8 Bipolar Transistors

The *bipolar junction transistor* (BJT) is among the most versatile of all semiconductor devices. In addition to its obvious applications as a voltage or current amplifier, it can also serve as the basis of voltage and current references, oscillators, timers, pulse shapers, amplitude limiters, nonlinear signal processors, power switches, transient protectors, and many other types of circuits. There are also certain applications for which bipolar transistors are ill-suited, the most important of these being low-power digital logic. Most logic is now constructed using *complementary metal-oxide-semiconductor* (CMOS) circuitry. Bipolar transistors remain important for constructing most analog circuits, although many of these circuits now contain CMOS elements as well.

Much of the information required to understand the operation and construction of bipolar transistors does not appear in elementary texts. This chapter opens with a review of several of these topics, including beta rolloff, avalanche breakdown, thermal runaway, and device saturation. The remainder of the chapter covers the design of small-signal bipolar transistors. This information lays the foundation for the more specialized topics covered in Chapter 9.

8.1 TOPICS IN BIPOLAR TRANSISTOR OPERATION

Figure 8.1 shows a simple model of an NPN transistor. Diode D_1 represents the base-emitter junction of the transistor. Current-controlled current source I_1 models the minority carrier current flowing across the reverse-biased base-collector junction. The current through I_1 equals the current through D_1 multiplied by the transistor's *forward active current gain* β_F. In terms of terminal currents, this relationship becomes

$$I_c = \beta_F I_B \qquad [8.1]$$

Unlike an MOS transistor, the BJT requires a constant base current to sustain the flow of collector current. This base current represents unavoidable losses due to recombination in the neutral base and carrier injection from base to emitter.

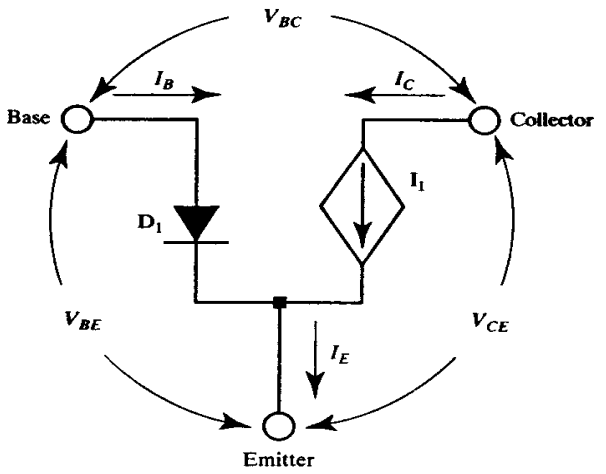


FIGURE 8.1 Simplified three-terminal model of an NPN transistor.

Because conduction requires constant base current, the bipolar transistor is often called a *current-controlled device.* This is somewhat misleading because the transistor can also be driven by a base-emitter voltage that forward-biases diode D_1 and provides base current to the transistor.

The base-emitter junction of a bipolar transistor is, for all intents and purposes, identical to the silicon junction diode discussed in Section 1.2.2. The base current of the transistor depends exponentially upon the base-emitter bias V_{BE}. If the forward beta remains constant and the collector-to-emitter bias V_{CE} suffices to maintain the transistor in the normal active region, then the collector current I_C also becomes an exponential function of V_{BE}. These relationships can be expressed by the following formulas:[1]

$$I_c = I_s e^{(V_{BE}/V_T)} \qquad [8.2]$$

$$V_{BE} = V_T \ln\left(\frac{I_c}{I_s}\right) \qquad [8.3]$$

The *emitter saturation current* I_S depends on several factors, including the doping profiles of the base and emitter diffusions and the effective area of the base-emitter junction. The *thermal voltage* V_T scales linearly with absolute temperature, and at 298K (25°C) it equals 26mV. The base-emitter voltage V_{BE} exhibits a negative temperature coefficient[2] of about –2mV/°C. This may seem small, but since collector current depends exponentially on base-emitter voltage, an increase of only 18mV in V_{BE} doubles the collector current. A 1°C temperature difference between two bipolar transistors will cause an 8% mismatch between their collector currents, corresponding to a temperature coefficient of 80,000ppm/°C! This enormous temperature coefficient has profound implications for the design of matched devices and power transistors.

---

[1] These formulas are slightly simplified; the actual Ebers-Moll equations include a term accounting for reverse conduction: $I_C = A_e J_S [\exp(V_{BE}/kT) - 1]$. The presence of the "–1" term has no significant effect upon conduction in the forward active region at any reasonable bias level and thus has been omitted. This simplified equation also neglects the ideality factor (or emission coefficient) η, which is very near unity for NPN transistor operating at moderate currents.

[2] Most of this temperature coefficient stems from the presence of V_T in the equation for V_{BE}, but the temperature coefficient of I_S also has some impact on V_{BE}.

8.1.1. Beta Rolloff

Elementary textbooks often assume that beta remains constant, but it actually varies considerably depending on collector current. Figure 8.2 shows typical beta curves for small signal NPN and lateral PNP transistors constructed in a standard bipolar process. The NPN beta remains relatively constant over a wide range of collector currents, which to some extent justifies the assumption of constant beta. At high current levels, typically beyond 5mA/mil^2 of emitter area (8μA/μm^2), the NPN beta begins to roll off. A similar but more gradual rolloff occurs at very low current levels, usually below 10nA/mil^2 (15pA/μm^2). High-current beta rolloff is caused by high-level injection, while low-current beta rolloff results from several mechanisms, including recombination within the depletion regions and at the oxide-silicon interface, and shallow emitter effects (Section 8.3.3).

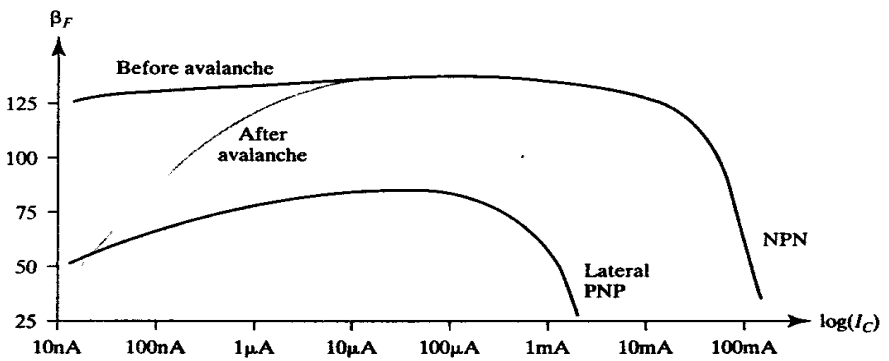


FIGURE 8.2 Beta versus collector current plots for small-signal NPN and lateral PNP transistors. The curve marked in gray shows the effect of emitter-base avalanche on NPN beta.

The beta curve of the lateral PNP differs considerably from that of the NPN. Not only does the lateral PNP have a lower peak beta, but it also exhibits more pronounced high-current and low-current beta rolloffs. Several factors account for the differences between the beta curves of the vertical NPN and the lateral PNP. The emitter of the PNP is much more lightly doped than that of the NPN, so the PNP exhibits a lower emitter injection efficiency that reduces its peak beta. The flow of carriers near the surface of the lateral transistor increases surface recombination and exacerbates low-current beta rolloff. The lightly doped base of the lateral PNP causes high-level injection to begin at relatively low current levels and thus accentuates high-current beta rolloff. The regions of high-current beta rolloff and low-current beta rolloff often overlap, causing a pronounced peaking of the beta curve. Transistors that exhibit such peaking are actually operating in high-level injection at the point of peak beta, complicating the design of certain types of circuits.

8.1.2. Avalanche Breakdown

The maximum operating voltages of a bipolar transistor are determined by the breakdown voltages of the base-emitter and base-collector junctions. Depending upon biasing conditions, several different breakdown voltages may be observed. The three most important of these are denoted V_{EBO}, V_{CBO}, and V_{CEO}. Each of these is measured between two of the transistor terminals with the third terminal left unconnected (*open*).

[3] I. Getreu. *Modeling the Bipolar Transistor* (Beaverton, Oregon: Tektronix, 1976), p. 48ff.

The breakdown voltage of the transistor base-to-emitter with collector held open is represented by V_{EBO}. For an NPN transistor constructed on a standard bipolar process, V_{EBO} equals approximately 7V. This breakdown voltage remains relatively constant over process and temperature, so NPN transistors biased into emitter-base avalanche form useful Zener diodes. Base-emitter avalanche rapidly degrades the beta of NPN transistors because hot carriers generated by the avalanche process induce the formation of recombination centers along the oxide-silicon interface.[4] These recombination centers increase the recombination current within the depletion regions. This current adds to the base current of the transistor and drastically decreases the low-current beta of the device (Figure 8.2). Although high-current betas are not affected to the same degree, one should avoid avalanching any NPN transistor not specifically intended to operate as a Zener diode.

The breakdown of the transistor collector-to-base with emitter left open is represented by V_{CBO}. This breakdown voltage depends on several factors relating to the base-collector junction, all of which are discussed in greater detail in Section 8.2. Both the base and the collector of most bipolar transistors are lightly doped, so V_{CBO} is usually quite large. For NPN transistors constructed on standard bipolar processes, it can range from 20V to 120V or more. Since collector-base breakdown is largely subsurface, it does not generate surface recombination centers and therefore does not affect beta. Both the V_{EBO} and the V_{CBO} of the lateral PNP transistor depend on the breakdown of the base-epi junction, so the beta of lateral transistors is largely unaffected by any form of avalanche.

The breakdown of the transistor collector-to-emitter with base held open is represented by V_{CEO}. This is substantially smaller than V_{CBO} due to an effect called *beta multiplication.* Low levels of impact ionization begin to occur at voltages well below the nominal breakdown voltage. Since the base terminal of the transistor is left unconnected, any avalanche injection into the base produces a corresponding (and larger) increase in collector current. The additional carriers transiting across the base-collector junction increase impact ionization and generate additional base drive. At a voltage equal to V_{CEO}, this positive feedback mechanism becomes self-sustaining and the transistor avalanches. The V_{CEO} of an NPN transistor usually equals about 60% of its V_{CBO} (Section 8.2.4).

Beta multiplication can be suppressed by connecting the base terminal to the emitter, holding the transistor in cutoff and preventing amplification of the collector-base leakage. The breakdown voltage collector-to-emitter with base shorted to emitter is denoted V_{CES}. This breakdown voltage can approach V_{CBO} as long as the base resistance of the transistor is relatively small and the extrinsic base terminal connects to a potential equal to or less than the extrinsic emitter potential. If for any reason current begins to pass through the transistor, the breakdown voltage will suddenly decrease to nearly V_{CEO}. This phenomenon, called *snapback,* occurs even if the extrinsic base and emitter terminals are shorted, since the necessary bias develops across the internal base resistance of the transistor. Figure 8.3 shows idealized curve tracer plots illustrating this phenomenon. As soon as the trace labeled V_{CES} exceeds 60V, it immediately snaps back to 43V. As the current through the transistor increases, the avalanche voltage begins increasing again. This results partially from extrinsic collector resistance and partially from high-current beta rolloff decreasing the beta multiplication effect. The V_{CEO} of many transistors also shows a small amount of snapback due to low-current beta rolloff. For example, the transistor in Figure 8.3 registers an initial V_{CEO} breakdown of 43V and snaps back to a

[4] A. McDonald, "Avalanche Degradation of h_{FE}," *IEEE Trans. on Electron Dev.*, Vol. ED-17, #10, 1970, pp. 871–878.

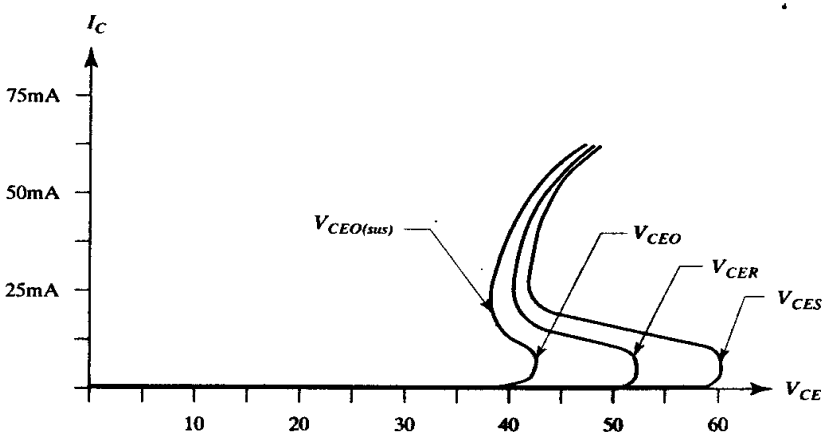


FIGURE 8.3 Idealized curve tracer plots of V_{CEO}, V_{CER}, and V_{CES} in an NPN transistor.

sustained V_{CEO} (sometimes called $V_{CEO(sus)}$) of 38V. Allowing for a small safety margin, this device would rate a V_{CEO} of approximately 36V.

The middle trace of Figure 8.3 represents the collector-to-emitter breakdown voltage of the transistor with a resistor connected between the base and emitter terminals (V_{CER}).[5] This condition lies between V_{CEO} and V_{CES} and shows the expected intermediate breakdown voltage along with a pronounced snapback.

8.1.3. Thermal Runaway and Secondary Breakdown

Bipolar transistors operating at relatively high power levels fall prey to a failure mechanism called *thermal runaway*. To illustrate how thermal runaway occurs, suppose that a large bipolar transistor suddenly begins to dissipate significant power. The center of the transistor quickly becomes warmer than its outer edges, causing the V_{BE} of the center to drop slightly. Due to the exponential character of the current-voltage relationship, a small change in base-emitter voltage produces a large change in collector current. Increased power dissipation occurs in the hotter portions of the transistor, leading to a further decline in V_{BE}. The region of the transistor that conducts current steadily shrinks as it grows hotter, until practically all of the current funnels through a very small, very hot area called a *hot spot*.

If the temperature in the hot spot reaches some 350 to 450°C, junction leakages become so large that the transistor essentially shorts out. Catastrophic failure occurs either when metallization is drawn through the contacts (as in a Zener zap structure) or when the silicon melts, cracks, or vaporizes. This type of self-destructive runaway does not always occur. Sometimes the increased current density causes beta to roll off far enough to stabilize the hot spot at a high, but not immediately destructive, temperature.[6] The presence of a "stable" hot spot dangerously overstresses a transistor and renders it vulnerable to electromigration, thermally accelerated corrosion, and various other long-term failure mechanisms.

A transistor that contains a stable hot spot often self-destructs during turn-off. Failure occurs due to an apparent avalanche of the collector-base junction, often at a voltage substantially below the rated V_{CEO} of the transistor. This unexpected

---

[5] V_{CER} has also been used to refer to the collector-to-emitter breakdown voltage with a reverse bias applied to the base-emitter junction. This mode of biasing is sometimes used in power circuits to raise the V_{CE} breakdown of the transistor above V_{CES}.

[6] P. L. Hower, D. L. Blackburn, F. F. Oettinger, and S. Rubin, "Stable Hot Spots and Second Breakdown in Power Transistors," *National Bureau of Standards*, PB-259 746, Oct. 1976.

reduction in avalanche voltage is called *secondary breakdown.*[7] It is a consequence of extremely high current densities within the transistor, due in this case to the presence of a stable hot spot. The velocity of carriers in the lightly doped collector increases in order to support the increasing current flow through this zone. Eventually the carrier velocity reaches its maximum limit (the *carrier saturation velocity*). Once this occurs, the electric field across the neutral collector intensifies, and the avalanche voltage of the transistor snaps back to a lower value, V_{CEO2}. If the voltage across the collector exceeds V_{CEO2}, the transistor avalanches and the resulting power dissipation quickly destroys the device.[8]

Secondary breakdown can also occur in transistors that have not experienced thermal runaway. During turnoff, the base lead withdraws charge from the neutral base. Charge removal begins in the portion of the base adjacent to the emitter periphery, and progresses inward toward the center of the transistor. As turnoff proceeds, conduction in the transistor collapses into an ever-shrinking area. This *emitter current focusing* causes the momentary appearance of extremely high current densities in small portions of the transistor. These current densities may become large enough to trigger secondary breakdown, especially if the transistor is conducting a large current at the time it is turned off.[9,10]

Thermal runaway and secondary breakdown can be avoided by restricting the operating conditions of the transistor. Figure 8.4 shows a graph illustrating the *forward-bias safe operating area* (FBSOA) of a typical bipolar transistor. The safe operating region is bounded by four separate curves.[11] The horizontal line represents the maximum current that the metallization and bondwires can safely carry without eventual electromigration failure. The vertical line represents the maximum voltage that can be placed across the transistor without fear of avalanche (usually assumed to equal $V_{CEO(sus)}$). A curve passing diagonally across the plot represents the maximum power that the device can dissipate without producing excessive temperatures

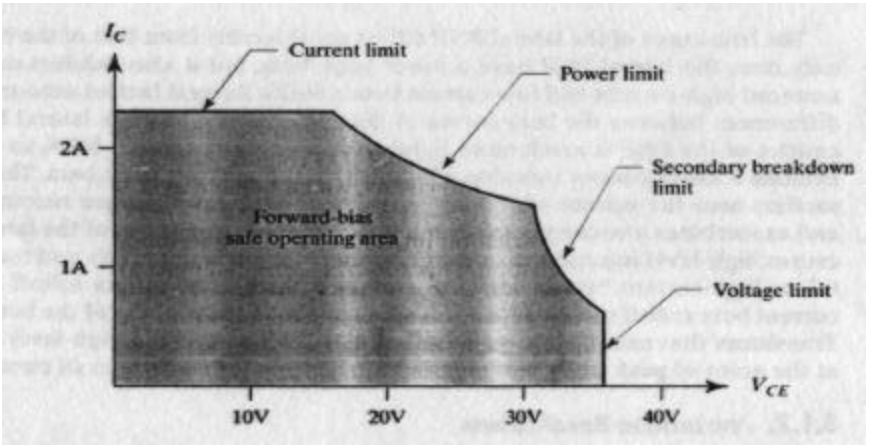


FIGURE 8.4 Forward-bias safe operating area (FBSOA) plot of a typical NPN power transistor.

[7] B. A. Beatty, S. Krishna, and M. S. Adler, "Second Breakdown in Power Transistors Due to Avalanche Injection," *Trans. on Electron Dev.*, Vol ED-23, #8, 1976, pp. 851–857.

[8] I. G. Kassakian, M. F. Schlect, and G. C. Verghese, *Principles of Power Electronics* (Reading, MA: Addison-Wesley, 1992), pp. 522–525.

[9] Kassakian, *et al.*, pp. 521–522.

[10] P. L. Hower and W. G. Einthoven, "Emitter Current-Crowding in High-Voltage Transistors," *IEEE Trans. on Electron Devices*, Vol. ED-25, #4, 1978, pp. 465–471.

[11] F. F. Oettinger, D. L. Blackburn, and S. Rubin, "Thermal Characterization of Power Transistors," *IEEE Trans. on Electron Devices*, Vol. ED-23, #8, 1976, pp. 831–838.

within the package. A fourth and final curve clips off the portion of the safe operating area where secondary breakdown may occur. A very robust transistor may not exhibit any FBSOA reduction due to secondary breakdown. A properly heat-sunk but poorly designed power transistor may lose a substantial fraction of its potential FBSOA to secondary breakdown. Section 9.1.2 discusses ways to increase the safe operating area of bipolar transistors.

8.1.4. Saturation in NPN Transistors

An NPN transistor enters saturation when both its base-emitter and base-collector junctions simultaneously forward-bias. Power transistors are often intentionally operated in saturation to reduce the collector-to-emitter saturation voltage $V_{CE(sat)}$ and to minimize power dissipation. Unfortunately, saturation also produces a whole host of problems. The unintentional saturation of bipolar transistors has probably caused more circuit malfunctions than any other device-related design flaw.

Saturation affects discrete and integrated transistors in different ways. The saturation of a discrete transistor merely prolongs its *turnoff time* (also called its *reverse recovery time*). As soon as the base-collector junction begins to forward-bias, minority carriers flow across it in both directions. In an NPN, holes flow into the collector and electrons into the base. A substantial population of excess minority carriers soon accumulates on both sides of the collector-base junction. Now suppose that the external circuit reduces the external base-emitter bias to zero. The transistor does not turn off immediately because the resistance of the neutral base impedes the withdrawal of the stored charge. The transistor continues to conduct until this charge is withdrawn or until it recombines. Saturation therefore increases the turnoff time by at least an order of magnitude. Fast bipolar logic usually incorporates anti-saturation clamps (as in LSTTL logic) or uses circuit topologies that are immune to saturation (as in ECL and DCML logic).

Saturation increases the reverse recovery time of an integrated bipolar transistor, but it also has other deleterious effects due to the presence of an additional junction between the collector and the isolation. Figure 8.5 shows a cross section of a typical NPN transistor fabricated on a standard bipolar process. The arguments presented for this structure apply equally to any other junction-isolated, vertical NPN transistor, including the CDI NPN of analog BiCMOS (Figure 3.48). These arguments do not apply to fully oxide-isolated transistors, which behave as if they were discrete devices.

The presence of junction isolation introduces a fourth terminal consisting of the P– substrate and the P+ isolation. This *substrate* terminal must always be biased to

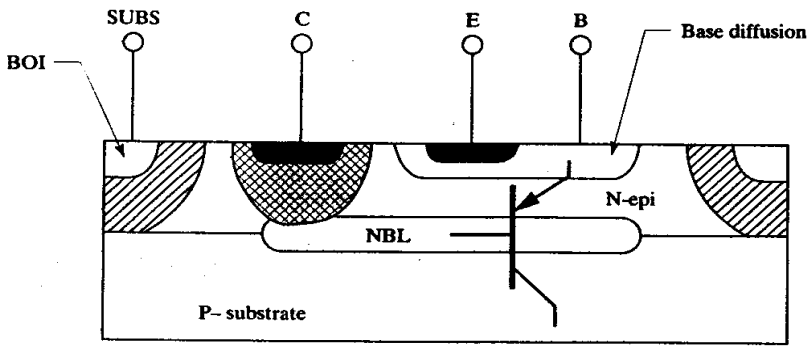


FIGURE 8.5 Cross section of an NPN transistor fabricated on a standard bipolar process, showing the parasitic PNP transistor.

a lower voltage than the collector terminal to avoid forward biasing the isolation junction. This reverse-biased junction isolates the transistor from the remainder of the integrated circuit as long as there are few minority carriers in the neutral collector. Junction isolation fails as soon as the base-collector junction begins to forward-bias. Many of the holes injected across the base-collector junction eventually diffuse across the collector to the collector-substrate junction. The electric field across this reverse-biased junction draws these holes into the substrate, where they become majority carriers.

Saturation can also be explained by imagining a PNP transistor superimposed upon the cross section of the NPN transistor (Figure 8.5). The collector-base junction of the vertical NPN also forms the base-emitter junction of this *parasitic PNP*. When the NPN saturates and forward-biases its collector-base junction, the parasitic PNP transistor turns on and diverts excess base drive into the substrate. This unexpected diversion of base drive has several unpleasant consequences. For one, an integrated NPN transistor cannot be driven as deeply into saturation as a discrete transistor can because the parasitic PNP steals its base drive as soon as it begins to saturate. Integrated NPN transistors therefore tend to have higher $V_{CE(sat)}$ voltages than equivalently constructed discrete transistors.

A saturating NPN also represents an unexpected source of substrate current that can potentially lead to substrate debiasing (Section 4.4.1). If the base drive to the transistor exceeds a few milliamps (as is often the case for power transistors), then special provisions may become necessary to prevent substrate debiasing. Substrate debiasing can be averted by adding guard rings to prevent the holes in the collector from reaching the substrate, or by designing the base-drive circuit to reduce the base-drive once the transistor begins to saturate (Section 9.1.3).

Saturation also causes a failure mechanism called *current hogging*. When a transistor saturates, some of its base current flows through the base-collector junction rather than the base-emitter junction. This diversion of current reduces the base-emitter voltage.[12] Many circuits connect the base-emitter junctions of several transistors in parallel and expect the collector currents of these devices to track the areas of the respective base-emitter junctions. This relationship ceases to apply if one of the transistors saturates because this transistor experiences a drop in V_{BE} relative to the remaining transistors. The base current of the saturating transistor therefore increases at the expense of the other transistors. In more colloquial terms, the saturating transistor *hogs* the base drive.

Figure 8.6 shows a simple current mirror constructed from three NPN transistors. Current source I_{BIAS} feeds diode-connected transistor Q_1. All three transistors see the same base-emitter voltage and therefore draw the same collector currents, *providing that all three transistors operate in the normal active region*. To illustrate why this is the case, suppose transistor Q_3 saturates. Much of its base current diverts into parasitic PNP transistor Q_P, causing the base-emitter voltage of Q_3 to decrease. Eventually Q_3's V_{BE} drops to a point just sufficient to satisfy the *intrinsic collector current* I_3, restoring equilibrium. In this case, the intrinsic collector current equals the sum of the extrinsic collector current, I_{t3}, and the base current of Q_P. Transistor Q_2 sees the same base-emitter bias as Q_3, so its collector current, I_2, equals the intrinsic collector current, I_3. In summary, when one transistor in a mirror saturates, the extrinsic collector currents of all the other transistors decrease to equal the

[12] A vertical transistor is particularly susceptible to current hogging, because its heavily doped emitter region raises the base-emitter voltage at a given bias level a few tens of millivolts above the corresponding base-collector voltage, so current flows preferentially through the base-collector junction.

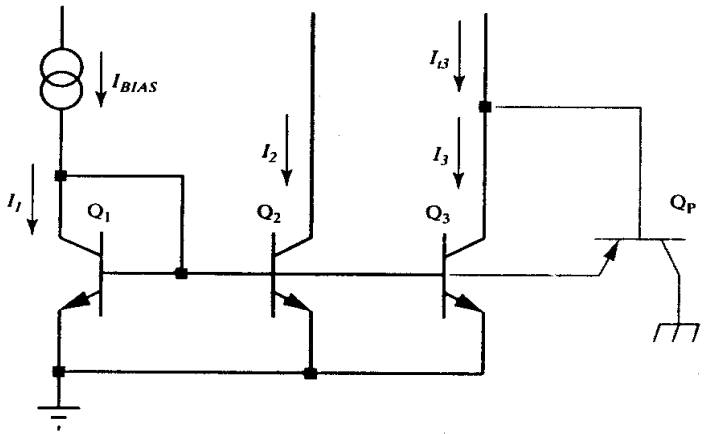


FIGURE 8.6 An example of a circuit that exhibits current hogging. Q_P represents the parasitic PNP transistor present in the structure of vertical NPN Q_3.

intrinsic collector current of the saturated transistor. This disturbs the balance of the circuit and often leads to serious malfunctions.

Circuit designers have developed several cures for current hogging, including base-side ballasting and Schottky clamps. *Base-side ballasting* requires the insertion of matched resistors into the base leads of each transistor (Figure 8.7).

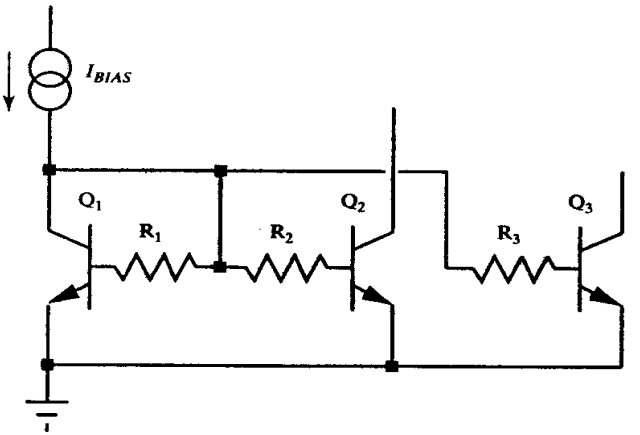


FIGURE 8.7 Base-side ballasting applied to the circuit shown in Figure 8.6.

The base ballasting resistors must ratio inversely to the emitter areas of their respective transistors. For example, if Q_2 has twice the emitter area of Q_1, then R_2 must equal half the resistance of R_1. The base ballasting resistors must match in order to maintain collector current matching. The base-side ballasting prevents current hogging by introducing localized negative feedback. If any one of the transistors begins to saturate, its base current will increase slightly. This causes a corresponding increase in the voltage drop across its base ballasting resistor, forcing a drop in the V_{BE} of this transistor. Typically, no more than 50 to 100mV appears across the base ballasting resistor of a saturating transistor, producing a correspondingly small disturbance in the base-bias currents.

A diffused base ballasting resistor must not occupy the tank of the NPN transistor it protects because then the ballasting resistor can forward-bias into the tank

(Figure 8.8). In effect, the parasitic PNP transistor simply moves from one point in the structure to another.

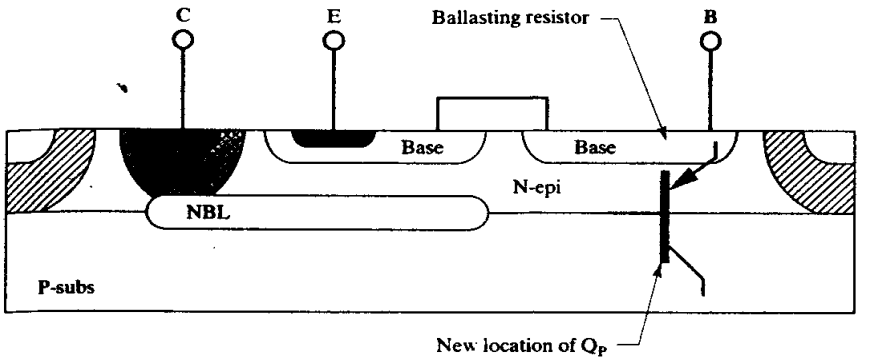


FIGURE 8.8 A base-side ballasting resistor becomes ineffective when merged into the same tank as the NPN it protects.

A clamping diode connected across the base-collector junction of a transistor will prevent it from entering saturation. In order for the clamping diode to function properly, it must have a lower forward voltage than the base-collector junction. Only a few types of Schottky diodes, most notably those constructed using platinum or palladium silicides, have the necessary characteristics. Figure 8.9A shows the connection of a Schottky clamp diode in parallel with the base-collector junction of an NPN transistor. The resulting *Schottky-clamped NPN* is often represented by the symbol of Figure 8.9B. The Schottky clamp works by providing an alternate path for current that would otherwise flow through the forward-biased base-collector junction. Because the diode prevents the base-collector junction from conducting, a Schottky-clamped NPN neither injects appreciable minority carriers nor experiences the prolonged turn-off times characteristic of saturation. The Schottky clamp does not prevent the base current from increasing, but it does prevent it from exceeding the collector current the transistor would normally conduct. Schottky-clamped transistors are used extensively in switching circuitry and bipolar logic to eliminate saturation-induced propagation delays. Section 10.1.3 discusses the layout of Schottky-clamped NPN transistors in further detail.

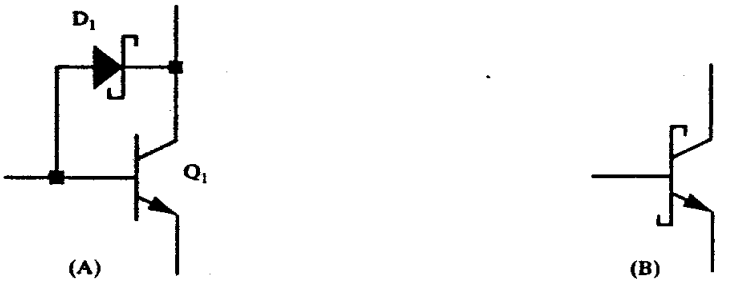


FIGURE 8.9 (A) Schottky-clamped NPN transistor and (B) its conventional schematic symbol.

Saturating NPN transistors can also cause problems when merged with other devices. When a transistor saturates, its forward-biased base-collector junction injects minority carriers into its collector. These carriers can be collected by the reverse-biased junctions of other devices merged into the same tank. Sneak currents

associated with minority conduction may cause circuit malfunctions or even catastrophic latchup. Section 13.2.1 discusses these difficulties in greater detail.

8.1.5. Saturation in Lateral PNP Transistors

Figure 8.10 shows a cross section of a lateral PNP transistor constructed in a standard bipolar process. An epi tank forms the base of the device while a small plug of base diffusion placed in the center of the tank acts as its emitter. The collector consists of an annular base diffusion surrounding the emitter. A substantial fraction of the total emitter-base junction area consists of sidewalls in close proximity to the surrounding collector. Most of the holes travel laterally from the sidewalls of the emitter to the sidewalls of the collector. Some holes are either injected from the bottom surface of the emitter, or diffuse away from the surface. These errant minority carriers can flow across the reverse-biased junctions isolating the transistor from the isolation and substrate. Parasitic PNP transistor Q_{P1} represents the undesired flow of holes down to the substrate, while Q_{P2} represents the flow of holes to the sidewalls of the tank. Figure 8.10 shows both of these parasitics superimposed upon the cross section of the transistor, along with the desired lateral PNP transistor Q_L.

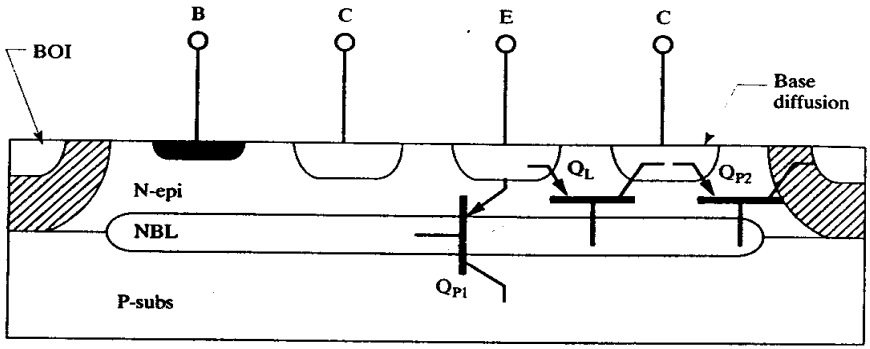


FIGURE 8.10 Cross section of a lateral PNP transistor constructed on a standard bipolar process showing parasitic substrate PNP transistors Q_{P1} and Q_{P2}.

A significant fraction of the total emitter current will be lost if measures are not taken to block the flow of holes to the substrate. The *collector efficiency* η_C of a lateral PNP equals the ratio of the current collected by lateral transistor Q_L to the sum of the collector currents of Q_L, Q_{P1}, and Q_{P2}. In terms of collector, emitter, and base terminal currents I_C, I_E, and I_B, the collector efficiency η_C equals

$$\eta_c = \frac{I_c}{I_E - I_B} \qquad [8.4]$$

For reasons explained below, the addition of NBL causes the collector efficiency of a lateral PNP to rise from less than 0.1 to near unity.[13] While lateral PNP transistors can be constructed in CMOS processes, the absence of NBL makes them extremely inefficient. Standard bipolar and analog BiCMOS processes both incorporate NBL and generally provide excellent lateral PNP transistors.

NBL improves the collection efficiency of a lateral PNP by repelling minority carriers moving toward it. This repulsion occurs because of the presence of an opposing electric field at the interface between the heavily doped NBL and the lightly doped N-epi. The presence of this electric field can be explained as fol-

[13] Analog BiCMOS processes can create lateral PNP transistors with very narrow base widths, producing collector efficiencies of 0.3 or more without using NBL.

lows: The difference in doping concentrations between N-epi and NBL causes a similar difference in majority carrier concentrations. A diffusion current consisting of electrons flows from the NBL to the N-epi. This diffusion current must be balanced by an equal and opposite drift current. In order to set up this drift current, the NBL must become slightly more positive than the N-epi. The resulting electric field attracts electrons from the N-epi back into the NBL. This same electric field also causes holes to scatter off the N+/N– interface and to rebound back into the overlying epi.[14] The scattered carriers continue to wander randomly through the epi tank. Those moving downward are again repelled from the N+/N– interface, while those moving upward eventually reach the collector-base depletion layer of the lateral PNP. Most of these carriers are eventually collected by the lateral transistor, leading to a large increase in collector efficiency. A small fraction of the downward-moving holes have sufficiently large instantaneous velocities to surmount the opposing electric field and to actually enter the NBL. Most of the holes that enter the NBL recombine within it due to the large population of available majority carriers.

A few of the holes scattered from the NBL interface move laterally through the epi until they reach the isolation sidewalls. Although this may seem a serious problem, the dimensions of the transistor actually preclude any significant loss of current through lateral parasitic conduction. The cross section in Figure 8.10 has been exaggerated vertically to create a relatively compact drawing. The lateral distance to the sidewall is roughly an order of magnitude larger than the vertical separation between the upper edge of the NBL interface and the lower edge of the collector-base depletion region. At higher collector voltages the depletion region extends down to the N+/N– interface, closing off the path for lateral parasitic conduction to the sidewalls. Even at low collector voltages, this path is so long and so narrow that few minority carriers can traverse its entire length without coming into contact with the collector-base depletion region.

Lateral parasitic conduction increases dramatically during saturation. The collector possesses a large sidewall area facing the isolation across a relatively narrow gap. This geometry forms an efficient PNP transistor that activates as soon as the collector forward-biases, as happens in saturated transistors and those operated in reverse-active mode. When a lateral PNP transistor saturates, the emitter current remains unchanged. Whatever injected holes are not collected by the lateral transistor travel to the substrate instead. Therefore lateral PNP current mirrors continue to operate properly even if one or more of their transistors in the mirror saturate. All that happens is that the unused collector current flows to the substrate. This does not present a problem as long as the collector currents do not exceed a few milliamps. If substrate injection cannot be tolerated, then a lateral PNP transistor can be fitted with a Schottky clamp. Section 9.1.3 discusses two alternate methods of limiting saturation in lateral PNP transistors.

Split-collector lateral PNP transistors are seriously affected by the saturation of any one of their several collectors. The saturating collector segment re-injects holes from all of its junction surfaces. Some of these are collected by the isolation sidewall, but significant portions of the re-injected carriers are collected by the adjacent collector segments. Consequently, the saturation of one collector in a split-collector lateral PNP causes the current drawn by the adjacent collectors to increase.

[14] NBL is often said to "reflect" minority carriers. This choice of wording is somewhat misleading because the mean free path of the carriers is usually quite short relative to the dimensions of the transistor. Therefore the "reflected" carriers do not move far before their motion is randomized through lattice interactions.

8.1.6. Parasitics of Bipolar Transistors

An integrated bipolar transistor behaves quite differently from an idealized textbook device due to the many parasitic elements it contains. Perhaps the most important of these is the PN junction which isolates the transistor from the rest of the die. If this junction forward-biases, it will inject current into the substrate. Leakage currents and capacitive coupling can still occur even if the junction remains reverse-biased. The substrate connection must therefore be counted as one of the terminals of the integrated bipolar transistor. Figure 8.11B shows one conventional method of representing an NPN transistor with an explicit substrate connection. This symbol is often called a *four-terminal* NPN to distinguish it from the *three-terminal* NPN in Figure 8.11A. PNP transistors can also have both three-terminal and four-terminal symbols (Figure 8.11C and D). Although the three-terminal symbols do not show a substrate connection, one still exists as long as the process employs junction isolation. This implicit substrate connection can cause considerable trouble if the designer forgets its presence and assumes that the transistor is truly isolated from the remainder of the die. Many designers routinely use four-terminal symbols rather than three-terminal ones to avoid problems of this sort.

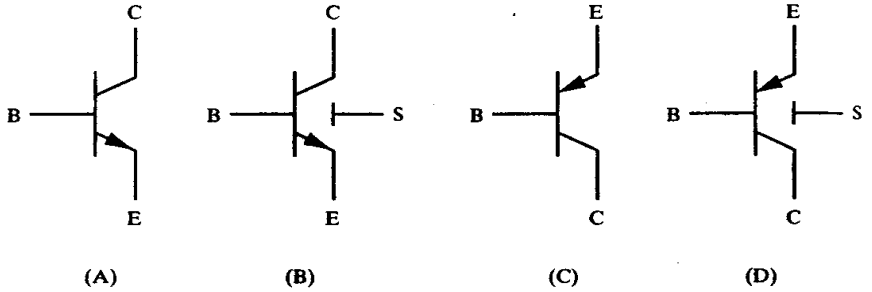


FIGURE 8.11 Symbols for three-terminal and four-terminal transistors (E: emitter, B: base, C: collector, S: substrate).

A complete model of the parasitics of a bipolar transistor includes a number of distributed effects, the discussion of which lies beyond the scope of this text. Figure 8.12 shows simplified parasitic models for a vertical NPN and a lateral PNP formed in a standard bipolar process. These same models also apply to most other junction-isolated transistors, including the vertical NPN and lateral PNP of analog BiCMOS.

The vertical NPN transistor model contains an ideal three-terminal NPN transistor Q_1 that models the intended functionality of the device. Transistor Q_2 represents the parasitic substrate PNP transistor that forward-biases when the vertical NPN transistor saturates (Section 8.1.4). Zener diodes D_{BE}, D_{BC}, and D_{CS} are not intended to model the behavior of the forward-biased junctions, as these are already subsumed into the behavior of transistors Q_1 and Q_2. The Zener diodes instead model the avalanche breakdown, leakage, and capacitance associated with the respective junctions. For example, diode D_{BE} models the base-emitter capacitance C_{BE} of the transistor as well as the base-emitter breakdown voltage V_{EBO}. Diode D_{BC} models the base-collector capacitance C_{BC} as well as the base-collector breakdown voltage V_{CBO}. Diode D_{CS} models the collector-substrate capacitance C_{CS} and the collector-substrate breakdown voltage. The collector-base capacitance C_{BC} and the collector-substrate capacitance C_{CS} are of concern because they limit the operating frequency of the transistor. The transistor switches faster if these junctions are made smaller. The avalanche voltages of the diodes D_{BE}, D_{BC}, and D_{CS} also set the operating voltage of the transistor, as discussed in Section 8.1.2.

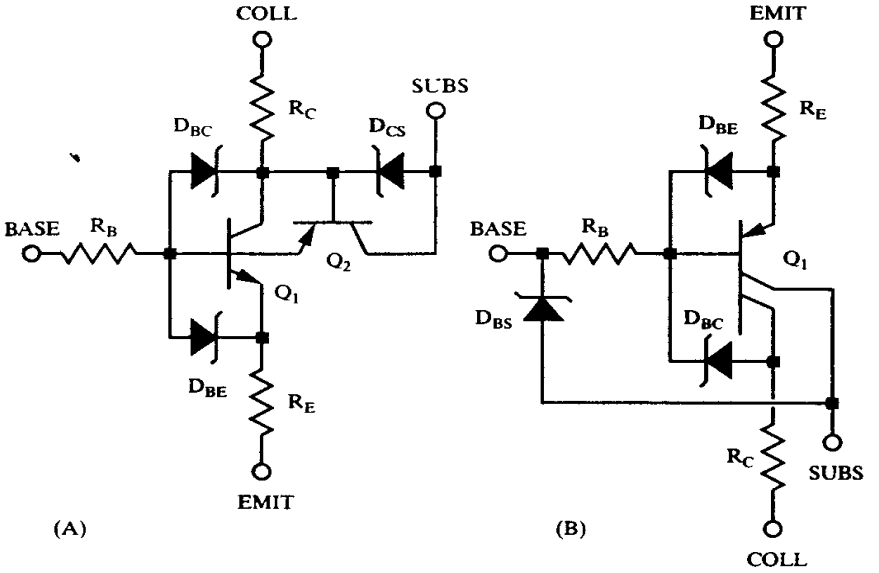


FIGURE 8.12 Subcircuit models for (A) vertical NPN transistor and (B) lateral PNP transistor.

Resistances R_E, R_B, and R_C model the Ohmic resistances of the emitter, base, and collector diffusions, respectively. Resistor R_E is usually quite small and is frequently neglected. Resistor R_B has a profound effect upon the switching speed of the transistor, as the base current must flow through a low-pass RC filter formed by R_B in series with capacitance C_{BE} of diode D_{BE}. The resulting low-frequency rolloff limits the maximum operating frequency of the transistor. Transistors with lower base resistances switch faster than those with high base resistances. The base resistance of a typical vertical NPN consists of the sum of the pinched base resistance beneath the emitter and the resistance of the extrinsic base underneath the contact. The presence of the pinched resistance causes the behavior of the base resistance to vary with bias in a very complex manner.[15] The collector resistance R_C limits the *saturation voltage* $V_{CE(sat)}$ that the transistor can achieve in saturation. Power transistors must have low collector resistances in order to minimize their saturation voltages at high currents. Also, the collector resistance of a transistor must not become too great or the transistor may intrinsically saturate even when the terminal voltages appear to indicate operation in the normal active region.

The lateral PNP transistor model of Figure 8.12B uses an ideal dual-collector PNP Q_1 to model the desired PNP transistor and its substrate PNP parasitic (Section 8.2.3). Some portion of the collector current flows through each of the two collectors. If the collector of the four-terminal transistor saturates, then most of the current instead flows to the substrate terminal (Section 8.1.5). Zener diodes D_{BE}, D_{BC}, and D_{BS} model the avalanche breakdown and capacitance effects of the three junctions of the lateral PNP transistor. Diode D_{BE} models the base-emitter junction's capacitance C_{BE} and its breakdown voltage V_{EBO}. Diode D_{BC} models the

[15] J. R. Hauser, "The Effects of Distributed Base Potential on Emitter-Current Injection Density and Effective Base Resistance for Stripe Transistor Geometries," *IEEE. Trans. on Electron Devices*, Vol. ED-11, #5, 1964, pp. 238–242.

base-collector junction's capacitance C_{BC} and its breakdown voltage V_{CBO}. The two breakdown voltages V_{EBO} and V_{CBO} are about equal because both junctions consist of the same diffusions. Capacitance C_{BC} is usually quite large due to the construction of the lateral PNP transistor, partly accounting for its poor frequency response. Diode D_{BS} models the base-substrate junction's avalanche voltage and its capacitance C_{BS}. This capacitance, which is also rather large, does not exist in a vertical transistor. The base-substrate capacitance of a lateral PNP also accounts for some of this transistor's slow frequency response.

Resistors R_E, R_B, and R_C model the Ohmic resistances of the emitter, base, and collector, respectively. Although both the emitter and the collector resistances may equal a few hundred Ohms, neither greatly affects the operation of the transistor. Resistor R_B is quite large because of the light doping of the N-epi tank. The presence of NBL does not completely counteract the effect of light doping because conduction in the lateral PNP transistor occurs near the surface, and the base current must consequently flow through virtually the entire thickness of the epi. This large base resistance forms an RC filter with capacitances C_{BC} and C_{BS}, accounting for the notoriously sluggish frequency response of lateral transistors.

8.2 STANDARD BIPOLAR SMALL-SIGNAL TRANSISTORS

Standard bipolar was originally used to construct digital logic circuits. Designers soon realized that this process could also fabricate analog integrated circuits such as voltage references, operational amplifiers, and comparators. None of these circuits operate at particularly high voltages or currents. Most of their transistors conduct, at most, a couple of milliamps. The design of these *small-signal transistors* emphasizes small size, high gain, and high speed at the expense of power-handling capability. Although designers may differ on exact definitions, most would probably agree that small-signal transistors handle currents of less than 10mA and power levels of less than 100mW. Transistors that exceed these limits begin to resemble power transistors more than small-signal devices (Section 9.1.2).

The standard bipolar process was optimized to fabricate NPN transistors, but its various diffusions can also create substrate and lateral PNP transistors. The design principles of small-signal transistors remain much the same regardless of the details of the process. The structure of any optimized bipolar transistor resembles that of a standard bipolar NPN. Most nonoptimized transistors resemble either the substrate or the lateral PNP. By studying the transistors available in standard bipolar, one can gain insight into how transistors are implemented in other processes.

8.2.1. The Standard Bipolar NPN Transistor

The standard bipolar NPN contains several features intended to optimize its performance. These include a heavily doped emitter; a precisely tailored base profile; a thick, lightly doped N-epi; a heavily doped NBL; and a deep-N+ sinker (Figure 8.13).

The emitter diffusion is heavily doped with phosphorus to maximize its emitter injection efficiency. The solid solubility of phosphorus in silicon allows doping levels exceeding 10^{20} atoms/cm^3, allowing the construction of a highly efficient emitter that takes full advantage of a carefully tailored base profile. Arsenic is sometimes added to the emitter diffusion to compensate for lattice strains induced by the heavy phosphorus doping. This precaution eliminates defects that might otherwise migrate into the neutral base and degrade the beta of the transistor.

The standard bipolar base diffusion has been tailored to provide a combination of high beta, high Early voltage, and moderate V_{CEO}. Light doping aids in main-

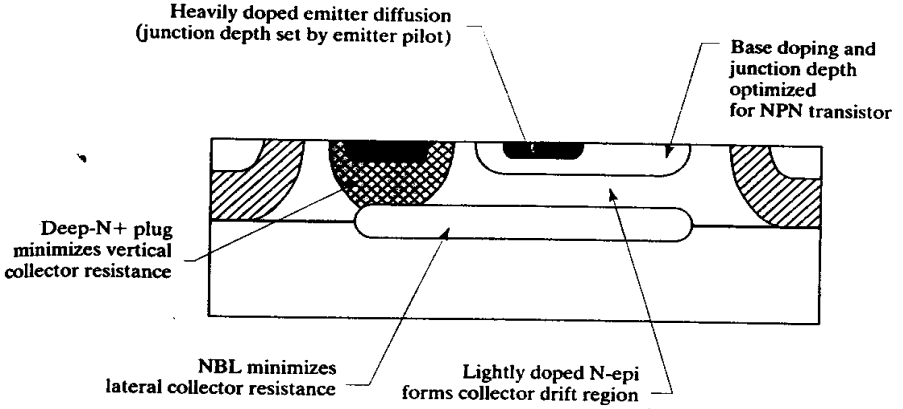


FIGURE 8.13 Key features of the standard bipolar vertical NPN transistor.

taining beta control because it allows the use of a wider, and therefore more controllable, neutral base region. A lightly doped base also increases the planar V_{CBO}, and by extension, the V_{CEO} of the vertical NPN. The standard bipolar base diffusion must still contain sufficient dopant to allow direct Ohmic contact since no shallow P+ diffusions exist in this process. Too low a surface doping concentration can also cause surface inversion and parasitic channel formation that can reduce the low-current beta of the transistor. Compromises between these conflicting requirements usually result in a base sheet resistance of 100 to 200Ω/□.

The NPN collector consists of three separate regions: a lightly doped N-epi, an N+ buried layer, and a deep-N+ sinker. The inclusion of a lightly doped layer adjacent to the collector-base junction increases V_{CEO} and Early voltage by allowing the formation of a wide depletion region extending primarily into the collector. This lightly doped *drift region* lies sandwiched between the base and a heavily doped *extrinsic collector*. In standard bipolar, the drift region consists of the remaining lightly doped N-epi beneath the base diffusion and above the NBL, while the extrinsic collector consists of the NBL and the deep-N+ sinker. The drift region, although lightly doped, does not impede the flow of collector current as long as it remains entirely depleted. The drift region depletes through at higher currents due to velocity saturation and at higher collector-to-emitter voltages due to the extension of the base-collector depletion region. Between these two extremes lies a range of collector voltages and currents that cannot entirely deplete the drift region, causing the effective resistance of the neutral collector to increase (an effect sometimes called *quasisaturation*[16]). Quasisaturation can be minimized by keeping the drift region as thin as possible consistent with the V_{CEO} rating of the process.

NBL creates a low-resistance path across the bottom of the transistor, but the current must still flow upward to reach the collector contact. The lightly doped epi separating the contact from the underlying NBL is highly resistive, so the inclusion of a deep-N+ sinker can reduce the total collector resistance by as much as an order of magnitude. For example, a typical minimum-size NPN transistor with NBL but without deep-N+ exhibits a collector resistance of about 1kΩ, while the same structure with the addition of a deep-N+ sinker has a collector resistance of about 100Ω. Power transistors always include deep-N+, but small-signal devices conducting no more than a few hundred microamps often omit it to save space.

[16] G. Massobrio and P. Antognetti, *Semiconductor Device Modeling with SPICE*, 2nd ed. (New York: McGraw-Hill, 1993), pp. 111–115.

Construction of Small-signal NPN Transistors

Small-signal NPN transistors usually employ square or rectangular emitters. Figure 8.14 shows two examples, both of which include the features discussed in the previous section. These transistors differ only in the placement of their base and emitter contacts. The structure of Figure 8.14A places the emitter between the collector and base contacts, forming a *collector-emitter-base* (CEB) layout. The structure of Figure 8.14B places the base contact between the collector contact and the emitter, forming a *collector-base-emitter* (CBE) layout. The CEB layout places the emitter and collector contacts closer together and therefore slightly reduces the collector resistance. All other factors remaining equal, the CEB layout will outperform the CBE layout. The difference is so small, however, that many designers use the two layouts interchangeably. The substitution of one style of transistor for the other often simplifies the lead routing of single-level-metal designs.

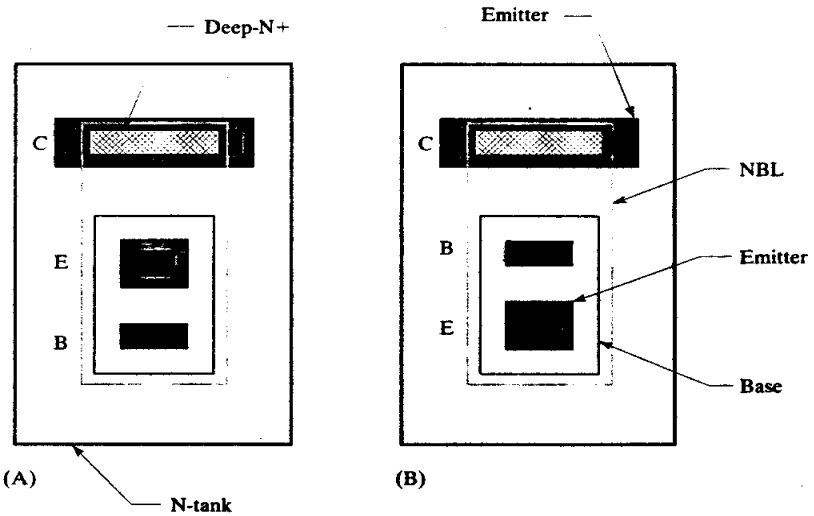


FIGURE 8.14 Two styles of NPN transistors: (A) collector-emitter-base (CEB), and (B) collector-base-emitter (CBE).

The size of an NPN transistor, or more precisely, the magnitude of its saturation current I_S, scales approximately linearly with drawn emitter area. Other factors that exert a lesser degree of influence on scaling include lateral conduction, current crowding, and emitter push. Carriers injected laterally from the emitter sidewalls suffer more recombination than those injected vertically because the laterally injected carriers must cross a wider and more heavily doped base region than the vertically injected ones. Surface states along the oxide-silicon interface also increase lateral recombination. Because of these effects, transistors with large area-to-periphery ratios usually have higher betas than those with smaller ratios. Measurements of two NPN transistors fabricated on a standard bipolar op-amp, one with a 1.0 × 1.0mil emitter and the other with a 1.5 × 3.5mil emitter,[17] showed that the smaller transistor exhibited a peak beta of 290 and the larger one a beta of 520. Most small NPN transistors employ either square or slightly elongated rectangular emitters to make efficient use of the available space while retaining a high area-to-

[17] B. A. Wooley, S.-Y. J. Wong, and D. O. Pederson, "A Computer-Aided Evaluation of the 741 Amplifier," *IEEE J. Solid-State Circuits*, Vol. SC-6, 1971, pp. 357–365.

periphery ratio (Section 9.2.1). Larger emitters also drive deeper into the base diffusion and thus reduce the neutral base width.

All of the other geometries that form the NPN transistor are placed relative to the emitter geometry, beginning with the emitter contact. This contact usually occupies as much of the emitter area as possible in order to minimize emitter resistance. The emitter diffusion should overlap the emitter contact equally on all sides to ensure uniform lateral current flow. The base diffusion must overlap the emitter on all sides sufficiently to prevent lateral punchthrough. To save space, the base is usually contacted along only one side of the emitter. The base contact can be elongated without enlarging the transistor, significantly reducing base resistance.

The outdiffusion of the P+ isolation determines the minimum tank overlap of base diffusion, which is among the largest spacings of the process. Up-down isolation can reduce this spacing by perhaps a third (Section 3.1.4). The base region occupies a tank contacted at one end by a deep-N+ sinker. Deep-N+ outdiffuses and must therefore reside some distance away from both the base and the isolation diffusions. Emitter coded over the deep-N+ increases the surface doping to ensure reliable Ohmic contact. Both the deep-N+ sinker and the emitter diffusion used to contact it can be elongated without increasing the size of the transistor. The deep-N+ sinker is often omitted from low-current devices to save space. Regardless of whether a sinker is used, the NBL should fill as much of the tank as possible to minimize the overall collector resistance. Minimum-geometry transistors often exhibit little overlap of NBL and deep-N+. As long as the drawn geometries touch one another, the collector resistance will be reduced sufficiently to allow low-current operation without significant collector voltage drops. If the transistor must conduct more than a few milliamps, then the transistor should be enlarged to allow the NBL to completely enclose the deep-N+ sinker.

Many variations of the standard bipolar NPN layout exist, especially for single-level-metal processes. The lack of second metal forces the designer to route leads through the transistor. Selective use of CEB and CBE layouts allows some rearrangement of the terminal ordering and often eliminates the need to jumper one or more signals. Transistors can also be stretched to allow one or more leads to route between their terminals. Stretched transistors exhibit increased base and collector resistances and capacitances that could interfere with proper circuit operation, so designers should avoid stretching devices whenever possible.

Figure 8.15A shows a typical *stretched-collector transistor*. The collector and base contacts have been moved apart to allow two leads to pass between them. This modification increases collector resistance and the collector-to-substrate capacitance.

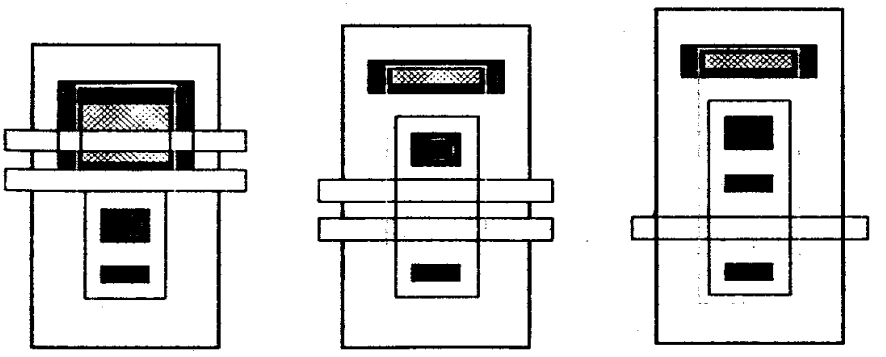


FIGURE 8.15 Three examples of stretched NPN transistors.

Nothing can be done to eliminate the added capacitance, but the increased resistance can be partially offset by elongating the deep-N+ and emitter diffusions surrounding the collector contact. In processes that employ a thin emitter oxide, leads routed over an extended emitter diffusion may represent an ESD vulnerability (Section 4.1.1), and only the deep-N+ and not the enclosing emitter diffusion should be elongated.

Figure 8.15B shows a *stretched-base transistor.* The base region of this transistor has been elongated to allow two leads to pass between the base and emitter contacts. The elongated base geometry causes an increase in base resistance and collector-base capacitance and a corresponding reduction in the frequency response. Stretched bases usually have a greater effect on circuit performance than stretched collectors, so they should be employed only when absolutely necessary, especially in high-speed signal paths. Figure 8.15C shows a different type of stretched-base transistor in which a lead tunnels through the base diffusion. This arrangement not only increases collector-base capacitance, but also inserts considerable resistance between the two base contacts. Assuming a base sheet of 160Ω/□, the resistance between the two base contacts of the illustrated transistor equals about 200Ω.

Double-level metal (DLM) virtually eliminates the need for stretched transistors. Not only does it eliminate stretched devices and tunnels, but it also reduces the die area because metal jumpers and vias can reside on top of other devices. DLM also allows the use of a small number of standardized layouts that can be fully characterized and modeled to enable more accurate simulation. The widespread implementation of multilevel metal systems has almost eliminated the use of stretched devices, but the technique still has merit for certain applications, for example in photodevices where metal-2 must act as a light shield, or in power devices where the upper metal layers are devoted to power routing.

Several approaches exist for scaling up NPN transistors, all of which seek to increase emitter area without reducing device performance. This turns out to be an elusive goal that requires different strategies for different applications. The naïve approach consists of enlarging the emitter while retaining the same overall geometry (Figure 8.16A). This yields a high beta due to an increased emitter area-to-periphery ratio, but it also entails several serious disadvantages. The lightly doped pinched base region beneath the emitter has a sheet resistance of about 2 to 10kΩ/□. This resistance introduces unwanted phase shifts and greatly slows transistor switching. More subtly, it causes nonuniform current flow at higher current levels. The portions of the emitter lying closest to the base contact experience the highest base-emitter bias and, consequently, inject more carriers than portions of the emitter far from the base contact. Figure 8.16B illustrates this effect, called *current*

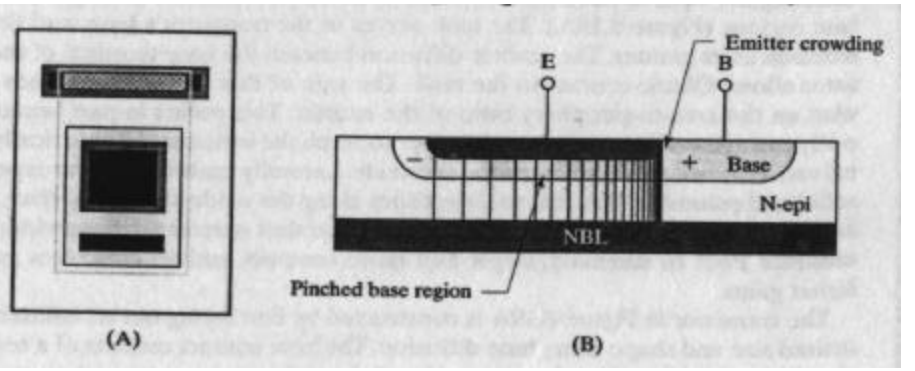


FIGURE 8.16 (A) Layout of a compact emitter NPN transistor and (B) a cross section of the active region of this device that illustrates the effects of emitter crowding.

crowding or *emitter crowding*. Its severity can be appreciated by remembering that a mere 18mV of debiasing doubles the emitter current.[18]

Current crowding reduces the active area of the emitter by forcing most of the conduction to occur near the base contact. This effect complicates device matching, reduces beta, and makes the transistor more susceptible to secondary breakdown. While current crowding is undesirable in small-signal transistors, it can actually enhance the robustness of properly designed power devices (Section 9.1.2).

An alternative style of layout uses long, narrow emitter stripes, or *fingers*.[19] Base contacts placed along both sides of each emitter finger help minimize base resistance, increasing switching speed and enhancing immunity to secondary breakdown (Figure 8.17A). The relatively small area-to-periphery ratio of this transistor yields a lower beta than the structure of Figure 8.16. This *narrow-emitter transistor* also becomes vulnerable to thermal runaway at emitter current densities of more than a few mA/mil^2 of emitter (Section 9.1.2).

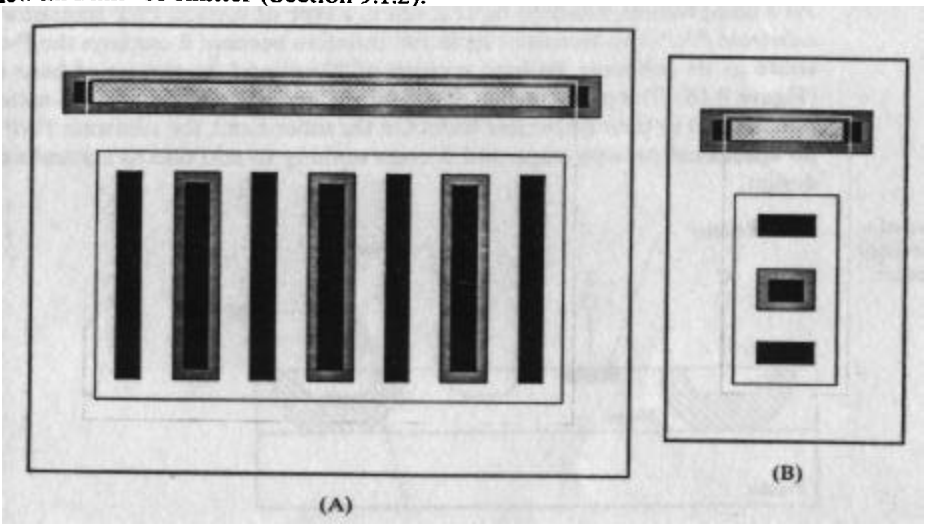


FIGURE 8.17 (A) The narrow-emitter transistor and (B) its equivalent minimum-size layout, the double-base transistor.

The compact emitter transistor in Figure 8.16 functions best in applications requiring high beta and moderate speeds. The narrow emitter transistor in Figure 8.17 provides superior frequency response, but inferior beta. A minimum-geometry transistor can also employ an emitter stripe paralleled by base contacts on either side, producing the somewhat misnamed *double-base transistor* (Figure 8.17B). This structure reduces the base resistance to approximately one-quarter that of the single-base layout of Figure 8.15.[20] Both the large-emitter and the narrow-emitter transistor function poorly at high emitter current densities, so neither is suitable for use as a power device (Section 9.1.1).

8.2.2. The Standard Bipolar Substrate PNP Transistor

Since NPN and PNP transistors differ only in doping polarities, one could theoretically create a PNP transistor by inverting all of the doping polarities of the standard

[18] R. J. Whittier and D. A. Tremere. "Current Gain and Cutoff Frequency Falloff at High Currents." *IEEE Trans. on Electron Devices*, Vol. ED-16, #1, 1969, pp. 39–57.

[19] A. B. Grebene, *Bipolar and MOS Analog Integrated Circuit Design* (New York: John Wiley and Sons, 1984), p. 76.

[20] A. B. Glaser and G. E. Subak-Sharpe, *Integrated Circuit Engineering* (Reading, MA: Addison-Wesley, 1977), p. 52.

bipolar process. The collector would then consist of a lightly doped P-type epi with the addition of a *P-buried layer* (PBL) and a deep-P+ sinker. The base would consist of a lightly doped N-type diffusion and the emitter of a heavily doped P-type one. The limitations of boron doping make it difficult to realize this structure. Buried layers should consist of slow-diffusing dopants such as arsenic or antimony. Boron diffuses relatively rapidly, so any attempt to fabricate a P-buried layer requires a radical process redesign to eliminate subsequent high-temperature drives. Standard bipolar NPN transistors also employ a heavily doped emitter to maximize emitter injection efficiency. The solid solubility of boron in silicon is only one-third that of phosphorus,[21] so the high-current performance of the PNP suffers accordingly. Even if these problems are somehow overcome, the mobility of holes in silicon is only about one-third that of electrons.

Standard bipolar cannot fabricate a fully isolated vertical PNP transistor. Although some processes do offer both vertical NPN and vertical PNP transistors, these *complementary bipolar* processes require several additional processing steps. As a compromise, standard bipolar offers a type of vertical PNP transistor called a *substrate PNP*. This transistor lacks full isolation because it employs the P-type substrate as its collector. Its base consists of N-epi and its emitter of base diffusion (Figure 8.18). The performance of the transistor suffers because these materials are not tailored to their respective tasks. On the other hand, the substrate PNP requires no additional process steps, and it costs nothing to add one to a standard bipolar design.

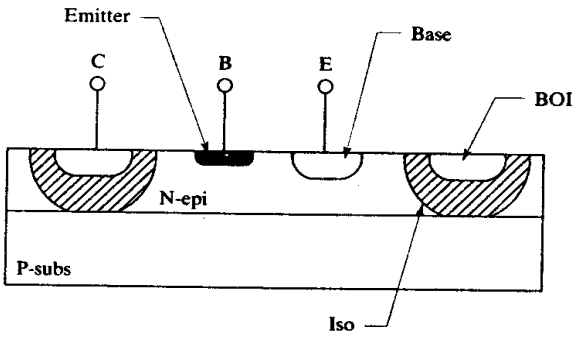


FIGURE 8.18 Cross section of a typical substrate PNP transistor fabricated in standard bipolar.

The various diffusions of the standard bipolar process take their names from the functions they perform in a vertical NPN transistor. These same diffusions play very different roles in a substrate PNP transistor. The *emitter* of a substrate PNP consists of *base* diffusion and the *base* consists of an N-tank contacted by means of *emitter* diffusion. As this example shows, one must pay close attention to the difference between the names of diffusions and the roles they play in a given structure.

The standard bipolar emitter diffusion typically has a dopant concentration in excess of 10^{20} atoms/cm^3, while the base diffusion rarely exceeds 10^{17} atoms/cm^3. The lighter doping greatly reduces the emitter injection efficiency of the substrate PNP. The still-lighter doping of the N-epi causes premature beta rolloff due to high-level injection. A substrate PNP typically retains a beta of 30 out to about 0.5mA/mil^2 (0.8μA/μm^2), while a vertical NPN transistor retains a beta of 150 out to perhaps 20mA/mil^2 (30μA/μm^2). The vertical resistance of the base diffusion exceeds that of the emitter diffusion by an order of magnitude, so a minimum-emitter

[21] F. A. Trumbore, "Solid Solubilities of Impurity Elements in Si and Ge," *Bell System Technical Journal*, Vol. 39, No. 1: 1960, pp. 205–233. Datapoints are taken at 1100°C.

substrate PNP typically exhibits 100Ω of emitter resistance, compared to the 10Ω of a comparable NPN transistor. The emitter resistance causes few problems at the low current densities characteristic of substrate PNP transistors. The increased emitter resistance even provides emitter ballasting, and this combined with high-current beta rolloff ensures a high degree of immunity to thermal runaway.

Standard bipolar employs a relatively lightly doped N-epi. In the absence of NBL, the epi-substrate junction diffuses upward during the long isolation drive. The base region of the substrate PNP is therefore both thin and lightly doped—much more so than the epi thickness might suggest. Most of the carriers flow vertically from emitter to substrate rather than laterally from emitter to isolation, so this transistor does not exhibit the difficulties associated with lateral current flow (Section 8.2.3). Substrate PNP transistors often have peak betas of 100 or more, but high-level injection causes their betas to peak at current densities of less than 1mA/mil^2 (1.5μA/μm^2).

The collector of the substrate PNP consists of a series combination of the P-substrate and the P+ isolation diffusion. The lightly doped substrate increases both the Early voltage and the V_{CEO} rating of the transistor, but strictly speaking it does not act as a drift region because it is not bounded by an adjacent P+ layer. Collector voltage drops rarely have much effect upon the substrate PNP itself, but they can cause debiasing of adjacent circuitry at higher current levels. Standard bipolar designs rarely experience objectionable levels of debiasing as long as the collector current in each substrate PNP does not exceed 1 to 2mA and the total collector current of all substrate PNP transistors does not exceed 10mA. A standard P+ isolation diffusion supplemented by base-over-iso can cope with current levels of this magnitude as long as substrate contacts are located near the substrate transistors. Any substrate PNP transistor that injects 1mA or more should have additional substrate contacts surrounding it (Section 4.4.1). Substrate PNP transistors become progressively more impractical as current levels increase; circuit designers may wish to consider substituting lateral PNP transistors for high-current substrate devices to minimize substrate debiasing. Alternatively, backside substrate contact can be employed to remove almost any amount of substrate current without debiasing.

Substrate PNP transistors operate primarily by vertical conduction and therefore scale with drawn emitter area. Several other factors influence the precise scaling of these devices, including outdiffusion, lateral conduction, and surface effects. As with vertical NPN transistors, precisely matched substrate PNP transistors must use identical emitter geometries.

Construction of Small-signal Substrate PNP Transistors

The simplest style of substrate PNP consists of an N-tank containing emitter and base regions (Figure 8.19A). The tank serves as the transistor's base, and the base diffusion as its emitter. The emitter diffusion beneath the base terminal of the transistor allows Ohmic contact to the tank. The gain of this transistor depends somewhat on the area-to-periphery ratio of the emitter. This occurs in part because laterally emitted carriers must travel further to reach the isolation than vertically emitted carriers must travel to reach the substrate. Laterally emitted carriers experience additional recombination due to defect sites along the oxide-silicon interface. Large base diffusions also drive deeper into the epi and thus narrow the base width of the substrate PNP. In summary, larger and more compact emitter structures produce higher gains.

The transistor in Figure 8.19A is constructed by first laying out an emitter of the desired size and shape using base diffusion. The base contact consists of a rectangle of emitter diffusion placed on one side of the emitter geometry and spaced away

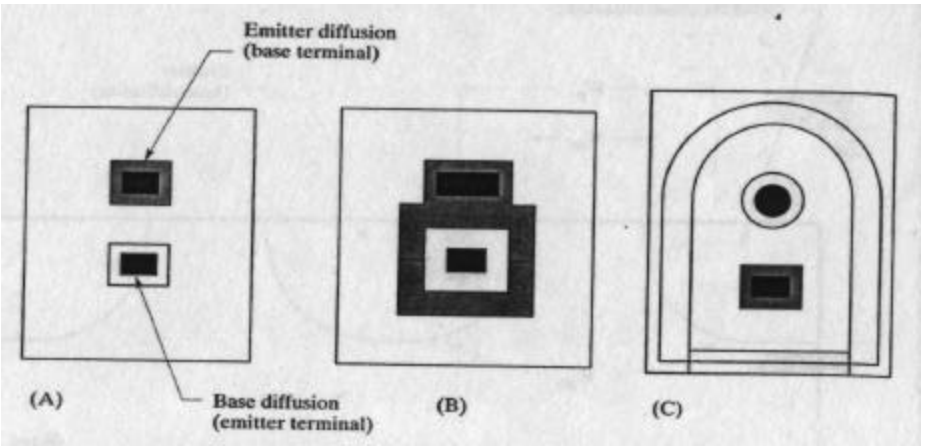


FIGURE 8.19 Examples of (A) standard, (B) emitter-ringed, and (C) verti-lat styles of substrate PNP transistor.

from it by the minimum allowed base-to-emitter spacing. The emitter diffusion need only be wide enough to contain a minimum-width contact. At least one substrate contact should reside near each substrate PNP. Ideally, this contact should abut the substrate PNP, but wiring constraints frequently force it to reside several mils away. Some separation between the substrate contact and the transistor can be tolerated as long as the collector current of the substrate PNP does not exceed a milliamp or two. Higher collector currents will probably debias the substrate unless substrate contacts are placed adjacent to the transistor.

Figure 8.19B illustrates another type of substrate PNP layout where the emitter consists of a rectangle of base diffusion surrounded on all sides by a thin ring of emitter diffusion. This emitter ring adjoins but does not overlap the drawn base diffusion. The emitter ring helps discourage lateral conduction by raising the voltage required to forward-bias the emitter sidewalls. The emitter-ringed transistor exhibits a higher beta at low current densities than the transistor in Figure 8.19A does. This advantage disappears at higher current densities because the base forward-biases into the emitter ring. Although this structure has some merit, most layouts do not employ it because it requires more room than the simple layout of Figure 8.19A, and because the latter structure has sufficient gain for most applications. If an emitter ring is used, then the spacing between the ring and the base diffusion's contact must equal the spacing between the emitter of a vertical NPN and its base contact. The other dimensions of this transistor follow much the same rules as apply to the structure of Figure 8.19A.

The device of Figure 8.19C is sometimes called a *tombstone PNP* or a *cathedral PNP* because of its distinctive tank outline. The same transistor is also called a *verti-lat PNP* because it attempts to capitalize upon both vertical and lateral conduction in the same device. The emitter of this transistor consists of a circular plug of base diffusion. The tank geometry passes around the emitter in a semicircular arc digitized around the same center as the emitter. A ring of base diffusion also surrounds the tank, overlapping into it as far as layout rules allow. This base ring helps counteract the high sheet resistance of the periphery of the isolation caused by its extensive outdiffusion. Lateral conduction in the transistor occurs outward from the plug of base diffusion serving as the emitter, to the ring of base diffusion serving as the collector. Conduction also occurs vertically downward from the emitter plug to the underlying substrate. This style of transistor requires much the same field plating as does a lateral PNP. The field plate should entirely cover the exposed surface of the N-epi at the semicircular end of the transistor, and it should extend toward the base

contact as far as the metal spacing rules allow. Failure to properly field-plate the transistor may cause unexpected leakage phenomena as well as a degradation of low-current beta (Section 8.2.3).

A verti-lat PNP should theoretically have a higher beta than a standard substrate PNP because its lateral conduction pathway has been optimized by minimization of the neutral base width and by the incorporation of a base field plate. In practice, the structure of Figure 8.19B generally outperforms the verti-lat transistor because it suppresses lateral conduction and relies instead on the more efficient vertical conduction.

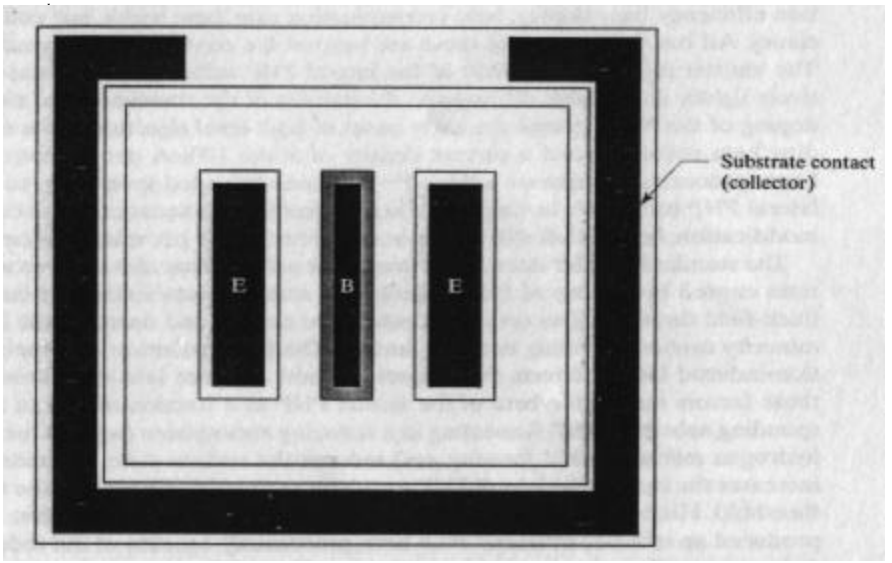


FIGURE 8.20 Higher-current substrate PNP transistor employing two wide emitter stripes and a single narrow base stripe.

Figure 8.20 shows the layout of a larger substrate PNP transistor. Like the large vertical NPN transistor in Figure 8.17A, this device employs interdigitated emitter and base stripes. Each emitter consists of a fairly wide (~25μm) strip of emitter diffusion. Since PNP transistors exhibit beta rolloff at relatively low current densities, hot spotting and secondary breakdown generally do not occur and the emitter width is limited only by the resistance of the pinched region beneath the base diffusion. Because this pinched sheet may exceed 10kΩ/□, emitter stripes more than 25 to 50μm wide are not advisable. The base contacts consist of thin stripes of emitter diffusion placed between adjacent emitter fingers. If desired, two additional stripes of emitter diffusion can be placed on the ends of the transistor to further reduce base resistance. The large substrate contact encircling the transistor helps limit substrate debiasing. The illustrated contact cannot deal with more than a few milliamps of substrate current without saturating; a much wider strip of contact would be required to deal with the maximum collector current of which this structure is capable. The substrate contact has been interrupted along the top of the transistor to allow the emitter and base leads to emerge on first-level metal. If double-level metallization is available, then the contact should form an unbroken ring around the transistor to minimize collector resistance.

8.2.3. The Standard Bipolar Lateral PNP Transistor

Although standard bipolar cannot fabricate an isolated vertical PNP, it does offer an isolated *lateral PNP* consisting of two separate base diffusions placed in a common

tank. One of these diffusions serves as the emitter and the other as the collector. When the emitter forward-biases, holes flow into the tank and travel laterally to the collector. Lateral transistors generally have slower switching speeds and lower betas than vertical devices. Although little can be done to boost switching speeds, proper design can substantially improve the beta. The layout designer can also vary the base width of a lateral PNP by moving the emitter and collector nearer together or further apart. A narrower basewidth produces a higher beta and a lower Early voltage, while a wider basewidth produces the opposite effect. The product of beta and Early voltage remains approximately constant regardless of basewidth.

The beta of a lateral PNP depends on at least five different factors: emitter injection efficiency, base doping, base recombination rate, base width, and collector efficiency. All but the last two of these are beyond the control of the layout designer. The emitter injection efficiency of the lateral PNP suffers from the use of a relatively lightly doped base diffusion as the emitter of the transistor. The even lighter doping of the N-epi causes the early onset of high-level injection and a corresponding beta rolloff beyond a current density of about 100μA per minimum emitter. Some processes incorporate a deep-P+ diffusion intended specifically to construct lateral PNP transistors having better high-current characteristics, but even with this modification, beta rolloff still begins at or below 500μA per minimum emitter.

The standard bipolar lateral PNP transistor suffers from elevated recombination rates caused by the use of (111) silicon. The same surface states that increase the thick-field threshold also act as recombination centers and decrease the lifetime of minority carriers traveling near the surface. The field oxidation also spawns oxidation-induced lattice defects that migrate a short distance into the silicon. Both of these factors reduce the beta of the lateral PNP to a fraction of that of the corresponding substrate PNP. Annealing in a reducing atmosphere (such as the nitrogen-hydrogen mixture called *forming gas*) reduces the surface state concentration and increases the beta of the lateral PNP transistor at the cost of reducing the thick-field threshold. Historically, the introduction of compressive nitride protective overcoats produced an increase in lateral PNP beta, presumably because of the reducing conditions prevailing during nitride deposition. Together with smaller feature sizes, reducing anneals increased the peak beta of standard bipolar lateral PNP transistors from less than 10 to more than 50.

The *drawn* basewidth of a lateral PNP equals the separation between the drawn base diffusions serving as its emitter and collector (Figure 8.21, dimension W_{B1}). The *actual* base width of the transistor is more difficult to determine because it depends on two-dimensional current flow. Near the surface, the actual base width W_{B2} is considerably less than the drawn base width W_{B1} due to outdiffusion. As one proceeds deeper into the silicon, the sidewalls of the emitter and collector curve away from one another. Carriers do not move in straight lines, and the greater distance traveled along alternative pathways also increases the *effective* basewidth of the transistor (W_{B3}). The effective base width of the transistor consists of a weighted average of all possible paths by which carriers might traverse the neutral base. The complexity of this problem precludes simple closed-form solution, and the solutions that exist provide little insight into transistor design.[22,23] Three general observations can still be made. First, the effective base width for purposes of computing punchthrough consists of the drawn basewidth minus twice the outdiffusion distance, or W_{B1}. Only

[22] D. E. Fulkerson, "A Two-Dimensional Model for the Calculation of Common-Emitter Current Gains of Lateral p-n-p Transistors," *Solid-State Electronics*, Vol. 11, 1968, pp. 821–826.

[23] K. N. Bhat and M. K. Achuthan, "Current-Gain Enhancement in Lateral p-n-p Transistors by an Optimized Gap in the n+ Buried Layer," *IEEE Trans. on Electron Devices*, Vol. ED-24, #3, 1977, pp. 205–213.

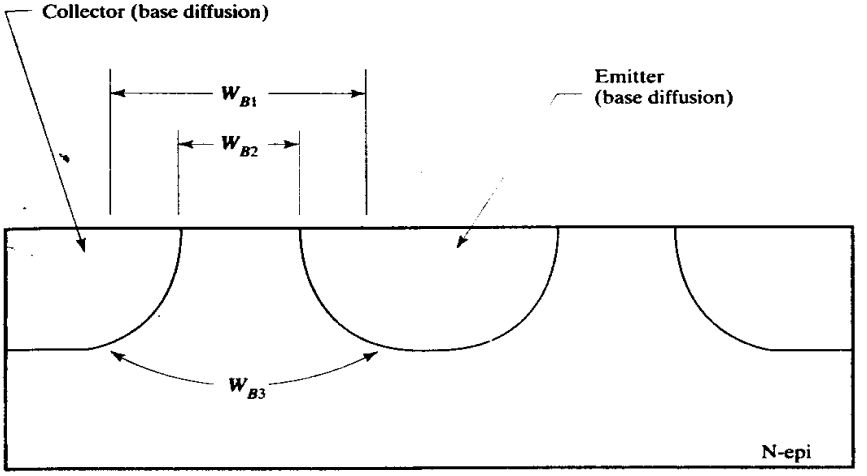


FIGURE 8.21 Cross section of a lateral PNP transistor depicting three different measures of neutral base width discussed in the text (drawn base width, W_{B1}, actual base width at the surface, W_{B2}, and effective base width beneath the surface, W_{B3}).

the smallest drawn basewidths exhibit any significant decrease in operating voltage due to punchthrough. Second, the effective base width for purposes of computing beta (W_{B3}) substantially exceeds the actual base width at the surface (W_{B2}) due to the contribution of subsurface conduction. This causes beta to scale more weakly with drawn base width than one might expect. If a transistor with a drawn basewidth of 8μm has a peak beta of 80, then one with a drawn basewidth of 16μm will have a peak beta greater than 40. Third, the Early voltage of a lateral PNP is inversely proportional to peak beta. Suppose a transistor with a peak beta of 80 has an Early voltage of 70V; a wider-base transistor exhibiting a beta of 60 should have an Early voltage of approximately (80/60)·70 = 93V.

Designers have long known that the gain of a lateral PNP appears to increase when NBL is placed underneath the emitter. This increase occurs because the presence of the NBL restricts the motion of holes to a smaller region of the N-epi and therefore reduces recombination. NBL also greatly increases collection efficiency by blocking the injection of holes into the substrate (Section 8.1.5). Because only a few of the carriers manage to surmount the built-in potential surrounding the NBL, the increase in base current caused by recombination in the NBL is quite small.

One might also expect some percentage of the emitted carriers to escape by traveling laterally beneath the collector diffusion. Experience shows that this does not occur to any appreciable extent while the transistor remains in the normal active region. The NBL updiffuses through the epi, establishing a retrograde doping profile that tends to drive the minority carriers upward along the built-in potential gradient. At the same time, the depletion region surrounding the collector reaches deep into the lightly doped N-epi and therefore intercepts a large fraction of minority carriers attempting to pass underneath it. Only a small percentage (usually less than 1%) of emitted carriers successfully escape the collector to reach the isolation sidewall. Transistors with shallow collector regions may experience greater sidewall losses (Section 8.3.2).

Construction of Small-signal Lateral PNP Transistors

The lateral PNP transistor traditionally consists of a small plug of base diffusion surrounded by a larger annular base region (Figure 8.22). The central base plug serves as the emitter of the transistor, while the surrounding base ring serves as its collector. This structure ensures that almost all of the carriers injected by the emitter are intercepted

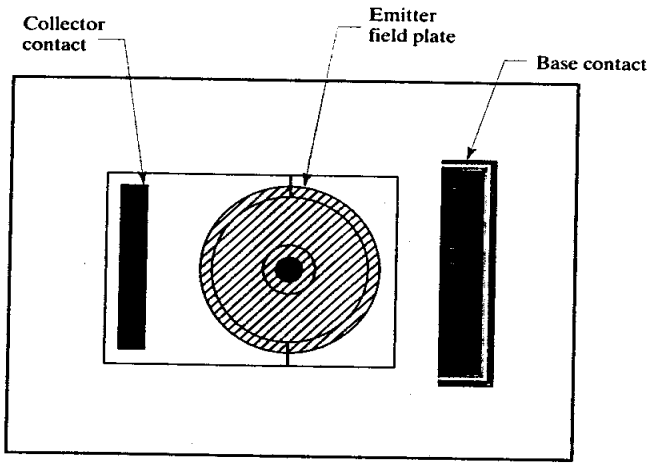


FIGURE 8.22 Layout of a lateral PNP transistor showing emitter field plate.

by the intended collector before they can reach the isolation. The apparent reverse beta of this style of lateral PNP transistor is always far smaller than its forward beta. In the reverse mode, the outer ring of base becomes the emitter and the majority of the injected carriers are injected toward the isolation sidewalls rather than to the small base plug in the center of the transistor. Thus, despite identical doping levels in emitter and collector, the lateral PNP transistor remains a profoundly asymmetrical device.

The circles and arcs used in constructing lateral PNP transistors are actually polygonal approximations. The number of segments in these polygons determines how closely they resemble ideal circles and arcs. The number of segments in the complete circle should be evenly divisible by four to ensure symmetry around both axes. Circular approximations with sixty-four sides are recommended for most applications. Annular geometries, such as the collector in Figure 8.22, usually consist of two matched halves to eliminate certain difficulties sometimes encountered during photomask generation.[24]

As stated previously, the peak beta of a lateral PNP depends inversely upon its effective base width. Carriers injected from the emitter sidewalls travel a shorter path than carriers injected from the bottom surface. Smaller emitters therefore have shorter effective base widths. Some designers mistakenly assume that the periphery-to-area ratio determines the beta of a lateral PNP, but the actual determining factor is the distance from the middle of the emitter to its periphery.

Practical lateral PNP transistors usually employ a circular emitter geometry just large enough to contain a minimum-size contact (Figure 8.22). The emitter contact should be made circular and concentric with the emitter to ensure that all portions of the emitter periphery are equidistant from the contact edge. If they are not, some portions of the emitter periphery would inject an undue share of the carriers. This nonuniform current flow can cause subtle mismatches, especially in split-collector transistors. The collector of the lateral PNP generally consists of a square of base diffusion having a circular hole in its center. The emitter occupies the center of this hole. The drawn base width equals the difference between the radius of the emitter and the radius of the collector opening. One end of the collector extends sufficiently to allow placement of a contact inside it. Although debiasing occurs along the collector periphery, this only results in a slight increase in the effective saturation voltage of the transistor due to premature saturation of the debiased portion of the col-

[24] The boundary of a so-called *semisimple figure* becomes coincident with itself at one or more points, while the boundary of a *simple figure* does not. Some pattern generation algorithms have problems with semisimple figures, so they are often avoided.

lector. If this increase in saturation voltage causes concern, then additional collector contacts can be added to reduce collector debiasing. For most applications, the additional contacts are not worth the space they consume.

The collector of the lateral PNP resides in the tank that forms its base. A strip of emitter diffusion added to one end of the tank serves as a base contact. A deep-N+ sinker is not normally required because the base current in the lateral PNP rarely exceeds a few tens of microamps. If space is at a premium, even a minimum-size base contact will suffice. On the other hand, a lateral PNP should always contain as much NBL as possible to minimize unwanted substrate injection.

Figure 8.22 shows a metal plate covering the exposed portions of the N-tank between the emitter and the collector. This *field plate* prevents unwanted surface inversion or accumulation from occurring due to charge spreading and field interactions with the adjacent collector. Without the field plate, the beta of the lateral PNP fluctuates depending on surface potentials.[25] If a negative charge accumulates at the surface, then the minority carriers move toward the oxide interface and the beta of the transistor decreases. Similarly, a positive potential repels the carriers from the surface and increases the beta. The presence of mobile ions greatly increases the magnitude of the instabilities caused by these surface charges. The use of a field plate connected to the emitter not only stabilizes the beta of the transistor, but also helps to increase it by repelling carriers from the oxide-silicon interface.

Lateral PNP transistors that lack field plates often exhibit collector-to-emitter leakage at voltages well below the thick-field threshold. These leakages result from parasitic channel formation. The geometry of the lateral PNP ensures that the parasitic channel will have a large W/L ratio, while the voltage differential present between collector and emitter attracts mobile ions and surface charges. Leakages have been observed in standard bipolar lateral PNP transistors at collector-to-emitter voltages of only 5 to 10V on a process having a thick-field threshold in excess of 40V. Properly designed field plates will completely eliminate these leakage currents. The field plate should connect to the emitter of the transistor and should entirely cover the exposed N-epi between the emitter and the collector. The field plate should overlap the periphery of the collector so that misalignment between the metal and the base diffusion cannot expose the epi surface. An overlap of 2 to 3μm (about 0.1mil) is more than enough, because the outdiffusion of the base helps shrink the size of the collector opening that the field plate must cover.

A minimum-emitter lateral PNP occupies considerably more space than a minimum-emitter NPN. A single lateral PNP can, however, be subdivided to form several smaller transistors sharing a common base and emitter. Figure 8.23A shows a simple example of such a *split-collector lateral PNP*. This transistor contains two partial collector segments, each occupying half of the emitter periphery. Since the emitter injects carriers uniformly in all directions, each collector receives half of the total injected current. This device therefore behaves as if it were actually two separate transistors, each having an effective emitter size one-half that of a normal lateral PNP. The transistor in Figure 8.23B carries the process of splitting the collector still further. Instead of two half-sized collectors, this transistor contains four quarter-sized ones. One can even construct a split-collector transistor containing segments of different sizes. For example, the transistor in Figure 8.23C includes three, one-sixth collectors and two, one-quarter collectors.

The multiple collectors will match one another within about ±1% as long as they all possess identical geometries placed symmetrically.[26] The split collectors in

[25] R. O. Jones, "P-N-P Transistor Stability," *Microelectronics and Reliability*, Vol. 6, 1967, pp. 277–283.

[26] B. Gilbert, "Bipolar Current Mirrors," in C. Toumazou, F. J. Lidgy, and D. G. Haigh, *Analog IC Design: The Current-Mode Approach* (London: Peter Peregrinus, 1990). p. 250.

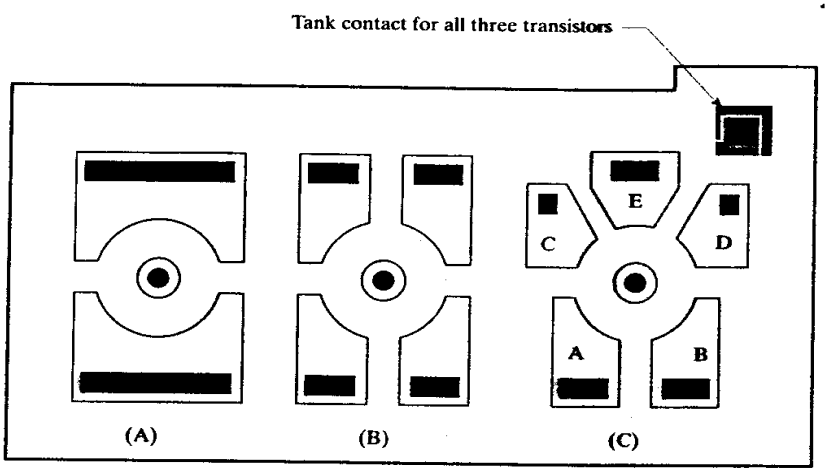


FIGURE 8.23 Examples of split collector transistors: (A) 1/2–1/2, (B) 1/4–1/4–1/4–1/4, (C) 1/6–1/6–1/6–1/4–1/4. The field plates have been omitted for clarity.

Figures 8.23A and 8.23B meet these criteria and therefore match fairly precisely. The split collectors in Figure 8.23C do not all have the same geometries and therefore do not all precisely match one another. Collectors A and B will match, as will collectors C and D. Collector E will not match collectors C and D because the former possesses a different geometry than the latter. Similarly, the ratio between collectors A and C will not exactly equal 2:3 because these segments have different geometries.

Split-collector laterals are frequently employed to construct current mirrors. A simple 1:1 current mirror can be constructed from a split collector transistor with two half-sized collectors. One collector connects back to the common base to form reference transistor Q_{1A} (Figure 8.24A). The other collector serves as the output transistor Q_{1B}. This arrangement saves considerable space, but it probably does not match as precisely as two separate lateral PNP transistors placed next to one another. Emitter degeneration cannot improve the matching of the split-collector mirror because the two transistors share a common emitter. Many schematic diagrams denote split collectors by placing multiple collector leads on a single base bar (Figure 8.24B). The ratio of the split collectors is indicated by values placed next to each collector lead. The passage of a base lead through the transistor does not indi-

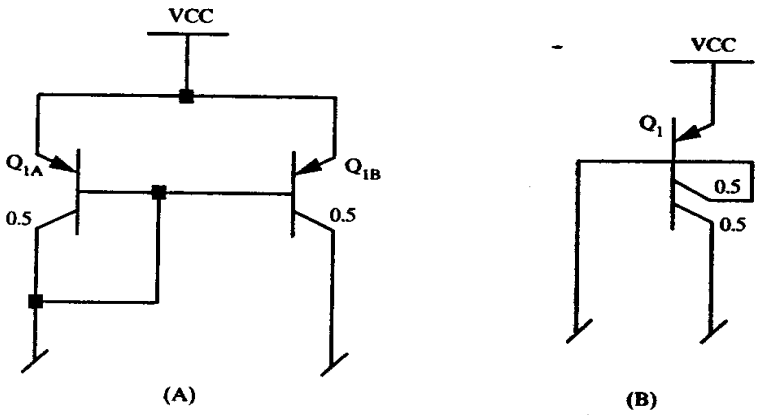


FIGURE 8.24 Schematic diagrams for 1:1 current mirrors constructed using split collector lateral PNP transistors: (A) conventional schematic diagram and (B) simplified schematic diagram.

cate the presence of multiple base terminals, but rather indicates two separate connections to the same base terminal. Using this style of schematic representation, the circuit of Figure 8.24A reduces to the considerably more compact, if less familiar, schematic of Figure 8.24B.

Some designers prefer to digitize lateral PNP transistors using a square emitter surrounded by a square collector ring. This style of lateral PNP is easier to digitize than the circular geometries previously discussed, but its base width increases slightly due to the greater length of the diagonal conduction paths and the poorer area-to-periphery ratio of the square emitter. Some designers fillet the four corners of the opening in the annular collector geometry so that the base width does not increase at the corners. Figure 8.25 shows examples of two types of square lateral PNP transistors.

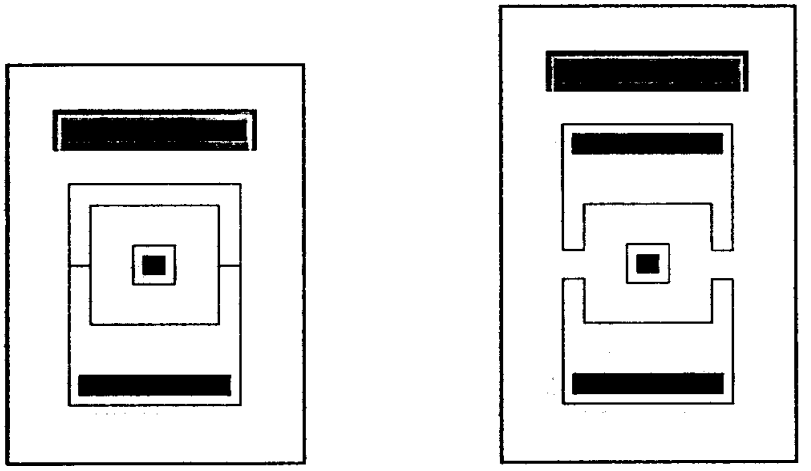


FIGURE 8.25 Square geometry lateral PNP transistors: (A) minimum-emitter and (B) 1/2–1/2 split collector. The field plates have been omitted for clarity.

The lack of radial symmetry in a square split-collector transistor prevents the use of anything except half and quarter collectors. Apart from this limitation, the same rules apply to the construction of square split-collector PNP transistors as to circular ones. In all cases, the emitter should be kept as small as possible and a field plate should entirely cover the exposed N-epi between the emitter and the collector.

Some designers claim that lateral PNP transistors scale with drawn emitter area, while others believe that they scale with drawn emitter periphery. The geometry of the emitter actually plays a more crucial role than either its area or its periphery. As pointed out previously, the effective base width is proportional to the distance from the center of the emitter to the collector periphery. The use of large square or circular emitters drastically reduces the beta of the transistor. An emitter elongated into a thin stripe can possess a very large area without requiring a proportionate increase in base width (Figure 8.26A). The *elongated-emitter lateral PNP* has an emitter geometry reminiscent of the outline of a hot dog, so it is sometimes called a *hot-dog transistor.*

The area and periphery of the elongated-emitter transistor both increase at about the same rate, so arguments about whether the transistor scales with area or periphery are largely academic. An elongated-emitter transistor will not match

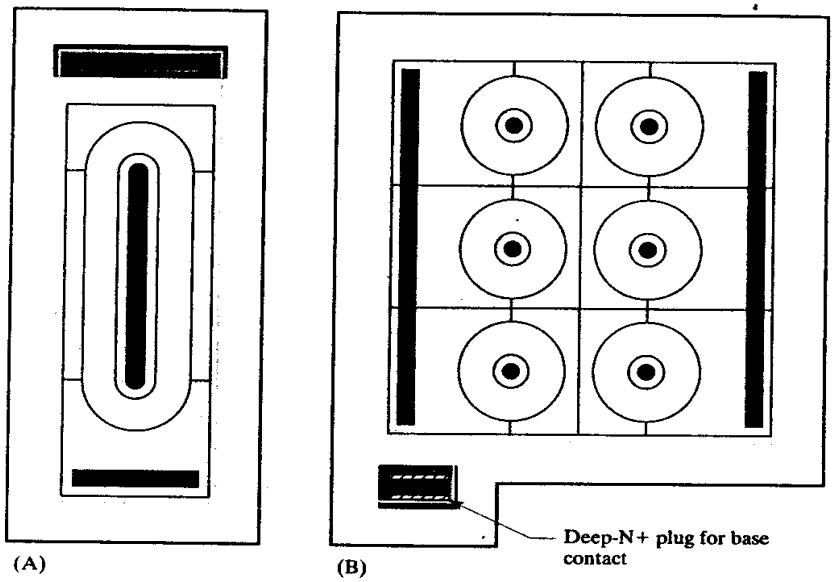


FIGURE 8.26 Higher-current lateral PNP transistors: (A) elongated-emitter or hot-dog transistor and (B) a small arrayed-emitter transistor.

a circular-emitter transistor because of junction sidewall effects, current crowding, high-level injection, and various other sources of mismatch. The beta of the elongated-emitter transistor decreases slightly as its emitter lengthens, although this decrease is much smaller than would occur if the emitter geometry was, for example, an enlarged square. This decrease results from the contribution to the effective base width from carriers moving down the length of the emitter stripe. The transistor in Figure 8.26A has a relatively large collector resistance that may disturb transistor matching and reduce beta at low collector-to-emitter voltages. If this causes concern, then the collector contacts can be placed along the longer sides of the collector to reduce the collector resistance.

Larger lateral PNP transistors can also be produced by arraying a multitude of minimum-size emitters (Figure 8.26B). Because the geometry of each emitter cell is exactly the same as that of a minimum-emitter device, the current-handling capability of this *arrayed-emitter transistor* is directly proportional to the number of emitters it contains. Large devices of this type often stagger alternating columns of emitters to achieve a tighter hexagonal packing arrangement.

Although scaling by area and scaling by periphery both give approximately the same results when applied to the lateral PNP transistors in Figure 8.26, scaling by periphery extends more naturally to the case of split-collector laterals. Each split collector functions as a separate transistor whose size depends on the fraction of the emitter periphery it subtends. Split-collector transistors employing elongated emitters also follow the same principle. Most designers scale lateral PNP transistors by drawn-emitter periphery, and split-collector transistors are assigned the appropriate fractions of the periphery of their shared emitter.

8.2.4. High-voltage Bipolar Transistors

The maximum operating voltage of a process cannot exceed that of its weakest device. In standard bipolar processes, the vertical NPN transistor usually breaks down before either the lateral or the substrate PNP. The NPN V_{CEO} thus determines the maximum operating voltage. The V_{CEO} of a vertical NPN transistor is related to the avalanche voltage of the planar base-collector junction V_{CBOP} by the equation[27]

$$V_{CEO} = \frac{V_{CBOP}}{\sqrt[n]{\beta_{max}}} \qquad [8.5]$$

where β_{max} represents the peak beta of the device and *avalanche multiplication factor n* usually lies in the range $3 < n < 6$. Low-beta devices have higher V_{CEO} ratings, but nominal betas below about fifty begin to restrict the usefulness of the device. Most processes therefore rely on a high base-collector planar breakdown voltage to obtain adequate V_{CEO} ratings. The width of the drift region determines V_{CBOP}, and thicker epi layers produce correspondingly higher breakdown voltages. The depth of the isolation diffusion must increase to keep pace with the epi thickness, but the other steps in the process remain the same. Manufacturers of standard bipolar processes usually offer several epi thicknesses corresponding to convenient operating voltages, such as 20V, 40V, and 60V. If a process offers a choice of voltage ratings, use the lowest one possible because the higher voltages require larger isolation spacings.

Parasitic channel formation and charge spreading become increasingly serious concerns at higher voltages. Charge spreading may cause low levels of leakage at operating voltages somewhat below the thick-field threshold V_{TF}. Circuits operating at low currents or employing precision matching are especially prone to leakage problems and therefore require careful field plating and channel stopping. At voltages exceeding the thick-field threshold, metallization can directly induce parasitic channel formation and complete field plates and channel stops become mandatory for proper circuit operation (Section 4.3.2). The exposed surface of the base region between the emitter and the collector of a lateral PNP transistor should always be field-plated regardless of operating voltages.

The breakdown voltage of any junction depends on its curvature: the sharper the curvature, the lower the breakdown voltage. All diffused and implanted junctions have a characteristic sidewall curvature. Deeper junctions have less sidewall curvature and therefore exhibit higher breakdown voltages. The effects of sidewall curvature can be quite dramatic. The observed breakdown voltage of a base-collector junction may equal only 60V, even though the planar avalanche voltage exceeds 120V. The observed breakdown voltages of shallow junctions often depend on geometry because the curvature of the corners of the diffusion exceeds the curvature of the sidewalls.[28] The observed breakdown voltages of such diffusions decrease slightly when the geometries contain acute angles. Deeper diffusions are less prone to this effect because outdiffusion rounds the corners off. Junctions with depths greater than 3μm rarely show any significant reduction in operating voltage due to the presence of 90° vertices. Base diffusions with a junction depth of about

---

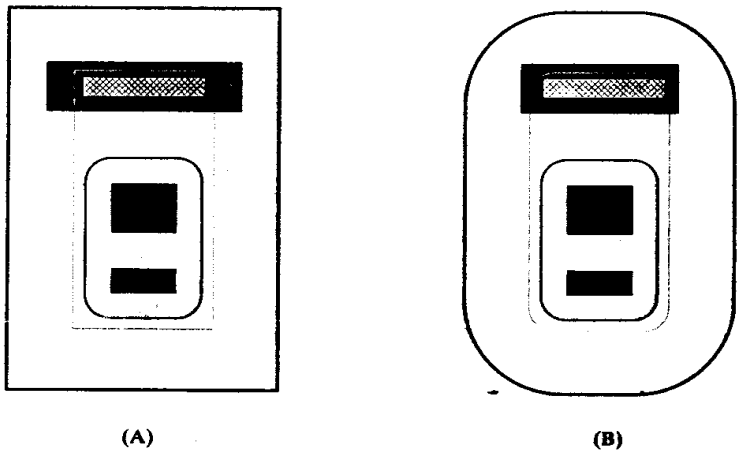
[27] A. Grove, *Physics and Technology of Semiconductor Devices* (New York: John Wiley and Sons, 1967), pp. 230–234.

[28] C. Basavanagoud and K. N. Bhat, "Effect of Lateral Curvature on the Breakdown Voltage of Planar Diodes," *IEEE Electron Device Letters*, Vol. EDL-6, #6, 1985, pp. 276–278.

2μm may experience a reduction of 5 to 10V in breakdown voltage due to 90° vertices, and relatively shallow HSR diffusions may experience even larger reductions.

Designers must sometimes push the operating voltage of base or HSR diffusions close to their respective limits. Under such circumstances, the 90° vertices of rectangular geometries become liabilities. The designer must round off each such corner with a small arc, or *fillet*, to achieve the full operating voltage. The radius of these fillets should exceed the junction depth of the diffusion. Fillets having a radius of 5μm (0.2 mil) serve admirably for both base and HSR diffusions. Both outside and inside corners should receive fillets. Circular geometries, such as fillets, sometimes produce false diagnostics during verification, so some designers use chamfers instead. A *chamfer* is a small diagonal facet drawn perpendicular to a line bisecting the vertex. Chamfers are not quite as effective as fillets because they still contain discernible vertices, although these are less acute than the original corner. If chamfers are used, their lengths should exceed the junction depth of the diffusion.

Figure 8.27 shows two examples of NPN transistors that incorporate fillets. The transistor of Figure 8.27A incorporates fillets only on the corners of the base diffusion. These suffice to obtain the full V_{CBO} and are all that are necessary. Some designers also fillet the transistor tanks, although the isolation diffusion is so deep that these fillets have little or no effect. Up-down isolation may occasionally benefit from fillets because it uses shallower isolation diffusions. Figure 8.27B shows an NPN transistor incorporating fillets on base, NBL, and tank geometries. The larger fillets are customarily drawn concentric with the base fillets to maintain constant spacings.



(A) (B)

FIGURE 8.27 NPN transistors incorporating high-voltage fillets (A) on base only and (B) on base, NBL, and isolation.

High-voltage layouts often require increased spacings between certain diffusions to accommodate wider depletion region widths. Spacings that frequently require modification include base–base, HSR–HSR, base–HSR, base–iso, and HSR–iso. Other spacings that might require adjustment include collector–base, collector–iso, NBL–iso, deep-N+–iso, and deep-N+–base (where the collector is defined as the emitter region around the collector contact). The need for increased spacings for higher-voltage diffusions causes some difficulties in verification. The simplest procedure, and one that will certainly produce a functional design, consists of applying the larger spacing rules to all diffusions. On the other hand, a considerable amount of space can be saved by applying the larger rules only to devices operating at higher voltages. This requires some means of distinguishing between high- and low-

voltage geometries during design rule verification. One technique consists of drawing a geometry on a special layer around low-voltage devices to distinguish them from high-voltage ones. Some designers prefer to code a figure around high-voltage devices rather than around low-voltage ones, but this practice is not recommended. The inadvertent omission of a low-voltage marker will produce errors that become apparent during verification, while the omission of a high-voltage marker may go undetected because it does not produce any design rule violations.

Designers sometimes attempt to push the voltage ratings of devices beyond their specified limits by using special circuit topologies. The most common example of this practice consists of pushing the V_{CEO} of a vertical NPN beyond its rated maximum by ensuring the device's base terminal never sees a high-impedance state. In effect, the circuit designer relies on the V_{CER} rating of the transistor rather than the V_{CEO}. This practice can cause yield problems because V_{CER} ratings are rarely specified or controlled by the wafer fab. Transistors that are pushed beyond their rated V_{CEO} may also latch up during turn-off due to the snap-back of V_{CER} to $V_{CEO(sus)}$ during conduction. Whenever possible, one should use a higher-voltage process rather than attempting to push the ratings of a lower-voltage one.

8.3. ALTERNATIVE SMALL-SIGNAL BIPOLAR TRANSISTORS

Many additional types of bipolar transistors exist. Process extensions allow standard bipolar to fabricate NPN transistors with extremely high betas and lateral PNP transistors with improved high-current performance. Analog BiCMOS offers bipolar transistors with reduced feature sizes that can equal or even surpass the performance of standard bipolar. Advanced bipolar and BiCMOS processes offer extremely high-speed transistors suitable for fast digital logic. These advanced transistors are also useful for constructing ultra-fast amplifiers and comparators. Even an N-well CMOS process can produce a low-gain substrate PNP useful for constructing voltage and current references for CMOS circuits. This section takes a look at all of these alternative devices.

8.3.1. Extensions to Standard Bipolar

The standard bipolar process was perfected in the late 1960s. Since that time, many other processes have been developed that offer superior device performance. Standard bipolar's only real advantages over these other processes are its simplicity and its low cost. Most process extensions proposed for standard bipolar have proven either too costly or too complex to justify their widespread adoption. Two options that have enjoyed some measure of popularity are the *super-beta NPN* and the *deep-P+ lateral PNP*.

Figure 8.28 depicts cross sections of a standard NPN transistor and a super-beta NPN. The super-beta device employs a deeper emitter diffusion that decreases the width of the neutral base to less than 0.1μm. Betas in excess of 5000 are possible,[29] but the thin, lightly doped base punches through at collector-to-emitter voltages of only 1 to 3V, and Early voltages lie in the same range. These limitations restrict super-beta transistors to a few specialized applications. They are, for example, useful for constructing the input differential stages of low-input-current operational amplifiers. Circuits that use super-beta transistors invariably employ regular NPNs as well, so the regular emitter diffusion must remain part of the process flow. Super-beta circuits

[29] W. M. Gegg, J. L. Saltich, R. M. Roop, and W. L. George, "Ion-Implanted Super-Gain Transistors," *IEEE J. Solid-State Circuits*, Vol. SC-11, #4, 1976, pp. 485–491.

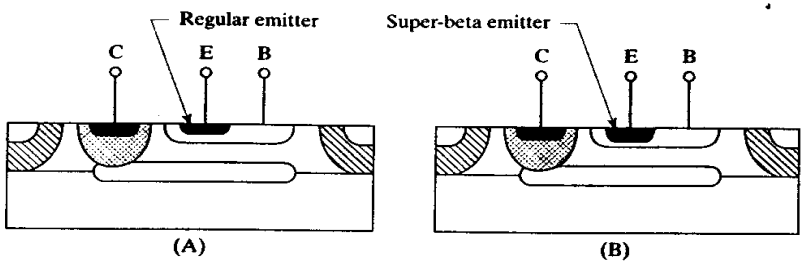


FIGURE 8.28 Comparison of representative cross sections of (A) a standard bipolar NPN and (B) a super-beta NPN.

have now largely been supplanted by BiFET and BiCMOS alternatives, but a few of the older products remain in production due to the low cost of this mature process technology.

Figure 8.29 shows cross sections of a standard bipolar lateral PNP and a corresponding deep-P+ lateral. The latter device employs a special *deep-P+* diffusion that is more heavily doped and deeper than the regular base diffusion. The increase in dopant concentration improves the emitter injection efficiency of the lateral PNP, while the deeper junction ensures that a larger percentage of emitter injection occurs from the sidewalls.[30] The high-current beta of a deep-P+ lateral does not roll off as quickly as that of a base lateral. Deep-P+ laterals can operate at current densities two or three times as large as base laterals. The beta for a typical emitter (10μm diameter) falls to half its peak value at around 200 to 500μA, as compared to 100 to 200μA for a base lateral. Although this increase in performance may seem relatively small, it can shrink the area required for a deep-P+ lateral PNP transistor to half that required by a base lateral, and it can be implemented at modest cost. Any design that contains a power lateral PNP transistor may benefit from the incorporation of deep-P+ lateral PNP transistors.

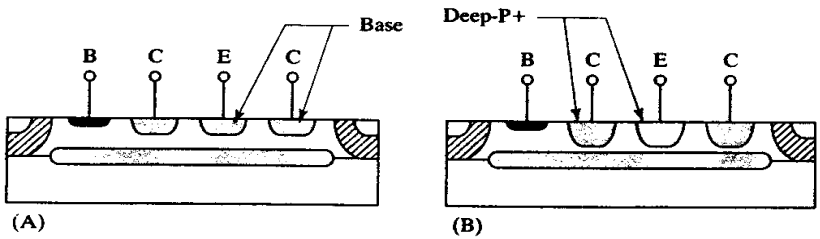


FIGURE 8.29 Comparison of representative cross sections of (A) a standard bipolar lateral PNP and (B) a deep-P+ lateral PNP.

8.3.2. Analog BiCMOS Bipolar Transistors

The analog BiCMOS process discussed in Section 3.3 employs a P-type (100) epi instead of the N-type (111) epi favored by standard bipolar. Analog BiCMOS components must therefore occupy N-wells isolated from one another by regions of P-epi. Except for the substitution of N-wells for N-tanks, the construction of bipolar devices in analog BiCMOS parallels that in standard bipolar.

An unexpected difficulty may arise from the use of N-well to form the collector of the NPN. The graded nature of the well causes the resistivity of its lowest portions to greatly exceed the resistivity of standard bipolar N-epi. The vertical resistance through the analog BiCMOS N-well is therefore much greater than the vertical

[30] B. Murari, "Power Integrated Circuits: Problems, Tradeoffs, and Solutions," *IEEE J. Solid-State Circuits*, Vol. SC-13, #3, 1978, pp. 307–319.

resistance through the standard bipolar N-epi. Unless the transistor includes a deep-N+ sinker, the vertical collector resistance will cause a soft transition from saturation to normal active operation similar to that caused by quasisaturation (Figure 8.30). This soft transition not only causes excessively high saturation voltages, but it also makes the transistor very difficult to accurately model. The inclusion of deep-N+ in the transistor eliminates the soft transition, but only at the cost of significantly increasing device area.

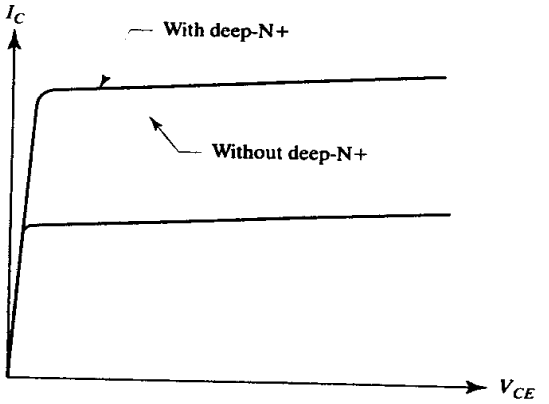


FIGURE 8.30 Comparison of saturation characteristics of CDI NPN transistors with and without the addition of deep-N+ sinkers.

The relatively shallow junction depth of the N-well limits the operating voltage of the CDI NPN transistor to a typical value of 15 to 20V. An alternative structure can provide higher voltages at the cost of reduced safe operating area and poorer Early voltage. Figure 8.31 illustrates the layout and cross section of this *extended-base NPN transistor.*

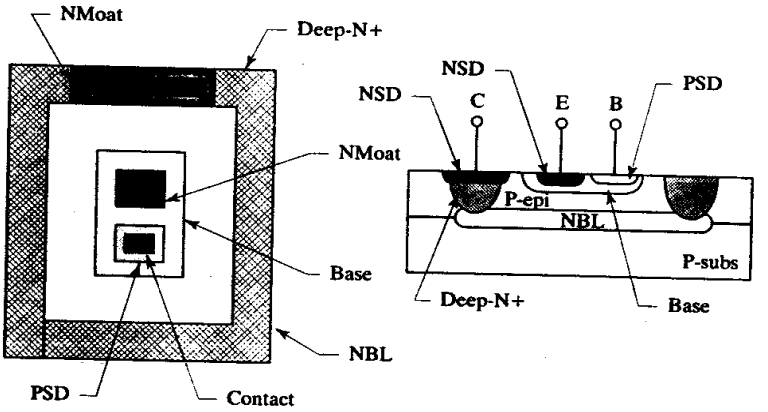


FIGURE 8.31 Layout and cross section of an extended-base NPN transistor.

Unlike the CDI NPN, the extended-base transistor does not employ N-well in its construction. The base of this transistor consists of a combination of a regular base diffusion and an isolated P-epi region, the latter lying sandwiched between the base diffusion and the underlying NBL. A ring of deep-N+ isolates the base from the surrounding P-epi and simultaneously allows contact to the NBL. The collector of this transistor consists of the NBL beneath its extended base and the deep-N+ ring surrounding it. The P-epi portion of the extended base acts as a drift region that greatly increases the operating voltage of this structure. The collector-base depletion region cannot penetrate far into the heavily doped NBL and instead moves

upward through the lightly doped P-epi. Because the depletion region intrudes primarily into the neutral base rather than into the collector, the Early voltage of this structure is somewhat lower than that of the CDI NPN. Although the additional dopant in the P-epi increases the Gummel number of the extended base, the increase is smaller than one might expect because the P-epi is very lightly doped in comparison with the base diffusion. The base diffusion terminates the drift region and restricts the upward growth of the collector-base depletion region, preventing base punchthrough. Multiple-well processes (Section 11.2.2) often use P-well as a substitute for the base diffusion in these devices.

The extended-base structure has a higher planar base-collector breakdown voltage than the CDI NPN and therefore also exhibits a higher V_{CEO} rating, typically ranging from 40 to 60V. The base diffusion can be eliminated entirely, resulting in an *epi-base transistor.* The elimination of the base diffusion greatly reduces the Gummel number of the device. The epi-base transistor offers a beta of several hundred at the cost of reduced operating voltage (due to punchthrough) and increased base resistance. One advantage of the epi-base transistor is that it does not require a separate base diffusion.

If the deep-N+ isolation ring is replaced by a similar N-well ring, the resulting device requires only one mask step that does not normally appear in the analog CMOS process flow. The resulting transistors can handle only relatively small currents due to their high collector resistances, but they still offer much better low-current beta characteristics and device matching than any other bipolar transistors compatible with CMOS processing (Section 8.3.3). Unfortunately, pure CMOS processes rarely offer an NBL layer suitable for the construction of an epi-base device.

The analog BiCMOS substrate PNP (Figure 3.50) employs an emitter constructed of PSD rather than base, because the larger junction depth of the base diffusion reduces the punchthrough voltage of the transistor. The performance of the PSD substrate PNP roughly equals that of a substrate constructed in standard bipolar. Since substrate PNP transistors inject current into the substrate, the designer must take precautions to avoid substrate debiasing (Section 4.4.1).

Lateral PNP transistors constructed in analog BiCMOS exhibit surprisingly high peak betas. The relative shallowness of the base diffusion allows the emitter and collector of the transistor to be placed in close proximity to one another. Simultaneously, the graded nature of the well (aided by the presence of a phosphorus channel stop implant) helps increase the punchthrough voltage near the surface where the base is narrowest. Most of the minority-carrier injection occurs deeper in the transistor, where the built-in potential of the base-emitter junction decreases due to the graded nature of both the well and the base diffusion, so the presence of higher surface doping levels does not unduly increase the Gummel number of the transistor. The graded nature of the well—again aided by the phosphorus channel stop—generates an electric field that forces minority carriers down and away from the oxide-silicon interface. Carriers overcoming this electric field still experience relatively low levels of surface recombination due to the low surface state charge of (100) silicon. Finally, the small feature sizes possible with the shallow diffusions and superior photolithography of analog BiCMOS allow the construction of very small emitters (typically 5μm in diameter). This increases the proportion of minority carriers injected from the emitter sidewall, while at the same time reducing the distance traveled by those carriers. All of these factors act in concert to raise the peak beta of the lateral PNP so much that it may exceed that of the CDI NPN. This high peak beta also helps extend the usable current range, allowing each minimum emitter to conduct as much as 100μA while retaining a beta of twenty. The smaller cell size of

the analog BiCMOS lateral PNP enables the construction of very area-efficient power lateral PNP transistors even without the aid of a deep-P+ diffusion.

The size of the emitter has a strong effect on the beta of an analog BiCMOS lateral PNP, primarily because carriers emitted from the bottom surface of the emitter must travel farther than those emitted from the sidewalls. Additionally, the graded well doping generates a weak electric field that causes minority carriers to drift toward the NBL/N-well interface. This field prevents the interface from reflecting minority carriers toward the collector as efficiently as would otherwise occur. Worse yet, the NBL layers used in analog BiCMOS processes are often relatively lightly doped to minimize lateral autodoping, while the wells in low-voltage processes are more heavily doped to prevent punchthrough. The reduced doping difference decreases the built-in field at the NBL/N-well interface, allowing carriers to penetrate into the NBL, where they recombine or travel to the substrate. Analog BiCMOS lateral PNP transistors should always employ minimum-size emitters to ensure the highest possible gain. Larger transistors should use arrayed emitters rather than elongated ones for the same reason.

PSD implants can also form the emitter and collector of a lateral PNP transistor. The shallowness of the PSD implant diminishes the size of the transistor, but it also reduces the collector efficiency. The source/drain implants are so shallow that a substantial percentage of the minority carriers travel underneath them and escape to the sidewalls. This problem is exacerbated by the recessed thick-field oxide and N-type channel stop implants that shadow the PSD collector, and by the graded nature of the well that imposes a downward drift on minority carriers. If PSD laterals must be used, their collector efficiency can be increased by widening the collector or by ringing the transistor with deep-N+.

Lateral PNP transistors constructed in analog BiCMOS do not require base field-plating unless the operating voltage of the transistor exceeds the thick-field threshold. The presence of the phosphorus channel stop implant produces a built-in potential that repels minority carriers from the surface, and the use of (100) silicon minimizes surface recombination experienced by minority carriers that do reach the surface. Leakages and beta variations are therefore much reduced in analog BiCMOS. The elimination of base field plates helps to reduce the overall size of the transistor, making the lateral PNP a more attractive component.

8.3.3. Bipolar Transistors in a CMOS Process

Straight CMOS processes offer few options for constructing bipolar transistors. The only bipolar device available in an N-well CMOS process is a substrate PNP. This transistor uses the same diffusions as the corresponding analog BiCMOS device, consisting of PSD for the emitter, N-well for the base, and the P-substrate for the collector. Unfortunately, this device rarely performs as well as its analog BiCMOS counterpart. The main reasons for the differences in performance are differences in N-well doping and the use of shallow clad moats.

The N-well doping profile of analog BiCMOS represents a compromise between the profiles best suited to bipolar transistors and those best suited to MOS devices. The CDI NPN requires a lightly doped well to act as its drift region. This same well forms a suitable base region for lateral and substrate PNP transistors, and it can serve as the backgate for relatively long-channel, high-voltage PMOS transistors. Modern CMOS processes usually have channel lengths that are much shorter than one micron. Higher well dopings are required to prevent punchthrough of these short channels. The heavier well doping increases the Gummel number of the substrate PNP and therefore reduces its gain. Many CMOS processes also alter the well dopant profile to increase the doping concentration beneath the surface in order to

prevent vertical punchthrough and to increase latchup immunity. The increased subsurface doping associated with such *retrograde* well profiles increases the Gummel number of the substrate PNP, reducing its beta still further.

The extremely shallow source/drain implants required to produce short-channel MOS transistors cause a decrease in the emitter injection efficiency of bipolar transistors, probably because of recombination at the contact and increased surface recombination effects. Clad moats are particularly troublesome because silicidation consumes some of the underlying silicon and further thins the already-shallow source/drain diffusions. The effects of reduced emitter injection efficiency include reduced peak beta and a more pronounced low-current beta rolloff, often resulting in a peaked beta curve reminiscent of that of a standard bipolar lateral PNP transistor.

The gain of substrate PNP transistors constructed in submicron CMOS can vary from less than one to more than fifty. Transistors with gains of less than five exhibit extremely poor matching. Small variations in beta produce corresponding variations in base current. Since the base resistance of these transistors is relatively large and the base current is comparable to the collector current, any variation in base current produces a corresponding variation in base-emitter voltage, V_{BE}. Collector current mismatches of ±3 to 5% are not uncommon in low-gain CMOS substrate transistors. This problem becomes most severe in devices whose beta lies near unity; devices with extremely small betas behave as diodes and therefore largely avoid this problem.

If at all possible, a silicide block mask should be coded over the emitter regions of the transistor. A small silicided area underneath each emitter contact is still necessary, but this does not degrade the gain of the transistor nearly as much as full moat cladding. The transistor should consist of minimum-width strips of NSD interdigitated with similar strips of PSD (Figure 8.32). This arrangement reduces the base resistance of the transistor and helps minimize the influence of base current variations on device matching. The collector current of this transistor should be kept as small as possible to further reduce voltage drops in the neutral base. Current densities of 1 to 10μA/mil^2 (1.5–15nA/μm^2) are generally advisable for transistors used as part of voltage or current references.

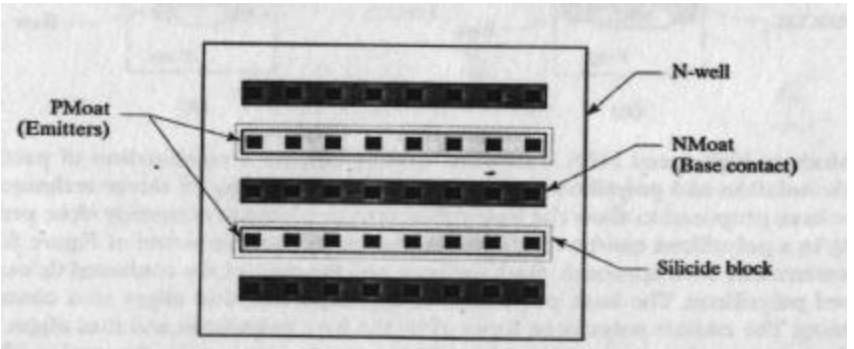


FIGURE 8.32 Layout of a substrate PNP transistor compatible with an N-well CMOS process.

Some designers have attempted to construct lateral PNP transistors in straight CMOS processes. These devices have generally proved unsatisfactory because of a combination of low gain and low collector efficiency. The effective beta of many of these devices does not exceed unity, and the effects of beta variation become so severe that even minimal matching cannot be maintained. The use of polysilicon to generate a self-aligned base region has enabled the fabrication of devices with

usable betas, but even these devices exhibit excessive beta variation.[31] A superior alternative to substrate and lateral PNP transistors can be fabricated with the addition of one process step. The presence of NBL allows the construction of epi-base transistors that use N-well isolation rings. These transistors cannot handle large currents because of their high collector resistance, but they greatly outperform the other transistors available in CMOS processing. Unfortunately, few pure CMOS processes offer any form of buried layer, and epi-base transistors are usually found only in stripped-down analog BiCMOS processes. Low-voltage, multiple-well processes can sometimes construct a *retrograde-well NPN transistor.* This device uses a retrograde N-well (or an N-well incorporating a punchthrough stop implant), as discussed in Section 11.1.2. The retrograde N-well is counterdoped with a shallower P-well to form the base region of the transistor. This device is otherwise similar to a P-well epi-base transistor (Section 8.3.2).

8.3.4. Advanced-technology Bipolar Transistors

All of the bipolar transistors discussed up to this point are relatively slow. Three factors limit switching speed: junction capacitance, base resistance, and neutral base width. The base-emitter, base-collector, and collector-substrate junction capacitances must be charged and discharged in order to switch the transistor. The resistance of the neutral base limits the charging and discharging rates of these capacitors and thereby restricts the maximum switching speed of the device. Assuming that these limitations can somehow be overcome, the transit time of minority carriers across the neutral base still sets a fundamental limit on the maximum switching speed of the transistor.

The simplest way to reduce junction capacitances consists of minimizing junction areas. Improved photolithography allows smaller drawn geometries, which translate directly into smaller junction capacitances and faster switching speeds. Further improvements can be achieved by redesigning the transistor to eliminate unnecessary overlaps and spacings. For example, the conventional transistor of Figure 8.33A requires the emitter diffusion to overlap the emitter contact to allow for misalignment. The *washed-emitter transistor* of Figure 8.33B eliminates this overlap, substantially reducing the size of the emitter. This in turn allows the use of a smaller base, leading to additional performance improvements.

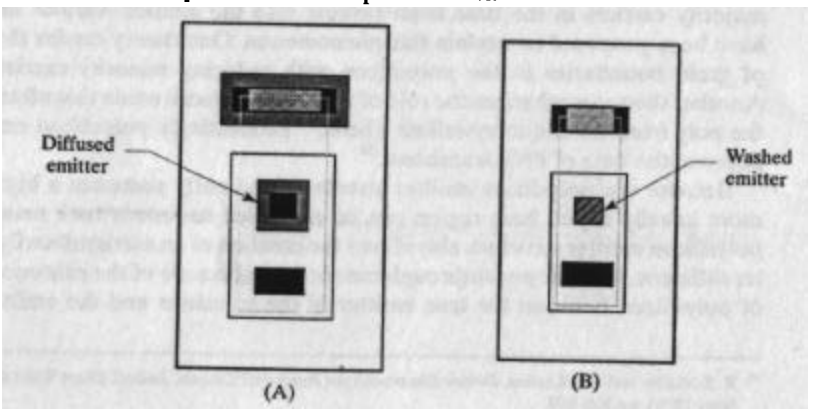


FIGURE 8.33 Comparison between (A) a conventional diffused emitter and (B) a washed emitter.

[31] E. A. Vittoz, "MOS Transistors Operated in the Lateral Bipolar Mode and Their Application in CMOS Technology," *IEEE J. Solid-State Circuits*, Vol. SC-18, #3, 1983, pp. 273–279.

A *washed emitter* consists of an emitter diffusion self-aligned to a contact opening by means of a special etching technique. The emitter is deposited using a conventional deposition and drive during which a thin emitter oxide forms. The emitter regions remain covered by photoresist during the contact oxide removal. After all of the other contacts have been opened and the photoresist has been removed, a brief, carefully timed etch strips the thin emitter oxide and forms emitter contacts. Outdiffusion of the emitter under the original oxide opening provides just sufficient overlap of emitter over contact to prevent base-emitter shorts.[32]

Even better results can be obtained by using a ***polysilicon emitter.*** Figure 8.34 shows an example of a CDI NPN fitted with a simple polysilicon emitter. This transistor is processed normally up through the completion of the base drive. Next, an oxide removal defines the extent of the drawn emitter. A layer of arsenic-doped polysilicon is deposited and patterned over the exposed emitter opening. A brief period of heating causes arsenic to diffuse from the polysilicon into the exposed monocrystalline silicon, producing an extremely thin and heavily doped emitter diffusion that self-aligns to the emitter oxide removal. This process has the same effect as using a washed emitter does, but it allows the formation of much thinner and more precisely controlled emitter diffusions.

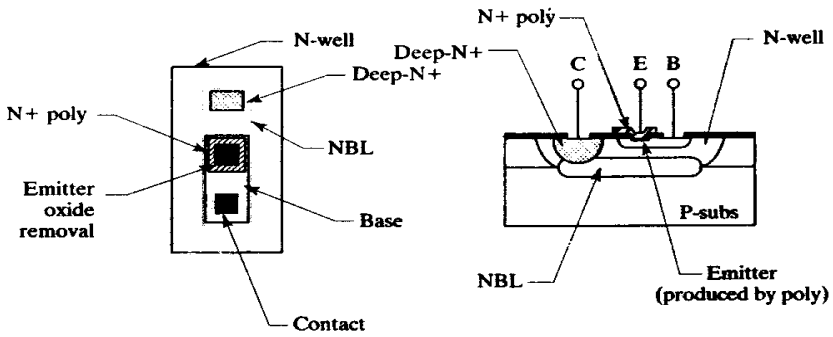


FIGURE 8.34 Layout and cross section of a CDI NPN transistor with a polysilicon emitter.

Polysilicon emitter transistors exhibit betas up to six times greater than those of equivalent diffused-emitter structures. The polysilicon emitter structure is believed to increase the emitter injection efficiency of the transistor by somehow preventing majority carriers in the base from flowing into the emitter. Various mechanisms have been proposed to explain this phenomenon. One theory credits the presence of grain boundaries in the polysilicon with reducing minority carrier mobility. Another theory emphasizes the role of the thin interfacial oxide that often separates the poly from the monocrystalline silicon.[33] Interestingly, polysilicon emitters also improve the beta of PNP transistors.[34]

Because the polysilicon emitter structure inherently possesses a higher beta, a more heavily doped base region can be employed to reduce base resistance. The polysilicon emitter structure also allows the creation of an extraordinarily thin emitter diffusion. Emitter punchthrough cannot occur because of the existence of a layer of polysilicon between the true emitter of the transistor and the emitter contact.

---

[32] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed. (New York: John Wiley and Sons, 1986), pp. 306–307.

[33] Z. Yu, B. Ricco, and R. Dutton, "A Comprehensive Analytical and Numerical Model of Polysilicon Emitter Contacts in Bipolar Transistors," *IEEE Trans. Electron Devices*, Vol. ED-31, 1984, pp. 773–784.

[34] C. M. Maritan and N. G. Tarr, "Polysilicon Emitter p-n-p Transistors," *IEEE Trans. on Electron Devices*, Vol. ED-36, #6, 1989, pp. 1139–1143.

Furthermore, the depth of the emitter junction can be controlled with great precision. These factors allow the use of a much thinner base than might otherwise be possible, and the resulting reduction in neutral base width translates into a smaller minority-carrier transit time and a much faster transistor. The thinner base also allows the use of a thinner epi, drastically reducing outdiffusion of deep-N+ and N-well and therefore greatly shrinking the size of the transistor. The benefits of polysilicon emitters become readily apparent if one compares the transistors of Figures 8.33 and 8.34.

Additional reductions in device area and junction capacitance can be achieved through the use of partial oxide isolation. LOCOS processing can oxidize entirely through a thin epitaxial layer to separate adjacent tanks without using isolation diffusions. The elimination of sidewall junctions not only reduces collector-substrate capacitance, but also shrinks tank dimensions. Figure 8.35A illustrates the layout and cross section of an NPN transistor in which the base diffusion abuts the oxide isolation. The elimination of base-isolation spacings produces a corresponding reduction in device area. Figure 8.35B shows a more radical structure in which the emitter also abuts the oxide isolation. Such *walled-emitter* structures generally require a more abrupt termination of the oxide region than conventional LOCOS processing can provide. Various modifications of the LOCOS technique can reduce the width of the bird's beak,[35] and special anisotropic etch techniques can entirely eliminate it.

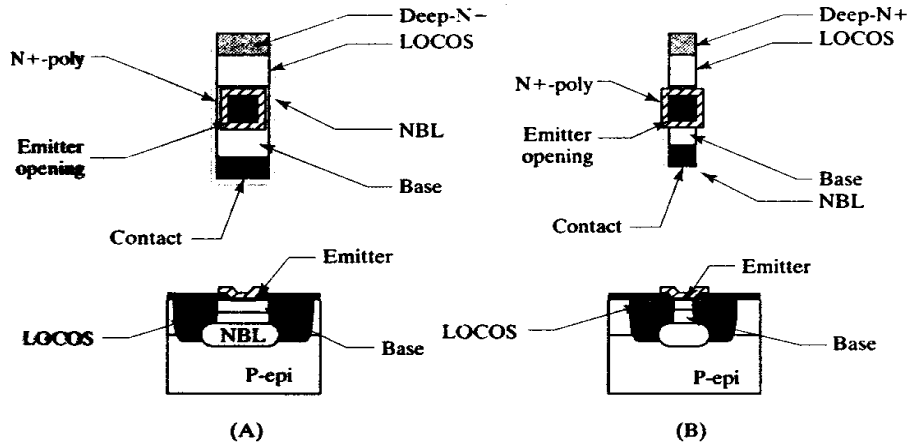


FIGURE 8.35 Partial oxide-isolated NPN transistors: (A) conventional, and (B) walled-emitter.

Modern, high-speed NPN transistors usually employ a combination of partial oxide isolation and polysilicon emitter technology. A variety of clever techniques have been proposed to allow the base contacts to be placed in extremely close proximity to a polysilicon emitter. The *super self-aligned* (SSA) structure of Figure 8.36 represents one such approach. Both the base and the emitter are contacted through doped polysilicon. The base polysilicon is deposited first and aligns to a contact opening. The emitter polysilicon forms after the base polysilicon and thus aligns to it. Both base and emitter ultimately self-align to the same contact opening. This structure can achieve switching times of less than 250pS, allowing the construction

[35] K. Y. Chiu, J. L. Moll, and J. Manoliu, "A Bird's Beak Free Local Oxidation Technology Feasible for VSLI Circuits Fabrication," *IEEE Trans. on Electron Devices*, Vol. ED-29, #4, 1982, pp. 536–540.

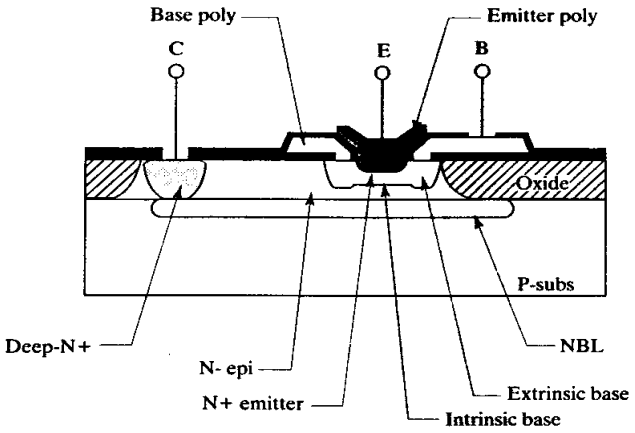


FIGURE 8.36 Cross section of a super self-aligned transistor.

of very fast logic circuits.[36] On the other hand, the complexity of this process limits its use to applications that demand speed at any price. Most analog BiCMOS processes target lower levels of performance to minimize processing costs. For example, the polysilicon-emitter CDI transistor in Figure 8.34 requires the same number of masks as a conventional CDI NPN, but it offers much smaller device sizes and lower junction capacitances.

8.4 SUMMARY

Bipolar transistors are extremely versatile devices, but they have their limitations. Improperly designed bipolar transistors may suddenly fail under heavy loads due to thermal runaway or secondary breakdown. Saturating bipolar transistors can also inject current into the substrate, debiasing adjacent circuitry and causing catastrophic latchup failures. These problems have caused circuit designers much anguish, but they can almost always be overcome by proper circuit design and device layout.

Bipolar transistors fall into one of two general categories: small-signal transistors and power transistors. Small-signal transistors are optimized for dense packing rather than for power handling capability. These devices are primarily used in analog signal processing and control circuitry, where they are particularly prized for their high transconductance and superior device matching. Standard bipolar processes offer relatively high-performance vertical NPN transistors plus several varieties of somewhat less useful PNP transistors. BiCMOS processes generally incorporate a similar selection of bipolar devices. The increasing popularity of analog BiCMOS ensures that bipolar transistors will remain an important part of analog circuit design for the foreseeable future.

The layout of bipolar transistors can be tailored to suit specific applications. Bipolar power transistors are frequently designed for improved immunity to thermal runaway and secondary breakdown. Small-signal transistors are frequently laid

[36] S. Konaka, Y. Yamamoto, and T. Sakai, "A 30-ps Si Bipolar IC Using Super Self-Aligned Process Technology," *IEEE. Trans. on Electron Devices*, Vol. ED-33, 1986, pp. 526–531. Also see T. Y. Chiu, G. M. Chin, M. Y. Lau, R. C. Hanson, M. D. Morris, K. F. Lee, M. T. Y. Liu, A. M. Voschenkov, R. G. Swartz, V. D. Archer, S. N. Finegan, and M. D. Feuer, "The Design and Characterization of Nonoverlapping Super Self-Aligned BiCMOS Technology," *IEEE Trans. Electron Devices*, Vol. 38, #1, 1991, pp. 141–150.

out to minimize device mismatches. The following chapter examines the techniques used to create these optimized bipolar transistor layouts.

8.4. EXERCISES

Refer to Appendix C for layout rules and process specifications.

8.1. What is the magnitude of the thermal voltage V_T at –55°C? At 125°C?

8.2. A lateral PNP transistor exhibits an emitter current of 110μA, a collector current of 98μA, and a base current of 7μA. What is the transistor's current gain and its collection efficiency?

8.3. Select base-ballasting resistors for the circuit in Figure 8.7. Assume I_{BIAS} = 100μA, and design the base-ballasting resistors so that a transistor in deep saturation will not consume more than 10% of I_{BIAS}. Assume that the V_{BE} of an NPN drops by a maximum of 100mV when it enters deep saturation.

8.4. Lay out the circuit in Figure 8.7 using the resistor values computed in Exercise 8.3 and the standard bipolar layout rules in Appendix C. Assume that all three transistors have minimum emitter areas. Use 6μm HSR resistors for R_1, R_2, and R_3. Place Q_1, Q_2, and Q_3 alongside one another and interdigitate the base ballasting resistors for proper matching. Justify your choice of tank biasing for R_1, R_2, and R_3.

8.5. Lay out a minimum-size standard-bipolar NPN transistor, including a deep-N+ sinker. Construct a similar device, but omit the sinker. Allow space for all necessary metallization. Assuming that the areas of the two devices equal the areas of their respective tanks, what percentage area reduction does the omission of deep-N+ produce?

8.6. Lay out a standard-bipolar, stretched-collector CEB transistor that allows one minimum-width lead to pass between its collector and emitter. The transistor should have a minimum-size emitter. Minimize the collector resistance, assuming the process uses a thick emitter oxide.

8.7. Lay out a standard-bipolar, narrow-emitter transistor having four minimum-width emitter fingers, each 100μm long. Place a deep-N+ sinker along one side of the transistor, and make sure that NBL fully encloses the sinker to minimize collector resistance. Include all necessary metallization to allow connection of the transistor into a circuit.

8.8. Construct minimum-size, standard-bipolar substrate PNP transistors using the standard, emitter-ringed, and verti-lat layout styles. Allow room for all necessary metallization. For the verti-lat transistor only, provide an emitter field plate overlapping the collector by 2μm.

8.9. Construct a split-collector lateral PNP transistor containing four quarter-sized collectors arranged around a minimum circular emitter. Allow room for all necessary metallization, including an emitter field plate overlapping the collectors by 2μm.

8.10. Construct a power PNP transistor consisting of sixteen minimum-size circular-emitter cells arranged in a 4 × 4 pattern. Include sufficient collector contacts to ensure that each cell resides adjacent to at least one collector contact. Include at least a minimum-size deep-N+ sinker in the base contact. Allow room for all necessary metallization, including emitter field plates overlapping the collectors by 2μm.

8.11. Construct a set of merged PNP transistors occupying the same tank. One transistor should have an elongated emitter 25μm long, while the second transistor should consist of two half-size collectors arranged around a minimum-size circular emitter. Allow room for all necessary metallization, including emitter field plates overlapping the collectors by 2μm.

8.12. Modify the layout in Exercise 8.11 to field plate the collector of the elongated transistor and one of the two collectors of the split-collector transistor. Overlap the field plate over the collectors by at least 6μm. Use flanges to elongate all channels as far as possible without enlarging the tank.

8.13. Construct a high-voltage NPN transistor using the standard bipolar layout rules. Assume the transistor uses a deep-N+ sinker and that it has a minimum-size emitter. Use 4μm fillets on the base diffusion and include all necessary field plates and channel stops.

8.14. Lay out an extended-base NPN transistor using the analog BiCMOS layout rules. Overlap the NBL 5μm into the deep-N+ isolation ring to ensure an adequate seal between the two. Construct a 64μm^2 emitter and allow for all necessary metallization.

8.15. Construct a CMOS substrate PNP transistor containing two emitter fingers that are each 30μm long. Assume that the process only silicides contacts and therefore no silicide block mask is required.