BIPOLAR TRANSISTORS

TRANSISTORS, BIPOLAR

The basic concept of the bipolar junction transistor *(BJT)* was patented by Shockley in 1947 (1), but the BJT was not experimentally realized until 1951 (2). Unlike the point contact transistor demonstrated earlier in 1947, the BJT can be completely formed *inside* the semiconductor crystal and thus it proved to be more manufacturable and reliable, and better suited for use in integrated circuits. In a real sense, the BJT was the device that launched the microelectronics revolution and, hence, spawned the *Information Age.* Until the widespread emergence of complementary metal oxide semiconductor (*CMOS*) technology in the 1980s, the BJT was the dominant semiconductor technology in microelectronics, and even today represents a significant fraction of the global semiconductor market.

At its most basic level the BJT consists of two back-to-back *pn* junctions (*p-n-p* or *n-p-n* depending on the doping polarity), in which the intermediate *n* or *p* region is made as thin as possible. In this configuration the resultant 3 terminal (emitter-base-collector) device exhibits current amplification (current gain) and thus acts as a "transistor" which can be used to build a wide variety of electronic circuits. Modern applications of the BJT are varied, and range from high-speed digital integrated circuits in mainframe computers, to precision analog circuits, to radio frequency (*RF*) circuits found in radio communications systems.

Compared to CMOS, the BJT exhibits higher output current per unit length, larger transconductance (g_{m}) per unit length, faster switching speeds (particularly under capacitive loading), and excellent properties for many analog and RF applications (e.g., lower 1/*f* and broadband noise). Today, frequency response above 50 GHz and circuit switching speeds below 20 ps are readily attainable using conventional fabrication techniques. The primary drawback of BJT circuits compared to CMOS circuits lies in their larger dc power dissipation and increased fabrication complexity, although in applications requiring the fastest possible switching speeds, the BJT remains the device of choice. Figure 1 shows unloaded emitter-coupled-logic (*ECL*) gate delay for today's technology and indicates that state-of-the-art BJT technology is rapidly approaching 10 ps switching times.

In this article we review the essentials of modern bipolar technology, the operational principles of the BJT, second-order high-injection effects, issues associated with further technology advancements, and some future directions. Interested readers are referred to Refs. 3–5 for review articles on modern BJT technology, and to Ref. 6 for an interesting historical perspective on the development of the BJT.

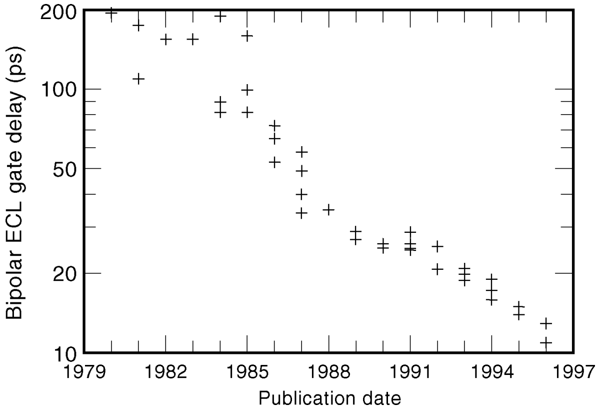


Figure 1. Unloaded emitter-coupled logic (ECL) gate delay (as a function of publication date) showing the rapid decrease in delay with technology evolution.

DOUBLE-POLYSILICON BIPOLAR TECHNOLOGY

In contrast to the depictions commonly found in many standard electronics textbooks, BJT technology has evolved radically in the past 15 years, from double-diffused, large geometry, non-self-aligned structures to very compact, self-aligned, "double-polysilicon" structures. Figure 2 shows a schematic cross section of a modern double-polysilicon BJT. This device has deep-trench and shallow trench isolation to separate one transistor from the next, a p^+ polysilicon extrinsic base contact, an n^+ polysilicon emitter contact, and an ion-implanted intrinsic base region. The two polysilicon layers (hence the name *double-polysilicon*) act as both diffusion sources for the emitter and extrinsic base dopants as well as low-resistance contact layers. In addition, to form the active region of the transistor, a "hole" is etched into the p^+ polysilicon layer, and afterwards a thin dielectric "spacer" oxide is formed. In this manner, the emitter and extrinsic base regions are fabricated without the need of an additional lithography step ("self-aligned"), thereby dramatically reducing the size of the transistor and hence the associated parasitic resistances and capacitances of the structure. The first double-polysilicon BJT structures appeared in the early 1980s (7, 8) and today completely dominate the high-performance BJT technology market. The reader is referred to Refs. 9–15 for specific BJT technology examples in the recent literature.

The doping profile from the intrinsic region of a state-of-the-art double-polysilicon BJT is shown in Fig. 3. The transistor from which this doping profile was measured has a peak cutoff frequency of about 40 GHz (14), and is typical of the state-of-the-art. The emitter polysilicon layer is doped as heavily as possible with arsenic or phosphorus, and given a sort rapid-thermal-annealing (*RTA*) step to out-diffuse the dopants from the polysilicon layer. Typical metallurgical emitter-base junction depths range from 25 to 45 nm in modern BJT technologies. The collector region directly under the active region of the transistor is formed by local ion-implantation of phosphorus. A collector doping of about 1×10^{17} cm^{-3} at the base-collector junction is adequate to obtain a peak cutoff frequency of 40 GHz at a collector-to-emitter breakdown voltage ($\mathrm{BV}_{\mathrm{CEO}}$) of about 3.5 V, consistent with the needs of digital ECL circuits. The intrinsic base region is also formed by low energy ion-implantation of boron. Resultant base widths range from about 60 nm to 150 nm at the state-of-the-art, with peak base doping levels in the range of $3\text{–}5 \times 10^{18}$ cm^{-3}. A traditional (measurable) metric describing the base profile in

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 2007 John Wiley & Sons, Inc.

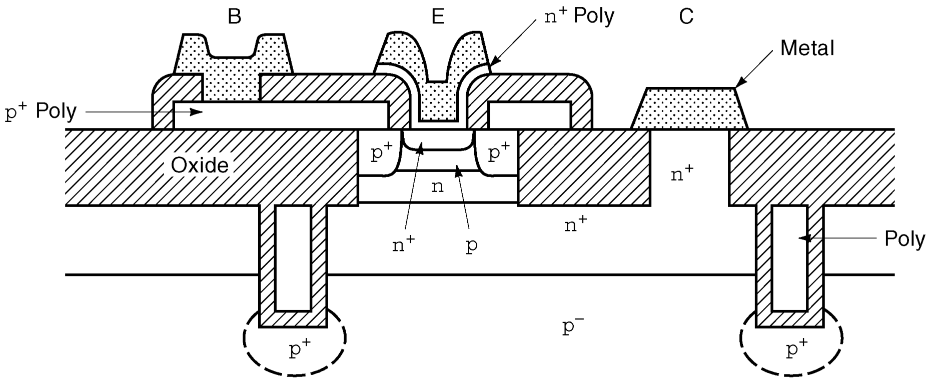


Figure 2. Schematic device cross section of a modern double-polysilicon self-aligned bipolar transistor.

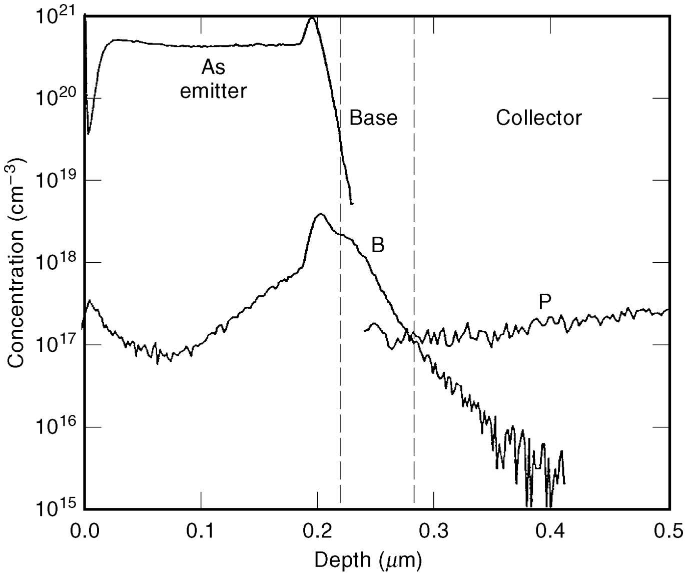


Figure 3. Measured secondary ion mass spectroscopy (SIMS) doping profile from an ion-implanted base bipolar technology with a 40 GHz peak cutoff frequency (14).

a BJT is the intrinsic base sheet resistance (R_{bi}), which can be written in terms of the integrated base doping (N_{ab}) according to

$$R_{bi} = \left[q \int_0^{W_b} \mu_{pb}(x) N_{ab}(x)\, dx \right]^{-1} \tag{1}$$

In Eq. (1), μ_{pb} is the position-dependent hole mobility in the base and W_{b} is the neutral base width. Typical R_{bi} values in modern BJT technologies range from 10–15 kW/*f*.

THEORY OF OPERATION

Basic Physics

The BJT is in essence a barrier-controlled device. A voltage bias is applied to the emitter-base junction such that we modulate the size of the potential barrier seen by the electrons moving from emitter to base, and thus can (exponentially) modulate the current flowing through the transistor. To best illustrate this process, we have used a 1-dimensional device simulator called SCORPIO (16). SCORPIO is known as a "drift-diffusion" simulator because it solves the electron and hole drift-diffusion transport equations self-consistently with Poisson's equation and the electron and hole current-continuity equations (see, for example, Ref. 6 for a formulation of these equations and the inherent assumptions on their use). These five equations, together with the appropriate boundary conditions completely describe the BJT.

Figure 4 depicts a "toy" doping profile of the ideal BJT being simulated. Both the layer thicknesses and doping levels are consistent with those found in modern BJTs, although the constancy of the doping profile in each region is idealized and hence unrealistic. Figure 5 shows the resultant electron energy band diagram of this device at zero-bias (equilibrium). The base potential barrier seen by the electrons in the emitter is clearly evident. The equilibrium carrier concentrations for each region are shown in Fig. 6. The majority carrier densities are simply given by the doping level in each region, while the minority carrier densities are obtained by use of the "law of mass action" according to the following:

$$p_{e0} = \frac{n_{ie}^2}{N_{de}}:\ n_{e0} = N_{de} \qquad \text{emitter} \tag{2}$$

$$p_{b0} = N_{ab}:\ n_{b0} = \frac{n_{ib}^2}{N_{ab}} \qquad \text{base} \tag{3}$$

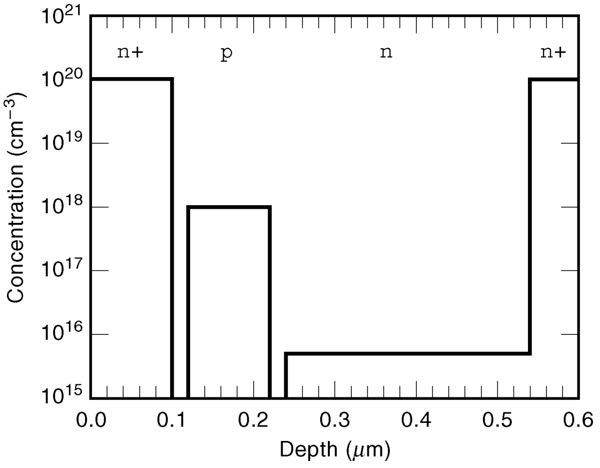


Figure 4. Doping profile of a hypothetical bipolar transistor used in the one-dimensional SCORPIO simulations.

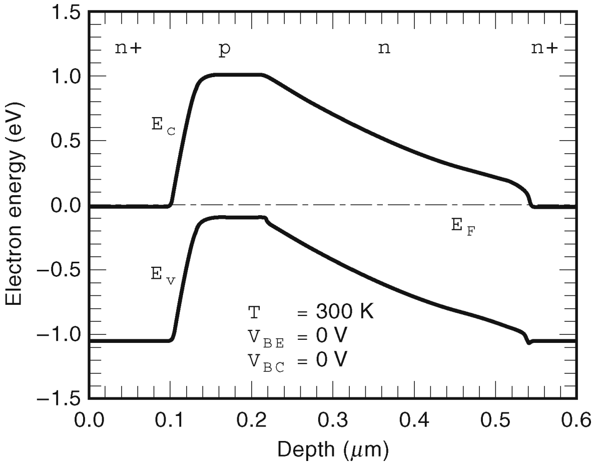


Figure 5. Simulated zero-bias energy band diagram of the hypothetical bipolar transistor depicted in Fig. 4.

$$p_{c0} = \frac{n_{io}^2}{N_{dc}}; \; n_{c0} = N_{dc} \qquad \textbf{collector} \tag{4}$$

In these equations, n_{io} is the intrinsic carrier density, the subscripts e, b, and c represent the emitter, base, and collector regions, respectively, N is the doping density, and,

$$n_{ie}^2 = n_{io}^2 e^{\Delta E_{ge}^{app}/kT} = N_C N_V e^{-E_g/kT} e^{\Delta E_{ge}^{app}/kT} \tag{5}$$

$$n_{ib}^2 = n_{io}^2 e^{\Delta E_{gb}^{app}/kT} = N_C N_V e^{-E_g/kT} e^{\Delta E_{gb}^{app}/kT} \tag{6}$$

where $\Delta E^{app}{}_{ge}$ and $\Delta E^{app}{}_{gb}$ represent the heavy-doping-induced apparent bandgap narrowing (17). The resultant collector current density (J_C) and base current density (J_B) from this structure are shown in Fig. 7. Observe that the BJT exhibits useful current gain ($\beta = J_C/J_B$) over a wide operating range.

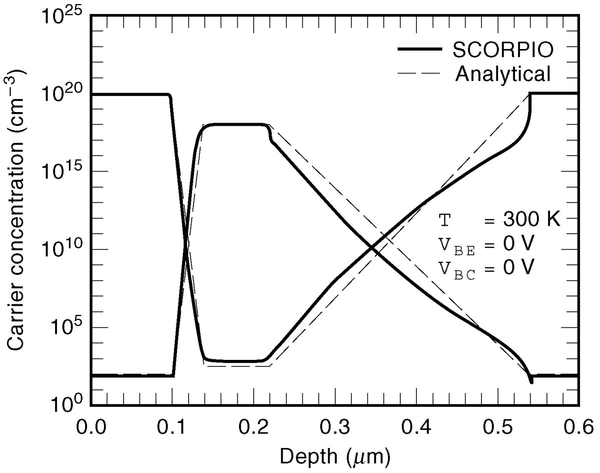


Figure 6. Simulated electron and hole concentrations of the hypothetical bipolar transistor depicted in Fig. 4. Also shown are analytical calculations.

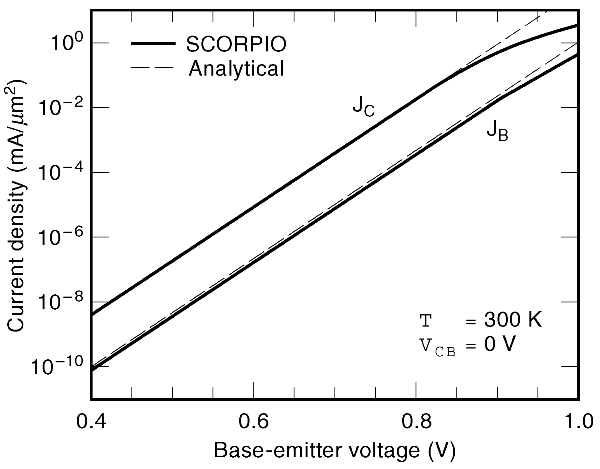


Figure 7. Simulated collector and base current densities as a function of emitter-base bias. Also shown are analytical calculations.

The basic operational principles of the BJT can be described as follows. If we imagine forward-biasing the emitter-base junction, and reverse-biasing the base-collector junction (i.e., forward-active mode), electrons from the heavily doped emitter are injected into and diffuse across the base region and are collected at the collector contact, thereby giving rise to a useful collector current. At the same time, if the base region is thin enough, the base current consists primarily of the back-injected hole current from base to emitter. Because the emitter is doped heavily with respect to the base, the ratio of forward-injected (emitter to base) electron current to back-injected (base to emitter) hole current is large (roughly equal to the ratio of emitter to base doping), and the BJT exhibits useful current gain. It is critical that the intermediate base region be kept as thin as possible because a) we do not want electrons traversing the base to have sufficient time to recombine with holes before they reach the collector contact, and b) the transit time of the electrons through the base typically limits the frequency response and, hence, the speed of the transistor. In the forward-active mode, a schematic representation of the magnitude of the various currents flowing in an ideal BJT is illustrated in Fig. 8 (6).

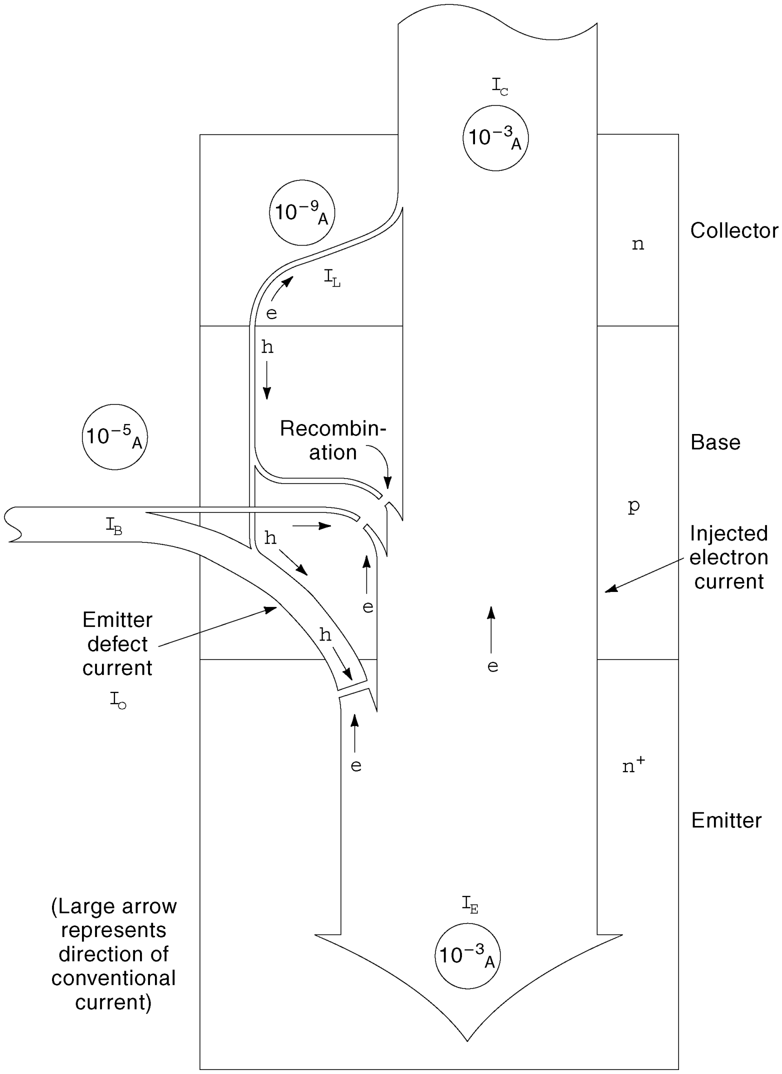


Figure 8. Schematic current flow distributions in a realistic bipolar transistor (6).

Current-Voltage Characteristics

For simplicity, we will limit this discussion to the currents flowing in the BJT under forward-active bias. Other bias regimes (e.g., saturation) are not typically encountered in high-speed circuits such as ECL. The reader is referred to Ref. 17–19 for a discussion of other operating regimes. In this case, for a BJT with a position-dependent base doping profile, the collector current density can be expressed as (20):

$$J_C = \frac{q[e^{qV_{BE}/kT} - 1]}{\displaystyle\int_0^{W_b} \frac{N_{ab}(x)\,dx}{D_{nb}(x)\,n_{ib}^2(x)}} \tag{7}$$

We see then that the collector current density in a BJT depends on the details of the base doping profile [more specifically the integrated base charge, and, hence, $R_{\rm bi}$ given in Eq. (1)]. The base current density can be obtained in a similar manner, except that the physics of the polysilicon emitter contact must be properly accounted for (21, 22). For the "transparent emitter domain" in which the holes injected from the base to emitter do not recombine before the reaching the emitter contact, the base current density can be written as

$$J_B = \frac{q[e^{qV_{BE}/kT} - 1]}{\displaystyle\int_0^{W_e} \frac{N_{de}(x)\,dx}{D_{pe}(x)\,n_{ie}^2} + \frac{N_{de}(W_e)}{S_{pe}n_{ie}^2(W_e)}} \tag{8}$$

where $S_{\rm pe}$ is the "surface recombination velocity" characterizing the polysilicon emitter contact (21). More detailed base current density expressions can be found in Refs. 21, 22. Observe that in this transparent domain, the base current density depends on the specifics of the emitter doping profile as well as the influence of the polysilicon emitter contact.

For position-independent base and emitter doping profiles, with no polysilicon emitter contact, Eqs. 8) simplify to their familiar forms

$$J_C \cong \left[\frac{qD_{nb}n_{ib}^2}{W_bN_{ab}}\right] e^{qV_{BE}/kT} \tag{9}$$

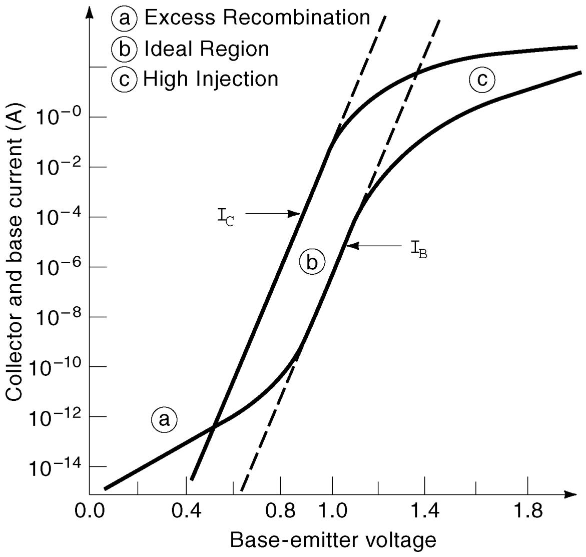


Figure 9. Schematic Gummel characteristics for a realistic bipolar transistor.

$$J_B \cong \left[\frac{qD_{pe}n_{ie}^2}{L_{pe}N_{de}}\right]e^{qV_{BE}/kT} \tag{10}$$

from which the ideal BJT current gain can be obtained

$$\beta \cong \frac{D_{nb}L_{pe}N_{de}}{D_{pe}W_bN_{ab}}e^{(\Delta E_{gb}^{app}-\Delta E_{ge}^{app})/kT} \propto \frac{N_{de}}{N_{ab}} \tag{11}$$

Thus, the current gain of the BJT depends on the ratio of emitter to base doping level. Given this fact, it is not surprising that the actual ratio of emitter to base doping level is typically found to be 100 (refer to Fig. 4), a common value for β in modern technologies. Note as well, however, from Eq. (11) that the ideal current gain in a BJT is reduced by the exponential dependence of the heavy-doping-induced bandgap narrowing parameters (the exponent is negative because the emitter is more heavily doped than the base). This latter dependence is also responsible for determining the temperature dependence of β in a BJT.

If one compares the measured *I-V* characteristics of a BJT with those expected from Eqs. (9) to (11), substantial deviations are typically observed, as depicted schematically in Figs. 9 and 10 (the dashed lines represent the ideal results). Referring to Fig. 9, at low current levels, base current nonideality is the result of emitter-base space-charge region recombination effects; at high current levels, the deviations are the result of various "high-injection" effects (discussed in what follows). Only over an intermediate bias range are ideal characteristics usually observed. Figure 11 shows typical measured *I-V* characteristics (a so-called "Gummel plot") from the same 40 GHz profile depicted in Fig. 3 (14). The inset of Fig. 3 shows the linear "output characteristics" of the BJT. The shape and doping level of the collector profile controls the breakdown characteristics of the device. In this case, the collector-to-emitter breakdown voltage (BV_{CEO}) is approximately 3.3 V, typical for a high-performance digital BJT technology.

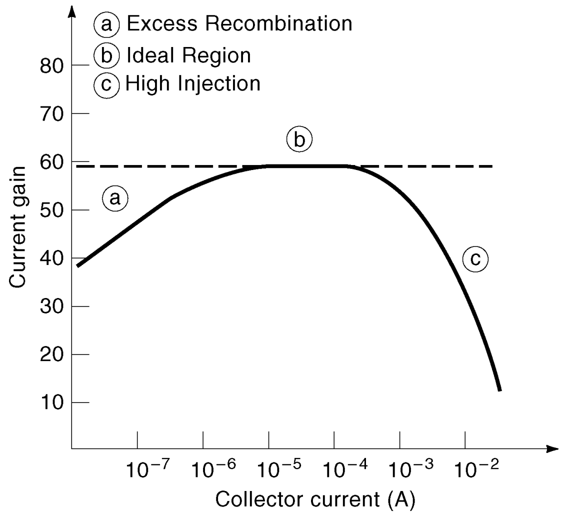


Figure 10. Schematic current gain versus bias for a realistic bipolar transistor.

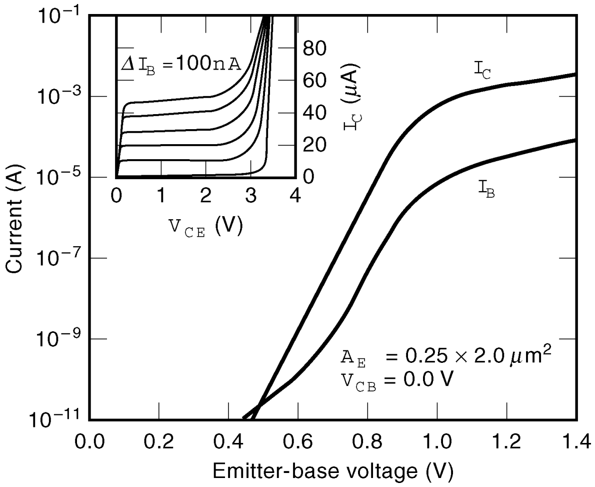


Figure 11. Measured Gummel characteristics for a scaled 0.25 μm double-polysilicon bipolar technology (14). Inset shows the common-emitter breakdown characteristics of the transistor.

Frequency Response

The frequency response of a BJT is determined by both the intrinsic speed of the carriers through the device (transit time), as well as the parasitic resistances and capacitances of the transistor. Two primary figures-of-merit are used to characterize the frequency response of a BJT, the unity gain cutoff frequency (f_T) and the maximum oscillation frequency (f_{max}). Using a small-signal hybrid-pi model both f_T and f_{max} can be derived (17), yielding

$$f_T = \frac{1}{2\pi\tau_{ec}} = \left[\frac{1}{g_m}(C_{be}+C_{bc}) + \tau_b + \tau_e + \tau_c\right]^{-1} \tag{12}$$

$$\begin{aligned} f_T = \Bigg[&\frac{kT}{qI_C}(C_{be}+C_{bc}) + \frac{W_b^2}{\eta\tilde{D}_{nb}} \\ &+ \frac{1}{\beta_{ac}}\left(\frac{W_e}{S_{pe}} + \frac{W_e^2}{2D_{pe}}\right) + \frac{W_{bc}}{2v_s} + r_cC_{bc}\Bigg]^{-1} \end{aligned} \tag{13}$$

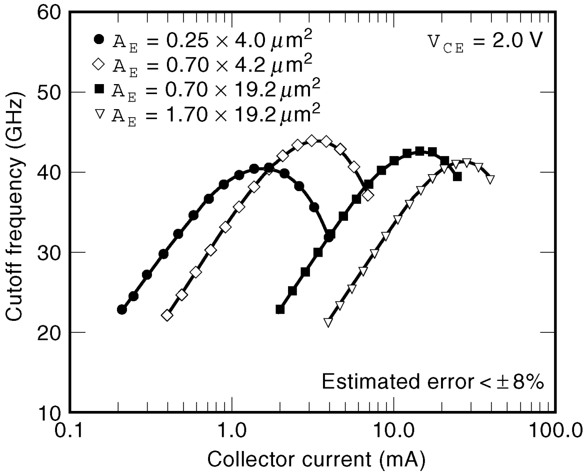


Figure 12. Measured cutoff frequency as a function of collector current for a scaled 0.25 μm double-polysilicon bipolar technology (14). Shown are a variety of device geometries.

and

$$f_{max} = \sqrt{\frac{f_T}{8\pi C_{bc} R_b}} \tag{14}$$

In Eqs. (12) to (14), g_m is the transconductance ($\partial I_C/\partial V_{BE}$), C_{be} and C_{bc} are the base-emitter and base-collector capacitances, τ_b, τ_e, and τ_c are the base, emitter, and collector transit times, respectively, v_s is the saturation velocity (1 × 10^7 cm/s), η accounts for any doping-gradient-induced electric fields in the base, and R_b is the base resistance; f_T and, hence, f_{max} is typically limited by τ_b in conventional Si–BJT technologies. A major advantage of ion-implanted base, double-polysilicon BJT technology is that the base width can be made very small (typically < 150 nm), and thus the intrinsic frequency response quite large. Figure 12 shows measured f_T data as a function of bias current for a variety of devices sizes for the doping profile shown in Fig. 3 (14).

ECL Gate Delay

Due to its nonsaturating properties and high logical functionality, the ECL is the highest speed bipolar logic family, and is in widespread use in the high-speed digital bipolar world. Figure 13 shows a simplified 2-phase ECL logic gate. A common large-signal performance figure-of-merit is the unloaded ECL gate delay, which can be measured using a "ring oscillator." A ring oscillator is essentially a delay chain of ECL inverters with output tied back to its input, thus rendering the resultant circuit unstable (Fig. 14). From the period of the oscillation (Fig. 15) the average gate delay can be determined for a given bias current. Multiple ring oscillators can then be configured to operate at various bias currents, and hence the "power-delay" characteristics of the BJT technology determined (average gate delay is plotted as a function of average power dissipation—or current in this case, because the supply voltage is constant). Figure 16 shows a typical measured ECL power-delay curve (14). A minimum ECL gate delay of 20.8 ps is achieved with this technology. Observe that the speed of the ECL gate becomes faster as the average switch current increases, until some minimum value of delay is reached. To better understand the functional shape of the power-delay curve, asymptotic expressions can be developed using a weighted time constant approach (23). Under low current (or power) conditions, the ECL gate delay is given by

$$\tau_{ECL}(\text{low-power}) \cong R_{CC}\sum_{k=1}^{n} a_k C_k \tag{15}$$

$$= \frac{V_L}{I_{CS}}[a_1 C_{bc} + a_2 C_{be} + a_3 C_{cs} + a_4 C_w + \cdots] \tag{16}$$

$$\propto \frac{1}{I_{CS}} \propto \frac{1}{\text{Power}} \tag{17}$$

while under high current (or power) conditions the ECL gate delay can be written as

$$\tau_{ECL}(\text{high-power}) \cong C_{diff}\sum_{k=1}^{n} b_k R_k \tag{18}$$

$$= \frac{q\tau_{ec} I_{CS}}{kT}[b_1 R_{bi} + b_2 R_{bx} + b_3 R_e + b_4 R_c + \cdots] \tag{19}$$

$$\propto I_{CS} \propto \text{Power} \tag{20}$$

In Eqs. (15) to (20), R_{CC} is the circuit pull-up resistor, V_L is the logic swing, a_k and b_k are delay "weighting factors," I_{CS} is the switch current, and C_{diff} is the transistor diffusion capacitance. We see then that at low currents, the parasitic capacitances dominate the ECL delay with a delay that is reciprocally proportional to the power dissipation, whereas at high currents, the parasitic resistances dominate the ECL delay, yielding a delay that is proportional to power dissipation. It is thus physically significant to plot the log of the ECL delay as a function of the log of the power (or current), as shown in Fig. 16. Also shown in Fig. 16 are large-signal circuit simulation results using the compact model depicted in Fig. 17, which confirm the stated dependence of delay on power.

HIGH-INJECTION EFFECTS

Substantial deviations from ideal behavior occur for BJTs operating at high current densities (as a rule of thumb, for $J_C \sim 1.0$ mA/μm^2 in a modern high-performance technology). This deviation from simple theory can be observed in the premature roll-off of both the current gain and the cutoff frequency at high current densities, as shown in Figs. 10 to 12. These so-called "high-injection" effects are particularly important because most high-performance BJT circuits will be biased at high current densities in order to achieve maximum transistor performance. High-injection in a BJT can generally be defined as that current density

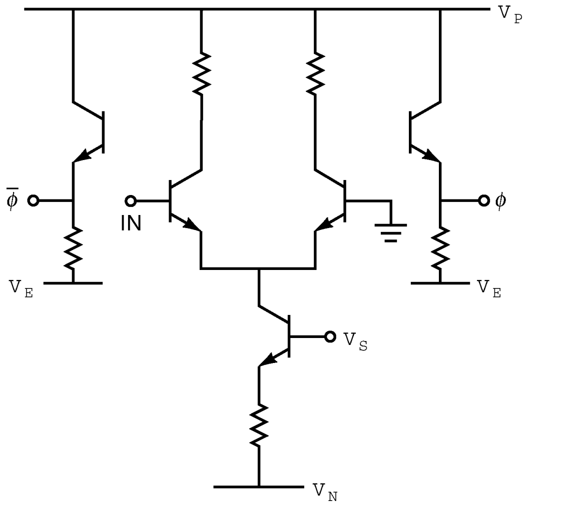


Figure 13. Circuit schematic of a two-phase emitter-coupled-logic (ECL) gate.

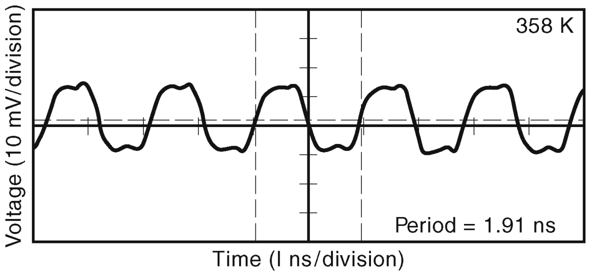


Figure 15. Measured output waveform from an ECL ring oscillator.

at which the injected minority carrier density (e.g., electrons in the base) becomes comparable to the local doping density. High-injection effects are generally the result of a number of competing physical mechanisms in the collector, base, and emitter regions, and are thus difficult to analyze together theoretically. In this work we will simply emphasize the physical origin of each high-injection phenomenon region by region, discuss their impact on device performance, and give some rule-of-thumb design guidelines. The interested reader is referred to Ref. 6 for a more in-depth theoretical discussion.

Collector Region

Collector region high-injection effects in BJTs can be divided into two separate phenomena: (1) Kirk effect, sometimes referred to as "base push-out" (24); and (2) quasi-saturation. The physical origin of the Kirk effect is as follows. As the collector current density continues to rise, the electron density in the base-collector space-charge region is no longer negligible, and modifies the electric field distribution in the junction. At sufficiently high current density, the (positive) background space charge due to the donor doping in the collector ($N+_{\text{dc}}$) is compensated by the injected electrons, and the electric field in the junction collapses, thereby "pushing" the original base region deeper into the collector (Figs. 18 and 19). Because both β and f_{T} depend reciprocally on W_{b}, this injection-induced increase in effective base width causes a strong degradation in both parameters. Approximate theoretical analysis can be used to determine the critical current density at which the Kirk effect is triggered, resulting in a BJT design equation

$$J_{\text{Kirk}} \cong q v_s N_{\text{dc}} \left(1 + \frac{2\epsilon V_{BC}}{q W_{epi}^2 N_{dc}}\right) \tag{21}$$

From Eq. (21) it is apparent that increasing the collector doping level is the most efficient method of delaying the onset of the Kirk effect, although this will have a detrimental impact on the BV_{CEO} and collector-base capacitance of the transistor. As the Kirk effect is typically the limiting high-injection phenomenon in modern high-performance BJTs, a fundamental tradeoff thus exists between peak f_{T} and BV_{CEO}.

The second major collector region high-injection phenomenon is called "quasi-saturation." At a basic level, quasi-saturation is the result of the finite collector resistance of the n-type epi-layer separating the base from the heavily doped subcollector in a BJT. At sufficiently high current levels, the IR drop associated with the collector epi becomes large enough to internally forward bias the base-collector junction, even though an external reverse bias on the collector is applied. For instance, for a collector resistance of 1 kΩ and a collector current of 2 mA, an internal voltage drop of 2 V is obtained. If the BJT were biased at a base-collector reverse voltage of 1 V, then the internal base-collector junction would be forward-biased by 1 V, artificially saturating the transistor. With both base-emitter and base-collector junctions forward biased, the dc signature of quasi-satuation is a strong increase in base current together with a "clipping" of the collector current. Dynamically, quasi-saturation has a strong negative impact on the f_{T} and, hence, circuit speed because excess minority charge is injected into the base region under saturation. Theoretically, quasi-saturation is difficult to model because the resistance of the epi layer is strongly bias-dependent and the collector doping profile in real devices is highly position-dependent. In a well-designed high-performance BJT, the

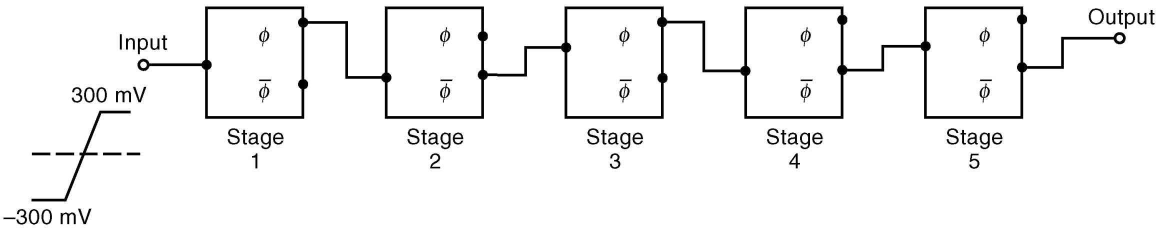


Figure 14. Schematic representation of an ECL ring oscillator circuit configuration.

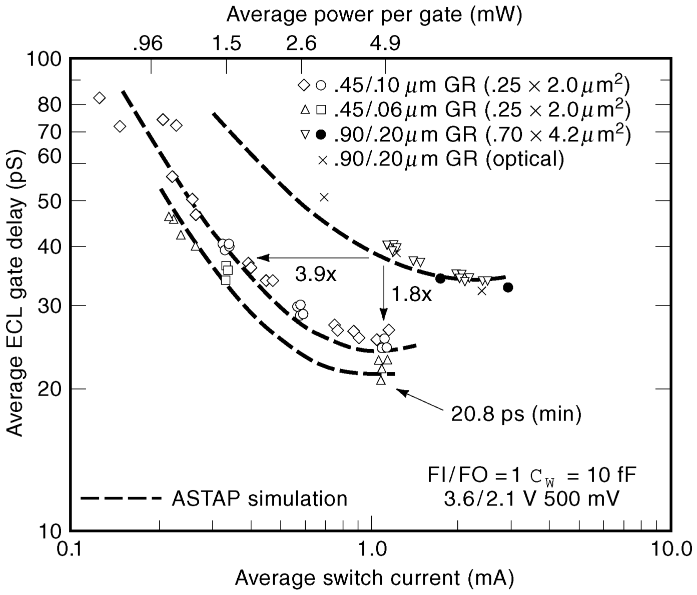


Figure 16. The ECL power-delay characteristics for a scaled 0.25 μm double-polysilicon bipolar technology (14). A minimum delay of 20.8 ps is achieved. The ECL circuits were operated on 3.6/2.1 V power supplies at a 500 mV logic swing. A fan-in (*FI*) and fan-out (*FO*) of one was used. The impact of transistor scaling from 0.90/0.20 μm lithography to 0.45/0.06 μm lithography is indicated. Also shown are circuit simulations calibrated to the data using a compact circuit model implemented in ASTAP.

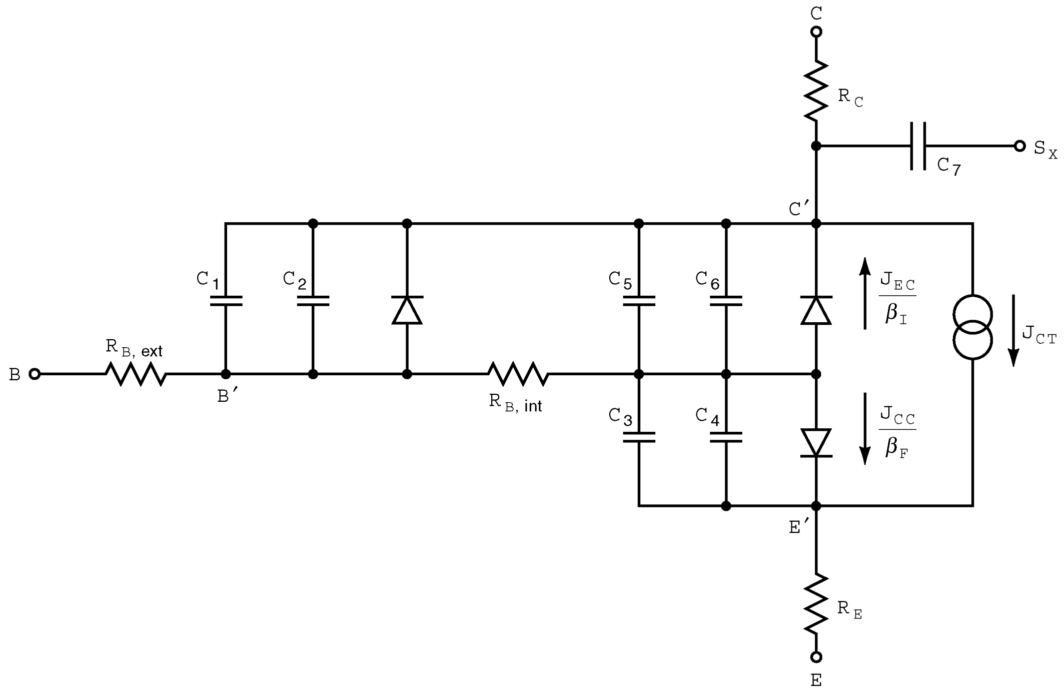


Figure 17. Compact circuit model used in the ASTAP circuit simulations.

Kirk effect is much more important than quasi-saturation.

Base Region

High-injection in the base region of a BJT leads to two major degradation mechanisms: (1) the Webster–Rittner effect (25, 26), sometimes known as "base conductivity modulation;" and (2) emitter current crowding. In the Webster–Rittner effect, the large electron density in the base region under high injection is no longer small compared to the doping in the base. To maintain charge neutrality in the neutral base, the hole density must therefore rise (refer to Figs. 18 and 19), changing the (low-injection) Shockley boundary condition at the emitter-base junction, and effectively doubling the electron diffusivity in the base. The result is a different voltage dependence of the collector current, which changes to one-half the slope of the exponential low-injection collector current according to

$$J_C\ (\text{Webster-Rittner}) \cong \left[\frac{q2D_{nb}n_{ib}(0)}{W_b}\right] e^{qV_{BE}/2kT} \tag{22}$$

This slope change of J_C has a detrimental impact on the current gain, although in practice for high-performance BJTs, the Kirk effect typically onsets before the Webster-Rittner effect because the base is much more heavily doped than the collector.

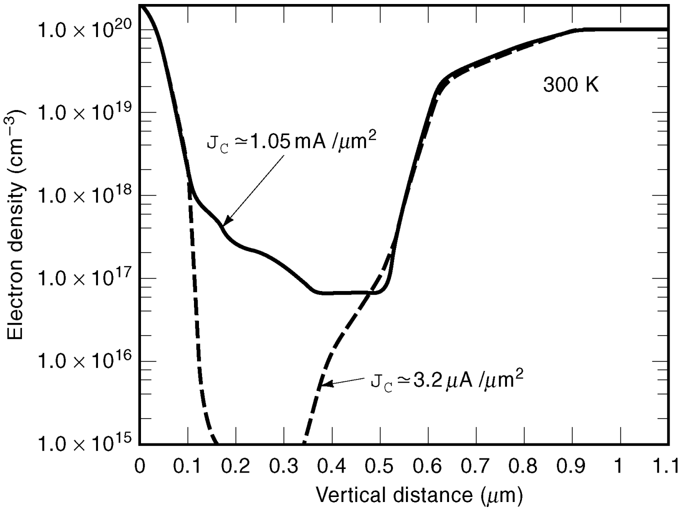


Figure 18. Simulated electron profile in a bipolar transistor at both low injection (3.2 μA/μm^2) and high injection (1.05 mA/μm^2).

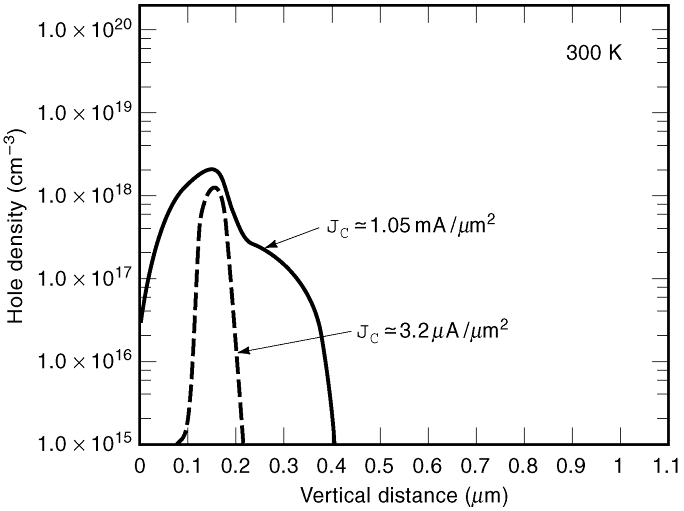


Figure 19. Simulated hole profile in a bipolar transistor at both low injection (3.2 μA/μm^2) and high injection (1.05 mA/μm^2). Observe that at high-injection levels the hole profile in the base exceeds the local doping level (as indicated by the low-injection result), and holes are present in the (n-type) collector region.

Emitter current crowding is the result of the finite lateral resistance associated with the intrinsic base profile (i.e., R_{bi}). Because the collector current depends on the actual base-emitter voltage applied at the junction itself, rather than that applied at the base and emitter terminals, large base currents flowing at high-injection levels can produce a lateral voltage drop across the base. This yields a lateral distribution in the actual base-emitter voltage at the junction, resulting in higher bias at the emitter periphery than in the center of the device. In essence, then, the collector current "crowds" to the emitter edge where the static and dynamic properties of the device are generally worse, and can even produce "thermal runaway" and catastrophic device burn-out. This is typically only a problem in large geometry power transistors, not high-speed digital technologies. In addition, as the base current is a factor of β smaller than the collector current, emitter current crowding is not generally a problem unless there is very large base resistance in the device.

Emitter Region

Because it is very heavily doped, the emitter region in modern BJTs always operate in low-injection. Thus, the only significant emitter region high-injection effect is the result of the finite emitter resistance of the transistor. Because polysilicon emitter contacts in fact exhibit reasonably high specific contact resistance (e.g., 20–60 Ω μm^2), however, emitter resistance (R_{E}) can be a serious design constraint. Emitter resistance degrades the collector and base currents exponentially as it decreases the applied base-emitter voltage according to

$$I_C = I_{C0}e^{q(V_{BE}-I_ER_E)/kT} \tag{23}$$

$$I_B = I_{B0}e^{q(V_{BE}-I_ER_E)/kT} \tag{24}$$

For instance, for a 1.0 μm^2 emitter area transistor operating at a collector current of 1.0 mA, a specific emitter contact resistance of 60 Ω μm^2 results in an emitter-base voltage loss of 60 mV, yielding a 10× decrease in collector current. Proper process optimization associated with the polysilicon emitter contact is key to obtaining a robust high-speed BJT technology, particularly as the emitter geometry shrinks.

SCALING ISSUES

Device miniaturization (“scaling”) has been a dominant theme in bipolar technology over the past 15 years, and has produced a monotonic decrease in circuit delay over that period (refer to Fig. 1). In general, optimized BJT scaling requires a coordinated reduction in both lateral and vertical transistor dimensions, as well as a change in circuit operating point (23). Unlike in CMOS technology, BJT circuit operating voltages (for conventional circuits such as ECL) cannot be scaled because the junction built-in voltage is only weakly dependent on doping. The evolution of BJT technology from nonself-aligned, double-diffused transistor structures to self-aligned, ion-implanted, double-polysilicon transistor structures was the focus for BJT scaling in the 1980s. During the 1990s more emphasis has been placed on vertical profile scaling and a progression towards both forms of advanced lithography (e.g., deep UV or electron beam lithography), low-thermal budget processing, and structural innovation to continue the advances in circuit speed over time.

Figure 20 represents an idealized ECL power-delay curve, and indicates the three principle regions that require attention during optimized scaling. In region (a), which is dominated by parasitic transistor capacitances [see Eqs. (15) to (17)], a reduction in lithography, and hence decrease in transistor size, is effective in reducing circuit delay at low current levels. Region (b) is dominated by the intrinsic speed of the transistor (i.e., τ_{ec}). Thinning the vertical profile, particularly the base width, is key to reducing the ECL delay at intermediate current levels. The evolution of ion-implantation has proven key to realizing viable sub-150 nm metallurgical base widths in modern BJT technologies. In region (c), the ECL delay is dominated by base resistance and high-injection roll-off of the frequency response of the device [Eqs. (18) to (20)]. Doping the base and collector regions more heavily is successful in improving the delay at very high current levels, although tradeoffs exist. For instance, doping the base more heavily decreases the peak f_T of the transistor (due to a lower electron mobility), and, hence, degrades the speed in region (b) at intemediate current levels. In addition, increasing the collector doping level to improve the high-injection performance in region (c) effectively increases the collector-base capacitance, degrading the ECL delay in region (a) at low-current levels. Optimized scaling is thus a complex tradeoff between many different profile design issues.

Clever solutions to certain scaling tradeoffs have emerged over the years, and include, for instance, the now pervasive use of the so-called Sul;elf-aligned, Implanted Collector (*SIC*) process. In an SIC process (see Ref. 10), phosphorus is implanted through the emitter window in the base polysilicon layer (either before or after sidewall spacer formation) to increase the collector doping level locally under the intrinsic device without increasing the collector-base capacitance in the extrinsic transistor.

Figures 21 and 22 show the results of a recent BJT lithographic scaling experiment (14). In this study a comparison was made between BJTs fabricated using three different lithographies (0.09 μm/0.20 μm–lithographic linewidth/lithographic overlay, 0.45 μm/0.10 μm, and 0.45 μm/0.06 μm). The latter two processes used advanced electron-beam lithography. As can be seen, the impact of scaling on device parameters is dramatic, resulting in an expected improvement in ECL delay across the entire power-delay characteristic, and a minimum ECL gate delay of 20.8 ps (Fig. 16).

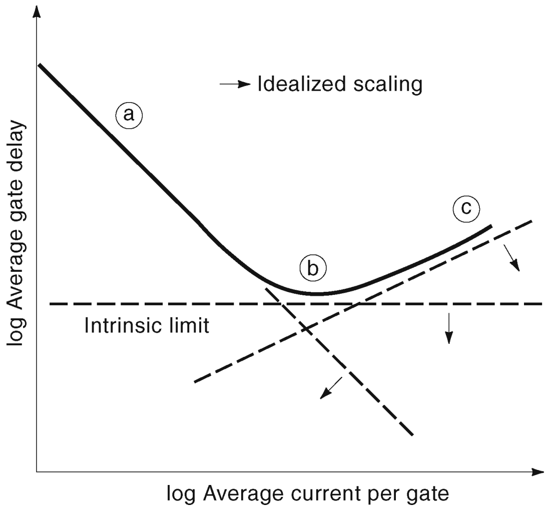


Figure 20. The ECL power-delay characteristics showing the impact of idealized scaling.

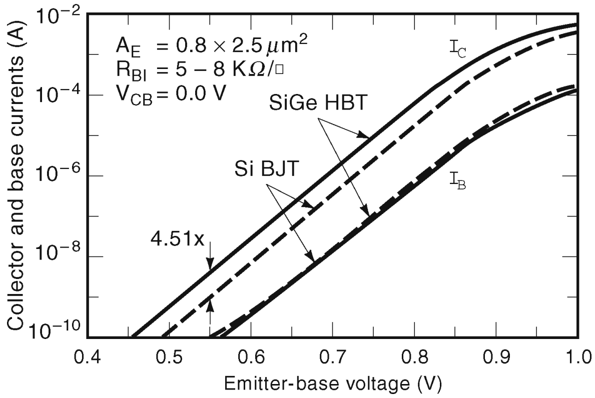
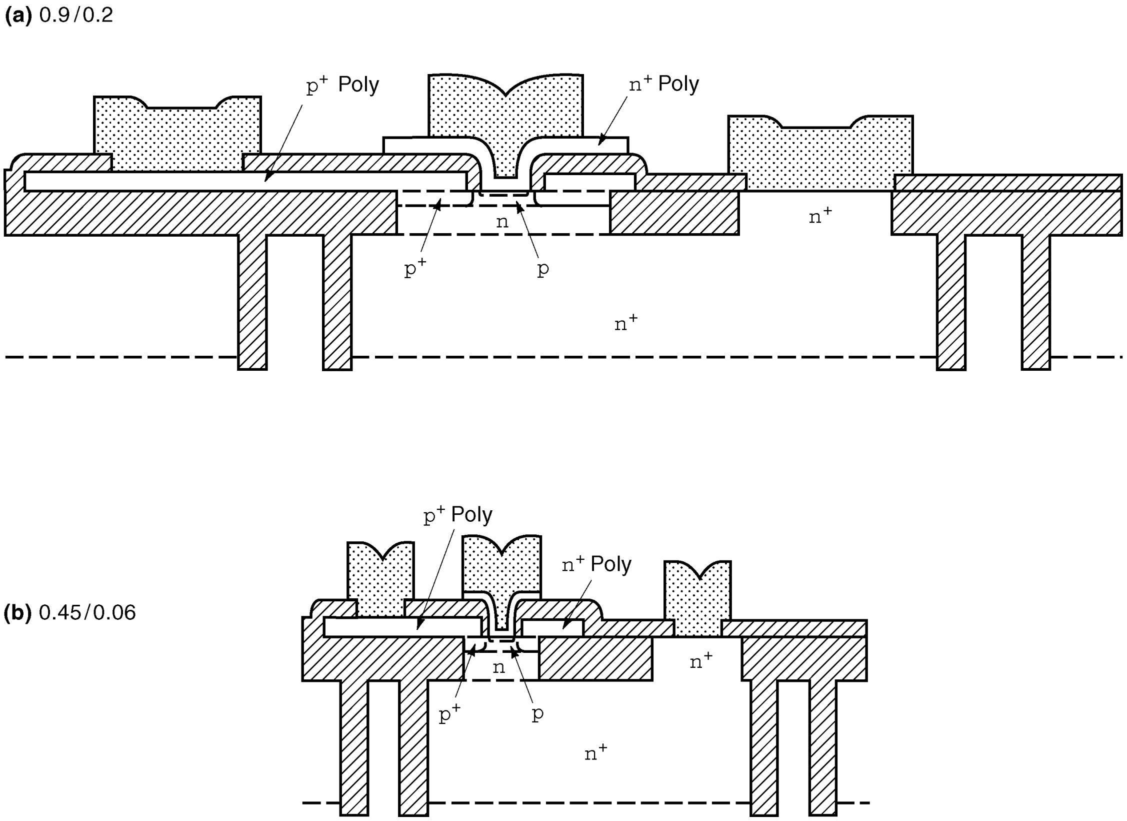


Figure 25. Measured Gummel characteristics for SiGe and Si transistors with comparable doping profiles. The expected enhancement in collector current (4.51×) can be observed.

Nonetheless, practical limits do exist for conventional ion-implanted, double-polysilicon BJT technology. Obtaining metallurgical basewidths below 80–100 nm with reasonable base resistance using low-energy ion-implantation is very difficult and places a practical limit of about 40–50 GHz on the resultant f_T of such transistors (see Fig. 12, which corresponds to the doping profile shown in Fig. 2). In addition, circuit operating voltages limit the useful BV_{CEO} of the transistor to about 3.0 V, and thus place a practical limit on collector doping levels of about 1×10^{17} cm^{-3} and a consequent maximum operating current density of about 1–2 mA/μm^2. The emitter junction depth (and, hence, the thermal process associated with the polysilicon emitter) is limited to about 25–30 nm, because the emitter-base space charge region must lie inside the single-crystal emitter region to avoid the generation/recombination centers asso-

Lithographic GR Min image/OL (μm)	0.90/0.20	0.45/0.10	0.45/0.06	
A_E (μm^2)	0.7× 4.2	0.25× 2.0	0.25× 2.0	
β_{MAX}	80	87	90	
R_{BI} (kΩ/□)	15.2	15.2	15.2	
ρ_{p+poly} (Ω/□)	176	176	176	
BV_{CEO} (V)	3.4	3.4	3.4	
BV_{CBO} (V)	11.2	11.3	11.4	
R_E (Ω)	143	101	93	
C_{CBi} (fF/μm^2)	1.13	1.13	1.13	
C_{CBx} (fF/μm^2)	0.70	0.70	0.70	
C_{EB} (fF/μm^2)	4.64	4.64	4.64	
Lumped ASTAP parameters, I_{CS} = 0.5 mA				
C_{CB} (fF)	10.2	3.5	2.7	(.2×)
C_{EB} (fF)	13.8	2.7	2.7	(.2×)
C_{CS} (fF)	11.7	7.0	5.6	(.4×)
R_{Bx} (Ω)	203	73	66	(.3×)

Figure 21. Comparison of measured device parameters as a function of scaling for: (1) 0.90/0.20 μm (lithographic image/overlay); (2) 0.45/0.10 μm; and (3) 0.45/0.06 μm transistors (14). Lumped ASTAP parameters are extracted from calibrated simulations of ECL ring oscillator data.


(a) 0.9/0.2

(b) 0.45/0.06

Figure 22. Scaled comparison of (a) a 0.90/0.20 μm (lithographic image/overlay) transistor with (b) a 0.90/0.06 μm transistor.

ciated with the heavily defective polysilicon region. More advanced profiles can be obtained using epitaxial growth techniques, as will be discussed in the next section.

FUTURE DIRECTIONS

Despite the continual improvements in speed that BJT technology has enjoyed over the past 15 years, and the inherent superiority of the analog and digital properties

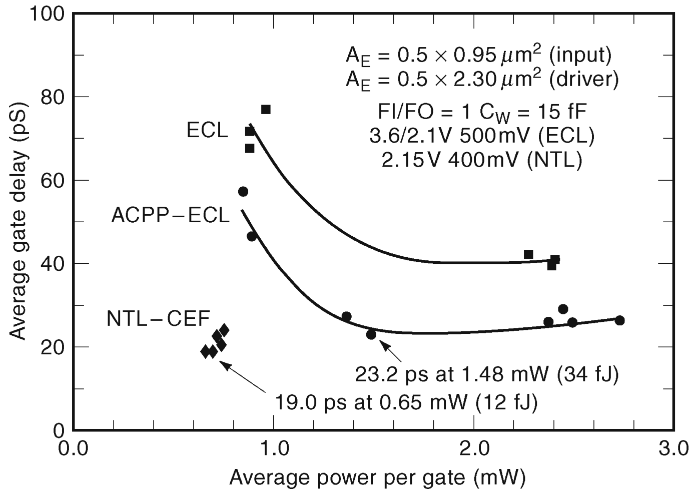


Figure 23. Measured power-delay characteristics from an advanced complementary bipolar technology (27). Three circuit families are compared: 1) conventional (*npn*-only) ECL; 2) complementary ac-coupled push-pull ECL (ACPP-ECL); and 3) complementary nonthreshold logic with complementary emitter-follower (*NTL-CEF*). The NTL-CEF circuit achieved a minimum power-delay product of 12 fJ.

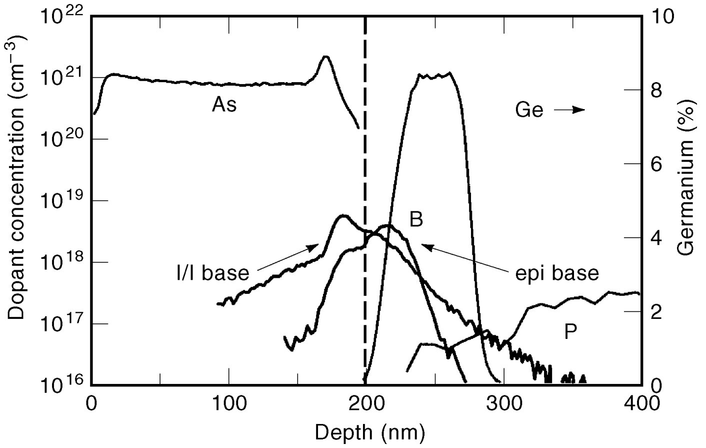


Figure 24. Measured secondary ion mass spectroscopy (SIMS) doping profile comparing a 60 GHz cutoff frequency epitaxial SiGe base bipolar technology with an aggressive (40 GHz cutoff frequency) ion-implanted (*I/I*) base bipolar technology.

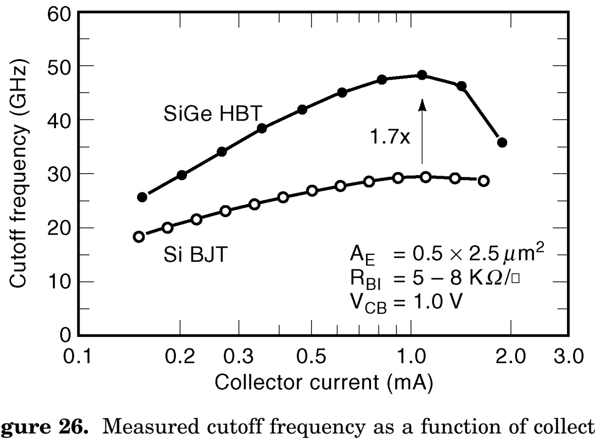


Figure 26. Measured cutoff frequency as a function of collector current for SiGe and Si transistors with comparable doping profiles. The expected enhancement in collector current (1.71×) can be observed.

of BJTs compared to field effect transistors (*FETs*), the world market for BJT ICs has steadily eroded. This is due to both the improved performance of FET technology as gate lengths are scaled into the submicron domain, the widespread emergence of CMOS with its low power-delay product, and the decreased cost associated with CMOS ICs compared to competing bipolar technologies. To confront this situation, many bipolar + CMOS (BiCMOS) technologies have been developed that seek to combine low-power CMOS with high-performance BJTs. The reader is referred to Ref. 4 for an examination of the process integration issues associated with modern BiCMOS technologies.

In addition, there are several areas of current research with the potential to extend BJT technology well into the 21st century; they include: (1) complementary bipolar technology; (2) SOI bipolar technology; and (3) silicon-germanium (SiGe) bipolar technology. Each of these three research areas seeks to improve either the power-dissipation associated with conventional BJT circuit families such as ECL, or improve the transistor performance to levels not possible in Si BJTs and thus to capture new and emerging IC markets.

Complementary Bipolar Technology

Complementary bipolar (C-bipolar) technology, which combines *n-p-n* and *p-n-p* transistors on the same chip, has been used for decades. In conventional usage, the *n-p-n* BJT is a standard, vertical high-performance transistor, while the *p-n-p* BJT is typically a slow-speed lateral device used only in analog circuits such as current sources where high-speed is unnecessary. Modern implementations of C-bipolar technology, on the other hand, combine a high-performance vertical *n-p-n* BJT and a high-performance vertical *p-n-p* BJT (see, for instance, Refs. 27, 28). The resulting IC technology, though inherently more complex than a traditional *n-p-n* only BJT technology, opens many new possibilities for novel high-speed, low-power circuit families. New C-bipolar circuit families such as ac-coupled push-pull emitter-coupled logic (*ACPP-ECL*) and nonthreshold logic with complementary emitter-follower (*NTL-CEF*) offer dramatic improvements in power-delay product compared to conventional ECL (Fig. 23).

Silicon-On-Insulator Bipolar Technology

Silicon-on-insulator (*SOI*) IC technologies have existed since the 1960s, but have emerged recently as a potential scaling path for advanced CMOS technologies. In SOI technology, a buried oxide dielectric layer is placed below the active Si region, either by ion-implantation (*SIMOX*) or by wafer bonding (*BESOI*). For the CMOS implementation, the active Si region is made thin, so that it is fully depleted during normal device operation, resulting in improved subthreshold slope, better leakage properties at elevated temperatures, and improved dynamic performance due primarily to the reduction in parasitic source/drain capacitance. Given this development, it is natural to implement a lateral BJT together with the SOI-CMOS to form an SOI-BiCMOS technology. While lateral BJTs are not generally considered high-speed transistors, the reduction in parasitic capacitance in the lateral BJT, together with clever structural schemes which allow very aggressive base widths to be realized, have resulted in impressive performance (29).

SiGe Bipolar Technology

Attempts to reduce the base widths of modern BJT technologies below 100 nm typically rely on epitaxial growth techniques. A recent high-visibility avenue of research has been the incorporation of small amounts of germanium (Ge) into these epitaxial films to tailor the properties of the BJT selectively while maintaining compatibility with conventional Si fabrication techniques. The resultant device, called an SiGe heterojunction bipolar transistor (HBT), involves introducing strained epitaxial SiGe alloys into the base region of the transistor, and represents the first practical bandgap-engineered device in Si technology (refer to Ref. 30–32, and references contained within, for reviews of SiGe HBTs).

Compared to an Si BJT with an identical doping profile, the SiGe HBT has significantly enhanced current gain, cutoff frequency, Early voltage (output conductance), and current gain Early voltage product, according to Refs. 31 and 32,

$$\frac{J_{C,SiGe}}{J_{C,Si}} = \frac{\beta_{SiGe}}{\beta_{Si}} = \gamma\eta\frac{\Delta E_{g,Ge}(\text{grade})/kTe^{\Delta E_{g,Ge}(0)/kT}}{1-e^{-\Delta E_{g,Ge}(\text{grade})/kT}} \tag{25}$$

$$\frac{\tau_{b,SiGe}}{\tau_{b,Si}} \propto \frac{f_{T,Si}}{f_{T,SiGe}} = \frac{2}{\eta}\left(\frac{kT}{\Delta E_{g,Ge}(\text{grade})}\right) \left[1-\frac{1-e^{-\Delta E_{g,Ge}(\text{grade})/kT}}{\Delta E_{g,Ge}(\text{grade})/kT}\right] \tag{26}$$

$$\frac{V_{A,SiGe}}{V_{A},Si} = e^{\Delta E_{g,Ge}(\text{grade})/kT}\left[\frac{1-e^{-\Delta E_{g,Ge}(\text{grade})/kT}}{\Delta E_{g,Ge}(\text{grade})/kT}\right] \tag{27}$$

$$\frac{\beta V_A|_{SiGe}}{\beta V_A|_{Si}} = \gamma\eta e^{\Delta E_{g,Ge}(0)/kT}e^{\Delta E_{g,Ge}\text{grade}/kT} \tag{28}$$

where $\Delta E_{\mathrm{g,Ge}}(0)$ is the Ge-induced band offset at the emitter-base junction, $\Delta E_{\mathrm{g,Ge}}(\text{grade}) = \Delta E_{\mathrm{g,Ge}}(W_{\mathrm{b}}) - \Delta E_{\mathrm{g,Ge}}(0)$ is the base bandgap grading factor, and γ, η are the strain-induced density-of-states reduction and mobility enhancement factors, respectively. With its improved transistor performance compared to Si BJTs and compatibility with standard Si fabrication processes, SiGe HBT technology is expected to pose a threat to more costly compound semiconductor technologies such as GaAs for emerging high-speed communications applications. Figure 24 shows a representative SiGe doping profile. Observe that the Ge is introduced only in the base region of the transistor. Experimental results comparing a SiGe HBT and a Si BJT having identical layout and doping profile are shown in Figs. 25 and 26 and indicate that significant enhancements compared to a comparably designed Si devices are possible. It is now clear that cutoff frequencies well above 300 GHz are possible using SiGe HBT technology, and thus SiGe represents the next evolutionary step in Si BJT technology.

BIBLIOGRAPHY

1. W. Shockley,US Patent 2,569,347, filed June 26, 1947, and issued September 25, 1951.
2. W. Shockley, M. Sparks, G. K. Teal, *pn* junction transistors, *Physical Review*, **83**: 151, 1951.
3. T. H. Ning, D. D. Tang, Bipolar trends, *Proc. IEEE*, **74**: 1669, 1986.
4. J. D. Warnock, Silicon bipolar device structures for digital applications: Technology trends and future directions, *IEEE Trans. Electron Devices*, **42**: 377, 1995.
5. T. Nakamura, H. Nishizawa, Recent progress in bipolar transistor technology, *IEEE Trans. Electron Devices*, **42**: 390, 1995.
6. R. M. Warner, Jr., B. L. Grung, *Transistors: Fundamentals for the Integrated Circuit Engineer*, New York, Wiley, 1983.
7. H. Nakashiba *et al.*, An advanced PSA technology for high-speed bipolar LSI, IEEE Trans. Electron Devices, 27: 1390, 1980.
8. D. D. Tang *et al.*, 1.25 μm Deep-grove-isolated self-aligned bipolar circuits, *IEEE J. Solid-State Circuits*, **17**: 925, 1982.
9. T. C. Chen *et al.*, A submicron high-performance bipolar technology, Symp. VLSI Technology Tech. Dig., 87, 1989.
10. S. Konaka *et al.*, A 20-ps Si bipolar IC using advanced super-self-aligned process technology with collector ion implantation, *IEEE Trans. Electron Devices*, **36**: 1370, 1989.
11. T. Shiba *et al.*, A 0.5 μm very-high-speed silicon bipolar technology U-grove isolated SICOS, *IEEE Trans. Electron Devices*, **38**: 2505, 1991.

12. V. de la Torre *et al.*, MOSAIC V—a very high performance bipolar technology, Bipolar Circuits and Technology Meeting Tech. Dig., 21, 1991.
13. J. D. Warnock *et al.*, High-performance bipolar technology for improved ECL power-delay, *IEEE Electron Device Letters*, **12**: 315, 1991.
14. J. D. Cressler *et al.*, A scaled 0.25 μm bipolar technology using full E-beam lithography, *IEEE Electron Device Letters*, **13**: 262, 1992.
15. T. Uchino *et al.*, 15-ps ECL/74 GHz f_T Si bipolar technology, Int. Electron Device Meeting Tech. Dig., 67, 1993.
16. D. M. Richey, J. D. Cressler, A. J. Joseph, Scaling issues and Ge profile optimization in advanced UHV/CVD SiGe HBTs, *IEEE Trans. Electron Devices*, **44**: 431, 1997.
17. D. J. Roulston, *Bipolar Semiconductor Devices*. New York: McGraw-Hill, 1990.
18. E. S. Yang, *Microelectronic Devices*, New York: McGraw-Hill, 1988.
19. R. F. Pierret, *Semiconductor Device Fundamentals*, New York: Addison-Wesley, 1996.
20. J. L. Moll, I. M. Ross, The dependence of transistor parameters on the distribution of base layer resistivity, *Proc. IRE*, **44**: 72, 1956.
21. A. Kapoor, D. Roulston, (eds.), *Polysilicon Emitter Bipolar Transistors*, New York: IEEE Press, 1989.
22. I. R. C. Post, P. Ashburn, G. Wolstenholme, Polysilicon emitters for bipolar transistors: A review and re-evaluation of theory and experiment, *IEEE Trans. Electron Devices*, **39**: 1717, 1992.
23. P. M. Solomon, D. D. Tang, Bipolar circuit scaling, Int. Solid-State Circuits Conf. Tech. Dig., 86, 1979.
24. C. T. Kirk, Jr., A theory of transistor cutoff frequency (ft) falloff at high current densities, *IRE Trans. Electron Devices*, **9**: 164, 1962.
25. E. S. Rittner, Extension of the theory of the junction transistor, *Physical Review*, **94**: 1161, 1954.
26. W. M. Webster, On the variation of junction-transistor current-amplification factor with emitter current, *Proc. IRE*, **42**: 914, 1954.
27. J. D. Cressler *et al.*, A high-speed complementary silicon bipolar technology with 12-fJ power-delay product, *IEEE Electron Device Letters*, **14**: 523, 1993.
28. T. Onai *et al.*, An npn 30 GHz, pnp 32 GHz f_T complementary bipolar technology, Int. Electron Device Meeting Tech. Dig., 63, 1993.
29. R. Dekker, W. T. A. van den Einden, H. G. R. Maas, An ultra low power lateral bipolar polysilicon emitter technology on SOI, Int. Electron Device Meeting Tech. Dig., 75, 1993.
30. J. D. Cressler, Re-engineering silicon: Si-Ge heterojunction bipolar technology, *IEEE Spectrum*, **49**: March 1995.
31. J. D. Cressler and G. Niu, *Silicon-Germanium Heterojunction Bipolar Transistors*. Boston, MA: Artech House, 2003.
32. J. D. Cressler (Editor), *Silicon Heterostructure Handbook—Materials, Fabrication, Devices, Circuits, and Applications of SiGe and Si Strained-Layer Epitaxy*. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2006.

JOHN D. CRESSLER
Georgia Tech,, Atlanta, GA