occur between the merged source/drain regions, whether these reside in a common well or in the epi. Any problems that occur usually stem from output transistors whose source/drain regions connect to pins other than supply or ground. Externally induced transients on such pins can forward-bias source/drain regions into backgates. All output transistors require their own minority carrier guard rings (Section 13.2). If possible, output transistors should not be merged with other transistors that do not connect to the same pin. For example, an output PMOS transistor in an N-well CMOS process should occupy its own well. On the other hand, the two PMOS transistors of a NAND gate both connect to the same output and can therefore occupy the same well.

   When constructing merged transistors, the designer should always provide as much backgate contact as possible to minimize debiasing should a merged

transistor forward-bias. If the process includes a suitable buried layer, then the well should contain as large an area of the buried layer as possible. Well contacts become more important as the size of the well and the number of transistors it contains increase. Very large MOS transistors often require integrated backgate contacts (Section 11.2.7).

Some circuits contain MOS transistors that regularly forward-bias into their backgate regions during normal operation. For example, some charge pumps contain devices that forward-bias during startup. These devices do not necessarily connect to pins and are not always easily identified. The circuit designer should clearly identify all such devices so that the layout designer can protect them using guard rings and separate wells.

2. *Diffused resistors in common tanks.*
   Diffused resistors are frequently merged in common tanks to save space. This practice allows the construction of compact resistor arrays that are less vulnerable to stress and thermal mismatch. Cross-injection cannot occur between the merged resistors as long as none of them forward-bias into the tank. Problems can occur if any of the resistors connect to a pin that is neither power nor ground. Such resistors should occupy their own tanks, and they may also require guard rings if the process is prone to latchup or if the transients occur during normal operation. Some circuits operate one or more resistors under biasing conditions that might cause minority carrier generation. The circuit designer should clearly mark each such resistor so that it can receive its own tank, along with any necessary guard rings. The layout designer should also be wary of merging noise-sensitive resistors with components carrying high-frequency signals, because of the potential for capacitive coupling. If in doubt, use separate tanks.
3. *Lateral PNP transistors.*
   Lateral PNP transistors sharing the same base connection can occupy the same tank. Many bipolar designs make extensive use of lateral PNP mergers. As long as none of the merged transistors saturate, their collectors act as P-bars isolating them from one another. Minority carrier cross-injection can occur if any of the transistors saturates. P-bars and N-bars (Section 4.4.2) can at least partially block cross-injection between adjacent lateral PNP transistors, but the safest solution consists of placing the saturating transistors in their own tanks.
4. *Split-collector lateral PNP transistors.*
   A split-collector lateral PNP is really a type of merged lateral PNP. A single split-collector transistor can perform the role of several ordinary lateral transistors while consuming much less die area. As long as none of the collectors saturates, few holes escape between the segments of a split-collector device. The saturation of any of the split collectors causes the currents flowing through the remaining collectors to increase. No way exists to block cross-injection in split-collector transistors, short of replacing the offending split-collector devices with separate transistors.
5. *Zener diodes.*
   Emitter-base Zener diodes can be merged with other components in a common tank as long as the tank voltage always equals or exceeds the voltage on the Zener's cathode. This condition ensures that the parasitic NPN transistor does not conduct. Series-connected Zener diodes can also occupy a common tank biased to a potential equal to or greater than that of the cathode end of the Zener string.

13.1.5. Devising New Merged Devices

Any imaginative designer will find many additional opportunities to merge devices. Before implementing a proposed device merger, determine whether it can pass the following three tests:

1. *Can any of the merged devices inject minority carriers into the shared tank or well?*
   If not, then the merged devices are safe from cross-injection and latchup. Possible sources of minority carriers include saturating bipolar transistors, forward-biased Schottky diodes, and any diffusion connecting to an external pin. Designers should be particularly wary of merging NPN transistors with other devices that can potentially inject minority carriers into the tank, because the resulting PNPN structure may latch. Potential sources of minority carriers should either reside in their own tanks or should be guarded by P-bars or N-bars unless the designer can show that cross-injection will not upset the operation of the circuit.
2. *Can any of the merged devices pull substantial current through the tank or well contact?*
   If so, then these devices may cause debiasing. Those that pull relatively small amounts of current (up to a few milliamps) rarely cause objectionable debiasing as long as the tank or well contact contains a plug of deep-N+. Higher-current devices require more extensive deep-N+ regions to prevent debiasing.
3. *Can noise coupling upset the circuit?*
   If the merged tank or well contains both noisy devices and noise-sensitive devices, then capacitive coupling between these can degrade circuit performance. The potential for noise coupling is particularly great if the noisy device pulls significant current through the tank contact.

13.2 GUARD RINGS

Of all the many types of failures that plague integrated circuits, none is so frustrating and so elusive as latchup. Devices that operate properly in one circuit latch up the moment they are inserted into another. Sometimes a device operates properly for hundreds or thousands of hours before it latches. Simulation rarely uncovers latchup problems, and neither do most forms of testing.

The most frequent causes of device latchup are external transients that pull device pins above supply or below ground. Common sources of such transients include low-level ESD events; momentary power interruptions; inductive kick-back from relays, motors, and solenoids; and inductive spiking of rapidly switched signals. Proper board-level design minimizes, but does not eliminate, these transients. Circuit designers must ensure that their designs can withstand at least moderate levels of transient injection without latching up or otherwise malfunctioning.

Power supply pins and substrate connections rarely trigger latchup, but any other pin (including grounds not connected to substrate) can cause problems. The designer should trace every lead from such a pin back through the circuit to determine whether or not it connects to any diffusions. Each diffusion connecting directly to the pin can inject minority carriers when the pin flies above supply or below ground. Diffusions connecting to pins through deposited resistors still pose a concern if the series resistance is less than about 50kΩ. Larger-value deposited resistors reduce injected currents so much that they no longer pose any significant threat.

Latchup can be suppressed by enclosing each vulnerable diffusion (or device) in a suitable minority carrier guard ring. Multiple diffusions connecting to a common pin can share a common guard ring. ESD devices residing around the periphery of

the die can often share a common guard ring separating the core of the die from the ESD devices and bondpads. Many older standard bipolar designs omitted guard rings from some or all pins, but new designs should not follow this practice because it sometimes results in costly redesigns.

13.2.1. Standard Bipolar Electron Guard Rings

Any tank connecting to a device pin can inject electrons into the substrate. No way exists to block these electrons from reaching the substrate. Guard rings do exist that can collect electrons that touch them, but nothing stops electrons from burrowing underneath the guard rings and escaping. The structure in Figure 13.8A is arguably the best electron guard ring that can be constructed in standard bipolar. It consists of a strip of deep-N+ residing in an N-tank augmented by both NBL and emitter.[3] This combination of diffusions forms the deepest possible guard ring and therefore collects the largest possible fraction of electrons. The presence of deep-N+ also helps prevent Ohmic debiasing. This guard ring would ideally connect to the highest available supply voltage to drive the depletion region as deeply as possible into the substrate. This style of guard ring also functions if it is connected to ground, but grounded guard rings are more susceptible to debiasing. Grounded guard rings are sometimes used to minimize power dissipation caused by minority carrier injection, which sometimes becomes a concern in high-current designs. If a grounded guard ring is used to minimize power dissipation, it can be supplemented by a second guard ring connected to the supply and placed outside of the grounded guard ring. This secondary guard ring will provide protection if the grounded guard ring saturates.

One can sometimes take advantage of adjacent tanks that connect to the supply. If these tanks are strategically situated between the point of minority carrier injection and adjacent sensitive circuitry, then they become very effective guard rings. All tanks used for this purpose should contain as much NBL as possible and should use deep-N+ sinkers to minimize debiasing. The efficiency of an electron-collecting guard ring also increases if it is placed next to the source of injected carriers to take advantage of the proximity effect.

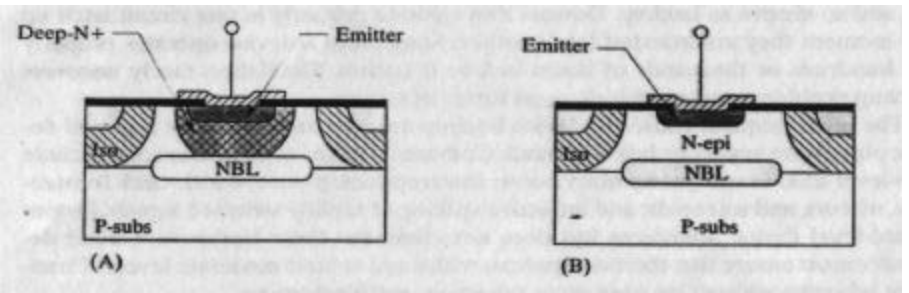


FIGURE 13.8 Electron-collecting guard rings for standard bipolar: (A) preferred structure and (B) alternate structure.

The guard ring in Figure 13.8B should only be used if deep-N+ is not available. The vertical resistance of the epi layer separating the NBL from the emitter diffusion makes this guard ring extremely vulnerable to Ohmic debiasing. This structure is still marginally effective as long as it connects to a power supply, but it is virtually useless when connected to ground.

---

[3] A similar guard ring for a BiCMOS process is presented in E. Bayer, W. Bucksch, K. Scoones, K. Wagensohner, J. Erdeljac, and L. Hutter, "A 1.0μm Linear BiCMOS Technology with Power DMOS Capability," *BCTM Proceedings*, 1995, pp. 137–141.

Electron guard rings constructed in standard bipolar are only marginally effective, yet they consume vast amounts of die area. Most designers omit these guard rings to save space and rely instead on large spacings and strategically placed hole guard rings to prevent latchup. These measures usually suffice for linear circuits such as operational amplifiers and voltage regulators. Devices that switch inductive loads are another matter entirely, as these loads can generate extremely energetic transients during normal operation. Even if these transients do not cause latchup, they can still inject noise into sensitive circuitry. High-frequency MOSFET gate drivers can also experience severe transients caused by resonance in the gate lead. The output circuitry of MOSFET gate drivers and inductive load drivers must be carefully shielded by electron guard rings to minimize noise coupling and latchup sensitivity.

Electron guard rings become much more effective if the process incorporates a P+ substrate. The P+/P– interface generates an electric field that traps most of the injected electrons in the P– epi. Those few that do penetrate into the P+ substrate quickly recombine. The P+ substrate makes deep guard rings such as the one in Figure 13.8A extremely effective, particularly if they are biased to a high-enough potential to drive a depletion region down to meet the P+ substrate.

13.2.2. Standard Bipolar Hole Guard Rings

Any P-type region can inject holes into a tank. Hole guard rings can prevent these carriers from flowing to adjacent P-type regions or to the sidewalls of the tank. Two types of hole guard rings exist: the *hole-collecting guard ring* and the *hole-blocking guard ring*. Figure 13.9A shows a typical hole-collecting guard ring deployed to prevent holes from reaching the sidewalls of a tank. The presence of NBL prevents the holes from flowing down to the substrate and instead forces them to flow laterally. The guard ring consists of a reverse-biased base diffusion surrounding the point of injection. This diffusion acts as the collector of a lateral PNP transistor. Any holes reaching the depletion region surrounding the guard ring are drawn into it. Hole-collecting guard rings are normally grounded to maximize the reverse bias between the tank and the guard ring. This not only drives the depletion region deeper into the tank but also minimizes the effects of Ohmic debiasing within the guard ring itself. Grounded hole guard rings tie to the same potential as the isolation system, so they can be merged to save space. Examples of such merged guard rings include the P-bar in Figure 4.24 and the grounded collectors of transistors Q_3 and Q_4 in Figure 13.7. A hole-collecting guard ring can also be tied to the tank potential, but this reduces its effectiveness and does not save any appreciable space.

Figure 13.9B shows a typical example of a hole-blocking guard ring.[4] This type of guard ring surrounds the point of injection with heavily doped N-type regions. The

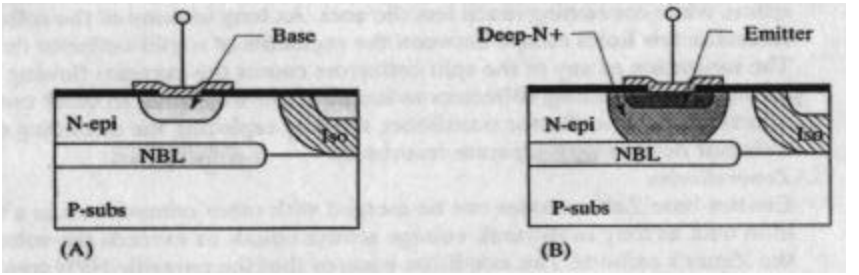


FIGURE 13.9 Hole guard rings for standard bipolar: (A) hole-collecting guard ring and (B) hole-blocking guard ring.

[4] *Ibid.*

N+/N– interface generates an electric field that acts as a barrier to the passage of holes. Any holes overcoming this barrier typically recombine before they can traverse the N+ region. The electron current required to sustain recombination flows through contacts to the N+ region. The structure in Figure 13.9B relies on NBL to block the downward travel of holes, and deep-N+ to block their lateral movement. In order to be fully effective, a hole-blocking guard ring must contain no gaps or holes. The only practical way to achieve this goal consists of entirely encircling the injector with a ring of deep-N+. Partial hole-blocking rings such as the N-bar in Figure 4.25 may allow a substantial fraction of the holes to escape around the gaps at either end. In some processes, these gaps can be eliminated by extending the deep-N+ bar into the isolation on either side of the tank. Most processes do not allow this configuration because of leakage across the isolation/deep-N+ junction. Both hole-collecting and hole-blocking guard rings can provide efficiencies of 95% or better in typical standard bipolar processes. A combination of both types of guard rings (Figure 4.24) can achieve an efficiency in excess of 99%.

Standard bipolar designs use relatively few hole guard rings because this process rarely requires them. Standard bipolar designs seldom experience latchup due to external transients because the deep-P+ isolation and the large spacing between components both help reduce the beta product of the parasitic SCR (Section 11.2.7). Hole injection into the substrate only becomes a problem if it overwhelms the capability of the substrate contact system. This is unlikely to occur because the grid of P+ isolation diffusion helps to magnify the effective area of substrate contacts, while the P– isolation helps to limit the maximum injected current and to contain substrate debiasing within relatively limited regions of the die. Hole guard rings are usually employed only to prevent cross-injection between merged components (Section 13.2.1). P-type regions connecting to external pins that are neither power supplies nor substrate ground are usually isolated by placing them in their own tanks. This practice requires about the same amount of space as the construction of hole guard rings and requires less effort.

13.2.3. Guard Rings in CMOS and BiCMOS Designs

CMOS designs are more prone to latchup than standard bipolar. This vulnerability results in part from the smaller dimensions of modern CMOS and BiCMOS processes and in part from differences between isolation systems. CMOS-derived processes usually substitute a lightly doped epitaxial layer for the vertical P+ isolation of standard bipolar. The light doping increases the gain of the lateral bipolar transistor formed across the isolation and makes it more probable that minority carrier injection will trigger SCR action. The light doping of the P-epi also makes it more difficult to extract substrate current. Most of these processes rely on the presence of a P+ substrate to reduce their vulnerability to latchup through the substrate, but scrupulous care must be taken to block lateral conduction by using guard rings.

Figure 13.10A shows an electron-collecting guard ring implemented in a CMOS process. This structure consists of an NMoat ring placed in the P-epi surrounding the source of injected electrons. The NSD implant is relatively shallow, so it can only intercept a fraction of the carriers. This type of guard ring relies on the P+ substrate beneath the wells to prevent minority carriers from bypassing the guard ring by burrowing through the substrate. Unfortunately, the presence of an electric field across the P+/P– interface tends to repel electrons from the substrate and to channel them laterally toward adjacent wells. This phenomenon makes it difficult to construct truly effective barriers against minority carrier injection into the substrate. Connecting the NMoat ring to a power supply rather than to ground helps only marginally, since the

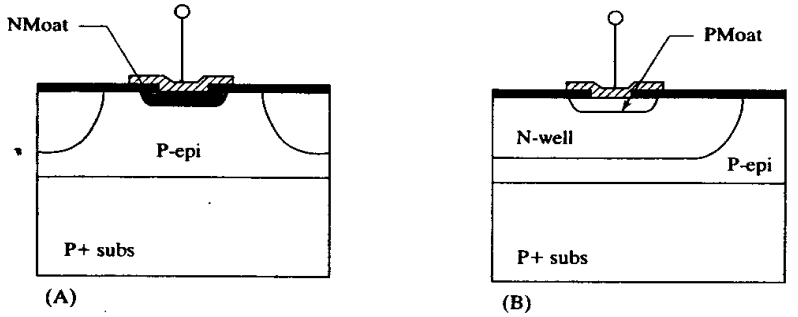


FIGURE 13.10 Minority-carrier guard rings for N-well CMOS: (A) electron-collecting guard ring and (B) hole-collecting guard ring.

added depth of the depletion region does not penetrate more than a small fraction of the epitaxial layer. In low-voltage CMOS processes where the NMoat resides in a P-well, the guard ring should connect to substrate potential rather than to a power supply. The increased surface doping of the low-voltage P-well reduces the width of the NSD/P-well depletion region and causes the electric field across it to increase. High electric-field intensities can trigger avalanche multiplication within the depletion region. The resulting debiasing actually improves the collection efficiency of the guard ring, but it may cause problems elsewhere on the die.

Figure 13.10B shows a hole-collecting guard ring implemented in a CMOS process. This guard ring consists of a ring of PMoat placed in the N-well around the source of injected holes. This type of guard ring is not very effective because most of the holes flow down to the substrate rather than laterally to the guard ring. Increasing the width of the guard ring does little to improve its effectiveness.

The minority carrier guard rings that can be constructed in a CMOS process usually have very limited effectiveness. The two types of guard rings tend to reinforce one another, so the best design practice consists of using both electron and hole collecting guard rings around every device that might inject minority carriers. Since CMOS devices do not inject any substantial level of minority carriers during normal operation, this requirement is satisfied if every device connecting to an output pin receives a guard ring. The designer should examine each pin that neither connects to a power supply nor to substrate potential. Each source/drain region connecting to such a pin requires a guard ring. PMOS transistors require hole-collecting guard rings even when placed in their own wells. NMOS transistors require electron-collecting guard rings. A combination of guard rings and backgate contacts should suppress most forms of latchup, but they may prove inadequate for handling the severe minority carrier injection problems associated with inductive kickback and resonance.

Analog BiCMOS processes normally include NBL and deep-N+. The presence of these layers allows the construction of deep electron-collecting guard rings similar to the one in Figure 13.8A. These guard rings are especially effective on designs using a P+ substrate because the built-in potential of the P-epi/substrate interface helps confine the electrons within the epi. A deep-N+ guard ring on a thin-epi P+ process may collect 90% or more of the electrons injected into the epi.[5]

Although analog BiCMOS supports the construction of hole-blocking and hole-collecting guard rings, these may not be as effective as their standard bipolar coun-

[5] R. R. Troutman, "Epitaxial Layer Enhancement of n-Well Guard Rings for CMOS Circuits," *IEEE Electron Device Letters*, Vol. EDL-4, #12, 1983, pp. 438–440.

terparts. Analog BiCMOS processes often use a lower NBL doping concentration to minimize lateral autodoping. The lower doping decreases the built-in potential at the NBL/N-well interface and allows more carriers to enter the NBL. The lighter NBL also decreases the Gummel number of the parasitic substrate PNP, so a substantial fraction of the carriers may actually penetrate through the NBL/N-well interface to the substrate underneath. This problem becomes even more acute on low-voltage processes using heavily doped wells, since the increased well doping further degrades the built-in potential at the NBL/N-well interface.

The efficiency of hole guard rings suffers if the NBL cannot efficiently block hole flow to the substrate. The addition of a hole blocking guard ring may actually increase substrate injection through a "leaky" NBL.[6] This seemingly paradoxical behavior probably results from a reduction in the effective volume of the N-well. The hole-blocking guard ring repels holes from the portion of the well it occupies, concentrating them within the remaining volume of the well. The higher concentration of holes near the NBL/N-well interface increases the injection rate of carriers into the substrate. This *reduction in volume effect* should not affect hole-collecting guard rings, but the presence of a "leaky" NBL still reduces their collection efficiency.

Analog BiCMOS designs also exhibit excessive substrate resistance. Even if the design uses a P+ substrate, the presence of a lightly doped P-epi makes it difficult to establish a low-resistance substrate contact. Even relatively low levels of substrate injection can produce substantial substrate debiasing. Substrate debiasing can be prevented by blocking minority carriers before they can reach the substrate through the use of hole guard rings. All high-current saturating NPN transistors should incorporate such guard rings to prevent substrate debiasing and noise coupling.

Analog BiCMOS designs sometimes use a P– substrate to avoid the necessity of growing two epitaxial layers. Designs constructed on a P– substrate are even more susceptible to latchup because electron guard rings no longer benefit from the presence of an electron barrier at the P–/P+ interface. Many designs can still achieve satisfactory levels of immunity to transient-induced latchup providing that every potential source of minority carrier injection is surrounded by a suitable guard ring. Even the most conservatively designed guard rings may prove unable to handle the severe minority carrier injection problems associated with inductive kickback and resonance. Such designs may require the use of a P+ substrate despite the additional cost associated with the second epitaxial deposition.

13.3 SINGLE-LEVEL INTERCONNECTION

Most modern processes offer at least two levels of metallization. Since leads can freely cross one another, the placement of components becomes constrained only by matching and packing. Routing almost never presents a problem as long as the designer leaves a little space between components. Given time, almost any designer can compress the wasted space out of such a layout to produce a reasonably densely packed design.

Interconnection becomes much more difficult if the process offers only one level of metallization. The lack of second metal makes it difficult to cross leads. Although low-value resistors can be inserted to create crossing points, these *tunnels* consume die area and add resistance and capacitance that degrades the performance of the circuit. A properly arranged layout contains surprisingly few tunnels. The compo-

[6] N. Gibson, unpublished report, 1998.

nents in such a layout are arranged to minimize the number of crossing points. Leads often route across resistors or between the terminals of transistors, and sometimes they even tunnel through tanks or base diffusions.

Single-level interconnection requires far greater skill and ingenuity than multi-level interconnection. The designer must be able to anticipate possible blockages between components and be able to mentally shuffle the components to clear the blockages. A move that clears one blockage often creates others. Skilled designers have a sort of "geometric intuition" that aids them in placing components and routing leads. This intuition seems largely an innate talent and not a learned skill. There are, however, a number of specific skills and techniques that can help any designer better cope with single-level-metal designs. Although these skills may not seem to have any application in modern multi-level-metal processes, many designs dedicate the upper metal layers for power routing, electrostatic shielding, or optical shielding. A skilled designer must therefore understand how to route designs with a minimum number of layers of interconnection.

13.3.1. Mock Layouts and Stick Diagrams

The greatest challenge of single-level routing lies in properly arranging the components to minimize the number of tunnels required. The presence of matched components often complicates this task so that even skilled designers have to try several arrangements before finding a suitable one. These trial arrangements usually take the form of rough sketches, or *mock layouts*, similar to that in Figure 13.11. The transistors appear in this sketch as rectangles with emitter, base, and collector marked. The resistors appear as strips with connections to either end. Dummies and resistor tanks do not appear in the sketch. The tank contacts are marked "TC." Merged devices occupying a common tank are shown abutting one another, as in the case of Q_3 and Q_4. Although crude, this sketch illustrates all of the important features of the proposed layout.

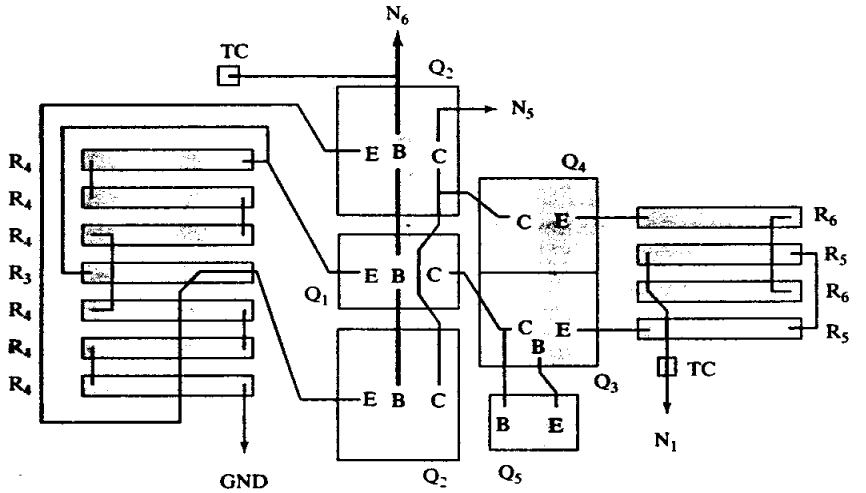


FIGURE 13.11 Mock layout for a portion of the circuit shown in Figure 14.2. The following components must match one another as accurately as possible: R_3–R_4, Q_1–Q_2, Q_3–Q_4, and R_5–R_6.

This particular layout contains a large number of matched devices. In order to obtain the best possible matching, each set of these components has been arranged symmetrically around one axis of the layout as advocated in Section 7.2.6. The axis of symmetry passes horizontally through the middle of the sketch.

Resistors R_3 and R_4 consist of 160Ω/□ base material and have values of 621Ω and 4kΩ, respectively. These resistors are not in simple integer ratio ($R_4/R_3 = 6.441$). This complicates, but does not prohibit, the construction of a sectioned and interdigitated array. Suppose that R_3 is taken as the unit resistor for the array. R_4 then requires a minimum of seven segments. Unfortunately, the centroid of a one-segment resistor cannot perfectly align to the centroid of a seven-segment resistor. In order to achieve a true common-centroid layout, R_4 must consist of eight partial segments. This arrangement is not particularly compact because all of the segments are relatively short (R_3 contains 3.88 squares). A better arrangement consists of six segments of 666.7Ω each for R_4 and a partial segment for R_3. The sliding contact in R_3 also allows the resistor ratio to be tweaked. R_3 occupies the center of the array to ensure that the centroids align (Section 7.2.6). The six segments of R_4 have been interconnected to cancel thermoelectrics (Section 7.2.7). Resistors R_5 and R_6 each consist of 18.75 squares of 160Ω/□ base diffusion. The mock layout shows that each resistor consists of two segments of 9.375 squares that are interdigitated to form a compact array.

Transistors Q_1 and Q_2 form a 6:1 ratioed pair. The layout of such transistors normally involves splitting the larger transistor into two halves placed on either side of the smaller transistor (Figure 9.19A). Transistors Q_3 and Q_4 are matched, minimum-size lateral PNP transistors. These transistors can reside in a common tank because they share the same base connection. They are placed side-by-side to improve matching and to simplify interconnection. P-bars and N-bars are unnecessary because neither transistor saturates in normal operation.

Although this circuit contains several crossing points, none requires a tunnel. The lead interconnecting the emitters of Q_2 routes through the resistor array R_3–R_4. The collector lead could follow the same route, but this requires separating the resistor segments. Instead, transistors Q_1 and Q_2 have been stretched to allow Q_2's collector lead to route between the base and collector of Q_1. This configuration actually requires little or no elongation of the transistor tanks because the presence of deep-N+ in Q_1 and Q_2 already necessitates a large base-to-collector spacing.

The mock layout in Figure 13.11 has not been drawn to scale, but the rectangles representing the components have roughly the same proportions as the components themselves. Sometimes designers carry this type of sketch one step further by using paper plots of the actual components. All of the components are plotted to the same scale, typically either 100:1 or 250:1. The individual components are cut out and shuffled about on a large sheet of paper until a suitable arrangement appears, and then the components are glued down and the interconnections are marked in pencil or ink. Cut-and-paste mock layouts (*paper dolls*) are especially useful for designing tight-packed layouts, since all of the dimensions of the components are to scale.

CMOS designers sometimes use another type of mock layout called a *stick diagram*. Although stick diagrams were originally intended to portray digital logic cells, they can represent analog circuitry. Figure 13.12 shows a schematic and a stick diagram of a CMOS NAND gate. The thick, black horizontal lines represent PMoat and NMoat regions. The NMoat usually lies at the bottom of the diagram and the PMoat at the top. The thick, gray vertical lines represent poly traces. A transistor forms wherever poly crosses either PMoat or NMoat. Contacts are represented by X-marks and metal leads by thin, black lines. Stick diagrams of analog circuits usually show NMoat and PMoat regions in different colors to aid in distinguishing them from one another. The names of nodes and devices may both appear on the stick diagram, and additional notations may be added to identify resistors and capacitors.

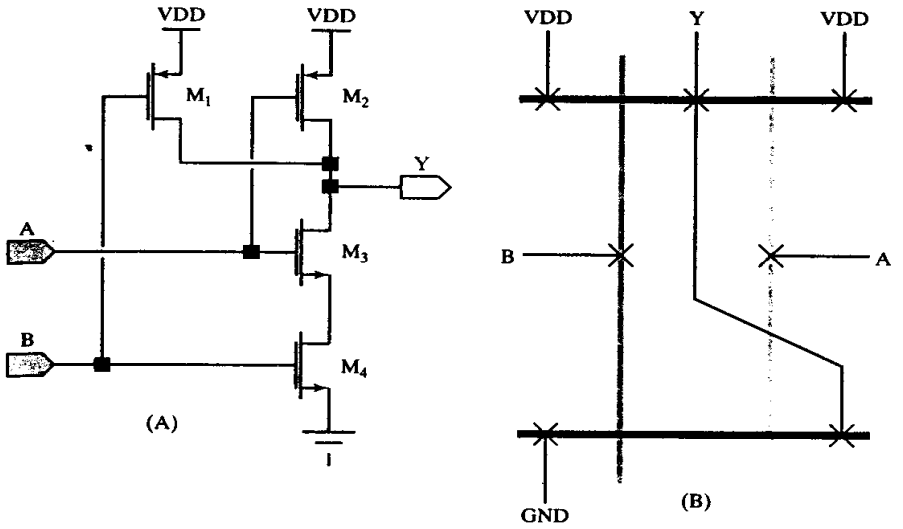


FIGURE 13.12 (A) Schematic and (B) stick diagram of a CMOS NAND gate.

13.3.2. Techniques for Crossing Leads

No matter how hard one tries to avoid them, most single-level metal layouts still require crossing points. The following rules summarize the techniques available for crossing leads using only one level of metal. These techniques were originally developed for standard bipolar designs, but they are also applicable to multiple-level metallization designs where the upper metal layers are required for power routing, electrostatic shielding, or some other purpose.

1. *Cross leads over resistors.*
   A lead routed across a resistor provides a crossing point without consuming additional area, but not every lead can safely cross every resistor. Some resistors require field plating that restricts or even prevents lead crossings. Other resistors are susceptible to noise coupling from overlying leads. Lightly doped materials such as 2kΩ/□ HSR may also experience voltage modulation effects.
2. *Rearrange the terminals of a device.*
   Crossing points can often be eliminated by altering the arrangement of device terminals. For example, the CEB layout for an NPN transistor places the emitter terminal between the collector and base, while the alternate CBE layout places the base terminal between the collector and emitter (Figure 8.14). The CBE layout exhibits slightly more collector resistance than the CEB layout, but the difference rarely affects circuit operation.
3. *Stretch devices to allow leads to pass through them.*
   Most types of devices can stretch to accommodate one or more leads between their terminals. Figure 8.15 shows three examples of stretched NPN transistors, and Figure 5.14 shows an example of a stretched HSR resistor. Stretched devices usually possess more parasitic resistance and capacitance than unstretched devices, so their use sometimes affects circuit operation. If one of a group of matched devices employs a stretched layout, then so should all the others.

4. *Connect signals through merged devices.*
   Certain types of devices lend themselves to use as tunnels. For example, the NPN transistor in Figure 8.15C contains a stretched base region with two contacts. This component actually merges an NPN transistor and a base tunnel into the same tank. A similar type of merger uses multiple tank contacts rather than multiple base contacts. These types of merged tunnels insert parasitic resistances and capacitances that can affect device operation. If large currents flow through the tunnel, the resulting debiasing may also affect circuit operation.
5. *Insert tunnels.*
   Tunnels, or *cross-unders,* are low-value resistors incorporated into the layout to allow leads to cross one another. Various types of tunnels exist, but all share similar disadvantages. They not only consume die area, but they also insert parasitic resistance and capacitance into the tunneled lead. The insertion of a tunnel into a high-current lead can cause excessive voltage drops and power dissipation. Tunnels can also upset matching by introducing voltage drops where they cannot be tolerated. Since the placement of tunnels affects circuit operation, the circuit designer must ultimately approve or reject each potential tunnel. When all of the tunnels have been placed, the circuit designer should add their resistances and capacitances to the circuit and resimulate to see if any critical parameters have shifted.
6. *Rearrange the bondpads.*
   If high-current leads must cross one another to reach their respective bondpads, consider rearranging the bondpads to eliminate the crossing point. Sometimes one bondpad arrangement may lend itself to interconnection much more readily than others.

The layout designer usually requires guidance from the circuit designer to determine what types of stretches and tunnels are allowable. One simple and effective means of communicating this information consists of an annotated schematic prepared by the circuit designer. This diagram only requires a few minutes to prepare, yet it can save many hours of layout effort. Table 13.1 describes a simple annotation scheme using different colors to highlight components and signals. The annotated schematic should also include lists of matched components, guard rings, and other special requirements that might influence the routing of the leads.

TABLE 13.1 A simple schematic annotation scheme.

Category	Precautions	Marking
Power leads	No tunnels allowed; leads must equal or exceed a certain width.	Highlight in **red**; mark width over lead.
Noisy leads	Do not cross sensitive devices.	Highlight in **yellow.**
Sensitive leads	Do not tunnel. Do not place substrate contacts in sensitive grounds leads.	Highlight in **green.**
Sensitive devices	Do not cross noisy leads over sensitive devices.	Highlight in **green.**

13.3.3. Types of Tunnels

Tunnels can be constructed from any diffusion having a relatively low sheet resistance. In standard bipolar, the candidates include base, emitter, deep-N+, and NBL. CMOS and BiCMOS processes often use gate poly jumpers instead of tunnels. The base diffusion usually has a sheet resistance of 100 to 200Ω/□, while the other three materials usually have sheet resistances of about 10Ω/□. Of these, only the base dif-

fusion can occupy a common tank with other components. Most standard bipolar designs contain a number of merged base tunnels and a few stand-alone tunnels of other types.

All tunnels add series resistance. In the case of base tunnels, this resistance usually equals a few hundred Ohms. Although this may not seem like much resistance, it is sufficient to cause some circuits to malfunction. A typical tunnel also exhibits a few hundred femtofarads of parasitic junction capacitance. Some high-speed circuits contain nodes that would be seriously slowed by even this small amount of capacitance. Any attempt to reduce the series resistance of the tunnel by widening also increases its shunt capacitance. Larger tunnels also become more vulnerable to junction leakage and to minority carrier collection.

The sheet resistance of the emitter diffusion is an order of magnitude lower than that of base. The simplest sort of emitter tunnel simply consists of a strip of emitter diffusion placed in a tank (Figure 13.13). The tank-substrate junction provides the necessary isolation between the signal and the underlying substrate. The tank requires no contact other than that provided by the tunnel itself. The addition of NBL does not significantly reduce the tunnel resistance or improve latchup immunity. NBL actually increases the parasitic shunt capacitance, so emitter tunnels generally omit it.

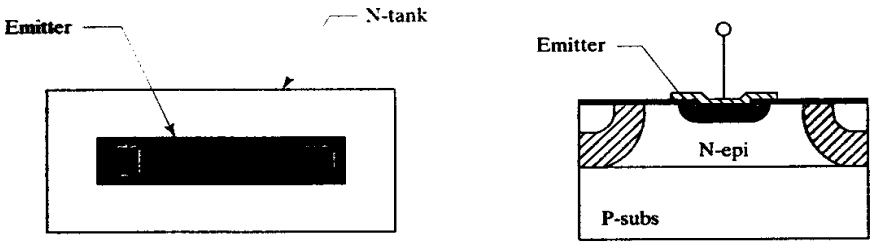


FIGURE 13.13 Layout and cross section of a conventional emitter tunnel.

The emitter-in-iso tunnel in Figure 13.14 saves considerable area by eliminating a tank geometry. The emitter diffusion can counterdope the isolation, but the breakdown of the resulting N+/P+ junction may equal only a few volts. Junctions with breakdown voltages this low tend to leak, but emitter-in-iso tunnels can still be used to route the substrate return line around the die.[7] Processes with emitter-iso breakdown voltages of 6V or higher can often employ emitter-in-iso tunnels to route other signals, but the capacitance of the emitter-iso junction is relatively large (typically

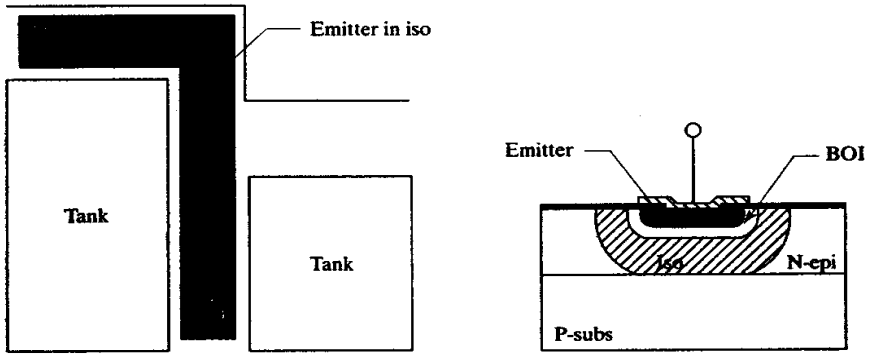


FIGURE 13.14 Layout and cross section of an emitter-in-iso tunnel.

[7] Davis, pp. 16–17.

1pF/mil^2, or 1.6fF/μm^2). Some circuits have taken advantage of this high capacitance to fabricate emitter-iso junction capacitors. These capacitors can occupy unused areas of the isolation, allowing the construction of large junction capacitors with little or no increase in die area.

Some applications require a resistance that is lower than emitter alone can provide. A combination of all available N-type regions (N-epi, NBL, deep-N+, and emitter) will produce a slightly lower sheet resistance (Figure 13.15). A stack of this sort usually has a sheet resistance of about 5Ω/□. NBL provides little benefit without deep-N+, so designs that do not use deep-N+ cannot make effective use of stacked tunnels.

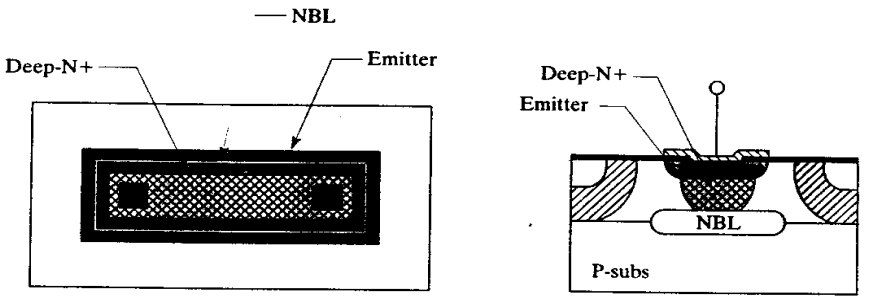


FIGURE 13.15 Layout and cross section of a low-resistance, stacked tunnel containing emitter, deep-N+, and NBL.

Several other types of tunnels are occasionally used. Of these, the deep-N+ tunnel merits at least a brief mention. These tunnels are sometimes used in place of emitter tunnels on processes with thin emitter oxides. An ESD strike could rupture the thin emitter oxide if an emitter tunnel were placed underneath a lead connecting to a bondpad. Deep-N+ tunnels are covered by a thick oxide that is virtually immune to ESD damage. There are few applications for deep-N+ tunnels in modern designs because the newer processes generally use thick emitter oxides, and because modern design practices dictate the use of ESD structures on virtually every pin.

13.4 CONSTRUCTING THE PADRING

The *padring* of an integrated circuit consists of scribe streets, pads, ESD structures, and guard rings. Each of these plays a vital role in determining the success or failure of the design. Many circuits have failed because of inadequate ESD protection, misplaced bondpads, or missing guard rings. The following sections provide guidance that should help the layout designer avoid most of these mistakes.

13.4.1. Scribe Streets and Alignment Markers

Scribe streets must surround the die to provide room for the passage of the sawblade used to separate the dice. The saw consumes a strip of silicon about 25μm (1mil) wide, but the scribe streets must be three or four times wider to provide room for misalignment. Oxide and nitride tend to crack during sawing, and metal clogs the sawblade; therefore most scribe streets consist of bare silicon. The edges of the die abutting the scribe street are often fitted with special structures called *scribe seals* to prevent contaminants from seeping underneath the exposed edges of the protective overcoat (Section 4.2.2). Additional structures may reside within the scribe street itself. Some fabs place *alignment markers* within the streets. These markers are used to align photomasks to previous steps of the process and are de-

stroyed during sawing. Sometimes arrays of test devices are also placed in the scribe streets. These devices can be used to evaluate the performance of the wafer before it is sawn and assembled. The test devices also provide a means of characterizing large numbers of devices for statistical device modeling. The test devices can be destroyed during sawing because they already will have been tested and so will have served their purpose.

Most wafer fabs specify the scribe streets required for their process. Sometimes the scribe street occupies a separate database prepared by the fab and sent to the mask shop independently of the main design. Alternatively, the scribe street structures may be provided to the layout designer to place around the edges of the main layout. Regardless of whether they appear in the layout, the scribe seals abut the die on all four sides. Since these seals usually incorporate substrate contacts, they can form a useful addition to the substrate contact system. A thin strip of metal placed around the edges of the pad ring will make contact to the scribe seal metallization. In addition to providing substrate contacts, the scribe seal metallization also provides a convenient method of routing the substrate potential around the periphery of the die. The width of the metallization in the scribe seal adds to the width of metal in the padring, producing a relatively wide lead that can conduct significant current without debiasing or electromigration failure (Figure 13.16). Because of its convenient placement around the periphery of the die, the substrate metallization often forms the return path for the ESD structures (Section 13.4.3).

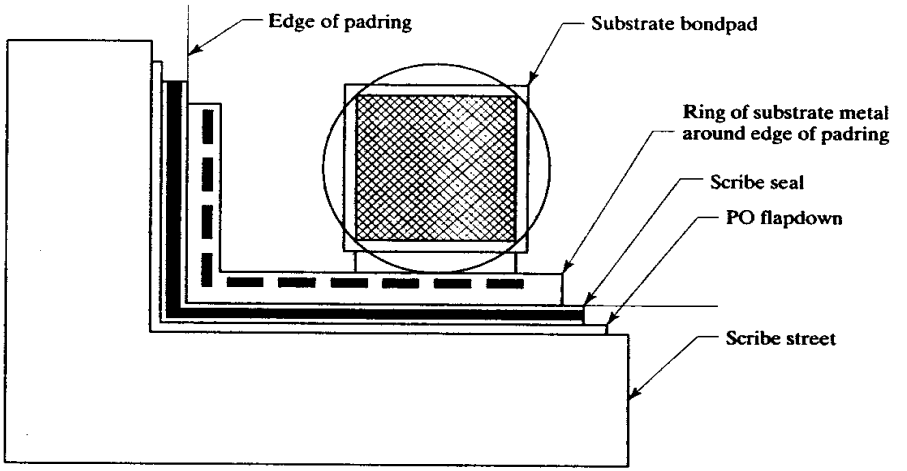
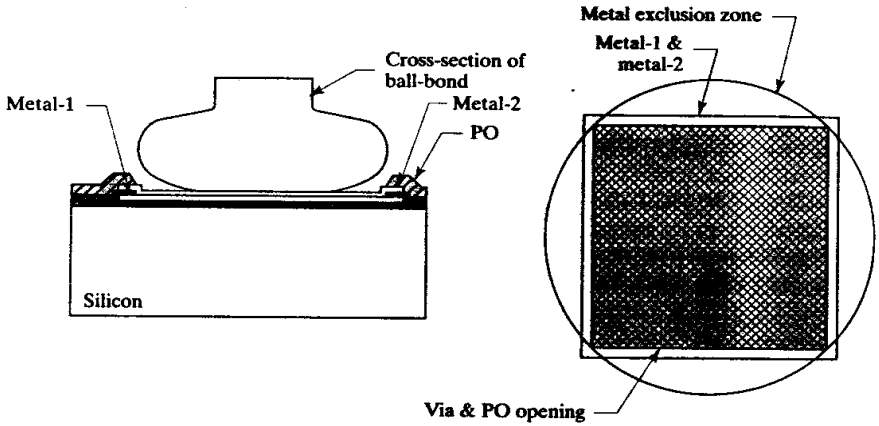


FIGURE 13.16 Diagram illustrating the relationship between the scribe street and the padring.

It is not possible to provide generalized guidance for constructing scribe seals, because every wafer fab has different requirements. The layout designer must seek the fab's guidance to determine the proper scribeline selection and placement. Some fabs may impose additional requirements, including limitations on the size and aspect ratio of the die. For example, some older steppers only accept dice whose dimensions are an integer number of mils. The efficient use of photolithographic equipment may also dictate that the die dimensions fall within a certain range of values. These issues should be resolved during die floorplanning so the designer can maximize the profitability of the design. These issues become much more difficult to resolve as the layout nears completion, so do not wait until the last moment to obtain the necessary information!

13.4.2. Bondpads, Trimpads, and Testpads

Most integrated circuits connect to the external world through bondwires. These wires consist of either gold or aluminum, and they range in diameter from 0.8 to 10.0 mils (20 to 250μm). The most common form of bonding uses gold wires approximately 1 mil (25μm) in diameter attached to the die by means of ball bonds. The ball bonding process uses a hydrogen flame to create a tiny gold ball on the end of the bondwire. A capillary tube presses this ball down against the exposed aluminum metal with enough force to cause the two metals to alloy together (Section 2.7.1). The bonding process deforms the soft gold ball into a pancake-like structure (Figure 13.17). The actual area of gold-aluminum alloying is usually about the same diameter as the original bondwire, but the metal pad on which the ballbond rests must be two to three times larger to account for the misalignments that inevitably occur during automated bonding. Each ballbond thus requires an exposed metal pad several mils across. These specialized structures are called *bondpads*.



IGURE 13.17 Cross section and ayout of a double-level-metal ondpad intended for allbonding.

The simplest conceivable bondpad consists of a square of metal placed underneath a matching opening in the protective overcoat (PO). The metal must overlap the opening sufficiently to seal the die against the ingress of mobile ions (Section 4.2.2), even if overetching and misalignment occur. If the process offers more than one level of metal, then the bondpad typically includes plates of each metal placed coincident to one another. The interlevel oxides should be removed from the area of the bondpad opening so the ballbond lands on the stacked metal layers. Figure 13.17 shows a typical double-level-metal bondpad constructed according to these principles. The bondpad consists of a square of metal-2 placed over an identical square of metal-1. These metal plates overlap a PO opening and a via, both of which also coincide to one another.

Dice are usually bonded using the smallest possible diameter of wire, since this enables the use of the smallest bondpads. Most assembly sites offer either a 1.0mil (25μm) or an 0.8mil (20μm) gold wire as a minimum wire diameter. A gold bondwire packaged in plastic can conduct approximately one amp of continuous current per mil of diameter (Section 14.3.3). Thus an 0.8mil gold wire can conduct a continuous current of about 800mA. Higher currents require either larger-diameter wires or multiple bondwires connected in parallel. Every pin of a typical package can accommodate two (or possibly three) bondwires. Some surface-mount packages are so small that they can only accommodate one bondwire per pin, and these same packages are usually so thin that they can only use the finest wire diameters. The designer

should verify that the package can handle the required number and diameter of bondwires during the earliest stages of floorplanning.

Small-diameter bondwires have relatively large resistances. The resistance of a bondwire can be estimated using the equation

$$R_w \cong \frac{\omega L}{D^2} \qquad [13.2]$$

where R_w is the resistance of the bondwire in Ohms, L is its length in mils, and D is its diameter in mils. The constant of proportionality ω equals approximately 1.1mΩ·mil for gold and 1.4mΩ·mil for aluminum.[8] Bondwires are typically about 30mils long, so a typical 1mil gold bondwire exhibits a resistance of 30mΩ. Equation 13.1 does not consider the resistance of the leadframe, nor the resistance of the bond contact, each of which may add a few additional milliohms of resistance. Gold bondwires are usually limited to a maximum of 2mils in diameter, but much larger aluminum bondwires are available. Although it is technically possible to bond a die using different types or diameters of wire, this requires a corresponding number of passes through the bonding equipment. The additional time and expense are rarely justifiable unless the design requires the use of extremely large-diameter wire.

Historically, gold ballbonds have required square bondpad openings about three times the diameter of the wire. The increasing precision of modern bonding equipment has enabled many assembly sites to accept smaller bondpads. The layout designer should obtain the current guidelines from the assembly site for the specific type and diameter of wire that will be used to bond the die. Aluminum wire must be wedge bonded rather than ballbonded, and this usually requires an elongated bondpad placed at a specific angle relative to the fingers of the leadframe. The rules for aluminum wirebonding can become quite complex, and the designer should seek guidance from the assembly site before attempting to lay out a design employing it.

The locations of the bondpads must simultaneously satisfy several conflicting requirements. The bondpads cannot lie too close to one another, or the capillary tube will damage one bond while placing the next. The bondwires must not pass too close to adjacent pads lest the capillary damage the wires while placing subsequent bonds. Long bondwires can short to one another or to adjacent bondpads due to a phenomenon called *wiresweep*. The injection-molding process forces molten plastic over the die, and the viscous drag of the plastic on the bondwires causes them to move. Larger-diameter wires are more rigid and can better resist wiresweep, allowing them to span larger distances. Wiresweep also makes it inadvisable to cross one bondwire over another. The best bonding arrangements consist of a ring of pads placed around the periphery of the die in locations that allow the shortest and most direct wirebonds. Parts packaged in ceramic or metal can ignore the limitations imposed on plastic packages due to wiresweep, but they must still follow the spacing rules required to prevent capillary damage.

It is often difficult to find a suitable bonding arrangement for a die having a large number of bonds. Some assembly sites provide software tools that can evaluate a proposed bonding arrangement to ensure manufacturability. Others require that any potential bonding arrangement pass through a review process in order to obtain production approval. The layout designer should always check to make sure that the bondpad arrangement meets the approval of the assembly site before beginning the top-level layout. If this is not done and the finished layout does not meet

[8] These values are only approximations based on the bulk resistances of gold and aluminum. The actual resistance of a bondwire is affected by impurities and work hardening.

the assembly site's requirements, substantial time and effort will be required to correct the problems.

The placement of bondpads also restricts the routing of adjacent metal leads. Misalignment or excessive bonding force may cause the bond to press against the protective overcoat adjacent to the bondpad opening. The resulting stress can crack the protective overcoat and can even damage underlying leads. Many assembly sites state that metal leads that do not connect to the bondpad must not pass within a certain distance of it. For ballbonds, this requirement usually takes the form of a circular exclusion zone centered on the bondpad. Many designers mark this exclusion zone with a circle placed on a special drawing layer. The layout in Figure 13.17 shows an example of one of these so-called *bondpad circles*. The assembly site will usually specify the dimensions of the exclusion zones. If no guidelines are available, assume that the bondpad circle passes through the four vertices of a square bondpad opening of minimum dimensions. Wedge bonding also requires exclusion zones, but the dimensions of these zones depend on the placement of the pads in relation to the leadframe.

There was a time when many assembly sites recommended placing tanks (or wells) underneath all bondpads and probepads. These tanks were generally left unconnected. They were intended to protect the die against shorts caused by the probe needles scratching through the bondpad metallization and field oxide during wafer-level testing. If such shorts occurred, they would connect the bondpad to a tank rather than to the substrate. This would—theoretically—prevent the device from failing. The placement of unconnected tanks under bondpads is actually a very questionable practice. If the bondpad shorts to the tank, then the tank may inject electrons into the substrate. This means that the tank requires an electron-collecting guard ring, which wastes considerable space. Most modern designs do not place tanks or wells under pads unless they form part of some adjacent device whose tank or well connects to the pad.

Some assembly sites also require that the bondpad for pin #1 be visually distinct from all of the others. This requirement originally arose from the limitations of early machine vision systems used in automated bonding equipment. Even though most modern machines no longer require a distinct pad #1, it still provides a convenient visual reference point for operators who must inspect the mounted dice. A variety of different techniques have been used to mark pad #1, the simplest of which notches the four corners of the protective overcoat (PO) opening (Figure 13.18A). These notches are usually about 0.5mil (~10μm) deep. If possible, the metal pattern should also contain notches corresponding to those of the PO opening on at least two corners of the bondpad. Another technique marks pad #1 with an octagonal PO opening (Figure 13.18B), while a third employs a circular opening (Figure 13.18C). The requirements of the assembly site may dictate a choice among these options. Otherwise, the designer should probably follow the conventions established by previous designs.

Since trimpads and testpads only allow access for probe needles, they need not follow all of the requirements that apply to bondpads. The size of trimpads and test-

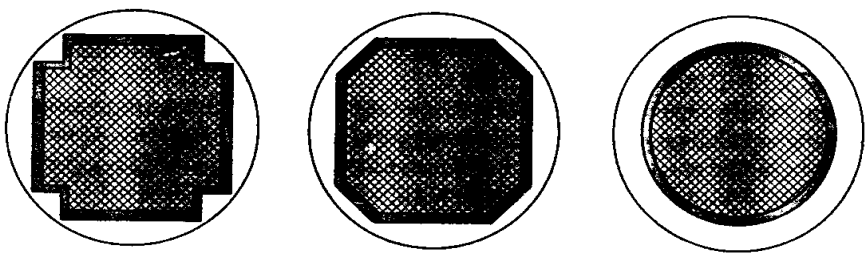


FIGURE 13.18 Three unique styles of bondpads sometimes used to identify pin #1.

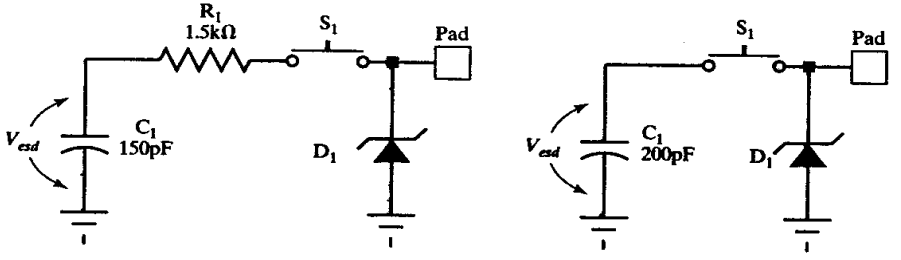
pads depends on the diameter of the probe needles and on the alignment tolerances of the probing equipment. These requirements usually prove somewhat less restrictive than those associated with bonding, so trimpads and testpads are usually smaller and more closely spaced than bondpads. The relatively high currents associated with trimming sometimes necessitate the use of larger-diameter needles, so trimpads may require larger PO openings than do testpads. Both types of pads are customarily placed around the periphery of the die to simplify the design of the probe card. Additional testpads are sometimes placed in the interior of the die, but these are usually removed before the design reaches production to minimize the ingress of contaminants.

Probing does not generate the same level of mechanical stresses as bonding, so many assembly sites allow the placement of trimpads and testpads over active circuitry. This concession virtually eliminates the area penalty associated with the use of small numbers of fuses or Zener zaps. Trimpads placed over active area consist of a square of top-level metal and an associated PO opening. They require neither vias nor lower-level metal, so they can reside over any portion of the circuit free of top-level metal. Some processes use a very thick layer of top-level metallization that provides enough mechanical compliance to dissipate the mechanical stresses that are induced during bonding. These processes may also allow the placement of bondpads over active area. This practice can result in substantially more compact designs, but only if the metallization system can support the resulting stresses.

13.4.3. ESD Structures

In addition to bondpads, the padring also contains structures intended to safely dissipate the energy of ESD events. These *ESD structures* must reside near their respective bondpads to minimize lead resistances and inductances that might otherwise interfere with their operation. The ESD structures usually require a low-impedance return path to the substrate terminal. The scribe seal metallization can provide this return path without consuming much additional die area. In order to take full advantage of the scribe seal, the ESD structures must either reside between the bondpads and the scribe seal, or between adjacent bondpads.

ESD structures are intended to limit the peak voltages seen at the bondpads in order to protect the remainder of the integrated circuit. The stresses imposed on ESD structures differ depending on the testing conditions chosen. The two most common tests are the *human-body model* (HBM) and the *machine model* (MM). The human body model charges a 150pF capacitor, C_1, to a known voltage, V_{esd}. The capacitor then discharges into ESD structure, D_1, through a 1.5kΩ series resistor, R_1 (Figure 13.19A).[9]



(A) Human-body model

(B) Machine model

FIGURE 13.19 Equivalent circuits for (A) the human-body model and (B) the machine model.

[9] T. M. Madzy and L. A. Price II, "Module Electrostatic Discharge Simulator," *EOS/ESD Symposium Proc. EOS-1*, 1979, pp. 36–40. MIL-STD-883 is identical to Madzy's model except that C_1 = 100pF. The 150pF value is now widely accepted.

The resistor absorbs most of the energy, but the ESD device must still conduct relatively large peak currents without failure. For example, a 2kV HBM strike produces a peak current of about 1.3A. Higher ESD voltages generate proportionately larger peak currents. Modern integrated circuits must generally pass 2kV HBM, while certain pins on selected devices must stand as much as 20kV HBM.

The machine model employs a 200pF capacitor C_1 charged to a specified voltage V_{esd}. This capacitor discharges directly through ESD device D_1 (Figure 13.19B). The peak currents flowing through the ESD device are primarily limited by the parasitic inductance of the external circuit, which is usually less than 500nH. The ESD structure must not only withstand the extremely high peak currents that result, but it must also dissipate the energy of the ESD strike without permanent damage. An ESD device can typically withstand only about one-tenth as much voltage in the machine model test as it can withstand in the human-body model test. Thus modern integrated circuits are often specified to withstand 2kV HBM and 200V MM.

The newest ESD test is called the *charged-device model* (CDM). This test involves charging the integrated circuit's package to a specified voltage relative to a ground plate, and then discharging one pin through a low-impedance probe to the plate.[10] The charged-device model generates even higher peak currents than the machine model, but the transients contain less energy because the capacitance of an integrated circuit package is much smaller than the 200pF capacitance specified by the machine model. The charged-device model is gradually replacing the machine model throughout the industry. A typical testing requirement is 2kV HBM and 1kV CDM.

The effects of ESD vary depending on the type of component involved.[11] PN junctions are usually destroyed by overheating. Since considerable energy is required to melt even a few cubic microns of silicon, diffused junctions are generally quite robust. An avalanching junction dissipates most of its heat within its depletion region. Since lightly doped junctions have wider depletion regions, they can dissipate more energy than can heavily doped junctions. A lightly doped junction also has additional series resistance that dissipates some of the ESD energy. Large junctions are more robust than small ones because they contain a correspondingly greater volume of silicon within their depletion regions. The collector-base and collector-substrate junctions of standard bipolar processes are so large and so lightly doped that they can withstand most ESD transients without damage. Base-emitter junctions are more vulnerable because of their smaller dimensions and heavier doping. The base-emitter junctions of NPN transistors are also susceptible to avalanche-induced beta degradation (Section 8.1.2). The shallow, heavily doped junctions used in CMOS processes are more easily damaged than the deep, lightly doped junctions of standard bipolar. CMOS transistors with silicided source/drain regions (*clad moats*) are particularly fragile because of the lack of ballasting and the presence of silicide immediately adjacent to the depletion region.

Thin insulating films, such as those used in MOS transistors and deposited capacitors, are extremely fragile. High voltages will rupture these films within nanoseconds. Even if the insulating film does not rupture, it may suffer degradation that causes it to fail during normal operation due to time-dependent dielectric breakdown (TDDB). These delayed failures are very difficult to detect during high-speed automated testing. The only proven way to prevent such failures is to limit the voltage across the di

[10] Y. Fukuda, S. Ishiguro, and M. Takahara, "ESD Protection Network Evaluation by HBM and CDM (Charged Package Method)," *EOS/ESD Symposium Proc. EOS-8*, 1986, pp. 193–199.

[11] A. Amerasekera, W. van den Abeelen, L. van Roozendaal, M. Hannemann, and P. Schofield, "ESD Failure Modes: Characteristics, Mechanisms and Process Influences," *IEEE Trans. on Electron Devices*, Vol. 39, 1992, pp. 430–436.

electric to safe values. Deposited resistors can also suffer dielectric breakdown through the thick-field oxide or the interlevel oxide (ILO), but hundreds of volts are required to rupture these layers. Any diffusion connected to a bondpad will avalanche long before the field oxide or the ILO ruptures, and will therefore protect them.

Early bipolar integrated circuits rarely incorporated any intentional ESD protection, but they generally withstood the rigors of ordinary handling because their junctions were sufficiently robust to absorb and dissipate low-level ESD strikes without damage. A few parts undoubtedly suffered damage during handling, but most of these failures were attributed to processing defects or infant mortality. CMOS integrated circuits proved much more fragile. Large numbers of early devices were destroyed through gate oxide rupture during normal handling. Once the mechanisms responsible for these failures were identified, designers began to recognize that even bipolar circuits were potentially vulnerable. A variety of protective structures were proposed, some of which worked and some of which did not. The reasons for success and failure were poorly understood, so ESD protection gained a reputation of being a "black art." This reputation is largely undeserved because ESD devices obey the same principles that govern other components. The following sections examine several types of ESD devices often used to protect analog integrated circuits. The strengths and weaknesses of each device are explained in terms of their structures and electrical properties. Using this information, the layout designer can construct ESD structures for a variety of applications.

Zener Clamp

The simplest ESD device consists of a Zener diode connected between the bondpad and the substrate return line (Figure 13.20A). Possible choices include the emitter-base Zener of standard bipolar and the NSD/P-epi and PSD/N-well Zeners of analog CMOS. An ideal Zener diode would impose a positive clamp voltage equal to its reverse breakdown voltage and a negative clamp voltage equal to its forward drop. Most Zener diodes contain enough internal series resistance to make the clamp voltages much larger than these ideal values. A minimum-size emitter-base Zener has from 100 to 300Ω of internal series resistance, and NSD/P-epi and PSD/N-well diodes have even more. These resistances actually increase the robustness of the Zener by spreading the ESD energy over a larger volume of silicon, but in so doing they cause the bondpad voltage to rise above the theoretical clamp voltage by some tens of volts. This consideration severely limits the usefulness of Zener diodes as ESD structures.

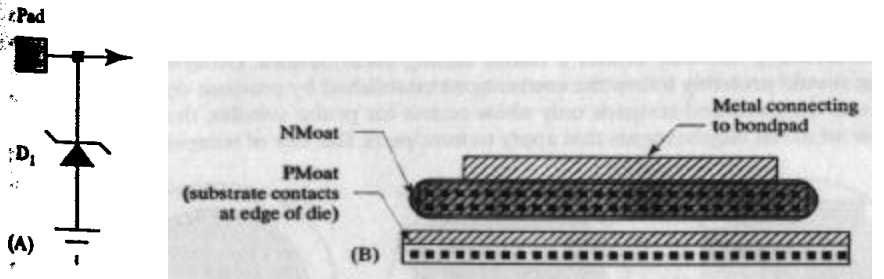


FIGURE 13.20 (A) Schematic diagram and (B) layout of the simple Zener clamp ESD circuit.

The source/drain regions of NMOS and PMOS transistors can sometimes protect themselves against ESD damage. Consider the case of a large NMOS transistor whose drain connects to a pin. Negative ESD transients forward-bias the NSD/P-epi junction. Most of the resulting voltage drop occurs within the P-epi. Positive ESD transients avalanche the NSD/P-epi junction, dumping energy into its depletion

region. Since the depletion region contains less silicon than the P-epi, NMOS transistors are more vulnerable to positive transients than to negative ones. A similar analysis shows that PMOS transistors are most susceptible to negative transients. If a pin connects to both PMOS and NMOS transistors, then both conduct some portion of the ESD pulse and the more fragile device will determine the circuit's survival or failure. Larger transistors are more robust than small ones because they can dissipate energy within a larger volume of silicon. A large number of small transistors usually provides the same degree of protection as a single large transistor. A 10V nonsilicided, single-diffused drain MOS device with a drawn drain area of about $1000\mu m^2$ will usually withstand 2kV HBM and 200V MM. Since PMOS and NMOS transistors avalanche under different biasing conditions, a large NMOS will not necessarily protect a small PMOS, or *vice versa*. If both PMOS and NMOS transistors connect to a pin, then the total PMOS drain area and the total NMOS drain area must both suffice to independently withstand the ESD strike.

Low-voltage CMOS processes are much more vulnerable to ESD damage than previous-generation, higher-voltage processes, due in part to the extreme shallowness of the low-voltage source/drain diffusions and in part to the higher backgate doping required to prevent punchthough. Both of these factors reduce the volume of silicon in the depletion regions. Clad-moat devices sometimes exhibit localized breakdown due to their lack of source/drain ballasting. Once localized breakdown occurs, ESD performance no longer improves with increasing device area. The ESD performance of clad-moat devices can be improved by removing silicide from the perimeter of the source/drain implants in order to provide a small amount of ballasting. Robustness can be further improved by increasing the overlap of the (unsilicided) source/drain diffusions over their respective contacts on any diffusions that might avalanche due to ESD.[12] Most researchers attribute the increased robustness to ballasting, but some studies suggest that minority carrier injection from the source/drain contacts may also play a role.[13]

If the area of the source/drain regions connected to a pad is insufficient to provide adequate ESD protection, then a dedicated ESD protection device can be connected to the pad. In an N-well CMOS process, an NSD/P-epi diode (often called a *thick-oxide device* in the literature) usually offers the best protection for a given die area. Figure 13.20B shows a typical layout for such a diode.[14] The elongated NMoat region is placed alongside a strip of substrate contacts forming part of the scribe seal. The close proximity of the substrate contacts minimizes the series resistance of the device. The aspect ratio of this device allows it to be placed between a bondpad and the scribe seal metallization, or between adjacent bondpads. The corners of the NMoat diffusion are filleted to prevent premature avalanche breakdown. The radius of these fillets should equal or exceed the junction depth of the diffusion. The overlap of the NMoat region over its contacts should exceed the minimum layout dimension by 1 to $2\mu m$ to provide additional ballasting. In clad-moat processes, the silicide block mask should block silicide for a distance of at least 1 to $2\mu m$ from the drawn junction. The diode should contain at least $500\mu m^2$ of NMoat, and it should be surrounded by an electron-collecting guard ring and by as many substrate con-

---

[12] T. L. Polgreen and A. Chatterjee, "Improving the ESD Failure Threshold of Silicided n-MOS Output Transistors by Ensuring Uniform Current Flow," *IEEE Trans. on Electron Devices*, Vol. 39, #2, 1992, pp. 379–388.

[13] T. J. Maloney, "Contact Injection: A Major Cause of ESD Failure in Integrated Circuits," *EOS/ESD Symposium Proc. EOS-8*, 1986, pp. 166–172.

[14] A somewhat similar diode appears in R. J. Antinone, P. A. Young, D. D. Wilson, W. E. Echols, M. G. Rossi, W. J. Orvis, G. H. Khanaka, and J. H. Yee, *Electrical Overstress Protection for Electronic Devices* (Park Ridge, NJ: Noyes Publications, 1986), p. 19.

tacts as area permits. This device provides a reasonable degree of protection for NMOS and PMOS source/drain regions that are not large enough to protect themselves. In some cases, a low-value series resistor may be required to ensure that the ESD current flows through the Zener clamp rather than through the protected device.

Two-stage Zener Clamps

Even a large protection Zener has an internal series resistance well in excess of 10Ω. A 2kV HBM strike produces peak currents of about 1.3A, which in turn produce voltage drops of tens of volts across the series resistance of the Zener. These voltages are sufficient to rupture a thin gate oxide. Although the Zener cannot protect the gate dielectric by itself, it can reduce the peak voltage of the ESD transient from thousands of volts to tens of volts. A second protection structure connected in series with the first can provide enough additional clamping to protect the thin gate oxide. The schematic diagram in Figure 13.21A shows the conceptual arrangement of the resulting *two-stage ESD clamp.* Zener diode D_1 clamps the pad voltage to a maximum of perhaps 100V. A second Zener, D_2, connects to the pad through a series limiting resistor, R_1. The presence of R_1 limits the current flow through D_2, enabling this second Zener to limit the voltage across the gate oxide to safe levels. In order for the circuit to function properly, the resistance of R_1 should equal at least several times the series resistance of D_2. A relatively small Zener diode may exhibit several hundred Ohms of internal series resistance, so R_1 typically equals several kilohms. The inclusion of this resistance limits the slew rate of the gate voltage, but this is often desirable since excessively large transient currents can damage gate dielectrics. R_1 also adds time delays of a few nanoseconds that may interfere with certain high-speed applications.

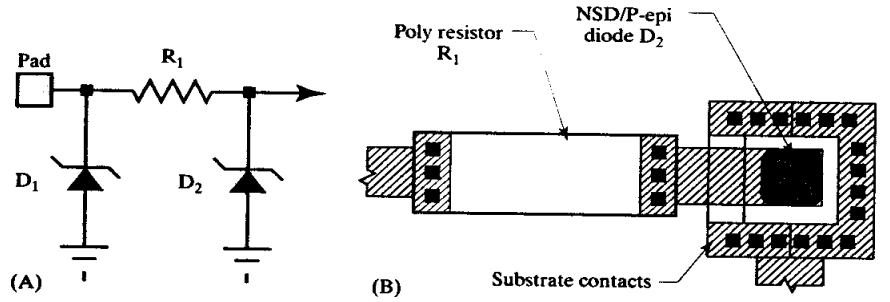


FIGURE 13.21 (A) Schematic diagram and (B) partial layout of a two-stage Zener clamp. The layout of primary protection diode D_1 is identical to that in Figure 13.20B.

Figure 13.21B shows one possible layout for the series limiting resistor R_1 and the secondary protection Zener D_2. Resistor R_1 consists of a wide strip of lightly doped polysilicon with multiple small contacts on either end. The relatively large physical dimensions of this resistor help to ensure that it can successfully dissipate the energy dumped into it during an ESD transient. Diffused resistors are more robust than poly resistors because they dissipate a portion of their energy through a distributed avalanche mechanism, so many authors suggest using them instead of poly resistors.[15] Poly resistors will survive 2kV HBM and 200V MM providing their resistance equals at least several hundred Ohms, they are at least 5 to 8μm wide, and each end of the resistor includes at least 6 to 8 minimum contacts (fewer contacts are required for higher-value resistors). Regardless of the type of resistor chosen, it should not include

[15] A. R. Pelella and H. Domingos, "A Design Methodology for ESD Protection Networks," *EOS/ESD Symposium Proc. EOS-7,* 1985, pp. 24–40.

any bends because current focuses near the inner corner of the bend, producing a hot spot that will fail before the remainder of the resistor. Zener diode D_2 consists of a relatively small plug of NMoat placed inside a concentric ring of substrate contacts. These substrate contacts not only minimize the series resistance of the Zener but also help minimize substrate debiasing near the secondary protection device. If these substrate contacts were omitted, or if they were located further away from the secondary Zener, then the large transient currents flowing through the primary Zener D_1 could induce tens of volts of substrate debiasing near D_2. This debiasing would add to D_2's clamp voltage, potentially resulting in the destruction of the gate oxide that the structure is intended to protect. If possible, the secondary Zener should be placed 50 to 100μm away from the primary one. One common arrangement places D_1 between the bondpad and the scribe R_1 alongside the bondpad, and D_2 on the inner side of the bondpad. Both D_1 and D_2 should be enclosed within an electron-collecting guard ring. The charged device model generates such extreme currents that the secondary protection device is often placed next to the source of the transistor that is to be protected to prevent substrate debiasing from generating destructive voltage drops.[16]

Two-stage ESD structures similar to that in Figure 13.21 have successfully protected MOS gates on many moderate-voltage CMOS processes. Since this type of ESD structure contains too much series resistance to allow its use on anything other than a high-impedance input terminal, it is often called an *input ESD device*. A similar two-stage ESD circuit can be constructed for some types of low-impedance applications, such as for the protection of the outputs of relatively small CMOS logic gates. This alternative type of structure uses a primary Zener diode D_1 identical to that employed in the input ESD circuit. The series resistance R_1 is reduced to 50 to 500Ω, and the source/drain diffusions of the output MOS transistors serve as the secondary Zener diode, D_2. Although the source/drain junctions avalanche, the presence of the series resistance limits the current to safe levels. Larger output transistors can employ proportionally smaller series resistors. This type of ESD structure is sometimes called an *output ESD device*. These devices have successfully protected even minimum-area source/drain implants.

Sometimes a circuit includes both source/drain diffusions and gate electrodes connected to the same bondpad. A combination of input and output ESD circuitry will successfully protect this circuit. A single primary protection device connects from the bondpad to the substrate return line. Two separate limiting resistors are required: a low-value one for the source/drain implants and a much higher-value one for the gate electrodes. The source/drain implants serve as their own secondary protection, but the gate electrodes require the addition of a secondary Zener.

Buffered Zener Clamp

The availability of bipolar transistors allows the construction of very robust ESD circuits. Figure 13.22A shows a ***buffered Zener clamp*** that uses an NPN transistor to reduce the effective series resistance of a Zener diode. Emitter-base Zener D_1 provides base drive to a much larger NPN transistor Q_1. This transistor multiplies the current through the Zener by its own effective beta. The positive clamp voltage of this structure equals the sum of an emitter-base breakdown and a diode drop. Assuming a standard bipolar V_{EBO} of 6.8V, the positive clamp voltage lies near 8V. The collector-substrate junction of Q_1 clamps negative ESD transients to one diode drop (plus substrate debiasing).

---

[16] L. R. Avery, "ESD Protection Structures to Survive the Charged Device Model (CDM)," *EOS/ESD Symposium Proc. EOS-9*, 1987, pp. 186–191.

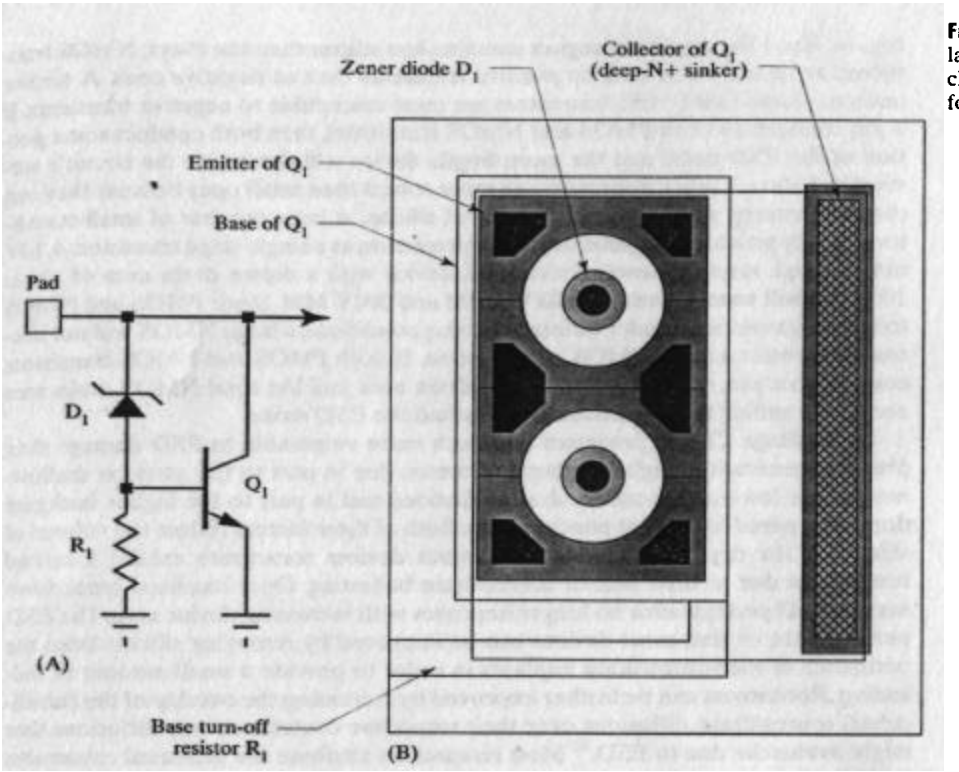


FIGURE 13.22 Schematic and layout of a buffered Zener clamp (metallization omitted for clarity).[17]

The positive clamp voltage of the buffered Zener remains roughly constant as long as the voltage across the collector resistance of Q_1 does not exceed about 7V. Even relatively small NPN transistors have collector resistances of less than 10Ω, so it is quite possible for a buffered Zener clamp to protect a gate dielectric without requiring a secondary protection device. If necessary, the clamp voltage can be increased by the inclusion of a second Zener diode, or by the addition of one or more diode-connected transistors in series with the Zener. The maximum clamp voltage this structure can safely support equals the $V_{CEO(sus)}$ of the NPN transistor. If one attempts to obtain a higher clamp voltage, the NPN transistor will avalanche and snap back to $V_{CEO(sus)}$. This type of snapback characteristic forms the basis of the V_{CES} clamp discussed in the next section.

The buffered Zener clamp dissipates most of its energy in its large base-collector depletion region. An NPN transistor with an emitter area of 300 to 600μm^2 will usually provide 2kV HBM and 200V MM protection. Larger NPN transistors can provide protection against proportionately higher ESD voltages. The ultimate limits of this structure are probably determined more by the metallization and the bondwires than by the ability of the silicon to absorb ESD energy.

All of the components of the buffered Zener clamp can occupy a common tank. In the structure in Figure 13.22B. the emitter of the power transistor Q_1 consists of a series of annular emitter geometries. Inside each hole are smaller plugs of emitter diffusion that form the cathodes of Zener diode D_1. Both the emitters of Q_1 and the

[17] M. Corsi, R. Nimmo, and F. Fattori, "ESD protection of BiCMOS Integrated Circuits which need to operate in the Harsh Environment of Automotive or Industrial" *(sic)*, *EOS/ESD Symposium Proc. EOS-15*, 1993, pp. 209–213.

cathodes of D_1 are enclosed in a common base region, a portion of which extends out into the isolation to form base resistor R_1. All of these merged devices reside within a common tank having a single shared deep-N+ sinker. The buffered Zener clamp operates as follows: when the cathodes of Zener diode D_1 avalanche into the shared base region, they inject holes into it. These holes forward-bias the base-emitter junction of Q_1 and cause the large NPN transistor to conduct. The ESD transient only lasts for a few hundred nanoseconds, which is not enough time for a hot spot to form and collapse. Since thermal runaway cannot occur, the shape of the emitter diffusion can be tailored to improve performance in other ways. The annular shape of the emitters of Q_1 ensures rapid and even turn-on of all portions of Q_1's emitter. Base resistor R_1 holds Q_1 off during normal operation; its value is not particularly critical to the operation of the whole device.

Although Figure 13.22B illustrates a buffered Zener clamp for a standard bipolar process, these structures are actually better suited to the protection of analog BiCMOS circuitry, because the smaller spacings of the latter process reduce the size of the structure to the point where it can reside in the pad ring. This structure has successfully protected the gate oxide of a 20V analog BiCMOS process against 2kV HBM and 200V MM ESD strikes. The extremely low series resistance of the buffered Zener often eliminates the need for a secondary breakdown ESD device for gate oxides with rupture voltages of 20V or more. Lower-voltage gate oxides usually require a secondary protection structure similar to that in Figure 13.21A. The buffered Zener clamp can be adapted to higher-voltage applications by inserting additional Zener diodes or diode-connected transistors in series with the anode of Zener D_1. These additional devices can also be merged into a common tank with the other portions of the ESD circuit.

V_{CES} *Clamp*

Figure 13.23A shows an ESD circuit that uses the collector-base breakdown of an NPN transistor to clamp positive ESD transients. The initial breakdown voltage of this circuit equals the V_{CES} rating of transistor Q_1. Once conduction has begun, it does not cease until the voltage across the transistor drops below $V_{CEO(sus)}$. These two thresholds are sometimes called the *trigger voltage* (or *strike voltage*) and the *sustain voltage*. A typical 40V standard bipolar transistor has a nominal trigger voltage of about 65V and a nominal sustain voltage of about 45V. The snapback from the higher trigger voltage to the lower sustain voltage decreases the voltage drop across the NPN and helps reduce the energy dissipated in the transistor. Despite the rela-

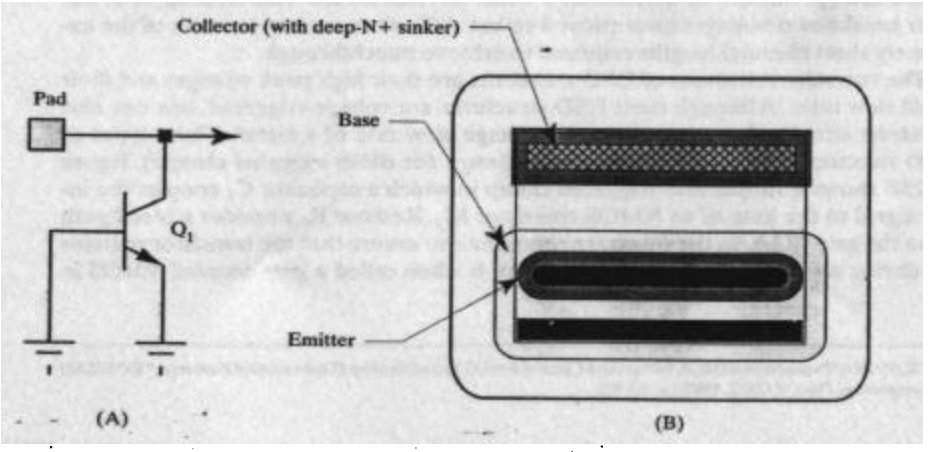


FIGURE 13.23 (A) Schematic diagram of the V_{CES} clamp and (B) a layout of a suitable NPN transistor.

tively high breakdown voltage of this structure, it is easily capable of withstanding 2kV HBM and 200V MM ESD strikes. An emitter area of about 300 to 500μm^2 provides this level of protection in standard bipolar. Larger emitter areas provide proportionately higher levels of ESD protection. This structure has successfully served as the primary protection device for a 20V analog BiCMOS gate oxide against 2kV HBM and 200V MM ESD events.[18]

ESD devices with snapback characteristics cannot safely protect low-impedance pins operating at or beyond their sustain voltage. If a transient triggers snapback, and the external circuit can supply enough current to sustain conduction, then the ESD device will continue to conduct indefinitely. The resulting power dissipation quickly overheats and destroys the integrated circuit. If the external circuitry cannot provide enough current to sustain conduction, then the ESD device can protect a pin even if it operates at a voltage in excess of the device's sustain rating. This type of application should not be contemplated unless the designer has full characterization data for the ESD device and can confidently state that the application will never deliver enough current to sustain conduction.

Figure 13.23B shows a typical layout of an NPN transistor used as a V_{CES} clamp. The filleted corners on the base diffusion raise the strike voltage of the device slightly, and also help make conduction slightly more uniform. Many designers also fillet other diffusions as shown in the illustration, but these fillets make no substantial difference in the operation of the transistor as a V_{CES} clamp. The transistor can be made more rugged by increasing the overlap of the base and emitter diffusions over their respective contacts by 1 to 2μm. This precaution is particularly useful if the process does not include silicide or refractory barrier metal, since pure aluminum contacts are more vulnerable to alloying failures than other types of contacts.

V_{CES} *Clamp*

If the emitter and collector terminals of a bipolar transistor are swapped, the device will continue to operate as a bipolar transistor. When such a transistor is biased into conduction, it is said to operate in the *reverse active,* or *inverse active,* mode. The collector-base junction forward-biases and injects minority carriers into the base, which are collected by the emitter-base junction. An NPN transistor operated in reverse active mode has a very low beta because the substitution of the lightly doped collector for the heavily doped emitter drastically reduces emitter-injection efficiency. The heavily doped base-emitter junction also avalanches at a much lower voltage than the lightly doped collector-base junction. Because of this reduction in breakdown voltage, a transistor operated in reverse active mode makes an excellent low-voltage ESD device. Suppose the transistor in Figure 13.23B is used as an ESD clamp with its emitter connected to a bondpad and its base and collector connected to ground (Figure 13.24A). The trigger voltage of the V_{ECS} *clamp* equals the V_{EBO} of the NPN transistor, and its sustain voltage equals approximately 60 to 80% of this voltage. Since analog BiCMOS NPN transistors typically have a V_{EBO} of 8 to 10V, these devices can serve as primary protection devices for pins that do not operate at more than 5V. This type of ESD device has a very low series resistance due to the presence of the NBL and the deep-N+ sinker. Devices with emitter areas of less than 600μm^2 have successfully withstood 2kV HBM and 200V MM ESD strikes, and somewhat larger emitter areas have protected circuits to 10kV HBM. These devices present a low-impedance path to negative ESD transients as well as positive

[18] J. Z. Chen, X. Y. Zhang, A. Amerasekera, and T. Vrotsos, "Design and Layout of a High ESD Performance NPN Structure for Submicron BiCMOS/Bipolar Circuits," *International Electron Devices Meeting Proc.*, 1995, pp. 337–342.

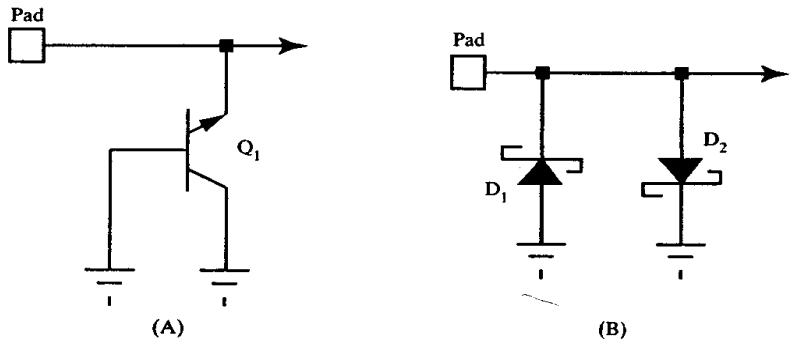


FIGURE 13.24 Schematic diagrams of (A) the V_{ECS} clamp and (B) the antiparallel-diode clamp.

ESD transients, a benefit that few other ESD devices offer. The V_{ECS} clamp does not require electron-collecting guard rings because its tank does not connect to a pin. This structure does contain a parasitic substrate PNP transistor that can inject large majority-carrier currents into the substrate, so the clamp should be surrounded by as many substrate contacts as possible to minimize debiasing in the surrounding substrate.

The V_{ECS} structure is an extremely useful device for protecting low-voltage pins. Higher-voltage ESD circuits can be created by stacking V_{ECS} clamps in series, but this arrangement increases both the area required for the device and its series resistance. A buffered Zener or a V_{CES} clamp will probably provide better performance than stacked V_{ECS} clamps.

Antiparallel Diode Clamps

Many integrated circuits require multiple ground pins. Sometimes all of the ground pins connect to the substrate, but some are frequently separated to minimize noise coupling and substrate injection. Those ground pins that do not connect to substrate require some form of ESD protection. The most common configuration consists of a pair of back-to-back (or *antiparallel*) diodes. One can employ diode-connected NPN transistors, diode-connected substrate PNP transistors, or even Schottky diodes (Figure 13.24B). Since the voltage drop across these devices is relatively small, they experience much less internal heating than do other types of ESD devices, and therefore they can be made somewhat smaller. Schottky diodes with areas of several thousand square microns will generally provide protection against 2kV HBM and 200V MM ESD strikes, and diode-connected transistors can be made even smaller than this. All of these structures require enclosure within electron-collecting guard rings.

Additional ESD Structures for CMOS Processes

The first ESD structures for CMOS processes relied on Zener clamps, but these were found to have objectionably large series resistance. A variety of other structures have been proposed. A lateral NPN transistor can be constructed by placing two NMoat regions next to one another. The NSD diffusions act as the collector and the emitter of the transistor, while the P-epi separating them acts as its base. If one of the two NMoat regions connects to the bondpad and the other to the substrate return line, then the bipolar transistor will operate as a V_{CES} clamp. Historically, most of the ESD devices constructed along these lines also included a gate electrode on top of the thick-field oxide separating the two NSD diffusions. Some designers connected this gate electrode to the bondpad, while others connected it to the substrate return. Since the thick-field threshold usually exceeds the breakdown of the NSD/P-epi junction, the gate electrode has little effect regardless of its connection. Most de-

signers included it anyway and treated the resulting device as an NMOS transistor (Figure 13.25A). Since the gate dielectric consisted of thick-field oxide, these devices were popularly called *thick-field transistors*.

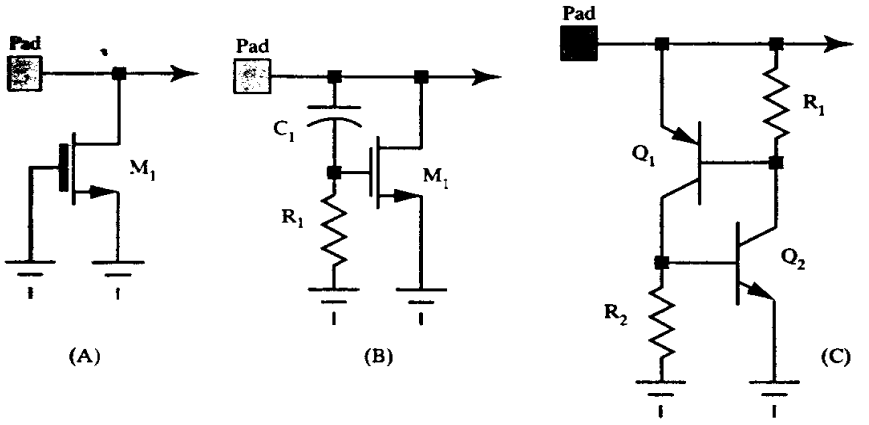


FIGURE 13.25 Three types of CMOS ESD: (A) the thick-field device, (B) the rate-triggered NMOS clamp, and (C) an SCR clamp.

The thick-field transistor has proved something of a disappointment. Its shallow, heavily doped junctions are prone to overheating during ESD transients, and incremental junction damage often causes excessive leakage, even if the device does not catastrophically fail. Ironically, the large series resistance of the NSD/P-epi Zener improves its robustness by spreading energy dissipation throughout a larger volume of silicon. The lower series resistance of the thick-field device actually renders it more fragile than an NSD/P-epi Zener of similar size. The snapback characteristic of the thick-field device also causes problems. The sustain voltage of this structure is usually about 60% of the NSD/P-epi breakdown voltage, a value that is often less than the maximum operating voltage of the NMOS transistor. When thick-field devices are used to protect low-impedance pins, the operating voltage of the pin must never exceed the sustain voltage of the thick-field device. Before using thick-field transistors, sample devices must be constructed and characterized to ensure that they can withstand ESD strikes, and to determine the actual sustain voltage of the structure. These devices should not be used unless such data is available, because they are often relatively marginal devices. A number of variants on the thick-field transistor have been developed, including thin-oxide transistors that break down by punchthrough instead of avalanche.[19] These devices do not exhibit snapback, but their breakdown voltages have proved rather difficult to control because of the extremely short channel lengths required to achieve punchthrough.

The two salient features of ESD transients are their high peak voltages and their rapid slew rates. Although most ESD structures are voltage-triggered, one can also construct circuits that respond to the voltage slew rate of a signal. These types of ESD structures are called *rate-triggered clamps* (or *dV/dt-triggered clamps*). Figure 13.25B shows a simple rate-triggered clamp in which a capacitor C_1 couples the input signal to the gate of an NMOS transistor M_1. Resistor R_1 provides a bleed path from the gate of M_1 to the substrate return line to ensure that the transistor remains off during normal operation. This structure is often called a *gate-coupled NMOS* in

[19] J. K. Keller, "Protection of MOS Integrated Circuits From Destruction by Electrostatic Discharge," *EOS/ESD Symposium Proc. EOS-2*, 1980, pp. 73–80.

the literature.[20] An ESD strike will couple sufficient energy through C_1 to turn M_1 on, allowing it to absorb the remainder of the ESD energy. In order for M_1 to properly clamp the ESD transient, it must have an $R_{ds(on)}$ of no more than a few Ohms. This usually results in a rather large device, and a variety of other rate-triggered devices have been proposed as smaller alternatives.

Rate-triggered devices have become more popular as the operating voltages of CMOS processes have decreased below 5V. The avalanche voltage of the source/drain junctions cannot drop below about 5V without causing excessive junction leakage due to tunneling. The gate oxides of low-voltage CMOS processes are therefore difficult, if not impossible, to protect using avalanche-triggered structures. Rate-triggered devices are susceptible to false triggering caused by rapidly slewing signals. Ironically, most digital inputs and outputs qualify as rapidly slewing signals. Rate-triggered devices should not be used unless sufficient characterization data are available to determine whether or not they will be triggered by the slew rates encountered in normal operation.

Many low-voltage processes use some variant of the *silicon-controlled rectifier* (SCR) for ESD protection.[21,22] An SCR is best described as a pair of back-to-back bipolar transistors, connected as shown in Figure 13.25C. In CMOS processes, these two bipolar transistors are both parasitic lateral devices. Substrate PNP Q_1 consists of a PSD diffusion inside an N-well placed in the P-epi. Lateral NPN Q_2 consists of the N-well, the P-epi, and an adjacent NSD diffusion. R_1 represents the resistance of the N-well, and R_2 represents the resistance of the P-epi and the substrate. This circuit is triggered into conduction by the collector-base avalanche of either Q_1 or Q_2. Suppose Q_2 avalanches first. Carriers injected into the base of Q_2 cause it to conduct. Q_2 now pulls current from the base of Q_1, causing it to turn on and provide additional base drive for Q_2. Each of the two transistors now provides base drive to the other, and conduction continues until the input voltage drops so low that R_1 and R_2 can extract more current than the transistors can supply. If R_1 and R_2 are both large, then the SCR may have a sustain voltage of less than 2V. Smaller values of resistance will produce higher sustain thresholds, but the relationship between resistance and sustain voltage is difficult to predict. In practice, a multitude of SCR structures are constructed and measured to determine which geometry gives the desired strike and sustain voltages.

The SCR clamp is surprisingly robust. Under HBM conditions, its low sustain voltage forces the external 1.5kΩ resistor to dissipate almost all of the ESD energy. SCRs can also dissipate remarkable amounts of energy during machine-model testing, perhaps due in part to the wide depletion region at the N-well/P-epi junction. A properly constructed SCR clamp can often withstand several times as much ESD energy as other CMOS ESD structures.

The trigger voltage of a simple SCR clamp is often too high to protect the gate dielectric of a low-voltage CMOS process. Rate-triggered SCR clamps have been developed that include capacitors connected from the pad to the base of Q_2, or from the base of Q_1 to ground. Rapidly slewing transients cause these capacitors to trigger the SCR. Rate-triggered SCR circuits can provide excellent protection, but the

---

[20] C. Duvvury and C. Diaz, "Dynamic Gate Coupling of NMOS for Efficient Output ESD Protection," *International Reliability Physics Symposium Proc.*, 1992, pp. 141–150.

[21] L. R. Avery, "Using SCR's as Transient Protection Structures in Integrated Circuits," *EOS/ESD Symposium Proc. EOS-5*, 1983, pp. 177–180.

[22] J. Z. Chen, A. Amerasekera, and T. Vrotsos, "Bipolar SCR ESD Protection Circuit for High-Speed Submicron Bipolar/BiCMOS Circuits," *IEDM*, 1995, pp. 337–340.

only way to ensure that they will operate properly is to construct and characterize them before using them in a design.

13.4.4. Selecting ESD Structures

Pins connected directly to the substrate, or connected only to relatively robust diffusions, can usually survive without the addition of dedicated ESD structures. Most other pins require some form of ESD protection. The following guidelines offer some specific advice for several commonly encountered situations:

1. *Pins connecting to base or emitter diffusions.*
   The relatively low sheet resistances of base and emitter diffusions render them vulnerable to ESD damage. Larger diffusions may spread the energy over sufficient area to protect themselves, but localized heating often damages smaller diffusions. The minimum diffusion area capable of self-protection depends on process parameters and testing conditions, but it is probably safe to say that a $500\mu m^2$ $160\Omega/\square$ base diffusion will survive 2kV HBM and 200V MM. Smaller diffusions should include some form of ESD clamp that avalanches before the base diffusion, such as a V_{CES} clamp or a V_{ECS} clamp. Series limiting resistors are rarely necessary because either the diffusion or the region enclosing it is usually quite resistive.
2. *Pins connecting to the emitters of NPN transistors.*
   The emitters of vertical NPN transistors are vulnerable to avalanche-induced beta degradation. If possible, the circuit should be designed to eliminate any direct connection between an emitter and a bondpad other than substrate ground. Otherwise, an ESD clamp device must be connected to the bondpad and a series resistance of several hundred Ohms placed between the bondpad and the emitter. The circuit designer must consider the impact of this resistance on circuit operation. The emitters of power NPN transistors sometimes operate at substrate potential, but return through a separate pin. In this case, an antiparallel diode clamp will provide adequate protection without requiring the insertion of any series resistance.
3. *Pins connecting to CMOS gates.*
   CMOS gate dielectrics are so fragile that they usually require some form of two-stage ESD protection. The primary protection device need only limit the voltage at the pad to a few hundred volts. The secondary ESD protection device should clamp the gate voltage to no more than 75% of the oxide rupture voltage. If the secondary ESD device returns through the substrate, then its clamp voltage must include any substrate debiasing generated either by itself or by the primary device. The series limiting resistor between the primary and secondary devices should have a resistance several times larger than that of the secondary protection structure. The resistor may consist either of a diffusion or of polysilicon, but poly resistors should be at least 5 to 8μm wide and should contain at least six or eight contacts at either end to help prevent excessive localized heating. Resistors used in ESD devices should not contain any bends, as these generate localized hot spots that may fail before the remainder of the resistor. Zeners used as secondary protection devices may require series limiting resistors of several kilohms. The secondary protection device and limiting resistor can be omitted if the primary protection device can clamp the voltage at the pad to approximately 75% of the gate oxide rupture voltage. The high currents generated by machine-model testing make this very difficult to achieve, but V_{ECS} clamps have successfully protected a 20V gate oxide against 2kV HBM and 200V MM. CDM testing will almost certainly

require secondary protection, and these devices frequently have to reside near the device to be protected in order to prevent substrate debiasing from developing excessive voltage drops. Some low-voltage CMOS processes have oxide rupture voltages below the trigger voltages of conventional avalanche-triggered ESD structures, in which case rate-triggered devices or SCRs must be used.

4. *Pins connecting to moat regions.*
   Some types of moat regions will protect themselves against ESD, while others will not. Silicided moats almost always require some form of additional ESD protection, as do moats with breakdown voltages of less than 5 to 8V. Nonsilicided moats of transistors with breakdown voltages of 10V or more will probably protect themselves against 2kV HBM and 200V MM transients, providing the total drawn area of each type of moat diffusion exceeds $500\mu m^2$. A large NSD diffusion will not necessarily protect a small PSD diffusion, or *vice versa.* The exact moat areas required to provide self-protection vary depending on processing parameters and testing conditions. Small moat regions, particularly clad ones, generally require some form of additional ESD protection. A single-stage ESD circuit will suffice if this structure can clamp the voltage at the bondpad to less than the avalanche voltage of the moat diffusions. V_{ECS} clamps and buffered Zener clamps can sometimes provide this level of protection, but Zener clamps usually have too much internal series resistance. A series limiting resistance of a few hundred Ohms enables the use of a Zener clamp as a protection device for small moat regions. Large silicided moats often exhibit localized breakdown due to lack of ballasting. Consider using a silicide block mask to remove the silicide from the periphery of moat regions connected to bondpads. The unsilicided moat periphery slightly increases the resistance of the transistor, but one can compensate by increasing the size of the device.
5. *Pins connecting both to moat regions and to CMOS gates.*
   The moats may serve as a primary protection device if they are sufficiently large; otherwise a primary protection device must be connected to the bondpad. Small moats, or ones made especially vulnerable by silicidation, may require a series limiting resistor of 50 to 200Ω. Unless the primary protection device has a very low series resistance, it cannot protect the gates without the addition of a secondary protection device. A resistor of several hundred Ohms to several kilohms should be connected between the pad and the gates, and a suitable secondary protection structure should be placed after this resistor. This structure now has separate conduction paths for gates (which require large series resistances) and moats (which do not).
6. *Pins connecting only to polysilicon.*
   The voltages generated during human-body model testing are sufficient to rupture the thick-field oxide and interlevel oxide surrounding polysilicon resistors and leads. If a bondpad does not directly connect to any diffusion, then the voltages across the oxide surrounding the polysilicon may rise to destructive levels. An N-well geometry placed beneath the bondpad and connected to it by means of a ring of NMoat contacts encircling the bondpad will provide adequate protection while consuming very little die area.
7. *Pins connecting to capacitors.*
   Thin oxide or nitride dielectrics require the same type of protection as gate dielectrics. Junction capacitors usually contain a thin, heavily doped diffusion that requires protection similar to an emitter region.

8. *Pins connecting to Schottky diodes.*
Field-plated Schottky diodes should not operate in avalanche breakdown because their depletion regions are very thin and are located immediately adjacent to a silicide layer. Large Schottky diodes can be protected by adding a field-relief guard ring that avalanches before the Schottky contact. Smaller Schottky diodes may be protected by a large area of moat diffusion, forming part of another device connected to the same pin. If no suitable moat region exists, then a field-relief guard ring and a series resistance of a few hundred Ohms should provide adequate protection.
9. *Bondpads operating at substrate potential but not connected to substrate.*
These bondpads are usually ground returns isolated from substrate to minimize noise coupling. ESD protection is not required if these pads are bonded to the same pin as the substrate pad. Otherwise, an antiparallel diode clamp connected between the pad and the substrate return will provide sufficient protection for most applications.
10. *Multiple bondpads connecting to the same pin through multiple bondwires.*
Many dice use multiple bondwires attached to a common pin. If two or more bondpads connect to the same pin through separate bondwires, then only one of these pads requires a primary ESD device. Series limiting resistors and secondary protection devices must be placed on every bondpad requiring them, as secondary protection placed on one bondpad cannot protect circuitry connected to another bondpad.
11. *Test pads and probe pads.*
Test pads and probe pads normally do not require ESD protection because they are encapsulated within the package and therefore do not experience ESD transients.

When placing ESD structures, always remember to include any necessary guard rings and substrate contacts. Of the types of ESD structures just discussed, only the V_{ECS} clamp does not require guard rings. The guard rings should be placed in the padring during its construction. They require so much room that they are often very difficult to add later.

13.5 EXERCISES

Refer to Appendix C for layout rules and process specifications.

13.1. Lay out three minimum-size standard bipolar NPN transistors without deep-N+ sinkers. Place the transistors side-by-side as shown in Figure 13.1A, and measure the area of a rectangle enclosing all three devices. Now lay out a merged device similar to that in Figure 13.1B. Assume the area consumed by the merged device equals the area of its tank. What is the ratio of the area of the merged device to the area of the three separate devices?

13.2. Describe the risks posed by each of the following mergers:
a. Two HSR resistors placed in the same tank, one of which connects to a bondpad.
b. A base resistor merged in the same tank as a Schottky diode.
c. A lateral PNP transistor merged with an NPN transistor.
d. A junction capacitor merged with a Darlington NPN transistor pair.
e. Two substrate PNP transistors merged in the same tank.

13.3. Propose measures that will minimize the risks associated with each of the mergers in Exercise 13.2.

13.4. Lay out a merged Darlington NPN transistor capable of conducting 100mA. Use standard bipolar layout rules and assume a maximum emitter current density of $8\mu A/\mu m^2$. Use a wide-emitter, narrow-contact structure with an emitter overlap of contact of $6\mu m$ for the power device, and connect a $5k\Omega$ base turnoff resistor

between its emitter and collector. Size the predrive transistor, assuming that the power transistor has a minimum beta of 20 at 100mA. Assume that the predrive transistor does not require a minimum-width base turnoff resistor. Include all necessary metallization.

13.5. Assume that the Darlington transistor in Exercise 13.4 must operate at voltages exceeding the thick-field threshold of the process. Modify the layout to include all necessary field plates and channel stops.

13.6. Lay out the totem pole driver circuit in Figure 13.26 using standard bipolar layout rules. Use wide-emitter, narrow-contact transistors with an emitter overlap of contact of 6μm for Q_1, Q_2, and D_1. The power supply voltage VCC can exceed the thick-field threshold of the process, and both OUT and PGND may go below substrate potential during switching transients. The leads connecting to VCC, OUT, and PGND must have widths of at least 15μm. Include all necessary metallization and label all leads and devices.

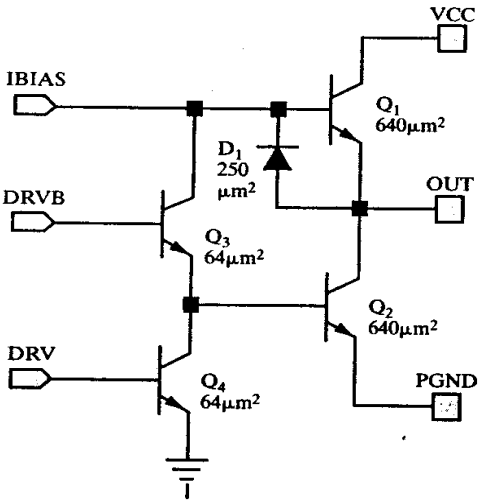


FIGURE 13.26 Schematic of totem pole output stage. All dimensions are emitter areas.

13.7. Modify the MOS operational amplifier in Exercise 12.16 to provide latchup protection, assuming that only the output pin (OUT) connects directly to a bondpad. Make the circuit as robust as possible without using either deep-N+ or NBL.

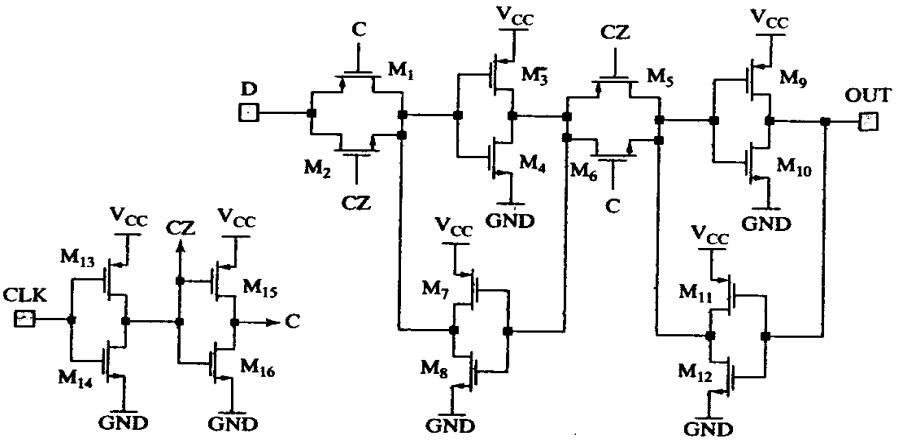


FIGURE 13.27 Schematic of D-type flip-flop.

13.8. (A) Draw a stick diagram of the flip-flop of Figure 13.27. Use a single VDD bus across the top of the cell and a single GND bus across the bottom of the cell. All NMOS transistors are 4/3 and all PMOS transistors are 7/3. Do not use any metal-2 within the cell. Poly can be used to route gate leads, and short poly jumpers can be placed in source and drain leads if necessary. (B) Following the stick diagram as closely as possible, lay out the flip-flop using the CMOS rules in Appendix C. Label all leads and devices.

13.9. Construct a padring for an analog CMOS layout. The die must be exactly 110mils wide by 86mils high, including scribe. The lower left-hand corner of the die must reside at the origin (0, 0). Denote the extents of the die using a rectangle on a layer called BOUNDARY. The scribe streets lie along the left side and bottom of the die, and each is exactly 110μm wide. Denote the locations for the scribe by drawing two rectangles on the BOUNDARY layer. Draw substrate metal coincident with the edge of the padring, as shown in Figure 13.16. This metal should include a 12μm strip of metal-1 and a 12μm strip of metal-2 that exactly coincide with one another. Place PMOAT underneath the substrate metallization. Provide contacts between the substrate metallization and the PMOAT, and vias between the two layers of substrate metallization.

13.10. Construct bondpads for the padring in Exercise 13.9. Assume that the pads require a square nitride opening 75μm across, and that they must contain both metal-1 and metal-2. The two metal geometries should exactly coincide. Place a single, large via 75μm across coincident to the nitride opening. Denote the metal exclusion zone using a circle with a diameter of 90μm centered on the bondpad opening. Place eight of these bondpads in the padring of Exercise 13.9 in the following locations: top left, top center (2), top right, bottom left, bottom center (2), and bottom right. The pad at the bottom right connects to pin #1. Choose a suitable way of denoting this pad, and modify the layout accordingly. Number the pads in counterclockwise order and label each.

13.11. Construct a Zener clamp similar to that in Figure 13.20 using the CMOS layout rules in Appendix C. The NMoat region should have a total area of at least 650μm^2, and should contain two rows of contacts. Overlap NMoat over contact by 3μm and fillet both ends of the diode. Place two of these Zener clamps in the layout in Exercise 13.10 to protect pins 3 and 5. These clamps should reside between the respective pads and the substrate metallization. Include all necessary guard rings.

13.12. Construct a two-stage Zener clamp using the diode of Exercise 13.11 as a primary protection device. The series resistor should equal 1kΩ and have a width of 8μm. Place at least eight contacts on either end of the resistor. The secondary protection structure requires an NMoat area of 100μm^2. Overlap the NMoat over the contacts by at least 3μm and encircle the device with a ring of substrate contacts. All metallization within the ESD structure should have a width of at least 6μm, and if vias are used in any of these leads, use a minimum of eight vias. Place two of these clamps in the layout of Exercise 13.11 so that they protect pins 1 and 7. Include all necessary guard rings.

13.13. Construct a V_{ECS} clamp using the analog BiCMOS layout rules in Appendix C. Assume that the clamp requires an emitter area of 350μm^2. Instead of a single contact for the emitter, use two rows of minimum contacts. Overlap the emitter over the contact by 4μm and fillet both ends of the emitter. The deep-N+ sinker should completely encircle the clamp to form a hole-blocking guard ring. Compare the area of this structure to the area consumed by all of the components of the two-stage Zener clamp in Exercise 13.12 (excluding guard ring).

13.14. Lay out a thick-field transistor for use as an ESD structure in a CMOS design. The source and drain consist of strips of NMoat, each 20μm long and just wide enough to contain two rows of minimum contacts. Overlap NMoat over the contacts by 2μm and include fillets on both source and drain. Separate the two moat regions by 6μm and place a metal-1 plate over the region between the source and drain to act as a "gate." Connect this gate to the source. Compare the area consumed by this clamp to the area consumed by the Zener clamp in Exercise 13.11.