9 Applications of Bipolar Transistors

The bipolar transistor possesses two key advantages over its MOS counterpart: higher transconductance and superior device matching. These advantages translate into faster circuits that consume less power and offer higher precision. Many high-performance operational amplifiers and comparators use bipolar circuitry to minimize input offset and maximize output drive. Some families of high-speed logic also employ bipolar transistors as output drivers. Voltage regulators and references almost always use bipolar circuitry to obtain precise temperature-invariant voltages. Almost all of the highest-speed and highest-accuracy integrated circuits rely on bipolar circuitry in one form or another.

The *transconductance* of a bipolar transistor equals the ratio of the change in collector current to the change in base-emitter voltage. A high transconductance produces a large change in collector current for a small change in base-emitter voltage. The transconductance of a bipolar transistor is directly proportional to emitter current and is independent of emitter area, so even a small bipolar transistor can provide a large transconductance if it receives enough current. MOS circuitry dominates low-power design because MOS transistors retain moderate transconductances at very low currents. As the current levels increase, bipolar transistors become increasingly attractive. A micropower amplifier probably uses all-CMOS circuitry to conserve power, but a high-drive amplifier probably incorporates a bipolar output stage to reduce output impedance and minimize standby currents. The bipolar transistors in this output stage must handle high currents while dissipating large amounts of power. Small-signal transistors, even ones with enlarged emitters, perform poorly in power applications, so a variety of specialized layouts have been developed for this purpose.

The high transconductance of bipolar transistors also improves their base-emitter voltage matching. An untrimmed differential input stage constructed using bipolar transistors can routinely achieve three-sigma input offset voltages of less than ±1mV over temperature. Only exceptionally well constructed (and very large)

MOS input stages can hope to rival this performance.[1] Ratioed bipolar transistors can also generate very accurate voltage differentials, which form the basis of most voltage and current references. MOS references, even carefully constructed ones, rarely perform as well as their bipolar counterparts.

Although bipolar transistors offer distinct advantages over MOS, many designers remain reluctant to use them. Bipolar transistors can fall prey to a number of failure mechanisms that rarely affect MOS designs. The problem of saturation in bipolar transistors has no direct equivalent in MOS design. Improperly constructed bipolar transistors frequently self-destruct under heavy loads, while MOS transistors rarely do so. Carelessly matched bipolar transistors are far more vulnerable to thermal gradients than similar MOS devices. This chapter explains how to retain the unique advantages of bipolar transistors while avoiding their many pitfalls.

9.1 POWER BIPOLAR TRANSISTORS

The previous chapter discussed the layout of small-signal transistors. These devices usually employ minimum-area emitters to conserve space. These small emitters are acceptable because small-signal transistors rarely conduct more than a fraction of a milliamp. Transistors conducting larger currents experience beta rolloff unless their emitter areas increase in proportion to their emitter currents in order to maintain a constant emitter current density. The beta of a typical vertical NPN transistor begins to roll off at current densities of about 1mA/mil^2 (1.5 μA/μm^2) of emitter. To conserve space, power transistors generally operate at lower betas than their small-signal counterparts. A beta of ten is often chosen as a minimum acceptable limit for high-current operation. Power NPN transistors can usually handle 5 to 10mA/mil^2 (8 to 15μA/μm^2) before their beta drops below ten. PNP transistors cannot handle more than a small fraction of this current density. Although substrate PNPs may retain a beta of ten up to current densities of 1mA/mil^2, substrate injection usually limits them to maximum currents of a few milliamps. Lateral PNP transistors rarely achieve more than 250μA/minimum emitter. Most high-current circuits avoid the use of PNP transistors entirely, even if this eliminates otherwise-attractive circuit topologies.

Small-signal transistors can handle up to about 10mA and 100mW without any precautions. Beyond this point they become increasingly vulnerable to failure mechanisms caused by high currents and high power dissipation. These problems become especially acute in transistors handling currents in excess of 100mA or dissipating power in excess of 500mW. Such transistors require specialized layouts to protect them from thermal runaway and secondary breakdown. With careful layout, one can successfully integrate transistors capable of conducting 10A and dissipating 100W. Power transistors of this magnitude require so much die area that they completely dominate the layout of the integrated circuit. The cost of constructing a large integrated power device greatly exceeds the cost of purchasing an equivalent discrete device. Very high-power or high-current devices also require special packages that do not readily accommodate large numbers of pins. Most integrated power transistors conduct less than 2A and dissipate less than 10W. Power lateral PNP transistors require enormous amounts of die area, and few designs incorporate PNP transistors conducting more than 500mA. The vast majority of power bipolar transistors are therefore power NPN devices.

Many different power NPN layouts have been proposed. Each offers its own unique combination of advantages and disadvantages. No single structure outper-

[1] H. C. Lin, "Comparison of Input Offset Voltage of Differential Amplifiers Using Bipolar Transistors and Field-Effect Transistors," *IEEE J. of Solid-State Circuits*, Vol. SC-5, #6, 1970, pp. 126–129.

forms all the others in all applications. In order to make an intelligent choice, the designer must understand the mechanisms that cause power transistors to fail.

9.1.1. Failure Mechanisms of NPN Power Transistors

The three most common problems encountered in the design of power bipolar transistors are emitter debiasing, thermal runaway, and secondary breakdown. These three problems all result from the large currents and the high power dissipations typical of power devices. None of these mechanisms cause much trouble in small signal transistors, but all impose significant constraints on power transistor design.

Emitter Debiasing

The term *emitter debiasing* refers to a nonuniform current distribution that may develop in a power bipolar transistor due to voltage drops in the extrinsic base and emitter, and in their respective leads. The high transconductance of bipolar transistors makes these devices very susceptible to changes in base-emitter bias. Small voltage drops down the base or emitter leads can radically redistribute current flow through the transistor. Some portions of the transistor may conduct little or no current, while others conduct far more current than they were designed to handle. The overloaded portions of the transistor become vulnerable to thermal runaway and secondary breakdown.

Figure 9.1 shows an example of emitter debiasing occurring between the separate emitter fingers of a power transistor. In the accompanying schematic, transistors Q_1 to Q_4 represent the four emitter fingers and resistors R_1 to R_3 represent the resistance of the metal leads connecting the fingers together. Assume that each emitter

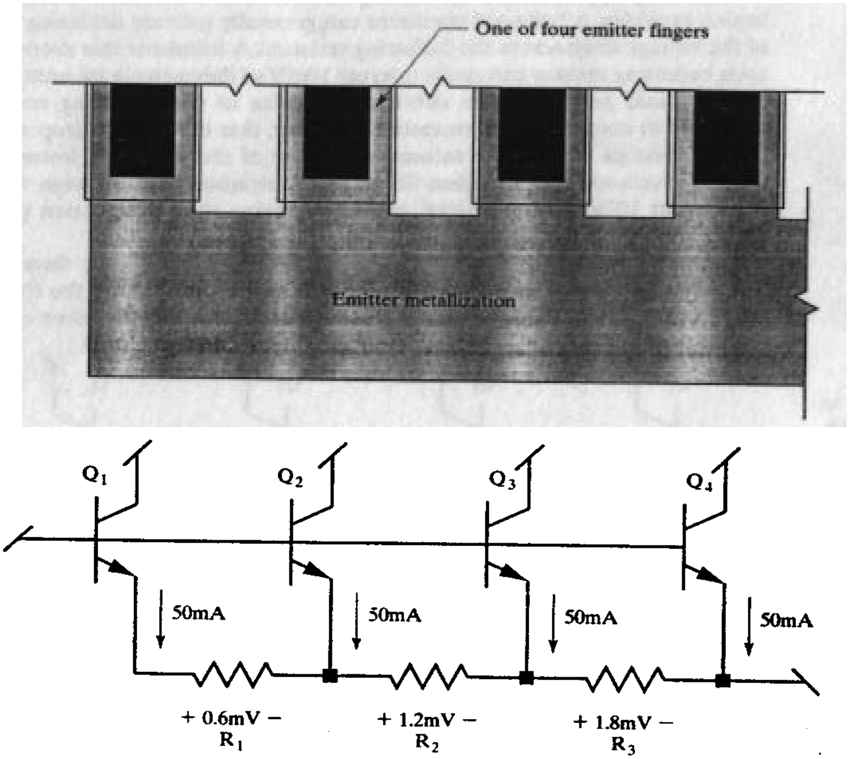


FIGURE 9.1 The layout and equivalent schematic of a power transistor having four emitter fingers. The values listed on the schematic follow the computations described in the text.

finger conducts 50mA, and that each resistor consists of one square of 20kÅ aluminum with a sheet resistance of 12mΩ/□. The total drop across the three resistors equals 3.6mV. The ratio of emitter currents η between two transistors whose base-emitter voltages differ by a voltage ΔV_{BE} equals

$$\eta = e^{\Delta V_{BE}/V_T} \qquad [9.1]$$

where V_T represents the thermal voltage of silicon (approximately 26mV at room temperature). In this example, the ratio of currents equals 1.15, so the rightmost finger Q_4 would conduct about 15% more current than the leftmost finger Q_1. Analog BiCMOS processes encounter even more severe debiasing problems because they use thinner metallization (typically 10kÅ).

The previous example illustrates the severity of emitter debiasing—relatively small currents flowing through short, wide leads still cause 3.6mV of debiasing. A technique called *emitter ballasting* can greatly reduce the impact of debiasing. Emitter ballasting requires the insertion of resistors into each emitter lead (Figure 9.2). These resistors are typically sized to provide a voltage drop of 50 to 75mV at full rated current. For example, emitter fingers conducting 50mA each might employ 1Ω ballasting resistors. The addition of these ballasting resistors forces the emitter current to redistribute about equally between the emitter fingers. If any emitter finger attempts to draw more than its fair share of current, then the voltage drop across its ballasting resistor increases. This limits the amount of current that can flow through this emitter finger. Voltage drops between ballasted emitters appear primarily across the ballasting resistors rather than across the base-emitter junctions of the transistors. Thus 3.6mV of debiasing between two emitters ballasted with 1Ω resistors would result in a 1.8mA current increase in one emitter and a 1.8mA current decrease in the other. These numbers are only approximations, but they serve to show how much benefit emitter ballasting provides. A ballasted transistor can generally tolerate debiasing equal to 25% of the voltage drop across the ballasting resistors. A transistor that drops 50mV across each ballasting resistor can easily tolerate 10mV of debiasing in its emitter leads. If the layout would produce more debiasing, the size of the ballasting resistors can be increased to compensate. Remember, however, that the voltage drop across the ballasting resistors adds to the saturation voltage of the transistor, lowers its effective transconductance, and increases its power dissipation. If the design requires more than about 100mV of ballasting, consider altering the metallization pattern or the aspect ratio of the transistor.

Emitter debiasing can also develop within a single emitter finger (*intrafinger debiasing*). Voltage drops accumulate as the current flows along the finger. One end of the emitter finger sees a larger base-emitter voltage and therefore conducts more

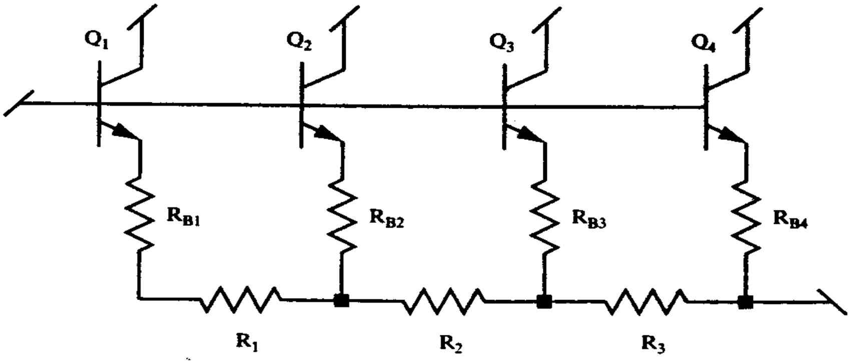


FIGURE 9.2 Connection of ballasting resistors to the segmented power transistor in Figure 9.1; R_{B1} to R_{B4} are the ballasting resistors for the emitter fingers Q_1 to Q_4.

current than the other. Debiasing along a long emitter finger can actually become a more serious problem than debiasing between separate fingers. For a narrow emitter finger like that in Figure 9.3, the voltage drop from one end to the other should not exceed 5mV. Assuming that an emitter lead of constant width runs down the emitter finger, and assuming that equal currents flow into the emitter lead along each increment of its length, then the total voltage drop from one end of the emitter contact to the other end equals

$$\Delta V_{BE} = \frac{LR_s I_E}{2W} \qquad [9.2]$$

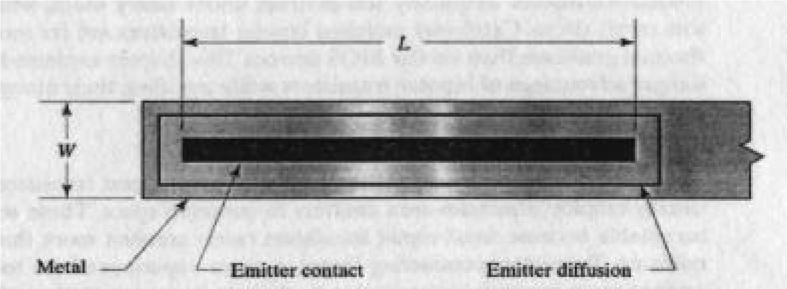


FIGURE 9.3 A sample emitter finger layout showing measurements L and W used in Equation 9.2.

where R_s represents the sheet resistance of the metallization, W is the width of the emitter lead. L is the length of the emitter contact, and I_E equals the total current flowing out of the entire emitter finger (Figure 9.3). For example, suppose an emitter finger conducts 50mA along a lead 300μm long by 30μm wide constructed from 12mΩ/□ aluminum. Equation 9.2 indicates that the debiasing along this emitter finger equals 9mV, which exceeds the maximum suggested debiasing of 5mV. Although this computation does not consider the redistribution of emitter current in response to debiasing, it still demonstrates the severity of the debiasing problem.

Several options exist for reducing intrafinger debiasing. The emitter fingers may be shortened and widened. This not only minimizes the finger length but also allows the use of wider metal leads. Alternatively, the transistor may employ a larger number of shorter emitter fingers of the same width as the originals. A ballasting technique also exists that applies to individual fingers (Section 9.1.2), but it can only provide a limited amount of ballasting, which may not suffice to compensate for poor emitter finger design.

Thermal Runaway and Secondary Breakdown

Both thermal runaway and secondary breakdown result from an intensification of current flow through portions of the power transistor. In the case of thermal runaway, the current flow localizes in response to increasing temperature. Suppose that one portion of the power transistor becomes slightly warmer than the rest. The V_{BE} required to maintain constant collector current drops by 2mV/°C, so a temperature rise of only a few degrees results in significant emitter debiasing. Almost all of the current flows through the hottest portion of the transistor, raising its temperature still further. In a matter of milliseconds, the region of conduction collapses to a tiny *hot spot* comprising only a few percent of the transistor's area. Perhaps beta rolloff can limit the collapse of the hot spot sufficiently to prevent catastrophic device failure, or perhaps not. Even if the hot spot stabilizes, the transistor will be so severely overstressed that it will become vulnerable to other failure mechanisms such as secondary breakdown, electromigration, and thermally accelerated corrosion.

Since thermal runaway involves emitter debiasing, ballasting resistors can provide some measure of protection against it. If each finger of a multi-emitter transistor has its own ballasting resistor, then a hot spot that develops in one finger cannot steal current from the other fingers. Even in a worst-case scenario in which hot spots develop in all of the fingers, each hot spot absorbs only a fraction of the total current. Usually 50mV of ballasting suffices to control thermal runaway, but more ballasting is sometimes necessary to offset voltage drops in the emitter metallization system.

Hot spots can still develop in individual emitter fingers even if all of the fingers have ballasting resistors. If each finger has its own ballasting resistor, then the current drawn by any one hot spot will decrease as the number of emitter fingers increases. Distributed emitter ballasting (Section 9.1.2) also provides some measure of protection against hot spot formation. Extremely demanding applications may require a combination of distributed emitter ballasting and individual ballasting resistors for each emitter finger.

Secondary breakdown occurs when the emitter current density in a transistor exceeds a critical threshold value J_{crit}. Beyond this point, the sustained collector-to-emitter breakdown voltage $V_{CEO(sus)}$ snaps back to a new, lower value called the *secondary breakdown voltage* V_{CEO2} (Section 8.1.3). A transistor is most vulnerable to secondary breakdown when it is in the process of turning off. The collector-to-emitter voltage across the transistor rises as the emitter current through the transistor decreases. Secondary breakdown occurs if the collector-to-emitter voltage exceeds V_{CEO2} while the emitter current density exceeds J_{crit}. Once avalanche begins, the base drive circuit can no longer turn the transistor off, and the transistor is soon destroyed by overheating or metallization failure.

Transistors driving inductive loads are extremely vulnerable to secondary breakdown. Consider a power transistor Q_1 driving a high-side inductive load L_1 (Figure 9.4A). As soon as Q_1 begins to turn off, the inductive kick-back of L_1 drives the collector voltage V_{CE} upward until recirculation diode D_1 begins to conduct (Figure 9.4B). The collector voltage reaches its maximum value almost immediately, long before the emitter current drops to zero. Secondary breakdown will occur if the collector voltage exceeds V_{CEO2} while the emitter current density exceeds J_{crit}.

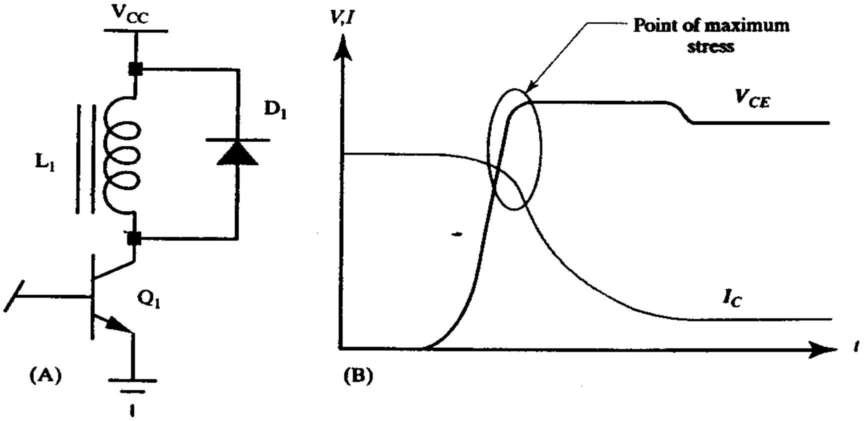


FIGURE 9.4 An example of (A) a bipolar transistor driving an inductive load and (B) the waveforms associated with the turn-off interval of the transistor.

Conservative design rules dictate that power transistors operate at an emitter current density of no more than 5 to 10mA/mil^2 (8 to 15μA/μm^2). These current densities lie well below those required to trigger secondary breakdown in order to provide a safety margin for emitter debiasing, emitter current focusing, and thermal gradient formation.

9.1.2. Layout of Power NPN Transistors[2]

Over the years, a number of alternative layouts have been proposed for NPN power transistors. Each layout has certain strengths and weaknesses, so knowledge of several different types of layouts will aid the designer in choosing the best style for any given application. Any layout can be scaled by adding or removing emitter sections or by connecting several power devices in parallel.

A transistor used in a *linear-mode* application remains in the forward active region for long periods of time. Linear transistors must withstand large collector-to-emitter differentials while simultaneously conducting large collector currents. Such a transistor must cover enough area to allow for heat dissipation. As a general rule, linear-mode transistors should not dissipate more than 100mW/mil^2 (150μW/μm^2) of emitter nor conduct more than 5mA/mil^2 (8μA/μm^2). These are conservative guidelines and, with sufficient ballasting and heatsinking, it is possible to successfully operate transistors at several times these stress levels. Still, one should generally follow these conservative guidelines unless empirical measurements show that a transistor can safely operate at higher stress levels.

Transistors used in *switched-mode* applications operate either in cutoff where no current flows, or in saturation where collector-to-emitter differentials remain small. Switching transistors dissipate power only during brief switching intervals. The average power dissipation of switching transistors remains relatively small, so they rarely experience hot spot formation or thermal runaway. On the other hand, switching applications generate many opportunities for emitter current focusing during turn-off. Conservative designs should not exceed an emitter current density of 10mA/mil^2 (16μA/μm^2) to ensure that emitter current focusing does not trigger secondary breakdown.

Transistors that drive purely capacitive loads such as MOS gates conduct current only as infrequent short-duration pulses. Such *pulsed-mode* transistors typically conduct large currents for a few hundred nanoseconds, then rest for a few microseconds before conducting again. Pulsed-mode transistors are unharmed by emitter-current focusing because the external capacitive load will quench conduction regardless of whether secondary breakdown occurs. Pulsed-mode transistors are also immune to thermal runaway because hot spots cannot form and localize in less than a few microseconds.[3] Most pulsed-mode applications rely on high-current beta rolloff and collector resistance to limit conduction. This practice is acceptable as long as the pulse duration does not exceed 1μS, the intervals between pulses are no shorter than 250nS, and the *average* emitter current density does not exceed 10mA/mil^2 (16μA/μm^2). The metallization for pulsed-power transistors should be designed following the electromigration rules for intermittent currents described in Section 14.3.3.

The Interdigitated-emitter Transistor

The oldest style of power transistor, the *interdigitated-emitter transistor,* remains in use because it can operate at higher speeds than any other style of bipolar power transistor. Figure 9.5 shows an interdigitated-emitter transistor constructed using a single-level-metal standard-bipolar process.

This transistor consists of a number of emitter fingers, each having its own dedicated emitter ballasting resistor. The ballasting resistors are all formed from a single

[2] F. W. Trafton, "High Current Transistor Layout," unpublished manuscript, 1988.

[3] H. Melchior and M. J. O. Strutt, "Secondary Breakdown in Transistors," *Proc. IEEE,* Vol. 52, 1964, p. 439.

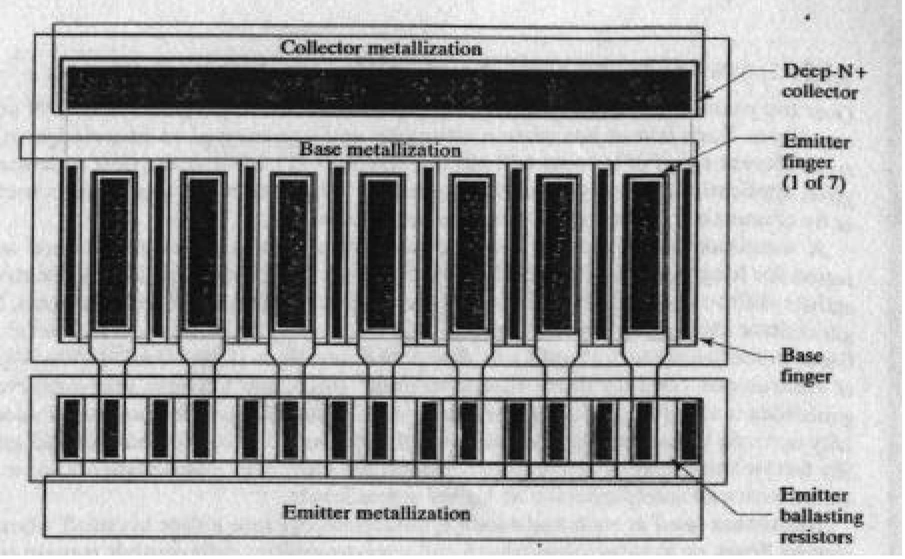


FIGURE 9.5 An example of an interdigitated-emitter power transistor. Each emitter finger has a separate ballasting resistor. Metallization is shown in gray for emphasis.

strip of emitter diffusion placed in a separate tank. The emitter diffusion is not isolated from the tank because small current leakages from one finger to another cause no harm. Each emitter finger connects to two ballasting resistors placed in parallel, each consisting of about one square of emitter. Assuming a minimum emitter sheet resistance of 5Ω/□, this provides a ballasting resistance of 2.5Ω per finger. This resistance will provide 50mV of ballasting at 20mA of emitter current.

The interdigitated-emitter transistor is extremely vulnerable to intrafinger debiasing. The voltage drop down each emitter finger computed using equation 9.2 should not exceed 5mV. A large number of short emitter fingers are preferable to a small number of long fingers. The width of the emitter fingers also affects performance. Widening the emitter fingers also widens the pinched base regions underneath them. The resulting increase in base resistance causes the transistor to exhibit slower switching and increased emitter focusing. The fastest and most robust designs incorporate minimum-width emitter fingers, but it is difficult to place enough metal on narrow emitter fingers to prevent them from debiasing. Double-level metal helps, but narrow fingers still do not make efficient use of available area. Most designers compromise on an emitter width of 0.3 to 1.0mil (8 to 25μm). In this style of transistor, the emitter contacts are always made as large as possible to reduce emitter resistance.

Base contacts along either side of each emitter finger reduce the base resistance and enable faster switching. Base contacts placed on either end of the emitter array ensure that the end fingers turn off as quickly as the others. If these end contacts were omitted, the end fingers would turn off more slowly than the others. This could lead to emitter current focusing and secondary breakdown during turnoff. Minimum-width base contacts help conserve space, and relatively few designs require more base metallization. Power transistors operating at high current densities may, however, experience enough beta rolloff to necessitate wider base metallization. Computation can show whether or not any particular design experiences significant base-lead debiasing. Designs exhibiting more than 2 to 4mV of base debiasing should be redesigned to reduce base metallization resistance. The comb-style base metallization in Figure 9.5 exhibits much less metallization resistance than the

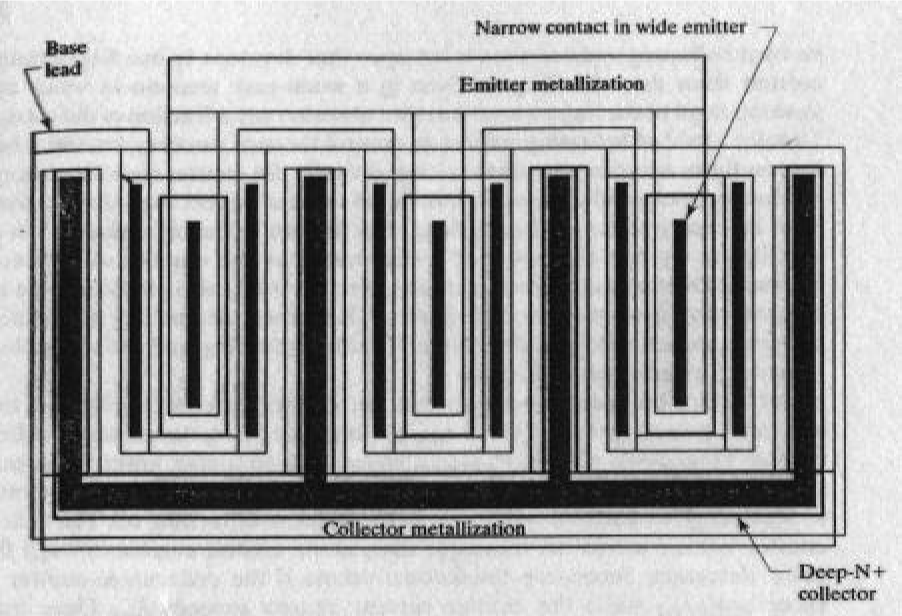


FIGURE 9.6 An example of wide-emitter narrow-contact power transistor. Emitter ballasting resistors, although not shown, can easily be added. Metallization is shown in gray for emphasis.

serpentine metallization in Figure 9.6. Unfortunately, many single-level-metal designs do not lend themselves to the use of comb-style base metallization.

The transistor in Figure 9.5 contains deep-N+ only along one side. This may suffice for a linear-mode device operating at a collector-to-emitter voltage differential of a half-volt or more. A switching transistor is a different matter, as their efficiency is determined by their collector-to-emitter voltage drop in saturation (their *saturation voltage*). If the saturation voltage is too large, then the transistor dissipates too much power. At high currents, the collector resistance of a switching transistor equals the sum of its vertical deep-N+ resistance and its lateral NBL resistance.[4] The vertical resistance of the deep-N+ sinker can be reduced by increasing its area. The sinker should not be less than 10μm wide to ensure that outdiffusion does not dilute its doping and increase its vertical resistance. The lateral NBL resistance can be reduced by contacting the NBL along a longer periphery or by decreasing the distance between the active portions of the device and the sinkers. Placing sinkers along both sides of the transistor reduces the NBL resistance by a factor of four, and an unbroken ring of deep-N+ around the transistor reduces it still further. The NBL should extend to the outer edge of the deep-N+ sinker to ensure a low-resistance connection between the two.

An unbroken ring of deep-N+ around a power transistor also forms a hole-blocking guard ring that helps control substrate injection during saturation. When an NPN transistor saturates, all of its unused base drive flows to the substrate. The guard ring does not reduce the amount of base drive consumed by the transistor, but it does prevent the majority of it from flowing to the substrate. Section 9.1.3 discusses several techniques for limiting the base current consumed during saturation.

[4] The drift region usually contributes little or no resistance since it depletes through under the influence of reverse bias or velocity saturation.

The Wide-emitter Narrow-contact Transistor

The interdigitated-emitter transistor uses relatively narrow emitter fingers to reduce base resistance and to control emitter crowding. This structure's low base resistance allows it to operate at higher frequencies than any other. Unfortunately, the narrow emitters are quite prone to emitter crowding. Emitter debiasing causes conduction to concentrate at the exit end of each finger, while thermal gradients focus conduction into the middle of the transistor. In either case, current tends to localize at one point in each emitter finger. Ballasting resistors can help ensure that the fingers conduct equal currents, but they cannot prevent intrafinger debiasing. Even well-ballasted interdigitated-emitter transistors tend to develop hot spots at higher current densities.

If each emitter finger is divided into a large number of individually ballasted sections, then no one portion of the emitter finger can contact more current than any other. Although it is generally not feasible to segment an emitter finger in this manner, placing a narrow emitter contact in a wide emitter finger provides similar benefits.[5] Figure 9.6 shows the resulting *wide-emitter narrow-contact transistor.*

The use of a wide emitter finger and a narrow contact produces the equivalent of a distributed network of ballasting resistors. This network consists partly of emitter resistance and partly of pinched base resistance. The emitter resistance is largest at the periphery of the emitter and smallest in the center directly beneath the narrow contact. Conversely, the base resistance is smallest at the periphery and is largest in the center directly under the emitter contact. These two forms of ballasting complement one another. At low currents, the base resistance is relatively insignificant, and current distributes uniformly across the width of the emitter finger. As the current increases, debiasing in the pinched base region causes conduction to move out toward the periphery of the emitter finger. The current must now flow through a larger emitter resistance. The resulting emitter voltage drops counteract the movement of current toward the emitter periphery. Together, the base-side and emitter-side distributed ballasting ensure that conduction occurs relatively uniformly across the entire width of the emitter finger. This type of emitter ballasting is distributed along the length of the emitter finger, so it protects all portions of the device against emitter debiasing and the formation of hot spots.

The emitter must overlap the contact by a distance sufficient to provide adequate ballasting. Typical wide-emitter narrow-contact structures employ emitter overlaps of 0.5 to 1.0mils (12 to 25μm). Larger overlaps unnecessarily slow the frequency response of the transistor, while smaller overlaps may not provide enough distributed ballasting to fully protect against thermal runaway and secondary breakdown. Transistors operating under extreme conditions often benefit from additional ballasting resistors inserted into the leads of each emitter finger, as illustrated in the interdigitated-emitter transistor in Figure 9.5.

Some designers employ a trapezoidal emitter contact tapering from a wide low-current end to a narrow high-current end. This design provides additional ballasting at the high-current end of the emitter finger to offset the effect of voltage drops in the emitter metallization. The lack of matching between metal and emitter sheet resistances makes this type of design problematic, and most wide-emitter narrow-contact transistors employ minimum-width contacts instead. One must not stretch the contact to the ends of the emitter finger, because this eliminates the ballasting at these points and renders them vulnerable to hot spot formation.

[5] A. B. Grebene, *Bipolar and MOS Analog Integrated Circuit Design* (New York: John Wiley and Sons, 1984), p. 510.

The transistor in Figure 9.6 employs multiple base regions with fingers of deep-N+ interdigitated between them. This structure minimizes collector resistance at the cost of increasing area and complicating lead routing. In single-level-metal layouts, the base lead must serpentine through the transistor between the collector and emitter metallization. The added length of the serpentined base lead can cause significant debiasing in the base metallization. Even with the distributed ballasting, the transistor should not contain metallization drops of more than a few millivolts. Base debiasing can be reduced by a factor of roughly four by connecting both ends of the serpentined emitter lead. Layouts employing double-level metal frequently use comb or grid arrangements to combat base metallization debiasing.

The wide-emitter narrow-contact structure is remarkably robust. The distributed emitter ballasting helps prevent thermal runaway and secondary breakdown within individual fingers, allowing the device to operate at higher current densities than comparable interdigitated structures do. Wide-emitter narrow-contact transistors that must operate under especially harsh conditions may benefit from the insertion of an additional 50 to 75mV of emitter ballasting in the leads of the individual emitter fingers. This structure does not switch as quickly as the interdigitated-emitter transistor, but the degradation is not as large as one might expect since considerable high-current conduction occurs along the emitter periphery.

The Christmas-tree Device

Another type of power transistor layout is nicknamed the *Christmas-tree device* because of the peculiar shape of its emitter geometry (shown in dark gray in Figure 9.7). Historically, this structure was widely used in linear applications because of its exceptional resistance to thermal runaway. It is rarely used for switching applications, because the same features that improve its immunity to thermal runaway degrade its ability to withstand emitter current focusing during turnoff.

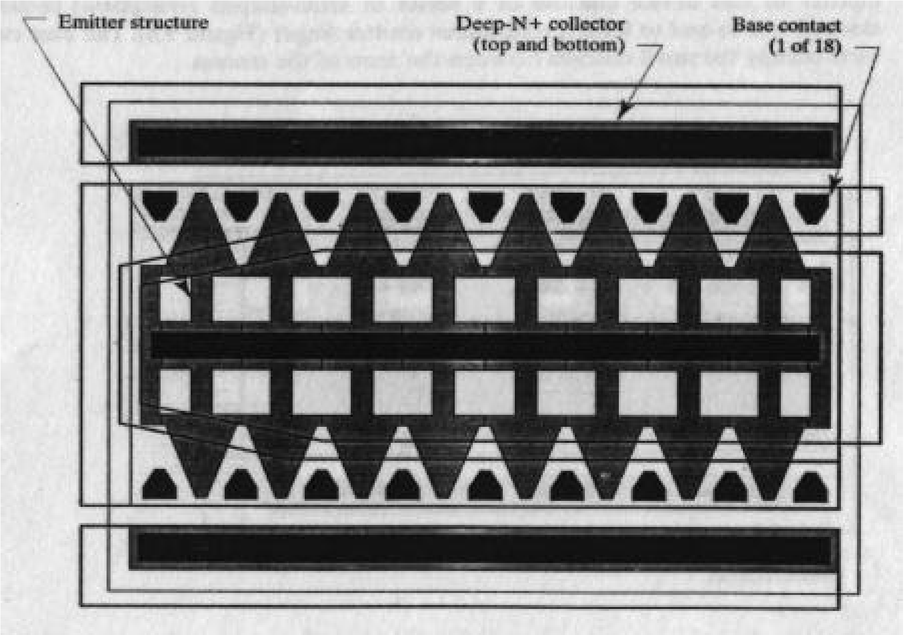


FIGURE 9.7 An example of a Christmas-tree power transistor. The base is shown in light gray and the emitter in dark gray in order to highlight the peculiar structure of the emitter.

The emitter of this transistor consists of a central spine surrounded by a complex branching structure of triangular prongs that give the transistor its picturesque name. Most of the conduction occurs in the triangular prongs along the emitter periphery. These connect to the central spine of the emitter through narrow emitter strips that act as ballasting resistors. At low currents, all portions of the emitter conduct. As the current increases, emitter crowding forces conduction out toward the periphery, causing current to flow through the ballasting resistors incorporated into the emitter structure. This device gains its resistance to thermal runaway from a large amount of distributed ballasting. Unfortunately, the great width of the emitter structure renders it vulnerable to emitter current focusing. As the transistor begins to turn off, the area of conduction retreats from the periphery toward the central spine. Because the spine represents only a small portion of the total emitter area, the emitter current density increases dramatically during the final stages of turnoff. This concentration of current flow can (and often does) trigger secondary breakdown. The wide-emitter narrow-contact structure exhibits superior immunity to secondary breakdown because the distance from the periphery to the center of the emitter is not as great and the effects of emitter focusing are not as dramatic.

The Christmas-tree device serves best in applications that dissipate large amounts of power, but where abrupt turn-off transitions never occur. Historically, this style of device was frequently chosen for the series-pass devices of linear voltage regulators and the output stages of audio power amplifiers. A number of variations on the Christmas-tree device have been developed in an attempt to minimize its vulnerability to emitter focusing while retaining its immunity to hot spot formation. None of these variants are as robust as the wide-emitter narrow-contact transistor, or its descendent, the cruciform-emitter transistor.

The Cruciform-emitter Transistor

The *cruciform-emitter transistor* represents an evolutionary development of the wide-emitter narrow-contact structure that seeks to incorporate additional emitter ballasting without rendering the device vulnerable to secondary breakdown. The emitter of this device consists of a series of cross-shaped (*cruciform*) sections stacked end-to-end to form a continuous emitter finger (Figure 9.8). The base contacts occupy the small notches between the arms of the crosses.

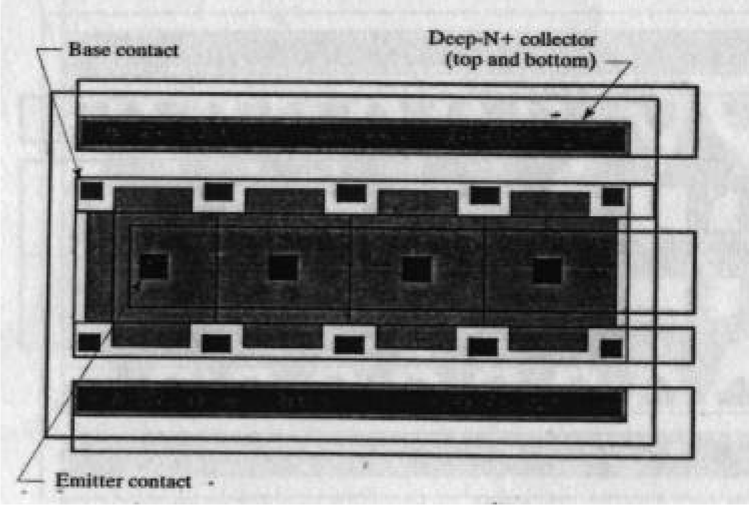


FIGURE 9.8 An example of a cruciform-emitter transistor. The base is shown in light gray and the emitter in dark gray for emphasis.

The width of the cruciform emitter has been increased to 3 to 5mils (75 to 125μm) to obtain additional ballasting. The narrow emitter contact has also been replaced by a series of small, square or circular contacts occupying the center of each cross. All of the emitter current must flow through these contacts, producing a distributed three-dimensional ballasting effect considerably more efficient than the two-dimensional ballasting generated by the wide-emitter narrow-contact structure. Consequently, the cruciform emitter combines the best features of the wide-emitter narrow-contact transistor and the Christmas-tree device. The cruciform transistor does not have quite the immunity to secondary breakdown that the wide-emitter narrow-contact transistor does. but it vastly outperforms the Christmas-tree device in this respect. The cruciform emitter transistor also makes extremely efficient use of space.

The cruciform structure suffers from two drawbacks. First, the small size of the emitter contacts renders them vulnerable to electromigration. All of the emitter current must cross the sidewalls of these contacts. and this produces very high localized current densities in the metallization. Even refractory barrier metal has its limits, which this transistor may exceed. Some designers replace the single contact in the center of each cruciform with an array of minimum contacts to increase the sidewall perimeter. Second. the compact design of the cruciform emitter can cause extreme localized heating at high power levels. Less area-efficient transistors are actually preferable to more compact ones from the standpoint of heat dissipation. If the heat produced by the transistor spreads over a wider area. then the thermal impedance between the transistor and the package decreases and the transistor can handle more power before it overheats. The cruciform structure is best suited for switching applications, as these are more strongly constrained by current-handling capability than by power dissipation.[6]

Power Transistor Layout in Analog BiCMOS

Any of the power transistors discussed above can also be implemented in analog BiCMOS. Figure 9.9 shows a BiCMOS version of a wide-emitter narrow-contact transistor. Double-level metallization allows the base contacts to completely encircle each emitter finger, whereas in a single-level-metal design they can only reach two or three sides of each finger. The complete ring of base contact helps ensure that all portions of the emitter periphery are equally active. An unbroken ring of deep-N+ sinker minimizes collector resistance and blocks substrate injection during saturation.

Figure 9.9B shows further details of the metallization system. Emitter current flows from the narrow emitter contacts to vias placed parallel to them. Passing up through these vias. the current reaches a metal-2 plate covering the top of the transistor. This plate minimizes emitter debiasing by reducing the metal-2 resistance to an absolute minimum. The resistance in the metal-1 plates actually serves as emitter ballasting and therefore is unobjectionable. The base metallization consists of a grid of first-level metal covering the base contacts. The base current exits the transistor through a metal-2 jumper placed between the emitter metal-2 plate and the encircling collector metal-2. If necessary, a second base lead can exit on the other side of the emitter plate. The collector metallization consists of a complete ring of metal-1 covering the collector contact and a U-shaped metal-2 plate covering the collector on three sides of the transistor. Vias along the inner edges of the collector contact allow current to flow through both levels of metallization. The collector lead can exit any side of the transistor except the side where the emitter lead exits. The

[6] A related structure called the *H-emitter transistor* is described in F. F. Villa, "Improved Second Breakdown of Integrated Bipolar Power Transistors," *IEEE. Trans. on Electron Devices*, Vol. ED-33, #12, 1986.

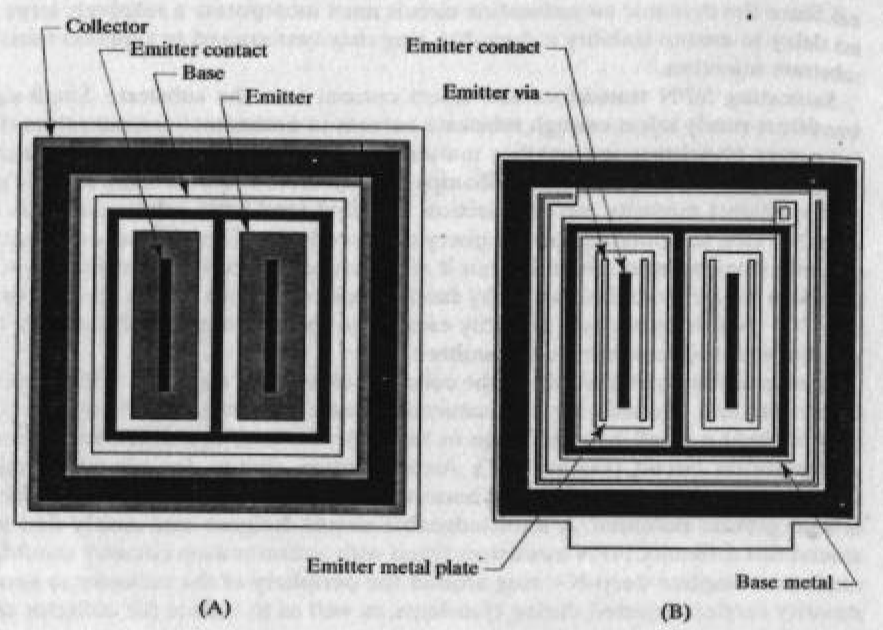


FIGURE 9.9 A wide-emitter narrow-contact transistor constructed in analog BiCMOS using double-level metal: (A) diffusions and (B) metal-1 pattern. The metal-2 pattern is not shown.

best arrangement places the collector lead diametrically opposite of the emitter lead. This minimizes the resistance of the collector metallization by ensuring that half of the current flows through the metal on either side of the transistor.

The structure in Figure 9.9 has been used to fabricate pulse-power transistors capable of operating at emitter current densities of more than 100mA/mil^2 (160μA/μm^2). This structure uses an overlap of emitter over contact of 8 to 12μm and a continuous ring of deep-N+ sinker at least 8μm wide. The NBL should completely overlap the deep-N+ sinker to minimize resistance and to ensure that no minority carriers can escape through a lightly doped portion of the extrinsic collector. This structure can continuously conduct in excess of 10mA/mil^2 (16μA/μm^2) and can operate as either a linear dissipative device or as a switching element. The distributed emitter ballasting inherent in the wide emitter fingers prevents hot spots from forming even at very high power levels. The use of a solid metal-2 plate to terminate the emitters helps minimize emitter debiasing, making individual emitter ballasting resistors unnecessary for all but the most demanding applications.

Selecting a Power Transistor Layout

All of the power transistor layouts presented in this section have their advantages and their disadvantages. The Christmas-tree device is best suited for linear applications that do not experience rapid switching transients. The interdigitated emitter transistor provides the best switching speeds and frequency response, but it requires individual ballasting resistors on each finger to avoid thermal runaway. The wide-emitter narrow-contact and cruciform transistors excel in switching applications. A wide-emitter narrow-contact transistor with ballasting resistors in each emitter finger is virtually immune to secondary breakdown at the voltages normally encountered in integrated circuit applications (10 to 40V). Some applications require that a small portion of the transistor's emitter be brought out independently to act as a sensing element. The interdigitated emitter structure offers the easiest insertion of a sense emitter and the best matching of the sense emitter to the remainder of the transistor. Table 9.1 summarizes these advantages and disadvantages.

	Interdigitated Emitter	Wide-emitter Narrow-contact	Christmas-tree Device	Cruciform Transistor
Thermal runaway	Good*	Good	Excellent	Excellent
Secondary breakdown	Fair	Excellent	Poor	Good
Frequency response	Excellent	Good	Fair	Fair
Compactness of layout	Poor	Good	Good	Excellent
Ease of emitter sensing	Excellent	Fair	Poor	Poor

* Assumes individually ballasted emitter fingers; otherwise *poor.*

TABLE 9.1 Comparison of four types of power NPN layouts.

9.1.3. Saturation Detection and Limiting

Both lateral PNP and vertical NPN transistors inject current into the substrate when they saturate. Substrate injection wastes supply current and may cause substrate debiasing and device latchup. Several techniques have been developed to suppress substrate injection, either by intercepting minority carriers before they reach the substrate or by preventing the transistor from saturating in the first place. Most of these techniques require specialized layouts.

The emitter of a lateral PNP continuously injects minority carriers into its tank. When the transistor saturates, most of this current flows to the substrate. Small-signal transistors rarely inject enough current to warrant concern, but a few designs incorporate large lateral PNP transistors that conduct tens or even hundreds of milliamps. Currents of this magnitude can easily produce enough debiasing to trigger latchup.

Figure 9.10A shows one way to prevent minority carriers from reaching the substrate. This transistor incorporates a continuous, unbroken ring of deep-N+ around the outside edge of its tank. This ring merges with the underlying NBL and completely

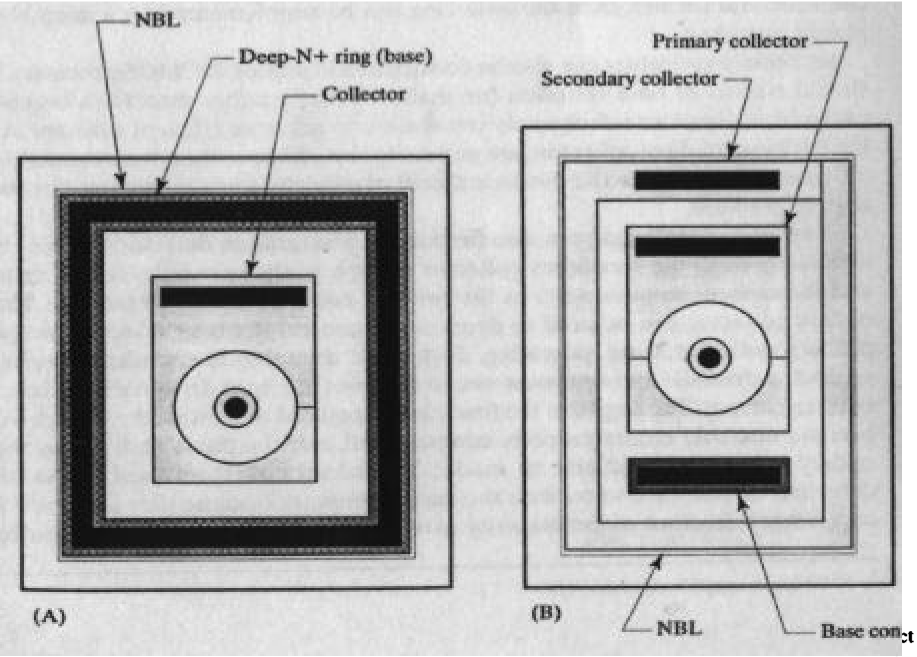


FIGURE 9.10 Two examples of lateral PNP transistors modified to minimize saturation: (A) transistor ringed with deep-N+ and (B) transistor with secondary collector.

encloses the base region of the lateral PNP within a wall of heavily doped N-type silicon. The built-in potential caused by the N+/N− doping gradient repels minority carriers. The few that actually penetrate the deep-N+ or the NBL usually recombine before reaching the isolation. Although this structure is compatible with BiCMOS processing, it may prove less effective in CDI processes. Most of these processes use relatively lightly doped NBL layers to reduce lateral autodoping during epitaxial deposition. The well doping gradient causes minority carriers to drift down toward the NBL/N-well interface, and a substantial fraction of these carriers may penetrate the NBL and enter the substrate. Measurements on a 7V analog BiCMOS process that uses relatively shallow, heavily doped wells yielded NBL penetration figures in excess of 10%.[7]

Figure 9.10B shows another method for preventing substrate injection. The illustrated device incorporates a ring of base diffusion completely encircling its primary collector. This ring acts as a *secondary collector.* As long as the primary collector does not saturate, few carriers can reach the secondary collector and it conducts little current. When the primary collector saturates, the carriers begin to flow to the secondary collector. So long as the secondary collector does not simultaneously saturate, it collects most of the carriers and prevents them from reaching the isolation sidewalls. The secondary collector is sometimes called a *ring collector* because it often takes the form of an unbroken ring enclosing the primary collector.

The secondary collector can perform one of several functions depending on how it is connected. If it connects to ground, then it returns any carriers it collects to the ground return line. When the transistor saturates, the emitter current flows to ground, but it does not pass through the substrate. This provides approximately the same electrical functionality as an unprotected lateral transistor without the risk of substrate debiasing. Alternatively, the secondary collector can connect to the base lead. When the transistor saturates, the carriers collected by the secondary collector add to the base current and cause the apparent beta to rapidly decline. This connection provides the same functionality as a deep-N+ ring, while consuming considerably less space. If the designer wishes to increase the efficiency of the ringed collector still further, then the base ring can be supplemented by a deep-N+ ring placed outside it.

Secondary collectors can also be constructed in analog BiCMOS processes. These should consist of base diffusion (or shallow P-well) rather than PSD, because the source/drain implant is frequently too shallow to act as an efficient collector. Analog BiCMOS secondary collectors are generally less effective than their standard bipolar counterparts due to the downward drift of minority carriers produced by the well doping gradient.

A secondary collector can also function as a saturation detector. Current begins to flow through the secondary collector as soon as the primary collector saturates, and the current stops as soon as the primary collector ceases to saturate. The secondary collector can be used to dynamically control the base drive to prevent the primary collector from saturating. Instead of dumping the unwanted current to ground, a *dynamic antisaturation circuit* throttles the base drive back to reduce the emitter current. The negative feedback loop required to control the base drive may become unstable unless properly compensated, and the phase shift across the secondary collector is difficult to model. Secondary collectors used for saturation detection do not have to encircle the entire transistor because they need only intercept a small fraction of the minority carriers to generate the necessary control sig-

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

nal. Since the dynamic antisaturation circuit must incorporate a relatively large signal delay to ensure stability, a deep-N+ ring may be required to suppress transient substrate injection.

Saturating NPN transistors also inject current into the substrate. Small-signal transistors rarely inject enough substrate current to necessitate antisaturation rings, but power transistors are another matter entirely. Any saturating NPN transistor that conducts more than a few milliamps of base drive requires some form of protection against minority carrier injection. The first (and best) solution consists of a deep-N+ ring surrounding the periphery of the collector. This ring not only contains minority carriers inside the tank, but it also reduces the collector resistance of the transistor at the same time. Minority carriers that recombine within the tank or the deep-N+ guardring become majority carriers in the collector, and from there they pass through the transistor to the emitter.

A base diffusion placed within the collector of an NPN transistor collects minority carriers and can also act as a saturation detector. Power switching transistors often include a small base diffusion in the collector tank connected to a dynamic antisaturation circuit (Figure 9.11). Antisaturation circuits for grounded-emitter transistors are difficult to construct because the secondary collector must operate at or near ground potential. A knowledgeable circuit designer can usually find ways around this difficulty. NPN transistors fitted with antisaturation circuitry should still employ a complete deep-N+ ring around the periphery of the collector to contain minority carriers injected during transients, as well as to reduce the collector resistance of the power transistor.

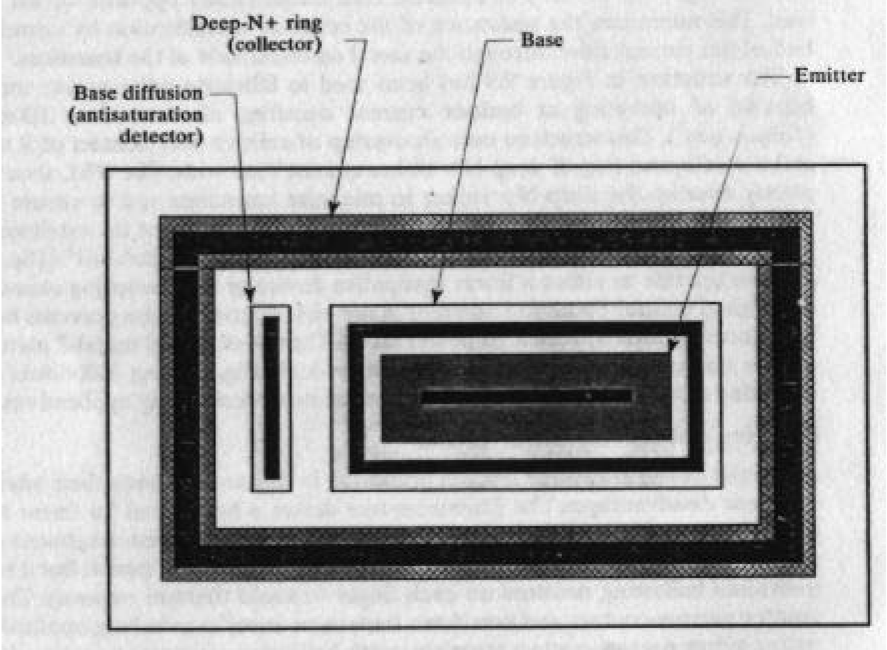


FIGURE 9.11 An NPN switching transistor incorporating both a complete deep-N+ ring and an outer collector that functions as a saturation detector.

No generally accepted symbols exist for transistors with secondary collectors or deep-N+ guardrings. Figure 9.12 shows a set of symbols that have achieved some degree of industry recognition. A thick base bar denotes a power transistor, or more generally, any transistor requiring special layout (A). A diagonal slash across the collector lead of an NPN indicates the presence of a deep-N+ sinker (B), as does

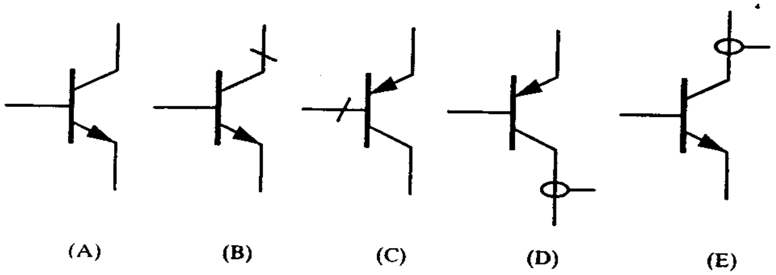


FIGURE 9.12 Proposed symbols for (A) a power NPN, (B) a power NPN with a deep-N+ collector, (C) a lateral PNP with deep-N+ in base, (D) a lateral PNP with a secondary collector, and (E) an NPN transistor with saturation detection.

the presence of a similar slash across the base lead of a lateral PNP (C). The addition of a small ring encircling the collector lead of a lateral PNP denotes the addition of a secondary collector[8] (D), while the addition of a similar ring around the collector lead of an NPN transistor denotes a base diffusion placed in the tank as a saturation detector (E). The majority of these circuits use NPN transistors because of their superior device characteristics.

9.2 MATCHING BIPOLAR TRANSISTORS

Many analog circuits require matched bipolår transistors. Current mirrors and current conveyors use them to replicate currents; amplifiers and comparators use them to construct differential input stages; references use them to produce known voltages and currents; and Gilbert translinear circuits use them to perform analog computations. All of these applications depend on precise matching of collector currents and base-emitter voltages, sometimes from transistors of the same size, and sometimes from ones of different sizes.

NPN collector currents scale approximately with drawn emitter area, but no model can precisely predict the influence of emitter geometry on matching. It is therefore very difficult to match transistors with different sizes and shapes of emitters. Most bipolar circuits employ simple integer ratios such as 1:1, 2:1, 4:1, or 8:1. These ratios are easily obtained by assembling multiple copies of an identical unit device. The same techniques can obtain virtually any ratio of small integer numbers, such as 2:3, 3:4, and 2:5. Ratios requiring more than eight or ten unit devices become increasingly impractical due to area requirements and the sensitivity of large devices to temperature gradients.

Two transistors having identical dimensions and operating at equal collector currents should theoretically develop exactly the same base-emitter voltage. In practice, small differences in the emitter saturation currents cause the two base-emitter voltages to vary slightly. The difference between the base-emitter voltages of two transistors operating at equal current densities is called the *offset voltage* ΔV_{BE} and can be computed from the following equation

$$\Delta V_{BE} = V_T \ln\left(\frac{I_{S1}}{I_{S2}}\right) \qquad [9.3]$$

where I_{S1} and I_{S2} are the emitter saturation currents of the two transistors. Given that the thermal voltage V_T equals 26mV at room temperature, a 1% mismatch in emitter saturation currents produces an offset voltage of 0.25mV. Saturation currents scale approximately with drawn emitter area, so a 1% variation in emitter area also produces a 0.25mV of offset.

[8] A different symbol for the ringed-collector PNP, as well as a unique application for this device, are found in H. Lehning, "Current Hogging Logic (CHL)—A New Bipolar Logic for LSI," *IEEE J. Solid-State Circuits*, Vol. SC-9, #5, 1974, pp. 228–233.

9.2.1. Random Variations

Random fluctuations in base doping and emitter junction area set the ultimate limits of vertical bipolar transistor matching. Other significant sources of random variation include recombination in the emitter-base depletion region and lateral injection across the base diffusion, both of which scale inversely with the emitter area-to-periphery ratio. Matched bipolar transistors therefore employ relatively compact emitter geometries. The three geometries favored for constructing matched vertical NPN transistors are squares, octagons, and circles (Figure 9.13). Each style of emitter has its proponents. The circle has the largest area-to-periphery ratio and therefore theoretically provides the best possible matching. However, circles are approximated as many-sided polygons during pattern generation. Squares require no such approximations, so many designers believe that they are rendered more precisely on the photomask. Octagons also require no approximations, and they possess slightly larger area-to-periphery ratios than do squares.

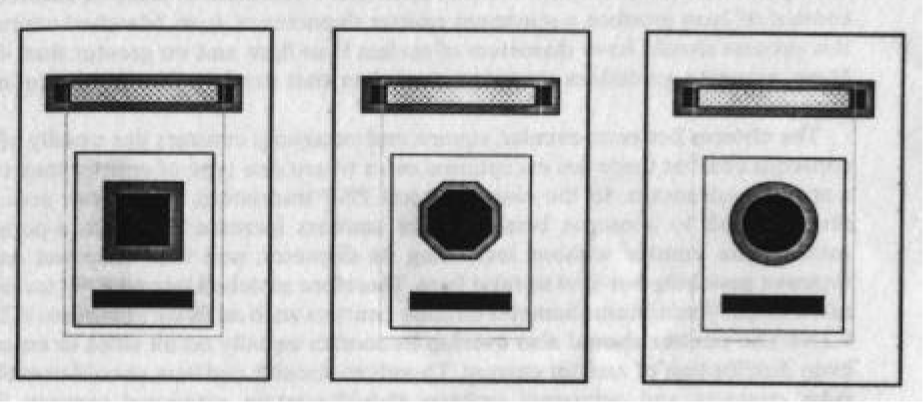


FIGURE 9.13 Examples of NPN transistors designed with square, octagonal, and circular emitters.

In practice, all three styles of emitters provide excellent matching. Although circles are approximated as polygons, identical circles produce identical polygons. The approximations involved in generating circular emitters therefore have little impact on their matching. Furthermore, the differences in area-to-periphery ratios among squares, octagons, and circles are relatively insignificant. The area-to-periphery ratio R_{AP} of any geometry can be determined using the following equation

$$R_{AP} = k_r\sqrt{A_e} \qquad [9.4]$$

where A_e represents the emitter area and k_r is a dimensionless constant equal to 0.250 for squares, 0.274 for octagons, and 0.282 for circles. Note that the area-to-periphery ratio is not itself a dimensionless quantity. For example, if the emitter area is measured in square microns, then R_{AP} will have dimensions of microns. Equation 9.4 shows that the reduction in peripheral effects gained by using a circular emitter can be equaled by simply increasing the area of a square emitter by 25%.

The mismatch between a pair of transistors due to peripheral and areal fluctuations has a standard deviation s that equals

$$s = \frac{1}{\sqrt{A_e}}\sqrt{k_a + \frac{k_r k_p}{\sqrt{A_e}}} \qquad [9.5]$$

where k_a and k_p are constants representing the contributions of areal and peripheral fluctuations, respectively. The contribution of the peripheral term decreases as the area increases. For sufficiently large emitters, the areal term dominates and the random mismatch becomes inversely proportional to the square root of emitter area

A_e. Most transistors with emitters that are more than two or three times the minimum diameter exhibit relatively little dependence on peripheral variation, again demonstrating the relative indifference of matching to geometric considerations (in particular, to the value of k_r).[9] Therefore, for most practical purposes, the standard deviation of the mismatch can be assumed to equal

$$s \cong \sqrt{\frac{k_a}{A_e}} \qquad [9.6]$$

Although large emitters exhibit less random mismatch than small ones, there are other factors to consider. Any increase in emitter size increases the spacing between the devices and therefore renders them more vulnerable to thermal and stress gradients. Large emitters also exhibit increased base pinch resistance. Because of these problems, one must avoid making matched emitters either too large or too small. As a general rule, the diameter of the emitter of a matched NPN transistor should not be less than twice nor more than ten times the minimum possible diameter. For example, a minimum contact width of 2μm and a minimum overlap of emitter over contact of 1μm produce a minimum emitter diameter of 4μm. Matched emitters in this process should have diameters of no less than 8μm and no greater than 40μm. More accurate guidelines require actual data that rarely exists for a production process.[10]

The choices between circular, square, and octagonal emitters are usually of little consequence, but there are exceptional cases where one type of emitter may confer a specific advantage. In the case of lateral PNP transistors, the emitter area must remain small to conserve beta. Circular emitters increase the area-to-periphery ratio of the emitter without increasing its diameter, and thus help not only to improve matching but also to raise beta. Therefore matched lateral PNP transistors often employ minimum-diameter circular emitters such as those in Figures 8.22 and 8.26B. The emitter should also overlap its contact equally on all sides to ensure an even distribution of emitter current. Therefore circular emitters should contain circular contacts, and octagonal emitters should contain octagonal contacts. These arrangements are not possible in processes that allow only minimum-dimension square contacts. In such cases, square emitters should be used rather than circular or octagonal ones.

Many circuits require matched transistors having unequal device areas. Although it is possible to connect identical unit transistors in parallel, the large amounts of die area required by collector isolation actually degrade matching by exacerbating the effect of thermal and stress gradients. Matched NPN transistors can occupy a common tank because the geometry of the collector has almost no effect on their matching (Figure 9.14A). The geometry of the base-collector junction also has relatively little impact on matching since most of the conduction occurs either directly underneath the emitter or immediately adjacent to it. Several emitters can therefore occupy the same base region (Figure 9.14B). The emitters must be placed far enough apart to prevent minority carriers that are injected by one from being collected by another.

[9] One study has shown that the area parameter alone provides a good fit to as small as 16μm^2: H.-Y. To and M. Ismail, "Mismatch Modeling and Characterization of Bipolar Transistors for Statistical CAD," *IEEE Trans on Circuits and Systems - I: Fundamental Theory and Applications*, Vol. 43, #7, 1996. pp. 608–610.

[10] The real problem lies not in finding experimental data, but in finding data collected from a properly designed experiment that has undergone proper statistical analysis. A few measurements made on a couple of devices constitute neither a properly designed experiment nor a properly analyzed set of data. Decisions made on such basis are unlikely to be any better than those based on the admittedly *ad hoc* rule given in the text.

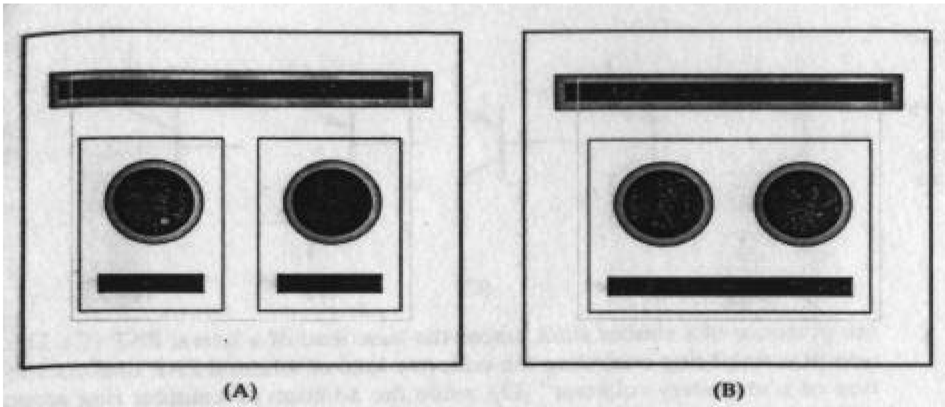


FIGURE 9.14 Two styles of multiple-emitter NPN transistors: (A) separate base regions in a common tank and (B) separate emitters in a common base region.

Similarly, the emitters should reside far enough inside the base diffusion to minimize lateral conduction that would otherwise produce mismatches between transistors having different numbers of emitters. These requirements can be met by increasing both the emitter-to-emitter spacing and the base overlap of emitter by 1 to 2μm. These increased spacings ensure that the individual emitters will not interact with one another or with the collector-base junction.

Matched lateral PNP transistors can also occupy a common tank to save area. Multiple emitters cannot occupy a single opcning in a collector geometry because each emitter would interfere with the flow of minority carriers from the others. Instead, each emitter must occupy its own collector opening, and all of these openings must have identical dimensions. The outside dimensions of the collector geometry and the size and shape of the tank have little impact on matching. Therefore matched lateral transistors usually consist of rectangular arrays of minimum emitters placed in a common collector region (Figure 8.26B).

9.2.2. Emitter Degeneration

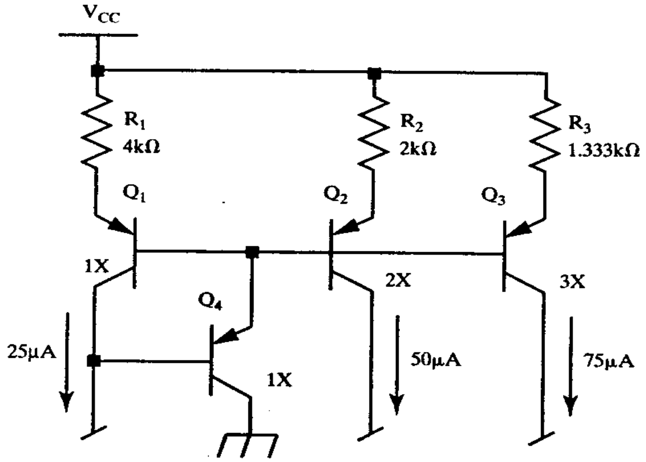
Regardless of the care taken in their construction, some types of bipolar transistors simply do not match very well. A technique called *emitter degeneration* can transfer the burden of matching from a set of bipolar transistors to a set of associated resistors. This technique will improve the overall matching of the circuit as long as the resistors match more precisely than the bipolar transistors. Emitter degeneration also increases the output resistance of bipolar transistors and therefore reduces the systematic errors due to finite Early voltages. The systematic mismatch in collector currents between two matched bipolar transistors operating at different base-collector voltages equals

$$\frac{I_{C1}}{I_{C2}} \cong 1 + \left(\frac{\Delta V_{BC}}{V_A}\right)\left(\frac{V_T}{V_T + V_d}\right) \quad [9.7]$$

where $\Delta V_{BC} = V_{BC1} - V_{BC2}$, V_A is the Early voltage of the transistors, V_T is the thermal voltage (26mV at 25°C), and V_d is the voltage developed across the degeneration resistors. Equation 9.7 is only valid as long as V_{BC1} and V_{BC2} are both much smaller than V_A. This equation indicates that 50mV of degeneration reduces the Early error by approximately a factor of three.

Figure 9.15 shows a lateral PNP current mirror consisting of three lateral PNP transistors Q_1 to Q_3. Each of these transistors has an associated emitter degeneration

FIGURE 9.15 Lateral PNP current mirror that incorporates emitter degeneration resistors.



resistor R_1 to R_3. The mirror also uses a *beta helper* transistor Q_4 to minimize the effect of low betas on matching. Transistor Q_4 does not need to match any of the other transistors, and it does not normally require emitter degeneration.[11]

In this example, transistors Q_1 to Q_3 have device sizes of one, two, and three, respectively. These sizes represent the number of unit emitters in each transistor. This dimensionless notation avoids any possible confusion between emitter periphery and emitter area, and simultaneously frees the circuit designer from worrying about exact layout dimensions. The values of resistors R_1 to R_3 are inversely proportional to the sizes of transistors Q_1 to Q_3. Each resistor therefore generates the same voltage differential, which in this case equals 100mV. This voltage represents the amount of degeneration applied to the transistors. Approximately 50 to 75mV of degeneration suffices to ensure that the resistors determine the matching of the mirror rather than the transistors. Few circuits require more than 100mV of degeneration as long as the ratio of the emitter areas of the transistors lies within ±10% of the desired value.

The improvement in matching obtained through emitter degeneration depends on how well the resistors match and on the nature of the mismatches between the bipolar transistors. Well-matched resistors vary no more than ±0.1%, while the currents of well-matched minimum-area emitters typically vary by ±1% or more. The matching of the degenerated transistors will approximately equal the matching of the emitter degeneration resistors. On the other hand, the area consumed by the resistors could also be used to increase the emitter areas. Increasing the emitter areas of matched NPN transistors usually proves more area-effective than adding emitter degeneration resistors, since well-matched resistors are not small. Emitter degeneration may sometimes provide better matching in the presence of large thermal gradients because resistors are less susceptible to these gradients than bipolar transistors.

Lateral PNP transistors might seem to benefit less from emitter degeneration than vertical NPN transistors do since much of the mismatch between lateral transistors stems from beta variations that remain unaffected by degeneration. Despite

[11] Sometimes a small resistor is added in series with the emitter of the beta helper as part of a frequency compensation network: this resistor does not take part in the matching of the mirror and its value is noncritical.

this, lateral PNP transistors are frequently degenerated because these devices use minimum emitters to maintain acceptable betas. Lateral PNP transistors also benefit from increased output resistance caused by emitter degeneration because these devices often have rather low Early voltages.[12] Emitter degeneration cannot improve the matching of split collector transistors because it is not possible to provide a separate emitter degeneration resistor for each split collector.

Large amounts of emitter degeneration are sometimes used to obtain noninteger ratios between transistors. The size of the transistors becomes relatively unimportant in the presence of 250 to 500mV of degeneration. For example, a 3.4:1 ratio can be obtained by ratioing a 3X transistor and a 10kΩ resistor with a 1X transistor and a 34kΩ resistor. This technique works equally well with both NPN and PNP transistors.

9.2.3. NBL Shadow

Surface discontinuities caused by oxidation during the NBL anneal propagate upward during epitaxial deposition to produce a surface discontinuity called the *NBL shadow*. A mechanism called *pattern shift* can displace the NBL shadow laterally by a distance of up to twice the epi thickness (Section 7.2.3). Mismatches can occur if the NBL shadow intersects the emitter of a vertical NPN transistor. Arrays of multiple-emitter NPN transistors are particularly vulnerable to pattern shifts perpendicular to their axis of symmetry. Consider the two transistors in Figure 9.16A. Pattern shift has displaced the NBL shadow toward the right, causing it to intersect the leftmost emitter of each device. Suppose that the affected emitters experience a 1% reduction in emitter area. Only one of Q_A's two emitters is affected, so its emitter area becomes 1.99. The sole emitter of Q_B is also affected, so its emitter area becomes 0.99. The new ratio between the two devices is 1.99:0.99, or about 2.01:1. This represents a mismatch of 0.5%, or an offset voltage of about 0.13mV.

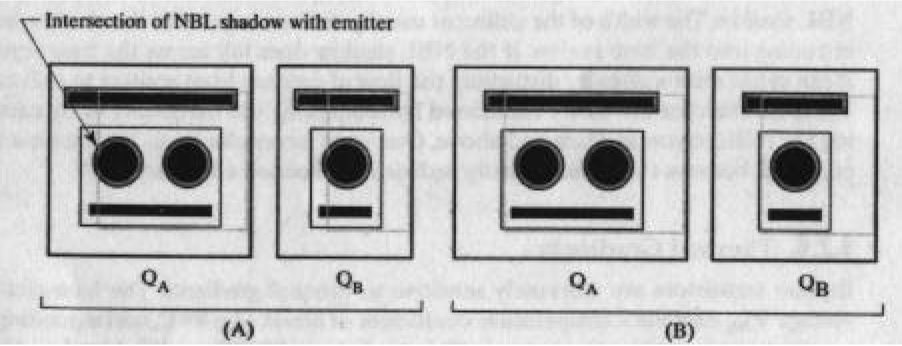


FIGURE 9.16 (A) NBL shadow causes mismatch between two transistors. (B) This mismatch is eliminated by oversizing NBL to prevent intersection of NBL and emitter.

Several ways exist to prevent pattern shift from causing mismatches. One approach consists of replacing the multiple-emitter transistor Q_A with two single-emitter transistors that are identical to Q_B. The NBL shadow will now intersect all three emitters in exactly the same manner, and the resulting systematic variations should cancel one another. Unfortunately, pattern distortion can also produce random variations in the NBL shadow that do not cancel one another. These can only be avoided by ensuring that the NBL shadow does not intersect the active area of the transistor, which, in the case of NPN transistors, is defined by the emitter diffusion.

[12] Gray, *et al.* conclude that NPN transistors derive more benefit from emitter degeneration than PNP transistors, but their arguments are flawed because they ignore the other factors discussed in the text: P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3rd ed. (John Wiley and Sons, New York: 1993), pp. 317–320.

If the direction of pattern shift is known, then the transistors can be laid out in a CEB array in which the main axis of symmetry lies parallel to the direction of pattern shift. In this way, the shift will displace the NBL shadow into either the collector contact or the base contact. The space required by these contacts usually suffices to prevent the NBL shadow from reaching the emitter. Pattern shift in (111) silicon usually occurs along the <211> axis. The direction of pattern shift in tilted (100) silicon depends on the direction of the tilt, which may vary from one manufacturer to the next. In order to properly orient the transistor array, the designer must determine the relationship between the X–Y coordinates of the layout and the wafer orientation. This information can usually be obtained by microscopic examination of a die. Planarization will obscure the NBL shadow, but a wafer can be removed prior to planarization for examination. Remember that the reticle array may differ from one device to another depending on the choices made during pattern generation, and that rotated or reflected reticle arrays will alter the apparent direction of pattern shift.

Sometimes the direction of pattern shift is not known, or layout considerations preclude the use of a specific orientation. In such cases, the overlap of NBL over emitter can be increased to prevent the NBL shadow from intersecting the emitter (Figure 9.16B). If no data exists on the magnitude of the pattern shift, then the designer should overlap the NBL over the emitter by 150% of the epi thickness.

Some designers eliminate the NBL shadow by omitting NBL from the matched transistors. Without NBL, the collector resistance of an NPN transistor can reach several kilohms. The V_{CEO} of the transistor may also diminish due to punchthrough of the lightly doped collector. CDI NPN transistors are particularly vulnerable to punchthrough due to the extremely light doping of the lowest portions of the N-well. One should not remove NBL from transistors unless characterization data indicates that the resulting device will function properly without it.

Lateral PNP transistors generally do not suffer from mismatches caused by the NBL shadow. The width of the collector usually suffices to prevent the shadow from intruding into the base region. If the NBL shadow does fall across the base region, it can cause mismatches by disturbing the flow of carriers from emitter to collector. These mismatches are easily eliminated by reorienting the transistors or by enlarging the NBL region as discussed above. One must never eliminate NBL from a lateral PNP, because this would greatly reduce its collection efficiency.

9.2.4. Thermal Gradients

Bipolar transistors are extremely sensitive to thermal gradients. The base-emitter voltage V_{BE} exhibits a temperature coefficient of about –2mV/°C, corresponding to a collector current temperature coefficient of about 80,000ppm/°C. Matched bipolar transistors are routinely expected to achieve offset voltages of less than ±1mV, corresponding to a temperature difference of only ±0.5°C. Temperature variations of this magnitude can easily occur in almost any integrated circuit.

Matched bipolar transistors are often used to construct differential pairs, ratioed pairs, and ratioed quads. A *differential pair* (also called a *diff* pair, an *emitter-coupled pair*, or a *long-tailed pair*) consists of two matched bipolar transistors whose emitters are connected as in Figure 9.17A. The input stages of amplifiers and comparators often consist of differential pairs whose collectors terminate into matched resistors or current mirrors. The input offset voltage of a bipolar amplifier or comparator depends largely upon the matching of the input differential pair. Various trimming schemes can reduce the random component of the offset voltage to a fraction of a millivolt. Trimming also minimizes the temperature coefficient of the offset

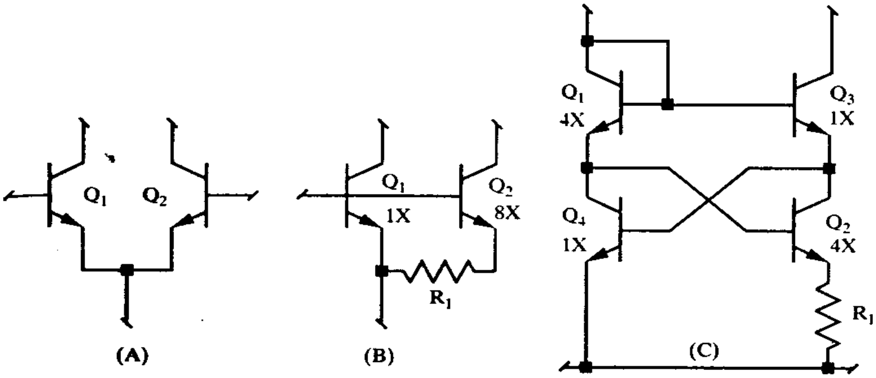


FIGURE 9.17 Three circuits containing matched NPN transistors: (A) differential pair, (B) ratioed pair, and (C) ratioed quad.

voltage as well as its absolute value,[13] so it is often used to minimize the vulnerability of high-gain amplifiers to a phenomenon called *thermal feedback*.

Thermal feedback occurs when one portion of a circuit influences another through thermal interactions rather than electrical ones. Changes in voltage or current within relatively high-power circuits (such as the output stages of an amplifier) produce localized temperature fluctuations that, in turn, generate small offsets between devices in the input stages of the circuit. The circuit then amplifies these offsets as if they were an electrical signal. The amplified offsets can produce further temperature variations, possibly even leading to oscillations. Since many amplifiers have voltage gains in excess of 10,000, even a very weak thermal interaction can cause significant thermal feedback. The frequency response of many commercial operational amplifiers contain low-frequency poles and zeros caused by this mechanism.[14] Thermal feedback can be minimized by increasing the separation of the input and output stages and by reducing the thermal sensitivity of the input stage. Many operational amplifiers place the input circuit on one side of the die and the output circuit on the other. Even so, thermal coupling remains a serious problem. The input differential pair of a high-gain amplifier should always be located and constructed to achieve the highest possible degree of matching in order to minimize its sensitivity to thermal variations.

A *ratioed pair* consists of two bipolar transistors whose emitter areas are in integer ratio. Assuming that the two transistors conduct equal currents, then their base-emitter voltages will differ by an amount ΔV_{BE} equal to

$$\Delta V_{BE} = V_T \ln\left(\frac{A_1}{A_2}\right) \qquad [9.8]$$

where V_T equals the thermal voltage (26mV at 25°C) and A_1 and A_2 represent the emitter areas of transistors Q_1 and Q_2.[15] The thermal voltage scales linearly with absolute temperature,[16] therefore ΔV_{BE} is a *voltage proportional to absolute*

[13] Mismatch in a differential pair has the same impact as deliberate ratioing of emitter areas; the ΔV_{BE} voltage so developed has a large positive temperature coefficient. Therefore minimizing the offset at one temperature also tends to minimize temperature variability.

[14] J. E. Solomon, "The Monolithic Op Amp: A Tutorial Study," *IEEE J. Solid-State Circuits*, Vol. SC-9, #6, 1974, pp. 314–332.

[15] This derivation ignores the ideality factor (or emission coefficient) η, which is usually very near unity for NPN transistors operating at moderate current levels.

[16] An *absolute temperature* is one measured with respect to absolute zero. The SI unit of absolute temperature is the Kelvin degree (K), which has the same magnitude as the Celsius degree (°C); 0°C ≅ 273K and 25°C ≅ 298K. The thermal voltage V_T equals kT/q, where k is Boltzmann's constant ($1.38 \cdot 10^{23}$J/K), T is the absolute temperature (in K) and q is the charge on the electron ($1.60 \cdot 10^{-19}$C).

temperature (VPTAT). The VPTAT produced by a ratioed pair of NPN transistors remains linear with temperature and independent of current over a remarkably wide range of operating conditions. A resistor connected between the emitters of a ratioed pair (as in Figure 9.17B) can transform this VPTAT into a *current proportional to absolute temperature*, or IPTAT.[17] IPTAT circuits form the basis of many precision voltage and current references.

In order for VPTAT and IPTAT circuits to operate properly, the ratioed pair must match very precisely. The most common ratio used in VPTAT and IPTAT circuits is probably 8:1, which produces a ΔV_{BE} of 54mV. A 1mV mismatch in such a circuit would produce approximately a 2% error in the voltage or current. A typical trimmed voltage reference must produce a voltage that varies no more than ±1% over all possible operating conditions. Poor layout often leads to excessive output voltage variation with input voltage (poor *line regulation*) or to excessive output voltage variation with output current (poor *load regulation*). Both of these problems usually stem, at least in part, from thermal feedback.

A *ratioed quad* is essentially a variation on the ratioed pair. The four transistors of the quad produce a VPTAT voltage that is usually imposed across a resistor to produce an IPTAT (Figure 9.17C). The VPTAT voltage ΔV_{BE} equals

$$\Delta V_{BE} = V_T \ln\left(\frac{A_1 A_2}{A_3 A_4}\right) \qquad [9.9]$$

where A_l to A_4 are the emitter areas of transistors Q_1 to Q_4, respectively. The ratioed quad can provide a much larger ΔV_{BE} than a simple ratioed pair because the sizes of Q_1 and Q_2 multiply together. Two 4X transistors can produce a ΔV_{BE} of 72mV, while two 8X transistors can generate 108mV.

The extreme thermal sensitivity of bipolar transistors requires that matched devices be laid out to cancel thermal gradients. Critical matched devices almost always employ common-centroid layout techniques similar to those discussed in Section 7.2.6. Differential pairs usually employ the two-dimensional common-centroid layout shown in Figure 9.18. This configuration is popularly called a *cross-coupled quad*.[18] Common-centroid layouts help minimize the impact of thermal variations, but they cannot completely cancel nonlinear thermal variations. The exponential nature of the I_C vs. V_{BE} relationship sharply limits the cancellation achievable from common-centroidal layouts. Compactness is thus a highly desirable property of matched bipolar arrays. Most of the more complex common-centroidal arrangements lack compactness and therefore serve less well than the simple cross-coupled quad.

The VPTAT voltage developed by a ratioed pair increases as the logarithm of the area ratio, while offsets produced by thermal and stress gradients increase roughly as the square root of the ratio. As the area ratio increases, a point is eventually reached beyond which mismatch increases more rapidly than VPTAT. For any given design, there exists a ratio that will provide optimal matching. This optimal ratio depends on many factors, but in most cases it probably lies somewhere between 8:1 and 16:1. Even-number ratios greatly simplify the task of constructing common-centroid layouts, and smaller ratios consume less space, so 8:1 is probably the most popular choice of ratio. Similar arguments lead to ratioed quads of 4:1:1:4.

[17] In actuality, the temperature coefficient of the resistor will distort the linearity of the IPTAT. Most circuits use the IPTAT current to regenerate a VPTAT across another resistor. If the two resistors have the same temperature coefficient, then this can be (and is) neglected.

[18] Grebene, pp. 348–349, 365.

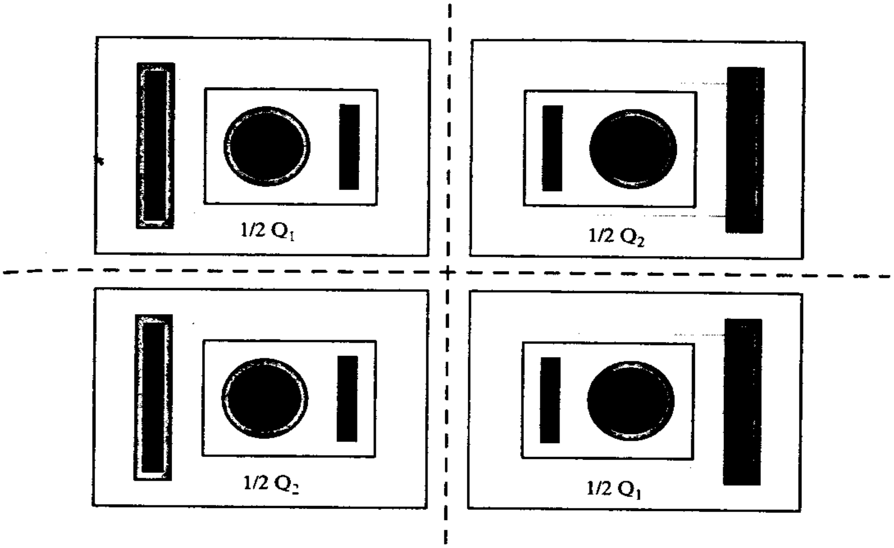


FIGURE 9.18 Example of cross-coupled bipolar transistors.

The layouts of ratioed pairs and ratioed quads must provide a high degree of symmetry in a compact layout. In a ratioed pair, one device usually possesses a single emitter and the other multiple emitters. One simple layout for a ratioed pair places the single-emitter device in the middle of two halves of the multiple-emitter device. The simplest structure embodying this arrangement is a one-dimensional common-centroid layout following the pattern ABA (Figure 9.19A). The placement of all of the emitters in a line generates a secondary axis of symmetry S_2 that enables the structure to reject thermal gradients perpendicular to the row of transistors, while the primary axis of symmetry S_1 rejects thermal gradients parallel to the line. The elongated shape of the array makes it more difficult to cancel gradients around S_1. Therefore arrays of this sort should be oriented so that the secondary axis S_2 lies parallel to the expected isotherms.

A more compact arrangement can be achieved for ratios that are multiples of 4:1. The emitters of the larger transistor Q_2 can then be arranged in two rows around an axis of symmetry S_2 (Figure 9.19B). Axis S_2 should also pass through the center of the single-emitter transistor Q_1. This arrangement is particularly beneficial for large ratios, such as 16:1, that would otherwise produce very elongated layouts. As before, the array should be oriented so that the secondary axis of symmetry S_2 lies parallel to the expected isotherms.

Ratioed quads are laid out as if they consisted of a pair of ratioed mirrors. Both the upper pair (Q_1 and Q_3 in Figure 9.17C) and the lower pair (Q_2 and Q_4) can employ layouts similar to those in Figure 9.19. Ideally, the two pairs should lie one above the other so that their primary axes of symmetry (S_1) coincide. This converts the entire arrangement into a two-dimensional common-centroid array. If for some reason this arrangement is not feasible, then each of the two pairs can be treated as an independent ratioed pair. A reasonable degree of matching will be achieved even if the two ratioed pairs reside some distance apart.

Some designers advocate laying out ratioed pairs using a circular array in which the smaller device occupies the center of a ring-shaped array of emitters forming the larger device. If the larger device contains a multiple of four emitters, then this arrangement will possess both horizontal and vertical axes of symmetry as well as a

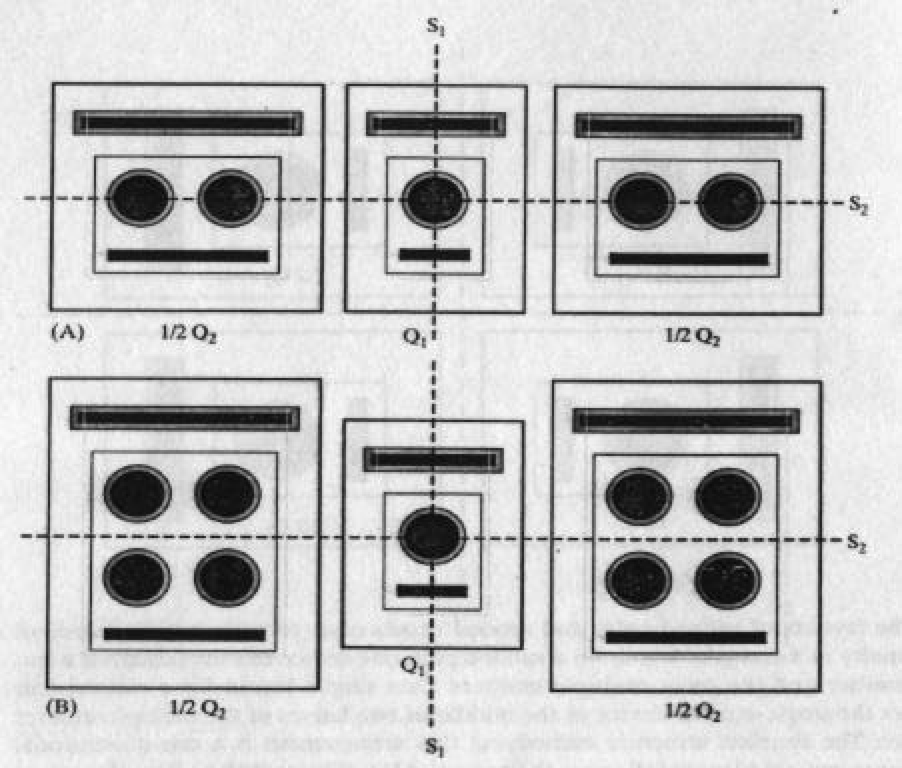


FIGURE 9.19 Two common-centroid layouts frequently used for constructing ratioed pairs.

number of subsidiary axes dependent on the total number of emitters. Although this arrangement has a high degree of symmetry, it consumes so much area that it is doubtful whether it actually matches as well as the simpler structures in Figure 9.19.

9.2.5. Stress Gradients

Mechanical stress can induce mismatches between bipolar transistors by altering their base-emitter voltages or by reducing their betas. The base-emitter voltage of a transistor depends on the bandgap voltage of silicon, which varies slightly under stress.[19] The degradation in beta under mechanical stress is principally due to mobility variations induced by piezoresistivity.[20] Together these effects can easily produce several millivolts of offset.

Assembled integrated circuits almost always experience a certain amount of stress. Plastic-molded devices are cured at elevated temperatures, and the mold compound elastically deforms as the assembled units cool. The resulting stresses become frozen into the assembled unit and do not dissipate over time. These residual stresses cause the base-emitter voltages of bipolar transistors to shift and can produce offsets between matched pairs of devices. These *package shifts* cannot be entirely trimmed out at wafer probe because they only appear after packaging. The residual stresses vary with temperature, so even post-package trimming cannot

[19] J. J. Wortman, J. R. Hauser, and R. M. Burger, "Effects of Mechanical Stress on p-n Junction Device Characteristics," *J. Applied Physics*, Vol. 35, #7, 1964, pp. 2122–2131.

[20] H. Mikoshiba and Y. Tomita, "Piezoresistance as the Source of Stress-induced Changes of Current Gain in Bipolar Transistors," *Solid State Electronics*, Vol. 25, #3, 1982, pp. 197–199.

entirely counteract package shifts. Package shifts therefore limit the minimum offset voltages obtainable between matched transistors.

Common-centroid layout techniques can greatly reduce the impact of stress on matched transistors. These techniques effectively move the matched transistors closer together and therefore equalize the stresses on them. Unfortunately, even the best common-centroid arrangements cannot cancel the higher-order components of the stress gradient. Matched transistors should therefore reside in low-stress regions of the die. Figure 9.20A shows the best locations for matched transistor arrays on a die fabricated in (100) silicon. These layouts assume that no significant sources of heat exist in the vicinity of the matched transistors. The best locations lie near the center of the die where the magnitude of the stresses reaches a broad minimum. The major axis of symmetry of the array S_1 should lie along one of the axes of symmetry of the die. This helps ensure that the isobars lie parallel to the secondary axis of the array S_2. If the matched transistors must reside along the side of the die, then they should be placed in the center of one side so that the primary axis of the array S_1 aligns to one axis of symmetry of the die. If the die is not square, then a location in the middle of the longer side is preferable to one in the middle of the shorter side. If possible, the transistors should reside at least 10mils (250μm) inside the edge of the die. Under no circumstances should critical matched bipolar transistors be placed near the corners of a die, as these experience excessively large stress gradients. In summary, the principles that guide the placement of matched bipolar transistors are analogous to the ones that apply to matched resistors, as discussed in Section 7.2.6.

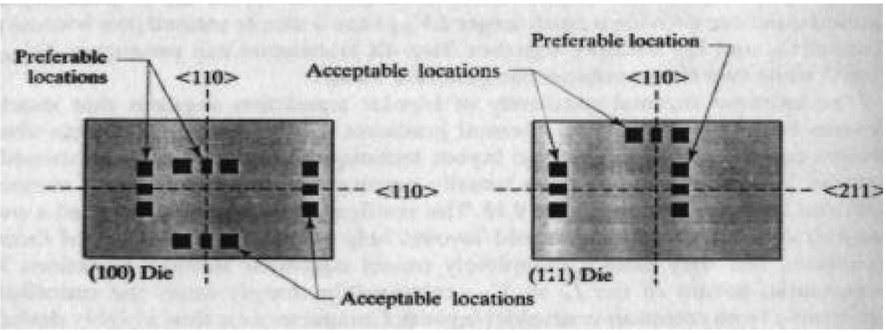


FIGURE 9.20 Locations for placing common centroid bipolar transistor arrays on (100) and (111) dice, in the latter case assuming an axis of symmetry exists in the stress distribution around the <211> axis (compare with Figure 7.16).

Figure 9.20B shows the best locations for critical matched transistors on a die fabricated in (111) silicon. Since isobars tend to lie symmetrically around the <211> axis, matched transistors ideally should reside around this axis instead of around the <110> axis. Locations near the center of the die will again give the lowest overall stress, although locations near either end of the die will provide acceptable matching. Less-critical matched transistors can also be placed on the <110> axis of symmetry. Again, matched transistors should reside at least 10 mils inside the die edge to avoid the increased stress gradients along the edges of the die, and matched transistors should never reside near the corners of the die because of the high stresses there.

Figure 9.21 shows the compromises required when the die contains both a significant heat source and critical matched transistors. Bipolar transistors are more susceptible to thermal gradients than to stress gradients, so the larger the distance between the heat source and the matched transistors, the better. The best compromise usually involves placing the heat source at one end of the die and the matched transistors at the opposite end. Consider increasing the aspect ratio of the die to 1.5:1, or even 2:1, since this helps increase the separation between the heat source

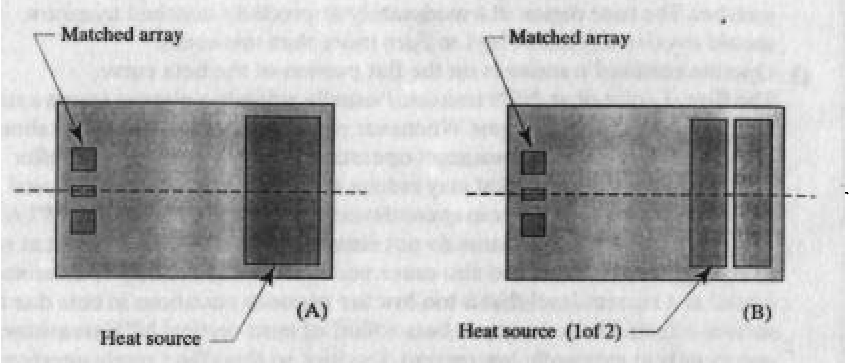


FIGURE 9.21 Layouts for matched transistor arrays on dice that contain one or more power devices.

and the matched transistors. The matched transistors should reside at least 5 to 10mils (125 to 250μm) away from the edge of the die opposite the heat source. This location should provide better overall matching than a location near the center of the die, even though the stresses are greater near the edges than at the center.

High-power integrated circuits may benefit from judicious application of emitter degeneration to transistors that would normally have sufficient area to match quite precisely on their own. High levels of power dissipation produce correspondingly large thermal gradients, especially in dice mounted on heat-sunk leadframes. These gradients have a much stronger effect on bipolar transistors than on passive components, so the matching of degenerated transistors usually proves superior to the matching of undegenerated ones. Emitter degeneration is particularly beneficial for matched transistors residing next to a large heat source; without degeneration such transistors often exhibit mismatches of many millivolts. Whenever emitter degeneration resistors are added to critically matched bipolar transistors, the resistors must be carefully laid out to ensure that they actually improve matching rather than degrade it. Poorly constructed degeneration resistors can easily ruin the matching of an otherwise well laid-out circuit.

9.3 RULES FOR BIPOLAR TRANSISTOR MATCHING

The previous section explained the mechanisms that cause mismatch in bipolar transistors. This section attempts to condense this information into a set of qualitative rules that will allow a designer to construct matched bipolar transistors with some degree of confidence, even if (as is usually the case) quantitative matching data is not available. For bipolar transistors constructed in a pure CMOS process, see Section 10.3.1.

The following rules use the terms *minimal, moderate,* and *precise* to denote increasingly precise degrees of matching. These terms should be interpreted as follows:

- **Minimal matching:** Typical three-sigma offset voltages of ±1mV or collector current mismatches of ±4%. This is suitable for constructing input stages of op-amps and comparators that must achieve three-sigma offsets of ±3 to 5mV without trim. It is also suitable for use in current mirrors for biasing noncritical circuitry.
- **Moderate matching:** Typical three-sigma offset voltages of ±0.25mV or collector current mismatches of ±1%. This level is suitable for use in ±1% bandgap references and in op-amps and comparators that must achieve ±1 to 2mV without trim. Since lateral transistors have difficulty maintaining this degree of matching, most untrimmed, moderately matched circuits use vertical NPN transistors instead.

- **Precise matching:** Typical three-sigma offset voltages of ±0.1mV or collector current mismatches of ±0.5%. This level of matching usually requires trimming or the addition of precisely matched degeneration resistors. Proper layout is still important because degeneration and trimming cannot entirely eliminate the effects of thermal gradients or package shifts. Lateral transistors cannot obtain this degree of matching unless they are heavily degenerated and the circuitry incorporates some means of base current cancellation. Circuits requiring precise matching usually employ heavily degenerated vertical NPN transistors.

9.3.1. Rules for Matching NPN Transistors

Vertical transistors inherently match better than lateral transistors because they are not subject to the vagaries of surface conduction. Most processes optimize the performance of their vertical NPN transistors at the expense of their lateral PNP transistors, which only strengthens the case for using NPN transistors. The following rules summarize the principles of designing matched NPN transistors:

1. Use identical emitter geometries.
   Transistors with different sizes or shapes of emitters match *very* poorly. Even minimal matching requires the use of identical emitter geometries. Matched transistors are therefore restricted to ratios of small integer numbers. The geometry of the base and collector regions matters much less than the geometry of the emitter region. Multiple emitters can thus reside in a common base region.
2. The emitter diameter should equal 2 to 10 times the minimum allowed diameter.
   The minimum diameter of the emitter equals the minimum contact width plus twice the minimum emitter overlap of contact. For example, a process having a minimum contact width of 2μm and a minimum overlap of 1μm has a minimum emitter diameter of 4μm. Matched emitters in this process should have diameters of 8 to 40μm. Emitter areas at the lower end of this range suffice for minimal matching. Moderate and precise matching generally require the use of larger emitters, but the presence of power devices may justify the use of smaller emitters to produce a compact structure that is less susceptible to thermal gradients.
3. Maximize the emitter area-to-periphery ratio.
   For a given emitter area, the transistor with the largest area-to-periphery ratio produces the best possible matching. Circular geometries provide the highest area-to-periphery ratios, but octagonal and square emitters are almost as good.
4. Place matched transistors in close proximity.
   Bipolar transistors are very sensitive to thermal gradients. Even minimally matched transistors should reside within a few hundred microns of one another. Moderately or precisely matched transistors should use common-centroid layout techniques to minimize the separation between the transistors.
5. Keep the layout of matched transistors as compact as possible.
   The use of common base and collector regions may cause slight mismatches, but the increase in compactness usually more than compensates. Layouts that arrange the emitters in tight clusters generally provide better matching than layouts that place them in a line. A pair of matched transistors of equal sizes should employ a cross-coupled layout.
6. Construct ratioed pairs and quads using even integer ratios between 4:1 and 16:1.
   Ratios that are too small or too large will match less well than those that lie within a certain range, typically between 4:1 and 16:1 for ratioed pairs and

between 4:1:1:4 and 8:1:1:8 for ratioed quads. The ratios used for quads tend to be smaller than those used for pairs, because quads develop larger VPTAT voltages for a given number of unit emitters.

7. Place matched transistors far away from power devices.
Power devices represent a significant threat to bipolar transistor matching. Minimally matched transistors should lie at least 10mils (250μm) away from major power devices (those dissipating 250mW or more) and should not reside adjacent to any power device dissipating more than 50mW. Moderately matched devices should lie at least 5 to 10mils (125 to 250μm) from any device dissipating more than 50mW and should be placed at the opposite end of the die from major power devices. Precisely matched devices should be separated as far as possible from any power device. Consider elongating the die to a 1.5:1 or even a 2:1 aspect ratio to increase the separation between precisely matched transistors and major power devices. Power dissipations of a watt or more generally preclude precise matching unless the transistors are heavily degenerated.
8. Place matched transistors in low-stress areas.
The presence of any significant heat source on the die precludes placing the matched transistors in the center, because they would lie too close to the heat source. In this case, moderately matched transistors should occupy the middle of the opposite end of the die from the heat source. Moderately matched transistors should not reside within about 10mils (250μm) of an edge of the die because stress levels increase near edges. Similarly, moderately matched transistors should be kept well away from the corners of the die where the stresses are greatest. Precise matching is very difficult to maintain in the presence of large thermal gradients.
9. Place moderately or precisely matched transistors on axes of symmetry of the die.
Moderately or precisely matched transistor arrays should be oriented so that their major axis of symmetry, S_1, coincides with one of the axes of symmetry of the die. If possible, matched arrays should reside around the <211> axis of a (111) die rather than the <110> axis.
10. Do not allow the NBL shadow to intersect matched emitters.
The NBL region of a moderately or precisely matched transistor should overlap its emitters by a distance sufficient to ensure that it does not intersect them. If the direction of NBL shift is unknown, allow adequate overlap of NBL on all sides of the emitter. If the magnitude of the shift is unknown, then overlap NBL over the emitter by at least 150% of the maximum epi thickness. Minimally matched transistors can forego this precaution because the impact of the NBL shadow is relatively small.
11. Place emitters far enough apart to avoid interactions.
If multiple emitters must occupy a common base region, then space them far enough apart to prevent their depletion regions from intersecting. If the layout rules specify the spacing between unconnected emitters, use this rule for matched emitters regardless of how they are connected. If no such rule exists. then the spacing between the matched emitters should exceed the minimum spacing by 2 to 3μm.
12. Increase the base overlap of moderately or precisely matched emitters.
If the base barely overlaps the emitter, misalignment can cause the lateral beta of a portion of the emitter periphery to increase enough to produce minor mis-

matches. The base region of a moderately or precisely matched transistor should overlap its emitter by 1 to 2μm more than minimum.

13. Operate matched transistors on the flat portion of the beta curve.
The β-*vs.*-I_C plot of an NPN transistor usually exhibits a plateau across a relatively broad range of currents. Whenever possible, matched transistors should operate on this plateau. Transistors operating at higher currents will suffer from high-level injection that may induce mismatches. In ratioed pairs and quads, high-level injection can cause deviations from the theoretical VPTAT voltages. Most NPN transistors do not enter high-level injection except at relatively high currents that can also cause undesirable self-heating. Transistors operated at a current level that is too low are prone to variations in beta due to surface effects. The low-current beta rolloff of most vertical NPN transistors occurs only at extremely low current densities, so this effect rarely interferes with device matching.

14. The contact geometry should match the emitter geometry.
A circular emitter should contain a concentric circular contact. Similarly, an octagonal emitter should contain an octagonal contact and a square emitter should contain a square contact. If the process allows only minimum-size square contacts, then use square emitters and square arrays of minimum contacts. The emitter contacts should fill as much of the emitter area as possible, except in cases where silicidation must be minimized to prevent beta degradation. These precautions help prevent interactions between the contact and the edge of the emitter from distorting the flow of emitter current.

15. Consider using emitter degeneration.
Minimally matched transistors will not normally benefit from emitter degeneration. Moderately matched transistors may benefit from degeneration in the presence of large thermal gradients. Precisely matched transistors are often degenerated, if for no other reason than to allow adjustment of their offset voltage by trimming the degeneration resistors. The degenerating resistors should develop at least 50mV for moderate matching and 100mV for precise matching. Emitter degeneration can also be used to match transistors with different emitter sizes or geometries. In this case, at least 200mV of degeneration should be employed for minimal matching and 500mV for moderate matching. This technique can achieve noninteger ratios between matched transistors. For example, a 1.64:1 ratio can be constructed from two transistors having equal emitter areas and a pair of emitter degeneration resistors with a ratio of 1:1.64.

9.3.2. Rules for Matching Lateral PNP Transistors

Lateral PNP transistors generally do not match as well as vertical NPN transistors. Their poorer matching is due partly to surface effects and partly to an inability to use large emitters. Emitter degeneration is frequently used to improve the matching of PNP current mirrors and whatever other circuits can tolerate its presence. The following rules summarize the principles of designing matched lateral PNP transistors:

1. Use identical emitter and collector geometries.
Both the emitter and the collector geometries affect conduction in lateral PNP transistors. Transistors with different emitter or collector geometries match very poorly. For minimal matching, only the size and shape of the inner periphery of the collector facing the emitter matters. For higher degrees of precision the entire collector geometry should be duplicated. The shape and size of the

tank (or well) are unimportant as long as none of the transistors saturate. If a transistor can saturate, it is safest to place it in its own tank. P-bar or N-bar isolation schemes (Section 4.4.2) should not be counted on to ensure complete isolation between matched devices. Shallow-collector transistors (such as analog BiCMOS devices constructed from PSD implants) should be placed in separate tanks or wells to minimize cross-injection caused by carriers passing underneath the shallow collectors.

2. Use minimum-size emitters for matched transistors.
   Larger emitters will degrade the beta of the transistor, and this effect probably hurts matching more than the increased area helps. Ratioed transistors should employ multiple copies of a minimum-emitter cell (Figure 8.26B).
3. Field-plate the base region of matched lateral PNP transistors.
   Field-plating ensures that electrostatic charges do not interfere with the flow of current across the neutral base. Improperly field-plated transistors are susceptible to long-term drifts that can play havoc with matching. Lateral PNP transistors constructed in analog BiCMOS processes that incorporate a channel stop implant across the neutral base do not require field-plating, because the channel stop performs this function. Still, the addition of field plates never hurts.
4. Split-collector lateral PNP transistors can achieve moderate matching.
   Moderate matching can be achieved only as long as all of the split collectors are identical copies of one another, and none of the collectors saturates. The presence of gaps between the collectors makes it impossible to accurately predict the division of current between split collectors of different sizes. The saturation of any split collector destroys the matching between the remaining split collectors. Split-collector laterals can be used to form very compact cross-coupled transistors that exhibit surprisingly precise matching.[21]
5. Place matched transistors in close proximity.
   Even minimally matched lateral PNP transistors should reside near one another to minimize the impact of thermal gradients. Moderately or precisely matched transistors may benefit from placement in a common base tank. If this is done, make sure that none of the transistors in the tank can saturate.
6. If possible, avoid constructing VPTAT circuits from ratioed lateral PNP transistors.
   An ideality factor ignored in the derivation of equations 9.8 and 9.9 becomes significant in high-level injection, where lateral PNP transistors usually operate. The VPTAT voltages developed by ratioed mirrors and quads often exhibit significant deviations from the values predicted by the equations due to the contribution of the ideality factor.
7. Place matched transistors far away from power devices.
   Minimally matched transistors should reside at least 10mils (250μm) away from major power devices and should not be placed adjacent to any device dissipating more than 50mW. Moderately matched devices should reside at least 5 to 10mils (125 to 250μm) away from any device dissipating more than 50mW, and they should be placed at the opposite end of the die from major power de-

[21] Gilbert reports ±0.1% typical matching from cross-coupled split-collector lateral PNP transistors: this is presumably a one-sigma value. See B. Gilbert, "Bipolar Current Mirrors," in C. Toumazou, F. J. Lidgy, and D. G. Haigh, *Analog IC Design: The Current-Mode Approach* (London: Peter Perigrinus, 1990), pp. 249–250.

vices. Precisely matched devices should be separated as far as possible from any power device. Devices that dissipate a watt or more generally preclude precise matching unless the matched transistors are heavily degenerated. Consider elongating the die to a 1.5:1 or even a 2:1 aspect ratio to increase the separation between precisely matched transistors and major power devices.

8. Place matched transistors in low-stress areas.
Precisely matched transistors should occupy the center of the die, but the presence of any significant heat source generally precludes placing the matched transistors in the center of the die. Moderately matched transistors should instead occupy the middle of the end of the die opposite the heat source. They should not reside within about 10mils (250μm) of an edge of the die, and they should be kept well away from the corners of the die.
9. Place moderately or precisely matched transistors on axes of symmetry of the die.
Moderately or precisely matched transistor arrays should be oriented so that their major axis of symmetry, S_1, coincides with one of the axes of symmetry of the die. If possible, matched arrays should be placed on the <211> axis of a (111)-oriented die.
10. Do not allow the NBL shadow to intersect the base region of a lateral PNP.
The presence of the surface discontinuity that causes the NBL shadow distorts the flow of current across the neutral base of the transistor. If the direction of NBL shift is unknown, allow adequate overlap of NBL on all sides of the base region. If the magnitude of the shift is unknown, then overlap NBL over the base region by at least 150% of the maximum epi thickness. The NBL shadow will have little or no effect on matching if it merely intersects the collector of the transistor.
11. Operate matched lateral PNP transistors near peak beta.
The β *vs.* I_C of a lateral PNP transistor usually exhibits a pronounced peak. Matched transistors should operate at or near this peak in order to minimize base current errors. Operating the transistor at either lower or higher current densities causes the beta to roll off and increases base current errors. Also, the nonidealities mentioned in rule 6 become increasingly important away from the point of maximum beta.
12. The contact geometry should match the emitter geometry.
A circular emitter should contain a concentric circular contact. Similarly, an octagonal emitter should contain an octagonal contact and a square emitter should contain a square contact. These precautions help prevent interactions between the contact and the edge of the emitter from distorting the flow of emitter current.
13. Consider using emitter degeneration.
Lateral PNP transistors usually benefit more from emitter degeneration than do vertical NPN transistors because of their lower Early voltages and the inadvisability of increasing their emitter areas. The degenerating resistors should develop at least 50mV for moderate matching and 100mV for precise matching. Emitter degeneration can also be used to match transistors with different emitter sizes or geometries. In this case, 200mV of degeneration should be employed for minimal matching and 500mV for moderate matching. This technique can also achieve noninteger ratios between matched transistors. Split collectors cannot be degenerated relative to one another because they share a common emitter.

9.4 SUMMARY

Bipolar power transistors are substantially more difficult to design than MOS power transistors. The negative temperature coefficient of V_{BE} makes bipolar transistors susceptible to thermal runaway, and current focusing during turnoff can destroy an otherwise robust transistor through secondary breakdown. Proper design can minimize these vulnerabilities. Bipolar transistors offer several unique advantages over MOS transistors: their high transconductance does not depend on large device areas or small channel lengths, and they also exhibit superior transient power handling capability due to the larger volume of silicon available to dissipate heat. Bipolar transistors perform exceptionally well as MOS gate drivers and as ESD protection devices (Section 13.4.3). Large lateral PNP transistors are incredibly robust. The large die area required to construct such a device spreads the heat dissipation over a corresponding volume of silicon, and the pronounced beta rolloff of the lateral PNP makes it almost impossible to destroy the transistor by excessive current conduction.

Bipolar transistors also exhibit better voltage matching characteristics than MOS transistors. The high transconductance of the bipolar transistor allows a single stage to generate higher gains and thus minimizes the number of matching transistors required. The emitter area of a vertical NPN transistor can be increased without impairing transconductance, while increasing the channel length of a MOS transistor rapidly decreases its transconductance and increases the required area. Matched MOS transistors almost always consume more area than matched bipolar transistors of similar precision. Properly ratioed bipolar transistors develop extremely accurate VPTAT voltages that form the basis of many voltage and current references. MOS transistors only develop VPTAT voltages when operated in subthreshold, a mode incompatible with high-temperature operation.

Although MOS transistors have supplanted their bipolar counterparts in many applications, bipolar transistors still have their advantages. The future of analog design appears to lie with analog BiCMOS processes that merge high-density CMOS with high-performance bipolar. These processes can provide high-density submicron CMOS logic in combination with precision analog functions currently achievable only through the use of bipolar transistors.

9.5 EXERCISES

Refer to Appendix C for layout rules and process specifications. For all power transistors, assume a minimum beta at full rated current of ten. Do not exceed an emitter current density of 5mA/mil^2 (8μA/μm^2) for linear-mode devices and 10mA/mil^2 (16μA/μm^2) for switched-mode devices.

9.1. What is the maximum current that can flow through a 100μm-long emitter finger with a metallization width of 12μm? Assume the metallization consists of 10kÅ of aluminum/copper/silicon alloy and that emitter debiasing must not exceed 5mV.

9.2. Construct an interdigitated-emitter power transistor using standard bipolar layout rules. The transistor is intended as a 500mA series-pass transistor for a linear regulator. Construct the transistor around a central spine of deep-N+ 20μm wide, and place banks of emitter fingers on both sides of this spine. Make the emitter fingers 20μm wide. Do not allow intrafinger debiasing to exceed 5mV, or base debiasing to exceed 3mV. Use emitter ballasting resistors that develop 50mV at full rated current. Include all necessary metallization.

9.3 Construct a wide-emitter narrow-contact power transistor for a lamp driver using standard bipolar layout rules. This switching transistor must handle 150mA of collector current and should contain as much deep-N+ as possible. Make all deep-N+ sinkers 16μm wide and all emitter fingers 24μm wide. Assume only one end of the base serpentine can be connected. Do not allow the sum of base metallization debiasing and emitter metallization debiasing to exceed 10mV.

9.4 Construct a cruciform-emitter power transistor for a relay driver using standard bipolar layout rules. This switching transistor must handle 700mA of collector current and should be ringed with a deep-N+ sinker 20μm wide. Maximize connection to the collector. Make the cruciform emitter sections 75μm wide and contact them with 10μm-diameter circular emitter contacts. Assume only one end of the base lead can be connected. Do not allow the sum of base metallization debiasing and emitter metallization debiasing to exceed 10mV.

9.5 Construct a wide-emitter narrow-contact transistor using analog BiCMOS layout rules. This gate driver transistor must conduct 500mA pulses. Assume that the transistor operates at a peak emitter current density of 75mA/mil^2. Use an emitter overlap of emitter contact of 8μm and completely ring the transistor with a deep-N+ sinker no less than 10μm wide. Maximize metallization of both emitter and collector. Since the layout rules do not allow strip contacts, use rows of minimum-width contacts instead. Make the narrow contact for the emitter out of two rows of minimum contacts. Include all necessary metallization.

9.6 Construct the relay driver transistor from Exercise 9.4 using analog BiCMOS layout rules. Maximize metallization to both emitter and collector. Devise a suitable replacement for the circular emitter contacts.

9.7 Construct the circuit shown in Figure 9.15 using standard bipolar layout rules. Transistors Q_1, Q_2, and Q_3 are minimum-area lateral PNP transistors having one, two, and three emitters, respectively, and transistor Q_4 is a minimum-area substrate PNP. Resistors R_1, R_2, and R_3 consist of 6μm-wide base resistors placed in a tank connected to V_{CC}.

9.8 Construct the circuit shown in Figure 9.17C using analog BiCMOS layout rules. Transistors Q_1, Q_2, Q_3, and Q_4 should use square emitters having a width of 10μm. Include as many emitter contacts as possible. Compute the value of resistor R_1 necessary to produce a current of 10μA, and lay this resistor out using PSD-doped poly-2 6μm wide. Take all necessary precautions to obtain optimal matching.

9.9 Lay out the Brokaw bandgap cell shown in Figure 9.22A using standard bipolar layout rules. Transistors Q_1 and Q_2 should employ circular emitters with diameters of 10μm. Resistors R_1 and R_2 should be constructed as an interdigitated array of base resistors in a common tank. This tank should connect to the base of transistors Q_1 and Q_2. Include all necessary interconnection and label all devices.

9.10 Lay out the simple operational amplifier shown in Figure 9.22B using analog BiCMOS layout rules. Transistors Q_1 to Q_5 should employ 5×5μm square emitters. Transistors Q_6, Q_7, and Q_8 should use 8×8μm square emitters. Cross-couple Q_4 and Q_5. Include all necessary interconnection and label all devices. The values on this schematic represent the areas, in μm^2, of the respective emitters.

9.11 Lay out the Gilbert multiplier core shown in Figure 9.23 using analog BiCMOS layout rules. Transistors Q_1 to Q_4, Q_6, Q_7, Q_9, and Q_{10} use 8×8μm emitters. Transistors Q_5, Q_8, and Q_{11} use 6×6μm emitters. Lay out all transistors for optimal matching. Transistors sharing a common collector connection can occupy the same tank. Include all necessary interconnection and label all devices. The values on this schematic represent the areas, in μm^2, of the respective emitters.

9.12 Suppose the Gilbert multiplier core in Exercise 9.11 forms part of a die with an area of 7.6mm^2 (not including scribe streets and seals), which also includes a power NPN transistor with an area of 4.3mm^2. Select an aspect ratio for the die and place rectangles representing the outline of the die and power transistor. Place the multiplier core at an optimal location for best matching.

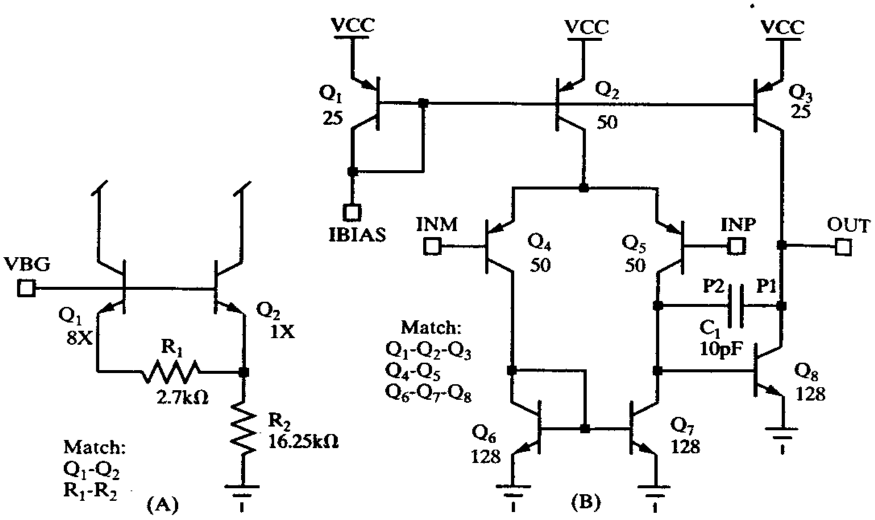


FIGURE 9.22 (A) Brokaw bandgap cell and (B) simple operational amplifier (B) for Exercises 9.9 and 9.10.

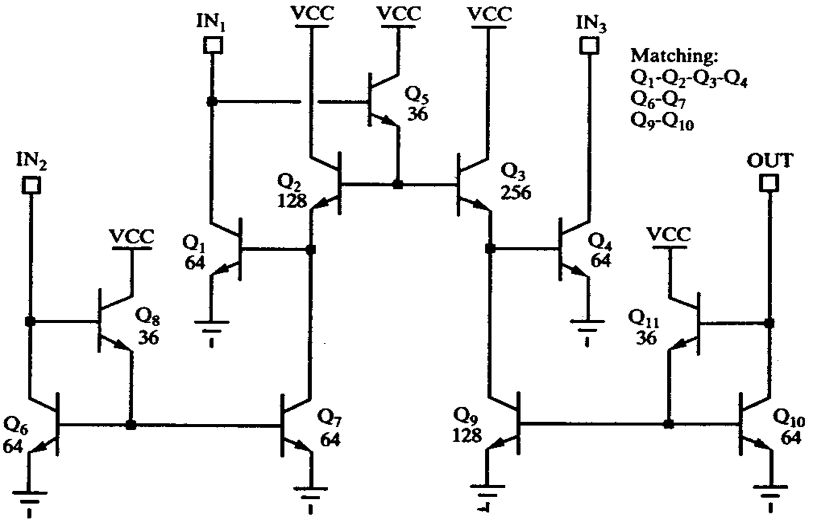


FIGURE 9.23 Gilbert multiplier core for Exercise 9.11.