Compound semiconductor heterostructure bipolar transistors

by S. Tiwari
S. L. Wright
D. J. Frank

This paper is primarily an overview of our work on the technology, material and electronic properties, and performance limitations of compound semiconductor heterostructure bipolar transistors. Graded-gap epitaxial n-type ohmic contacts and p-type shallow diffusion ohmic contacts are important in the fabrication of high-performance (Al,Ga)As/GaAs devices. In the device structure implemented, the presence of a wide-gap p-type (Al,Ga)As extrinsic base region at the surface suppresses surface recombination, thereby enhancing the current gain at small device dimensions. We discuss experimental and theoretical results concerning the limiting physical effects due to heterostructure design and intrinsic and extrinsic bulk phenomena of compound semiconductors, emphasizing the understanding developed and the discoveries made during the course of our efforts. As device speeds have increased with coordinated scaling, dispersive effects have become increasingly important. We show how these may be included by modifying the conventional quasi-static modeling of the bipolar transistor, in order to obtain a realistic simulation of fast switching transients. Finally, we discuss scaling of heterostructure bipolar transistors, and implications of the use of lower-bandgap materials and operation at cryogenic temperatures.

Introduction

The success of heteroepitaxial growth technologies in tailoring the composition and doping of morphological and pseudomorphological metallurgical systems has led to considerable progress in demonstrating the performance potential of the heterostructure implementation of the conventional bipolar transistor. Such implementations have been usually called heterostructure bipolar transistors (HBTs). Compound semiconductors, with their usually superior electron transport characteristics and greater freedom in selection of material structures, allow a significant redesign of the bipolar transistor with the intent of improving its speed and frequency limits. This paper is a retrospective of our

understanding of the technology and operational basis of compound semiconductor HBTs and a perspective of their future as the reduction in device dimensions continues.

In the HBT, the use of a heterostructure wide-gap emitter (or narrow-gap base) relaxes the design requirements of both the emitter and the base doping because it increases injection efficiency by decreasing the backward injection of majority carriers from the base. The injection efficiency increases exponentially by a scaling factor proportional to the difference in bandgap and inversely proportional to the thermal voltage. This allows the base doping to be increased, thus decreasing the base resistance. Shallow emitter devices with a large gain can thus be fabricated without the use of poly-emitters or their equivalents because of the suppression of injection from the base. Electron transport in compound semiconductors is also very rapid compared to that in silicon. All of these factors lead to a reduction in several of the delay components associated with transistor switching or frequency limits of the device. The heterostructure also allows a transistor to have a large gain when the emitter is relatively large in area and located toward the substrate of the device (such as in I^2L circuits), leading to high-speed circuits with more efficient packing and reduced power.

Several of these advantages were foreseen by Shockley in his original patent on the bipolar transistor, and were emphasized theoretically by Kroemer [1]. Experimental demonstrations of the device concepts were made in the following decade in both silicon and germanium using polysilicon or amorphous emitters made of other semiconductors. The demonstration by Dumke, Woodall, and Rideout [2] of IBM, however, was the first to use a single-crystal structure throughout the device and thus avoid interface recombination effects. They accomplished this by growing a GaAlAs emitter on a GaAs base using liquid-phase epitaxy. Further progress, however, had to await the development and maturing of materials growth techniques such as molecular beam epitaxy and metal-organic vapor-phase epitaxy.

In historical context, progress demonstrating improvements related to the above and to the maturing of the technology are best represented by chronological plots of relevant device and circuit figures of merit. **Figures 1** and **2** contain plots of such figures of merit, based on data obtained over the past ten years: the logic gate delay, expressed by the gate frequency f_g, which is related to the switching delay τ_D of unloaded ring-oscillators by $f_g = 1/2\tau_D$; the unity current gain cut-off frequency (f_T); and the unity unilateral power gain frequency (f_{max}) for (Al,Ga)As/GaAs HBTs. Figure 2 indicates the corresponding increase in the number of transistors in a circuit (a measure of integration). The

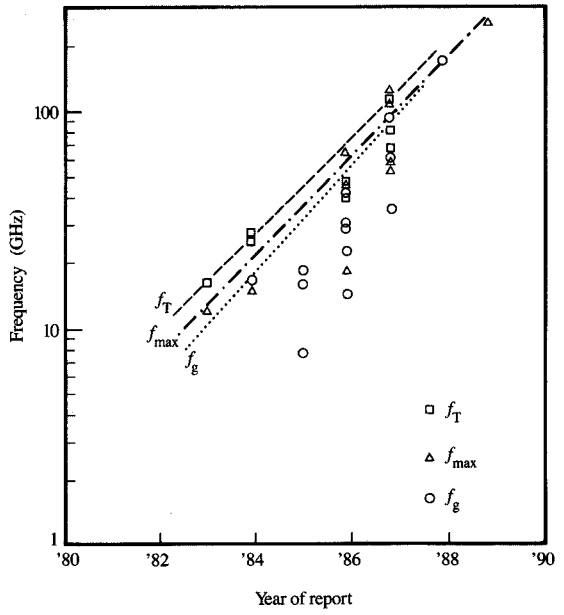

Figure 1

HBT figures of merit over the past ten years. Data are shown for the unity current gain cut-off frequency f_T, the unity unilateral power gain frequency f_{max}, and the gate frequency $f_g = 1/2\tau_D$, where τ_D is the ring-oscillator gate delay (the toggling speed of an unloaded logic gate). The latter represents the maximum functional speed of a logic gate driving another identical logic gate.

exponential increases in capability with time are similar to the trends of silicon technology. We do not know, as yet, of any fundamental problem that should limit growth in the integration level of heterostructure devices up to that of silicon bipolar transistors. To date, the performance figures of merit of HBTs are unparalleled. The shortest unloaded gate delay measured in HBTs is 1.9 ps [3]. The highest levels of integration which have been achieved are 13 000 transistors in a microprocessor [4] with an emitter-down circuit, and 5000 transistors with an emitter-up circuit [5], resulting from an IBM/Rockwell effort. These achievements represent the state of the art in performance and integration for the compound semiconductor technologies. The gate delays are the lowest achieved for any three-terminal semiconductor device operated at any temperature.

However, the technological limitations of compound semiconductors are accentuated in this inherently more complicated device, and additional physical effects occur because of the use of heterostructures and compound semiconductors. Our knowledge of minority-carrier

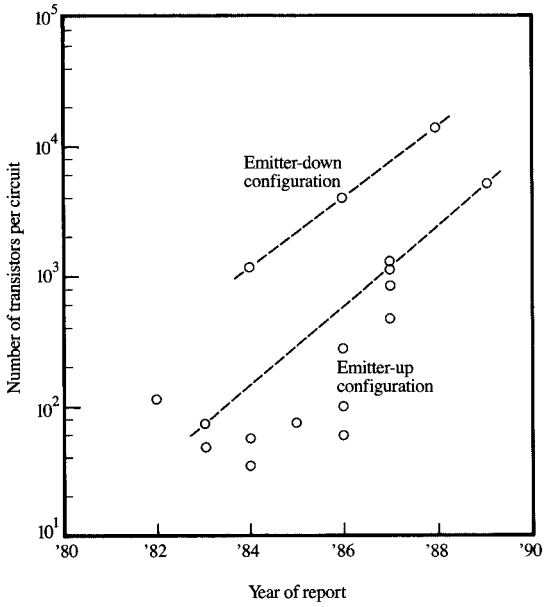

Figure 2

Integration level, in number of HBTs per circuit, utilizing emitter-up and emitter-down configurations, over the past ten years. The emitter-up configuration, more commonly employed, results in faster circuits.

behavior is also inadequate. This paper discusses these aspects of the subject, emphasizing our own contributions. This is followed by a discussion of further refinements in modeling to accurately reflect the physical basis of these fast devices, and potential developments that may result from further scaling of their dimensions, temperature of operation, and use of smaller-bandgap materials.

Ohmic contacts

Although compound semiconductor microwave devices have been of interest for more than 25 years, starting with Gunn diodes, and digital devices have been of interest for more than ten years, starting with MESFETs, an acceptable and reliable processing technology for large-scale integration has remained elusive. Among the prominent reasons for this are problems related to achieving ohmic contacts which are compatible with the processing requirements of high levels of integration. Bipolars and complementary FETs require both n- and p-type contacts in the same semiconductor medium. The ohmic contact problem in compound semiconductors is particularly difficult because of low solubility limits of dopants, increasing compensation and diffusion with increasing doping, rapid oxidation of surfaces, and Fermi-level pinning of the surface at nearly midgap in GaAs. A commonly used n-type contact based on the eutectic alloy AuGe is formed at 420–450°C. All subsequent processing must be at lower temperatures, and it is difficult to make these contacts sufficiently shallow for bipolar devices. The maximum achievable n-type dopant concentration in GaAs is low- to mid-10^{19} cm^{-3} for present growth techniques, and this limits the tunneling contact resistance which can be achieved with nonreactive metallurgies. Significantly higher p-type doping levels can be achieved by growth; hence, it is theoretically possible to achieve contact resistances less than 10^{-7} Ω-cm^2. In practice, high-level p-type doping during epitaxial growth causes anomalous redistribution during growth, and this also limits both device design and processing temperatures. The high-level p-doping can also be achieved by diffusion, but because of concentration-dependent diffusion, this usually results in the formation of a very deep contact. We first describe our approaches to solving the problems related to both n- and p-type contacts.

It is desirable to make use of ohmic contacts which withstand high-temperature processing, such as that which activates an ion implantation. This property greatly facilitates the formation of self-aligned structures, and virtually guarantees that the contacts will be stable during subsequent processing. Furthermore, it is desirable to be able to dry-etch the metal layer which is used in the ohmic contact. Both of these properties require the use of a refractory metal as the major constituent in the contact metallurgy. One of our earliest investigations [6] involved the creation of a refractory silicide-like n-type ohmic contact based on diffusion. We chose to use molybdenum–germanium because of earlier evidence of out-diffusion of gallium from GaAs into molybdenum at elevated temperatures. It was assumed that this would favor the formation of a heavily doped n-type layer from diffused germanium in vacant Ga sites. The elevated processing temperatures required an As overpressure to prevent decomposition of the surface, and we chose the simple technique of using InAs as a source of As at those temperatures. Early experiments were successful, and we also succeeded in fabricating heterostructure bipolar transistors 1.6 μm in minimum dimensions. The ohmic contacts, however, showed poor reproducibility, and it became clear from the work of Murakami et al. [7] that small amounts of In from the InAs played a central role in lowering the contact resistance. Murakami et al. have further exploited this in creation of refractory-metal ohmic contacts with small amounts of In [8], which appear to be particularly suitable for MESFET-like devices.

The role of indium in contact formation has been apparent for some time: Indium-alloy contacts to GaAs have been routinely employed over the last three decades as simple ohmic contacts for Hall-effect measurements to characterize epitaxial layers. However, due to melting during alloying, these contacts penetrate deep into the GaAs substrate, limiting the use of contacts with In as a major constituent to only the simplest applications. The use of In as a constituent in shallow contacts has been shown relatively recently. It is believed that during cooling of indium-alloy contacts, a thin interfacial layer of heavily doped (In,Ga)As is formed. It was first proposed, and demonstrated by Woodall et al. [9], that a suitably graded layer of (In,Ga)As can provide a low-resistance contact to GaAs. The graded-gap layer was grown by molecular beam epitaxy, following the growth of the n-GaAs layer to which contact was to be achieved. The (In,Ga)As layer was approximately 200 nm thick, and the top-layer metal was deposited after air exposure. Subsequent work by Wright et al. [10, 11] has shown that very low resistance contacts to GaAs can be achieved with very thin graded gap layers, approximately 10–20 nm thick. If the GaAs interfacial region is doped heavily enough, even nearly abrupt layers of n^+ InAs yield a low-contact resistance after short-term heat treatment.

We believe that there are several mechanisms which govern the formation of a low-resistance contact in such structures. First, the Fermi level is pinned at the interface between the (In,Ga)As layer and the top metal layer. The Fermi-level pinning position is a smooth function of the In mole fraction, varying from a barrier of 0.8 eV for GaAs to an estimated pinning position of −0.1 eV in the conduction band for pure InAs. In this manner, if the In mole fraction is high enough, the metal/(In,Ga)As interfacial barrier becomes low, and largely *independent* of the top contact metal. Second, if the doping in the GaAs and (In,Ga)As layers is high enough, the (In,Ga)As layer can be graded so that no barrier to electron flow remains. Third, it is believed that As-decorated dislocations pin the Fermi level in a manner similar to the free surface [12]. At low doping levels, the dislocation structure can trap enough electrons to form a barrier to current flow.

The best contact structures fabricated to date have incorporated a 20–30-nm graded-gap layer of (In,Ga)As, followed by a thin cap of InAs. Both top layers are heavily doped to 10^{19} cm^{-3}. A cross-sectional transmission electron microscope (TEM) image of such a structure is shown in **Figure 3**. Due to the large lattice mismatch (7% between GaAs and InAs), the dislocation density is extremely high. The strain fields associated with the dislocation arrays are striated in the direction of growth, as might be expected for the growth of a compositionally graded mismatched layer. Because of the lattice mismatch, it is difficult to grow smooth layers, at least at substrate temperatures which yield good electrical properties. Typically, this roughness is of the order of the average thickness of the mismatched layer itself, and prohibits fine-line lithographic processing. We have found that the use of modulated molecular beams during growth can produce smooth lattice-mismatched layers, forcing the growth mode toward two-dimensional growth [13].

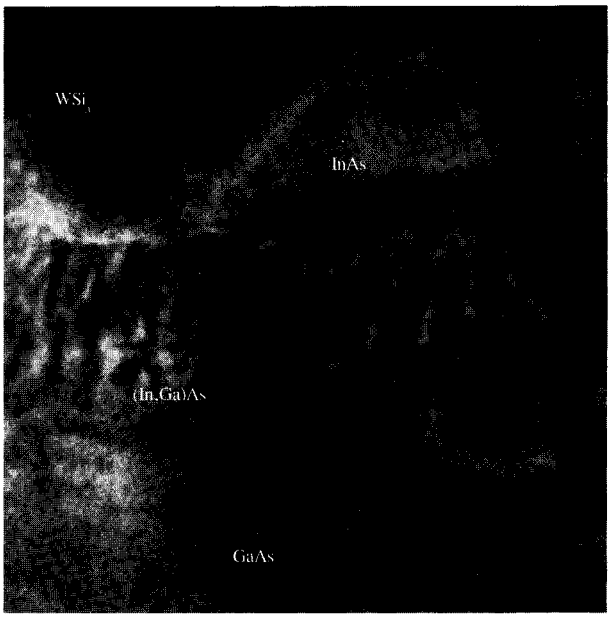

Figure 3

Cross-sectional transmission electron micrograph of an InAs n-type ohmic contact region prior to annealing. GaAs is the semiconductor to which the ohmic contact is to be made. (In,Ga)As is used as a graded intermediate layer, with InAs as the surface layer. WSi_x is the refractory metal which is deposited to complete the contact. Each spot in the lattice image portion of the micrograph corresponds to an anion–cation pair of atoms (lattice constant ≃2.8 Å).

Because of the relatively high melting point of (In,Ga)As alloys, this ohmic contact structure is stable during the rapid thermal treatment used to activate ion implantation. **Figure 4** is a TEM image of the same graded-gap contact of Figure 3, after an 18-second heat treatment at 850°C. There is relatively little structural change, and the contact resistance is slightly lower than the as-grown value. After heat treatment, the dislocation strain fields have a distinctive "swirl-like" character, probably indicating intermixing and/or phase separation on a nanometer scale.

Of the two n-type contacts in the npn bipolar transistor, the emitter contact is the most critical. The highest current density in the device is in this region, and the contact resistance to the emitter strongly degrades circuit performance. The requirements for a high-

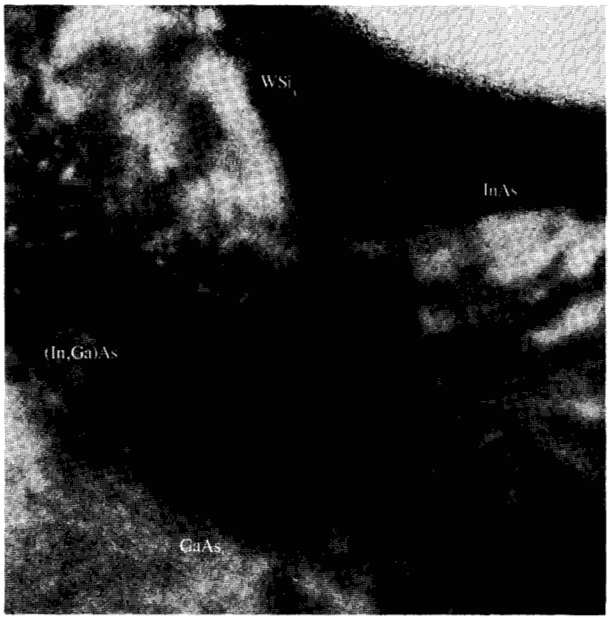

Figure 4

Cross-sectional transmission electron micrograph of the ohmic contact of Figure 3, after annealing for 18 seconds at 850°C.

performance HBT are severe: Even a relatively good contact resistance of 1×10^{-6} Ω-cm^2 will result in an emitter contact resistance of 100 Ω for an emitter area of 1 μm^2. The contact resistance in properly designed and grown graded-gap structures is lower than 1×10^{-7} Ω-cm^2, the lowest value which can be reliably measured by usual characterization methods. There is evidence that the actual contact resistance is much lower, and is limited by the top metal/InAs interface, in spite of the Fermi-level pinning in the conduction band of the InAs.

In our initial experiments with p-type contacts, we used magnesium ion implantion to create heavy doping [14], and good device characteristics were thus obtained. However, ion implantation, as a technique, suffers in two important respects for optimized device performance. First, reduction of the extrinsic base sheet resistance is achieved at the penalty of decreasing the lifetime in the extrinsic material, which lowers the gain. Second, activation in (Al,Ga)As and in GaAs shows dissimilar characteristics, with (Al,Ga)As requiring the use of higher temperatures and showing significantly poorer ohmic contacts. In order to maintain the gain of devices at small dimensions, carrier lifetime must be high. However, we recognized at the beginning of this effort that it is more important to place a larger-bandgap barrier layer at the extrinsic surface—a purpose well served by (Al,Ga)As, as discussed in the following section. The poorer ohmic contacts to (Al,Ga)As resulted in the abandonment of ion implantation in favor of another technique based on diffusion of Zn by rapid thermal annealing [15]. Diffusion of zinc, which is concentration-dependent, usually leads to large dopant tails but also results in high concentrations at the surface, permitting the formation of excellent ohmic contacts, even to (Al,Ga)As.

The technique of rapid thermal diffusion using Zn_3As_2 resulted in excellent ohmic contacts to both GaAs and (Al,Ga)As and limited the vertical and lateral diffusion.

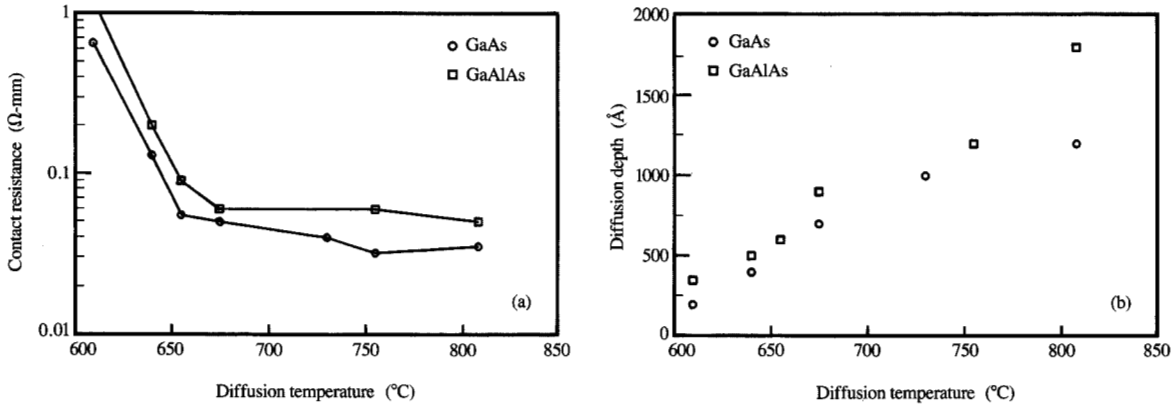

Figure 5

(a) Contact resistance of W(Zn) p-type ohmic contacts to GaAs and $Ga_{0.7}Al_{0.3}As$ as a function of the temperature at which the ≃6-second diffusions are performed. (b) Associated diffusion depth of the contacts as a function of temperature for the ≃6-second diffusion.

The latter is dependent on stress due to dielectric films at the interface, and can be suppressed. We found that this diffusion phenomenon also occurs through W films, only to discover later that Marinace [16] had previously observed this in 1970 in our own laboratory. This permits the advantageous *in situ* formation of a tunneling ohmic contact using a low-resistivity refractory metal. Since compound semiconductors do not exhibit a reproducible and controlled reaction with metals, such as formation of silicides with Si, this is a unique method that avoids interface problems related to surface oxides in the formation of a tunneling contact. Additionally, the process involves the use of diffusion, as for the p-type polysilicon base contact and diffusion step in silicon bipolar processing; it is therefore a prototypical equivalent of polysilicon or "polycide," and may have similar uses elsewhere.

The selectivity of this diffusion is accomplished by using the insulating layer of SiN or the $WSi_{0.6}$ layer that forms the emitter metal contact. The drawback of this technique is that the time associated with establishing the vapor pressure of Zn is about the same as that associated with diffusion through the W film itself. These diffusion times are a few tenths of a minute. A simple solution to this problem was to make use of a composite film of W and Zn in which a limited quantity of Zn was incorporated. Use of such a film (doped in excess of 2% for a few hundred angstroms, followed by undoped W film) results in contact resistances of mid-10^{-7} Ω-cm^2. **Figures 5(a)** and **5(b)** show the associated contact resistance and diffusion depth for GaAs and (Al,Ga)As as a function of rapid thermal processing temperature. Good contact resistances (<0.1 Ω-mm) and low diffusion depths can be achieved simultaneously. Such a contact has more general applications; e.g., it has also been applied to field-effect devices [17, 18].

Using such ohmic contacts, self-aligned HBT structures can be implemented with a base contact metal which surrounds the emitter and lowers the extrinsic base resistance. Because of self-alignment, the extrinsic base semiconductor resistivity requirements are relaxed. Also, reduced base–collector junction area is possible without sacrificing base contact resistance. The HBT structure that was implemented [18] is shown in **Figure 6**; associated device design parameters and resulting electrical parameters are listed in **Table 1**. The lumped resistance and capacitance values are modeled values obtained from microwave measurements; their partitioning into components is based on our estimates. In the table, the base diffusion capacitance is that associated with the minority carrier charge storage in the base; and the extrinsic and intrinsic collector capacitances are the space-charge-region capacitances associated with the base–collector junction in the intrinsic and extrinsic

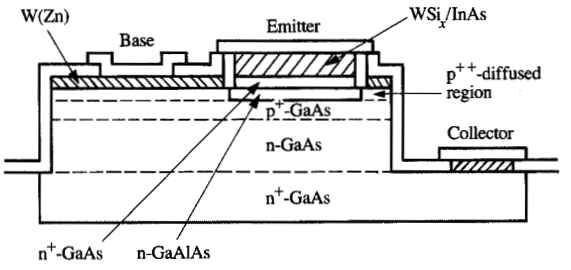

Figure 6

Cross-sectional schematic of self-aligned HBT described in the text.

Table 1 Nominal parameters for the HBT of Figure 6.

Design parameters	
Emitter dimensions:	$0.8 \times 1.6\ \mu m^2$
Emitter GaAs doping:	5×10^{18} cm^{-3}
Emitter GaAs thickness:	1200 Å
Emitter GaAlAs doping:	8×10^{17} cm^{-3}
Emitter GaAlAs thickness:	1200 Å
Emitter GaAlAs mole fraction:	0.22
Emitter GaAlAs grading length:	300 Å
Intrinsic base width:	1000 Å
Intrinsic base doping:	5×10^{18} cm^{-3}
Extrinsic base doping:	$\sim 1 \times 10^{19}$ cm^{-3}
Collector doping:	1×10^{17} cm^{-3}
Collector thickness:	2500 Å
Subcollector doping:	5×10^{18} cm^{-3}
Subcollector thickness:	2500 Å
Electrical parameters at an operating current of 0.5 mA	
Emitter resistance:	50 Ω
Emitter contact:	44 Ω
Semiconductor body:	6 Ω
Base resistance	40 Ω
Base via:	1 Ω
Base contact:	7 Ω
Base metal:	7 Ω
Extrinsic semiconductor body:	25 Ω
Intrinsic semiconductor body:	12 Ω
Collector resistance:	31.5 Ω
Collector contact:	14.0 Ω
Subcollector body:	11.5 Ω
Collector body:	6.0 Ω
Capacitances:	
Emitter capacitance:	11.6 fF
Base diffusion capacitance:	39.0 fF
Extrinsic collector capacitance:	14 fF
Intrinsic collector capacitance:	7.5 fF

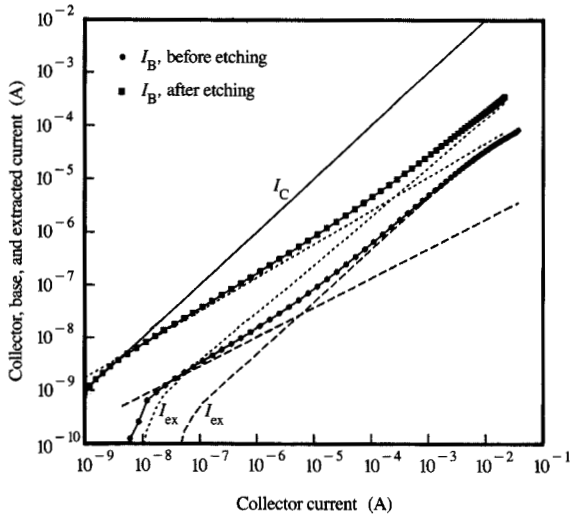

Figure 7

Collector, base, and extracted current vs. collector current, before and after etching of the surface passivation layer of an HBT. The surface passivation is accomplished by use of the barrier layer of p-type (Al,Ga)As at the surface of the extrinsic base region. I_C is the collector current, I_B is the base current, and I_{ex} is the exponential component of base current dominant in the high-current region. The ratio of emitter perimeter to area is 6.7×10^3 cm^{-1}.

part of the HBT. The n-type contact is based on InAs, and the p-type contact is based on Zn diffusion through the wide-gap n-type emitter region, which becomes p-type in the extrinsic base region. The parasitic resistances of the device are sufficiently low to allow operation at an emitter current density of mid-10^4 A-cm^{-2}.

We next discuss the physical effects and properties unique to the compound semiconductors and their HBT implementations, beginning with the important problem of surface recombination.

Surface recombination

One of the more important developments in (Al,Ga)As/GaAs HBTs has been the strong reduction of surface recombination effects [19] in these devices. GaAs exhibits a high surface recombination velocity because of Fermi-level pinning and poor lifetime at the surface. As a result, injected carriers in the base exhibit additional base recombination at the GaAs base surface. This is reminiscent of the behavior in germanium bipolar transistors, except that it is considerably stronger, and physically different because of the strong Fermi-level pinning in GaAs. As device dimensions decrease, these surface features assume increasing importance. They lead to a significant decrease in current gain. Use of the wider-gap p-(Al,Ga)As at the surface introduces an additional barrier ($\Delta E_g \sim 0.4$ eV at $x = 0.3$) which suppresses the injection of carriers to the surface. This has allowed fabrication of high-gain devices that also exhibit a much weaker dependence of current gain on device dimensions.

An estimate of this surface recombination can be obtained by removing this p-(Al,Ga)As region from an HBT device [20], as shown in **Figure 7**. After etching, the base current (I_B) increased by more than an order of magnitude, while the collector current (I_C) remained unchanged. The increase in base current is attributable to the surface recombination. It exhibits both a slow bias dependence at low biases [$\propto \exp(V/\sim 2kT)$] and a fast bias dependence at high biases [$\propto \exp(V/\sim kT)$], with different prefactors.

This experimentally observed change can be explained theoretically [20] and is a consequence of band-bending at the surface due to the presence of surface states. For both GaAs and (Al,Ga)As, under thermal equilibrium, the Fermi level is pinned at ~0.8 eV below the conduction band. **Figure 8** shows the conduction-band edge at and near the surface of an HBT with a graded heterojunction, at a base-emitter forward bias of 1.2 V. The electrons injected from the emitter into the base experience a lower barrier height at the surface because of the Fermi-level pinning. The injected electrons at the surface channel along the GaAs base surface and recombine. The recombining electrons *do not* come primarily from the quasi-neutral region of the base; instead, they are injected from the emitter junction along the surface. Fermi-level pinning plays a central role in this behavior. For the same recombination parameters, but in the absence of Fermi-level pinning, the surface recombination current is significantly smaller because in this case the recombining electrons originate only in the quasi-neutral region of the base. This potential barrier at the surface is a function of the bias and the design of the junction which determine the boundary conditions for the theoretical treatment leading to the exponential behavior which is observed.

An interesting consequence of the Fermi-level pinning is that abrupt barriers show lower surface recombination. **Figure 9** shows a similar conduction-band edge profile for an abrupt barrier device at the same 1.2 V base–emitter forward bias. As a result of the use of an abrupt heterojunction, even though the barrier at the surface "corner" is low compared to that in the bulk of the base, it is significantly larger than that in the graded heterojunction case. A higher barrier to the surface injection results in fewer electrons being injected into the surface channel and less recombination current.

Surface recombination is one example of an important minority-carrier feature that limits the operation of the

device. The rate at which minority carriers recombine in the quasi-neutral base and the way in which they drift and diffuse in the base are other examples of minority-carrier behavior which influence device performance. Bandgap narrowing is strongly dependent on doping level and is important to carrier transport. Accurate characterization and parameterization of these features, as a function of doping level, is very important to the modeling and assessing of the HBT. However, not enough is known about the transport and recombination of minority carriers under the heavy doping conditions employed in the base. This is probably because most prior work on minority carriers was stimulated by an interest in optical devices such as lasers and solar cells. Fortunately, the HBT is an excellent tool for characterization of these effects [21]. The following section summarizes the functional dependences of the more important parameters: minority-carrier mobility (and hence diffusivity using the Einstein relationship), minority-carrier lifetime, and bandgap narrowing as relevant to the base of an npn HBT.

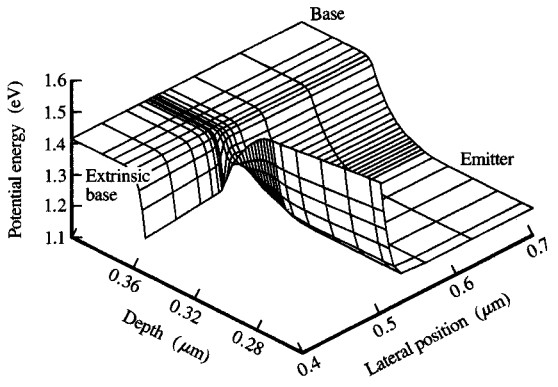

Figure 8

Conduction-band edge of an HBT with a graded heterojunction under a forward bias of 1.2 V, applied at the base–emitter junction. Band-bending at the surface is due to Fermi-level pinning.

Material parameters

In the absence of surface recombination, the current gain of an HBT at medium to high current densities is determined by the base transport factor $\beta = \alpha/(1-\alpha) \simeq 2(L_B/W_B)^2$ for $L_B \gg W_B$, where L_B is the diffusion length and W_B is the base width. Thus, the current gain may be used to determine the diffusion length of electrons in the base of an npn HBT. The diffusivity (or mobility) can be determined by measuring the frequency response of an HBT which has a very thick base, because in such a device the frequency response is limited by the base transport. For example, Nathan et al. [22(a)] were able to measure the electron mobility in a p-type base which was 1 μm thick. There also exist independent measurements at various other doping levels from which the minority-carrier mobility can be extracted [23, 24]. **Figure 10** shows measured values of the electron mobility in p-type GaAs together with a fit that we have found to be in agreement with modeled base time constants obtained from published mobility values and from microwave parameters we obtained. At 300 K this fit for minority-carrier mobility is given by the relation

$$\mu_n^p = \frac{8300\ \text{cm}^2\text{-V}^{-1}\text{-s}^{-1}}{\left(1 + \dfrac{N_A}{3.98 \times 10^{15} + N_A/641}\right)^{1/3}}, \qquad (1)$$

where N_A is the base acceptor doping in cm^{-3}. A subject of theoretical controversy has been whether the minority-carrier mobility should be larger or smaller than the majority-carrier mobility at the same dopant concentration. Experimental measurements on GaAs indicate that the minority-carrier mobility is smaller,

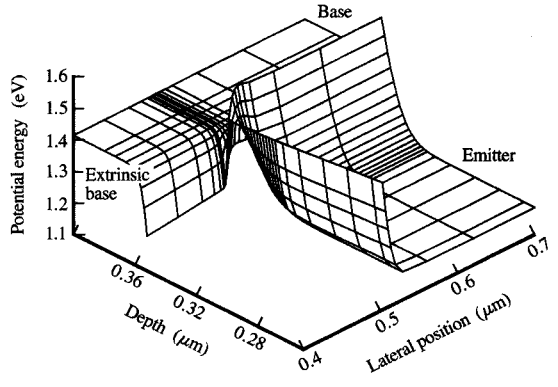

Figure 9

Conduction-band edge of an HBT with an abrupt heterojunction, under a forward bias of 1.2 V, applied at the base–emitter junction. As in Figure 8, band-bending at the surface is due to Fermi-level pinning.

presumably because electrons as minority carriers have a lower Fermi energy than electrons as majority carriers at the same doping level [25], leading to stronger impurity scattering.

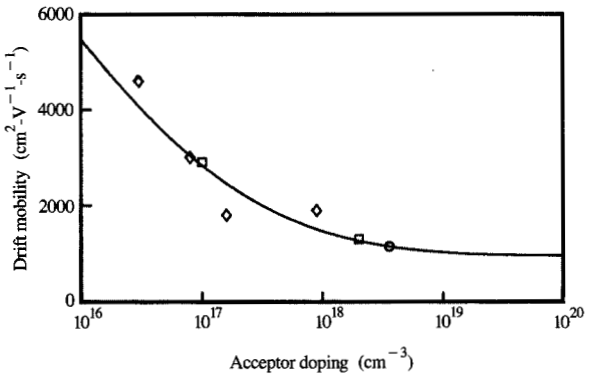

Figure 10

Drift mobility of minority carriers in p-type GaAs in the doping range of interest for an HBT. The open circles are from the work of Nathan et al. [22(b)], the open squares from time-of-flight measurements of Ahrenkiel et al. [23], and the open diamonds from photoluminescence-based measurements of Nelson [24].

Diffusion length is another important parameter because the gain of an HBT is limited by the base transport factor. **Figure 11** contains a compilation of measurements of the diffusion length in p-type GaAs, from our work and the literature. We are interested in the high doping levels that occur in the base of an HBT. The diffusion lengths represented by open circles were derived from published current gain data obtained from HBTs which were not passivated; hence their gain was degraded. The associated diffusion lengths were thus underestimated. The results represented by the solid circles are based on our measurements of surface-passivated HBTs and thus are higher than the other results. The results represented by open diamonds are based on luminescence measurements [24, 26] and hence are based on optical response. There is considerable scatter in the data, although the highest diffusion lengths as a function of doping show a monotonically decreasing trend. The large scatter at low doping levels is probably related to the various growth techniques employed, with liquid-phase epitaxy producing the largest diffusion lengths. Scatter is also introduced by different interpretations of the luminescence measurements to account for surface recombination. The scatter reduces with increasing doping level because of the decreasing importance of unintentional impurities and compensation of the material.

Using the mobility relationship of Equation (1), the electron lifetime can be obtained as a function of acceptor doping, as shown in **Figure 12**. The open circles represent values derived from published results on unpassivated bipolar transistors assuming no surface recombination; they are therefore underestimates. The highest doping level in the base [27] is 2×10^{20} cm^{-3}. At such a doping level, the lifetime is very low and the Fermi level may be unpinned. The effect of surface recombination relative to the quasi-neutral base recombination is small, and the lifetime is probably not a significant underestimate. The highest lifetime data displays a behavior similar to that of silicon [28]. At the highest doping levels, the lifetime is expected to be limited by Auger processes, and at lower doping levels by Shockley–Read–Hall processes and additionally by radiative recombination processes. In Figure 12, the two lines represent empirical fits to the behavior of the electron lifetime with doping in p-type GaAs for the two processes, which bound the lifetime data. The lifetime τ_n may be represented by

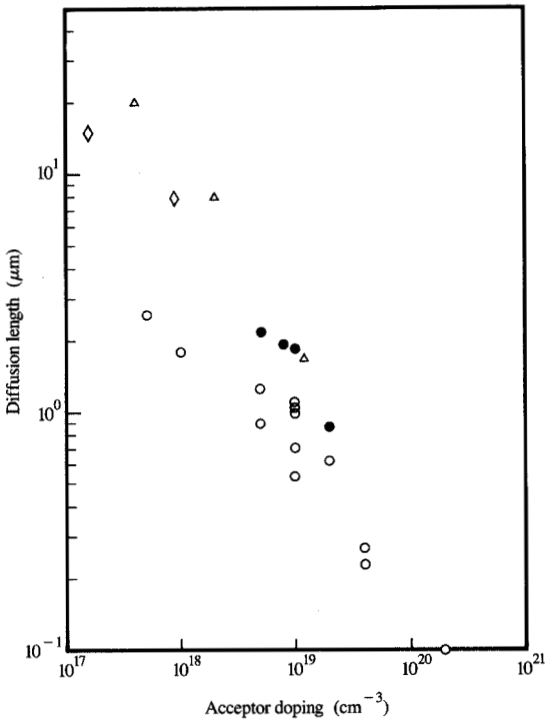

Figure 11

Diffusion length as a function of acceptor doping for GaAs in the doping range of interest for an HBT. The solid circles represent data obtained from our work. No surface recombination is assumed; therefore the diffusion length is underestimated. The open circles represent values derived from published bipolar device results [21]. The open diamonds are from photoluminescence-based measurements of Nelson [24]; the open triangles are from the work of Jastrzebski et al. [26].

$$\frac{1}{\tau_n} = \frac{1}{\tau_{SRH,Rad}} + \frac{1}{\tau_A} = \frac{N_A}{1 \times 10^{10}} + \left(\frac{N_A}{4 \times 10^{14}}\right)^2, \qquad (2)$$

where the three indicated lifetimes are expressed in units of seconds and the doping in units of cm^{-3}. The first term is due to limitation by Shockley–Read–Hall recombination and radiative recombination, and the second term is due to limitation by Auger recombination. This behavior is similar to that obtained by Henry et al. [29] for $Ga_{0.53}In_{0.47}As$ using measurements based on luminescence decay. The lifetime τ_A limited by Auger recombination is small and seems to occur at modestly high doping levels, perhaps as a result of Auger processes involving the split-off band.

We conclude our discussion of material properties with a discussion of bandgap shrinkage effects, which are so important in Si bipolar transistors. In contrast to the homostructure silicon bipolar, bandgap narrowing in HBTs is strongest in the base region. This reduces the electron injection barrier and allows formation of a larger collector current for a given bias level. It is beneficial because it reduces the junction voltage for constant-current operation, and hence the power dissipation. Following the earlier work in Si devices, the saturation collector current can be used to obtain the effective intrinsic carrier concentration (an electrical fitting parameter) because the Gummel number in the base is known. Either this increase in effective intrinsic carrier concentration can be used directly as a parameter in transistor-related calculations, or, as is quite often done, an effective bandgap narrowing can be derived using the low doping intrinsic carrier concentration ($n_{i0} = 2.25 \times 10^6$ cm^{-3}; this lumps together band-tailing, anisotropy, degeneracy, and other effects). **Figure 13** shows this effective bandgap narrowing as a function of acceptor concentration, derived by the above technique and verified using activation-energy measurements of junction transport. For reference, we show the optically derived fit of Casey and Stern [30], and the bandgap fit derived from the calculations of Bailbe et al. [31] of the intrinsic carrier concentrations in GaAs. Our data points (the solid circles) are more in agreement with the bipolar measurement of Klausmeir-Brown et al. [32] and also show good agreement with Bailbe's predictions. The bandgap narrowing, however, is unexpectedly significantly larger than the effective "optical" bandgap narrowing obtained by Casey and Stern. The large disagreement in the electrical and optical data

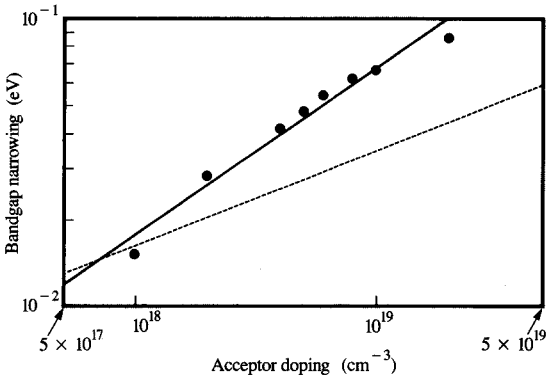

Figure 13

Effective electrical bandgap narrowing as a function of base doping; the narrowing can be regarded as a parameter associated with current transport across the base. The solid line is the fit to this narrowing. The dashed line shows the effective optical bandgap narrowing obtained from the work of Casey and Stern [30].

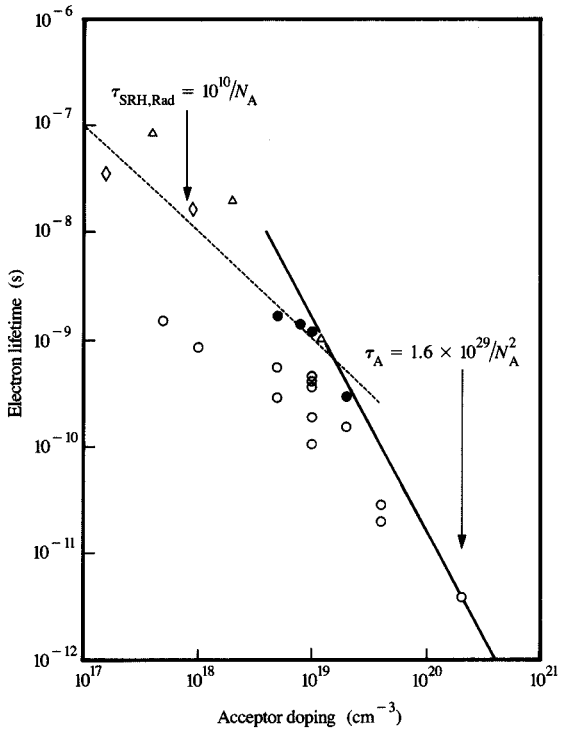

Figure 12

Electron lifetime in p-type GaAs. The solid circles represent data obtained from our work [21]. The open circles are values derived from published bipolar device results. No surface recombination is assumed; therefore the lifetime is underestimated. The open diamonds are from photoluminescence-based measurements of Nelson [24], and the open triangles are from the work of Jastrzebski et al. [26].

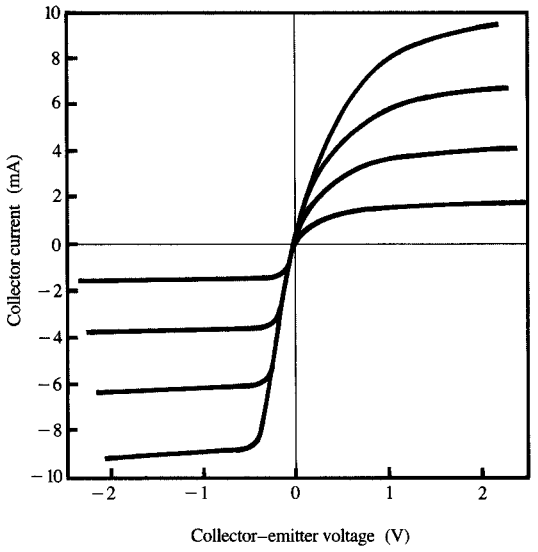

Figure 14

Output characteristics of a double-heterostructure HBT, showing symmetry of the dependence of collector current on collector–emitter voltage. The base current steps are each 20 μA; the device dimensions are 70 × 70 μm^2.

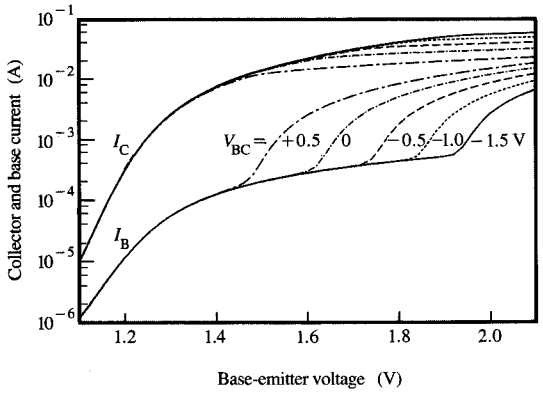

Figure 15

Collector and base current at a high current density, for an HBT with a wide-gap collector. Emitter dimensions are 4 × 12 μm^2.

presumably is because the former is a fitting parameter for conduction processes, and the latter is derived from optical measurements, assuming a specific band-tailing and screening behavior [25]. A similar discrepancy was observed in early Si measurements [33]. A fit for the bandgap narrowing is given by the relation

$$\Delta E_g = 2.0 \times 10^{-11} \times N_A^{0.50}, \quad (3)$$

where the units of bandgap shrinkage are in eV and the doping level units in cm^{-3}. At a base doping level of 5×10^{19} cm^{-3}, the total bandgap shrinkage is 0.17 eV, corresponding to an approximate 12% reduction in the junction turn-on voltage.

This reduction in turn-on voltage results from the reduction in base bandgap. In general, the current–voltage characteristics can also depend on the composition profile of the heterojunction itself. Abrupt heterojunction barriers contain a conduction-band discontinuity (e.g., 0.24 eV for $Ga_{0.7}Al_{0.3}As/GaAs$) which creates an excess conduction-band barrier. The latter limits the current and increases the turn-on voltage. As discussed in the next section, at high currents a graded heterojunction can also develop a barrier which limits current flow.

Heterobarrier design

The double-heterostructure HBT, which employs a heterostructure collector as well as an emitter, displays several interesting features. It is suited to minimizing the offset voltage, provided surface recombination is suppressed, and it permits symmetric operation. Injection efficiency is dramatically increased in heterojunction emitters by the presence of the hole barrier, even at low emitter doping. As a result, even at low collector doping levels, such a device can exhibit large current gains when operated in the reverse mode, as shown in **Figure 14**. For current injected in the base lead, the output current can be modulated with a significant gain for both polarities of emitter–collector bias. The device may be used in reverse mode and forward mode on the same chip, if higher integration and lower power density are desired, most likely at the cost of slower speed. An example of this is the combining of I^2L and ECL circuits on a chip to accomplish objectives similar to those of the BiCMOS technology. Since collector capacitance is reduced in "collector-up" structures, they need not be slow devices. Self-aligned collector-up devices should be quite fast and should exhibit improved radiation resistance because of the use of a common emitter ground plane.*

The use of the heterostructure in the collector without a systematic redesign can, however, lead to a surprising loss in speed and gain [34]. An example of such a loss in gain is shown in **Figure 15**, which contains plots of the collector current (I_C) and base current (I_B), at a high current density, for an HBT with a wide-gap collector.

* P. M. Solomon, IBM Thomas J. Watson Research Center, unpublished work.

The base current increases rapidly at a current density of 1×10^5 A-cm^{-2} because of unusual effects resulting from the use of the heterostructure collector. The change in the alloy potential at a graded heterojunction gives rise to a quasi-field similar to that due to an electrostatic potential. When the electrostatic fields in a heterostructure collector of GaAs/(Al,Ga)As are larger than this quasi-field, only the barrier to hole transport is significantly affected. However, at a high current density, the electrostatic field is reduced, and the field due to the alloy potential gradient may dominate in a part of the junction, resulting in a barrier to electron flow into the collector. This is similar to, but substantially smaller than, what happens at the extrinsic p^+ GaAs/(Al,Ga)As surface. The barrier results in increased accumulation of carriers in the base, and hence causes a decrease in the base transport factor and the gain. A similar effect occurs in the base–emitter junction, leading to a rapid saturation of the injected current density [34, 35]. This particular behavior controls the transport of carriers across a heterojunction in a way that is significant and that is uniquely different from that across a homojunction. It becomes even more important with scaling because of the increased current density in smaller devices, and the reduction in electrostatic fields if the doping levels of the emitter or collector are not increased. We have described the appearance of this barrier, at high current densities, in devices containing a graded-heterojunction collector. If an abrupt or semi-abrupt heterojunction with a conduction-band discontinuity were employed, the poor transmissive properties across the barrier would limit the current flow across it. These devices would similarly display poor speed and gain because of excess storage in the base at even lower current densities.

The high current density across a heterojunction presents a particularly interesting problem because it is concomitant with a smaller space-charge region (due both to the larger doping that a larger current capability requires and to the forward biasing of the junction). The smaller space-charge region and decreased electrostatic field lead naturally to a barrier as a result of the alloy potential in the region of the heterojunction. The grading length needed to prevent barrier-limited transport effects in the depletion approximation limit can be derived by consideration of the electrostatic and alloy potential variation. If the electrostatic field remains larger than the alloy field, the band-bending remains monotonic, and no barrier exists. Mathematically, this requirement can be summarized as

$$\frac{d\phi}{dz} < E_{cs}, \tag{4}$$

where E_{cs} is the local electrostatic field and ϕ the alloy potential. For a one-sided p–n junction such as in an HBT, this results in

$$W_g > \frac{\Delta E_g}{\sqrt{2kT(E_g - qV - 2kT)}} L_D, \tag{5}$$

where W_g is the grading width, ΔE_g the total bandgap change, q the electronic charge, E_g the bandgap in the base, V the applied bias, and L_D the extrinsic Debye length. The expression is derived assuming the depletion approximation, and hence is inaccurate at high forward-bias conditions, for which the mobile charge should be taken into consideration. Under low forward-bias conditions, a grading distance longer than L_D is sufficient for suppressing this barrier. At high forward-bias conditions, the barrier eventually appears and limits the flow of current. If the doping level were allowed to increase as the inverse of the square of the scaling factor, the maximum current would continue to increase, since L_D decreases in proportion to the scaling factor ($\propto 1/\sqrt{N_D}$). Therefore, this does not place a limit on the design of foreseeable devices. However, at current densities of 1×10^7 A-cm^{-2}, impractical at present, it would place restrictions [36].

High-speed and high-frequency modeling

During the early 1960s much work was published on the small-signal modeling of the bipolar transistor, with careful attention to dispersive effects and transit-time effects in the collector space-charge region. The modeling involved calculations in one dimension, while the actual transistors had very substantial parasitics associated with their lateral structure. These lateral features, characterized by the collector capacitance, the base resistance, etc., actually limited their operation. The success of the Ebers–Moll and Gummel–Poon quasi-static models in predicting their static characteristics and frequency limits ended any academic interest in the dispersive effects. In modern devices, these parasitics have been significantly reduced by a reduction in horizontal dimensions through self-alignment, and intrinsic capabilities have been improved by reducing vertical dimensions. Dispersive and collector transit-time effects have become increasingly important today. For example, an HBT with a 1.9-ps logic delay [3] should have transit-time components that are a major factor of the total delay, and should also have significant frequency-dispersion effects.

The usual quasi-static analysis (e.g., Ebers–Moll or Gummel–Poon) ignores much of this, by making single-pole or single-zero approximations of the network parameters of the device. For example, transit of the carriers in the base is included in the form of a base time constant which is modeled by emitter conductance and diffusion capacitance. This time constant is the correct average transit time of a carrier in the base under steady-

state conditions. It has usually been sufficient even at moderate frequencies or switching speeds of fractions of f_T, because until recently f_T has been mostly limited by parasitics. Since this is no longer true, consideration must be given to the question of the methodology of the inclusion of the nonnegligible base dispersion and collector signal delay.

We now show how to include these by modifying the conventional steady-state analysis. We first consider transport in the base transit as an example. Such transport leads to a "minimum"-type phase delay. A good approximation for the base transport factor α of a homogeneously doped base is

$$\alpha = \frac{1}{\cosh \xi_0} \simeq \alpha_0 \frac{\exp(-j\delta\omega\tau_B)}{1 + j\omega\tau_B}, \tag{6}$$

where

$$\xi_0 = \sqrt{\left(\frac{W_B}{L_B}\right)^2 + j\omega \frac{W_B^2}{D_B}},$$

α_0 is the low-frequency base transport factor, ω is the radial frequency of applied small signal, D_B is the diffusion length for electrons in the base, τ_B is the usual base time constant, and δ is an excess phase factor ($\delta \sim 0.22$) [37]. The equivalent expression for a nonhomogeneously doped base or an alloy-graded base with a built-in quasi-field is more complicated but has a similar excess phase delay. In quasi-static analysis, this phase factor is ignored; only the pole response is taken into account.

The importance of the phase factor can be gauged by an exact series solution using Laplace transforms. The collector current response to a sudden change ΔI_B in the input current [37] is

$$i_c(t) = \frac{2\Delta I_B}{\sqrt{\pi}} \sum_{i=1}^{\infty} \left[\frac{1}{i} \exp\left(\frac{iW_B}{L_B}\right) \operatorname{erfc}\left(\frac{iW_B}{2\sqrt{D't}} + \sqrt{\frac{t}{\tau_B}}\right), \right.$$
$$\left. + \frac{1}{i} \exp\left(\frac{-iW_B}{L_B}\right) \operatorname{erfc}\left(\frac{iW_B}{2\sqrt{D't}} - \sqrt{\frac{t}{\tau_B}}\right) \right], \tag{7}$$

where $D' = L_B^2/\tau_B$. In the case of large times ($t \gg \tau_B$), this expression reduces to a simple exponential response involving a single time constant. In the limit of small times, however, it becomes

$$i_c(t) = \Delta I_B \times \sqrt{2t/\tau_B} \exp\left(-\frac{\tau_B}{2t}\right) \quad \text{for } t \ll \tau_B. \tag{8}$$

For a time period of $\sim 0.22\tau_B$ the collector current is negligible, and then it initially rises more slowly than would a pole response. This occurs for the same reasons that led to the phase factor in the frequency response. If we are interested in a reasonable characterization of signal response at times of $\sim\tau_B$, the pole response using τ_B as a time constant is inadequate. Complete incorporation of this effect in the time-domain predictions of ASTAP [38] programs such as those based on Gummel–Poon models or Ebers–Moll models is very difficult if not impossible. However, we may incorporate simplifications of it into the models. For example, in an Ebers–Moll model that already includes most of the effect of the base transport by employing a diffusion capacitance and a junction resistance, we may include an excess RC delay term of the form $[1 - \exp(-t/\delta\tau_B)]$ as a simple modification; the excess phase-factor term then becomes an additional capacitance term.

The collector signal delay is a more serious problem, because it is relatively more important than the base phase-factor term. The transport factor ζ for small-signal analysis, assuming a constant (saturated) transit velocity, is of the form

$$\zeta = \frac{\sin(\omega\tau_c')}{\omega\tau_c'} \exp(-j\omega\tau_c'), \tag{9}$$

where $\tau_c' = W_c/2v_s$ is a collector signal delay time constant, W_c is the width of the collector space-charge region, and v_s is the carrier velocity. In the modern transistor, the collector depletion region becomes short enough that nonequilibrium overshoot effects make this something of an overestimate. The prefactor of Equation (9) is usually very close to unity at frequencies up to the cutoff frequency, and its effect is usually ignored. Including Equation (9) in Equation (6) leads to the following approximate time domain expression:

$$i_c(t) = \Delta I_B \frac{\alpha_0}{1 - \alpha_0} \cdot \left\{1 - \exp\left[\frac{-(1 - \alpha_0)t}{\tau_B + \delta\tau_B + \tau_c'}\right]\right\} \quad \text{for } t \gg \tau_B. \tag{10}$$

This delay can also be approximately included in Gummel–Poon or Ebers–Moll models by including an additional capacitance together with the emitter conductance to model the collector signal delay. The error incurred in using this approach increases at time scales close to τ_c'.

The small-signal analysis of the device can be performed much more rigorously than the time-domain analysis. The intrinsic model of the device can be represented in terms of y-parameters which can be derived from the transport equations of the device. In such an analysis one usually ignores the emitter transit time and related signal delay. For an HBT under high forward bias, this is about one third of the base transit time, and even smaller still than the collector transit time. The derivation of collector transit also includes the assumption of a constant limited velocity for transport.

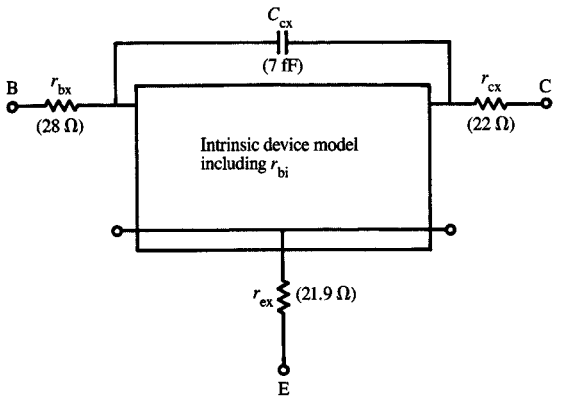

Figure 16

Small-signal model for intrinsic HBT device, lumped parasitic elements for its extrinsic portion, and network parameters.

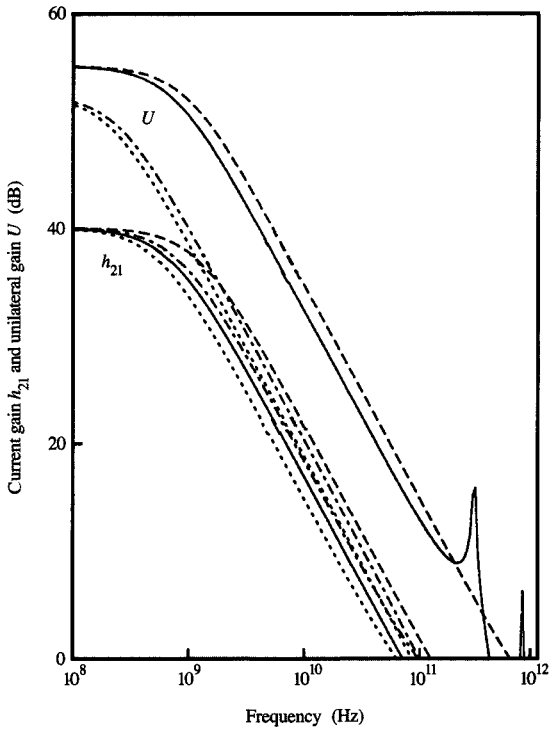

Figure 17

Effects of parasitics and collector signal delay on short-circuit current gain h_{21} and unilateral gain U of the HBT described in the text. The intrinsic device excludes the extrinsic lumped parasitics of Figure 16. Extrinsic device behavior that excludes the effect of collector signal delay is indicated by the dot-dash line; extrinsic device behavior that includes the effect of collector signal delay is indicated by the short-dash line; intrinsic device behavior that excludes the effect of collector signal delay is indicated by the wide-dash line; and intrinsic device behavior that includes the effect of collector signal delay is indicated by the solid line.

In such a device, velocity overshoot may occur because of the abrupt rise in electric field, and hence collector signal delays may be an overestimate. Fortunately, these emitter and collector transit approximations are in counterbalance, so presumably the net error is small. The y-parameters for the device in the common base mode [37] are

$$y_{ee} = g_e\xi_0 \coth(\xi_0) + j\omega C_e, \tag{11}$$

$$y_{ec} = -g_c\xi_0 \operatorname{csch}(\xi_0), \tag{12}$$

$$y_{ce} = -\zeta g_e\xi_0 \operatorname{csch}(\xi_0), \tag{13}$$

$$y_{cc} = \zeta g_c\xi_0 \coth(\xi_0) + j\omega C_c. \tag{14}$$

Here, $g_e = kT/qI_E$ and $g_c = g_e kT/qV_A$ are the dc conductances of the emitter and the collector, V_A is the Early voltage ($V_A = W_B\sqrt{2qN_B(N_B + N_C)(E_g - V)/\varepsilon_s N_C}$, with N_B the base doping and N_C the collector doping); C_e and C_c denote the transition capacitances. The behavior of a practical device with other parasitics can be obtained by converting these parameters to the common emitter parameters, and then adding the parasitics as shown in **Figure 16**. Intrinsic base resistance (r_{bi}) is in series with the base node; extrinsic collector capacitance (C_{cx}) is added from this base node to the intrinsic collector node; and the semiconductor body and the contact resistances are added in the form of extrinsic resistances at the base, emitter, and collector (r_{bx}, r_{ex}, and r_{cx}). In the figure, the values of the elements pertain to a device having a 0.07-μm-thick base and a wide-gap emitter that is 0.05 μm thick.

Figure 17 shows the effects of the extrinsic elements and collector transit time on the current gain (h_{21}) and the unilateral gain (U) of the device. The cutoff frequency f_T is the frequency at which $h_{21} = 1$, and f_{max} is the frequency at which $U = 1$. The cutoff frequency does not change appreciably as a result of the parasitics, but f_{max} is affected significantly by both the device parasitics and the transit-time effects. In particular, most of the resonance features due to the transit-time-induced negative-output conductances disappear in the extrinsic device, and the commonly used technique of extrapolating low-frequency behavior (a 20-dB/decade roll-off) continues to be adequate for both the calculations of current gain and unilateral gain. To show

the importance of collector signal delay, Figure 17 includes the same calculations performed without including the transit time τ_c'. The effect of the transit time is purely to decrease the magnitudes of f_T and f_{max}. Interestingly, one consequence of the parasitics for the small-signal behavior is that the commonly used expression for the figures of merit $f_{max} = \sqrt{f_T/8\pi r_b C_c}$ and $f_T = [2\pi(C_e + C_c)/g_e + \tau_b + \tau_c]^{-1}$ continues to be quite accurate because the parasitics continue to dominate.

Scaling

In assessing the future outlook for the HBT, we consider ways in which it can be scaled, taking into account the unique characteristics of compound semiconductors and heterostructures. Generally, there are three interdependent goals in scaling vertical and lateral device dimensions: higher speed, increased circuit density, and minimal increase in power density.

To achieve higher speed, one must both improve the intrinsic frequency response discussed in the previous section, and reduce the parasitic effects. The extraneous phenomena and the parasitic resistances and capacitances must either remain insignificant or scale down with device dimensions in order not to limit the device speed. If the parasitics are significant, the parasitic time constants must decrease at the same rate as the intrinsic device time constants. One particularly important parasitic phenomenon that must be suppressed is surface recombination, which lowers device gain. In the devices described here, this is accomplished by using p-(Al,Ga)As at the surface. This technique is scalable because the extent of the depletion region required at high p-type doping levels is less than 100 Å, even for a barrier of 0.3 eV at the dissimilar bandgap junction. In double-heterostructure devices, gain degradation may also occur through increased base storage. Preventing this requires increased doping of the collector.

Scaling of the intrinsic speed is closely coupled to the way in which the HBT scales with dimensions. Let λ be the scale factor by which all the dimensions are multiplied, including the base width, in order to produce a scaled-down device. To achieve scaling of the intrinsic speed by λ, the base time constant must also scale by at least λ while maintaining the device gain. The base time constant τ_B varies as $W_B^2/2D_B$, where D_B varies very slowly with N_B at the highest doping, while the lifetime in the base varies inversely with the doping ($\tau_n \propto N_B^{-1}$ to $\tau_n \propto N_B^{-2}$), and the current gain varies as $\beta \propto (D_B\tau_n/W_B^2)$. If the base width is scaled as λ, the desired scaling can be obtained by varying the base doping at levels between $\lambda^{-1.5}$ and λ^{-1}. This scaled increase in HBT base doping does not occur at a rate as high as in homojunction bipolar structures because the heavy doping causes a rapid decrease in lifetime.

For constant voltage scaling and constant power-supply voltages, the current levels remain constant, and current density varies as λ^{-2}. The collector and emitter doping levels also vary as λ^{-2} to permit high current transport without increased charge storage in the base, and high current injection. As remarked earlier, the decrease in L_D is sufficient to compensate for the increase in the alloy-grading field with reduced grading width. The increase in collector doping and reduction in lithographic dimensions change the collector capacitance and the collector signal delay as λ. The emitter signal delay and the emitter capacitance delay also change as λ.

Finally, to reduce the power and obtain other performance enhancements, we consider scaling with bandgap and temperature. This is feasible, since HBTs can readily be designed for operation with adequate speed at low temperatures, and with other compound semiconductors such as (Ga,In)As alloys. At constant speed, the changes in bandgap affect the power levels linearly; that part of the power dissipation which is proportional to the diode voltages decreases linearly with the decrease in base bandgap. Small bandgaps and the desire for lower interconnection resistance may necessitate lower-temperature operation. Decreasing the temperature allows a proportionate decrease in both logic swing and current drive, while maintaining constant speed. The smaller logic swing permits the use of lower supply voltages, resulting in greater than proportional decreases in power dissipation. Unfortunately, heat-removal capabilities decrease with temperature at roughly the same rate, resulting in no net change in the power required to maintain a constant operating temperature.

By grading the composition of the alloy in the base, the bandgap can be gradually decreased from the emitter junction to the collector junction. This grading results in the formation of a quasi-drift field for electrons, which enhances the speed of the HBT. An additional advantage of smaller-bandgap alloys is that they permit the use of a graded-gap base without a significant increase in power dissipation. Smaller bandgaps usually result in an increase in ionization coefficients. The applied biases which cause the ionization, such as the base-collector reverse bias, however, must exceed a threshold potential (typically a factor of 1 to 3 times the bandgap) before the ionization can begin to occur. These set a limit to the allowed bandgap for desired bias voltages, or the converse. A device with an InAs base and collector can sustain at most a couple of volts bias at the base–collector junction, for collector doping levels less than 3×10^{17} cm^{-3}. This is more than the supply voltages that would be needed for logic circuit operation at 77 K. The 0.6 V needed for the supply difference across the source follower, together with a logic swing of, e.g., 0.15 V,

suggests that for the same logic circuit delay, the power requirements of an InAs-based ECL gate at 77 K should be a factor of 10 lower than those of a GaAs-based ECL gate at 300 K, and about 2.5 times lower than those of a GaAs ECL gate at 77 K. This calculation assumes that the current and logic swing vary in proportion to the temperature, and that the supply voltages vary approximately as the sum of the bandgap and the logic swing.

The other major advantage of smaller-bandgap-based devices is related to transit. Since electron mobility improves significantly (the mobility in InAs is about five times that in GaAs, regardless of temperature), the transit can be correspondingly faster with the help of some quasi-drift field in the base. Since the *RC* time constants are relatively invariant with respect to the choice of materials, and an optimal design should balance all of these, the design constraints of the space-charge regions and the quasi-neutral base region should be more relaxed. For example, at a constant horizontal feature size of an HBT, it should be possible to use a significantly thicker base with the base resistance correspondingly reduced while still maintaining excellent base transit. Similarly, it should be possible to reduce the collector doping and capacitance without increasing the Kirk effect or the collector transit time. Finally, the smaller-bandgap materials can also exhibit large secondary-valley energy separations. For example, InAs has a Γ-to-L-valley energy separation which exceeds 1 eV, while GaAs has a Γ-to-L-valley energy separation of 0.3 eV. This should permit significant velocity overshoot in the graded base and in the base-collector space-charge region. These hot-carrier effects might thus make it possible to significantly decrease collector transit time, as time constants are reduced in submicron designs.

Estimates of the speed and power improvements of an HBT which should result from the use of several lower-bandgap materials combinations are shown in **Table 2**. In these approximate calculations, using drift diffusion transport without any drift effects in the base, the 77 K cutoff frequency and power dissipation are calculated for a number of materials combinations, assuming a collector current density of 5×10^4 A-cm^{-2} at $V_{BC} = 0$. The devices are assumed to be 1×2 μm^2 in emitter size with a 1000-Å base width, and designed to sustain the above current density. Majority-carrier transport parameters are employed where minority-carrier parameters are not known, and no bandgap narrowing is assumed. The objective of this simplistic calculation is to obtain an approximate relative comparison of performance at low temperatures. As can be seen, several of the materials combinations appear to be attractive candidates for high-frequency use at 77 K.

Table 2 Calculated unity current gain cutoff frequencies at 77 K for various materials combinations, assuming drift-diffusion transport, a 1000-Å base width, and a 1×2-μm^2 emitter.

Materials combination	*Base bandgap* (eV)	Power dissipation (mW)	f_T (GHz)
(Al,Ga)As/GaAs	1.48	1.42	85
InP,(Al,In)As/(Ga,In)As	0.77	0.75	78
(Ga,Al)Sb/GaSb	0.77	0.75	92
SiGe/Ge	0.66	0.61	67
(Al,Ga)Sb/InAs	0.41	0.38	113

Concluding remarks

This paper has presented a general discussion of technology and modeling questions which pertain to compound semiconductor HBTs, with emphasis on relevant work performed by the authors. The technology of compound semiconductor devices, particularly aspects related to achieving technologically acceptable ohmic contacts, has been less mature and reproducible than that of Si devices. Nevertheless, the developments described here indicate that for devices such as the npn HBT it should be possible to alleviate this by the use of *in situ* n-type contacts formed by using heteroepitaxy, and p-type contacts formed by using rapid thermal diffusion. We have discussed physical limiting effects related to the use of heterostructures and the use of compound semiconductors. We have also shown that accurate modeling should include non-quasi-static effects, have indicated how they may be included in models oriented toward circuit simulation in time and frequency domains, and have elucidated a number of scaling issues, including the implications of using smaller-bandgap materials.

Acknowledgments

The authors wish to acknowledge in particular the work of A. L. Ginzberg, R. F. Marks, and J. F. DeGelormo, as well as the support of numerous other colleagues. Some of the work reported here was performed while one of us (Sandip Tiwari) was on sabbatical leave at the Department of Electrical Engineering and Computer Sciences of the University of Michigan. The transmission electron micrographs were obtained by Yih-Cheng Shih (present address: AT&T Bell Laboratories, Reading, PA).

References

1. H. Kroemer, "Theory of a Wide-Gap Emitter for Transistors," *Proc. IEEE* **45**, 1535 (1957).
2. W. P. Dumke, J. M. Woodall, and V. L. Rideout, "GaAs–GaAlAs Heterojunction Transistor for High Frequency Operation," *Solid-State Electron.* **15**, 1339 (1972).
3. T. Ishibashi, O. Nakajima, K. Nagata, Y. Yamauchi, H. Ito, and T. Nittontomi, "Ultra-High Speed AlGaAs/GaAs

Heterojunction Bipolar Transistors for Self-Aligned HBTs," *IEDM Tech. Digest* **IEDM-88,** 826 (1988).
4. D. A. Whitmire, V. N. Garcia, and S. A. Evans, "A 32b GaAs RISC Microprocessor in GaAs HI2L," *ISSCC Tech. Digest* **ISSCC-88,** 34 (1988).
5. J. D. George, J. D. Harr, R. Young, C. J. Anderson, Y. H. Kwark, H. F. Basit, S. Fang, K.-C. Wang, P. M. Asbeck, M. F. Chang, R. Nubling, G. J. Sullivan, M. McDonald, C. Honaker, and T. McDermott, "A High Speed Gate Array Implemented with AlGaAs/GaAs Heterojunction Bipolar Transistors," *ISSCC Tech. Digest* **ISSCC-89,** 186 (1989).
6. S. Tiwari, T. S. Kuan, and E. Tierney, "Ohmic Contacts to n-GaAs with Germanide Overlayers," *IEDM Tech. Digest* **IEDM-83,** 115 (1983).
7. M. Murakami, W. H. Price, Y.-C. Shih, K. D. Childs, B. K. Furman, and S. Tiwari, "Thermally Stable Ohmic Contact to n-Type GaAs: I. MoGeW Contact Metal," *J. Appl. Phys.* **62,** 3288 (1987).
8. M. Murakami and W. H. Price, "Thermally Stable, Low-Resistance NiInW Ohmic Contacts to n-Type GaAs," *Appl. Phys. Lett.* **51,** 664 (1987).
9. J. M. Woodall, J. L. Freeouf, G. D. Pettit, T. Jackson, and P. Kirchner, "Ohmic Contacts to n-GaAs Using Graded Band Gap Layers of GaInAs Grown by Molecular Beam Epitaxy," *J. Vac. Sci. Technol.* **19,** 626 (1981).
10. S. L. Wright, R. F. Marks, S. Tiwari, T. N. Jackson, and H. Barrate, "In-Situ Contacts to GaAs Based on InAs," *Appl. Phys. Lett.* **49,** 1545 (1986).
11. S. L. Wright, E. Marshall, R. F. Marks, T. N. Jackson, S. Tiwari, and H. Baratte, "In-Situ Contacts to GaAs Based on InAs," *J. Vac. Sci. Technol. B* **5,** 777 (1987).
12. J. M. Woodall, G. D. Pettit, T. N. Jackson, C. Lanza, K. L. Kavanagh, and J. W. Mayer, "Fermi-Level Pinning by Misfit Dislocations at GaAs Interfaces," *Phys. Rev. Lett.* **51,** 1783 (1983).
13. S. L. Wright, R. F. Marks, E. D. Marshall, Y. C. Shih, and A. B. Young, "Growth and Characterization of In-Situ (In,Ga)As Ohmic Contacts to n-GaAs," *J. Cryst. Growth* **95,** 245 (1988).
14. S. Tiwari, J. C. DeLuca, and V. Deline, "Rapid Thermal Annealing of Zn^{64} and Mg^{24} Implants in GaAs," *Inst. Phys. Conf. Ser.* **74,** 83 (1984).
15. S. Tiwari, J. Hintzman, and A. Callegari, "Rapid Thermal Diffusion of Zinc and Use in p-Type Ohmic Contacts to GaAs and GaAlAs," *Appl. Phys. Lett.* **51,** 2118 (1987).
16. J. C. Marinace, "Diffusion of Zinc Through Films of Refractory Metals on GaAs," *J. Electrochem. Soc.* **117,** 145 (1970).
17. R. A. Kiehl, S. Tiwari, S. L. Wright, and M. A. Olson, "p-Channel Quantum-Well Heterostructure MI^3SFET," *IEEE Electron Device Lett.* **9,** 309 (1988).
18. S. Tiwari, A. Ginzberg, S. Akhtar, S. L. Wright, R. Marks, Y. H. Kwark, and R. Kiehl, "Heterostructure Devices Using Self-Aligned Diffused p-Type Ohmic Contacts," *IEEE Electron Device Lett.* **9,** 422 (1988).
19. S. Tiwari, "GaAlAs/GaAs Heterostructure Bipolar Transistors: Experiment and Theory," *IEDM Tech. Digest* **IEDM-86,** 262 (1986).
20. S. Tiwari, D. Frank, and S. L. Wright, "Surface Recombination in GaAlAs/GaAs Heterostructure Bipolar Transistors," *J. Appl. Phys.* **64,** 5009 (1988).
21. S. Tiwari and S. L. Wright, "Material Properties of p-Type GaAs at Large Dopings," *Appl. Phys. Lett.* **56,** 563 (1990).
22. (a) M. I. Nathan, W. P. Dumke, K. Wrenner, S. Tiwari, S. L. Wright, and K. A. Jenkins, "Electron Mobility in p-Type GaAs," *Appl. Phys. Lett.* **52,** 654 (1988); (b) M. I. Nathan, S. Tiwari, P. M. Mooney, and S. L. Wright, "DX Centers in AlGaAs p–n Heterojunctions and Heterojunction Bipolar Transistors," *J. Appl. Phys.* **62,** 3234 (1987).
23. R. K. Ahrenkiel, D. J. Dunlavy, D. Greenberg, J. Schlupmann, H. C. Hamaker, and H. F. MacMillan, "Electron Mobility in p-GaAs by Time of Flight," *Appl. Phys. Lett.* **51,** 776 (1987).
24. R. J. Nelson, "Measurement of 100 Micron Minority Carrier Diffusion Lengths in p-GaAs by a New Photoluminescence Method," *Inst. Phys. Conf. Ser.* **45,** 256 (1979).
25. E. O. Kane, "Band Tails in Semiconductors," *Solid-State Electron.* **28,** 3 (1985).
26. L. Jastrzebski, J. Lagowski, H. C. Gatos, and W. Walukiewicz, "Minority Carrier Lifetime in GaAs at Elevated Temperatures: Implications for Solar Cell Performance," *Inst. Phys. Conf. Ser.* **45,** 437 (1979).
27. J. L. Lievin, C. Dubon-Chevallier, F. Alexandre, G. Leroux, J. Dangla, and D. Ankri, "GaAlAs/GaBeAs Heterojunction Bipolar Transistor Grown by Molecular Beam Epitaxy," *IEEE Electron Device Lett.* **EDL-7,** 129 (1986).
28. M. S. Tyagi and R. Van Overstraeten, "Minority Carrier Recombination in Heavily-Doped Silicon," *Solid-State Electron.* **26,** 577 (1983).
29. C. H. Henry, R. A. Logan, F. R. Merritt, and C. G. Bethea, "Radiative and Nonradiative Lifetimes in n-Type and p-Type 1.6 Micron InGaAs," *Electron. Lett.* **20,** 358 (1984).
30. H. C. Casey and F. Stern, "Concentration-Dependent Absorption and Spontaneous Emission of Heavily Doped GaAs," *J. Appl. Phys.* **47,** 631 (1976).
31. J. P. Bailbe, A. Marty, and G. Rey, "Influence of Degeneracy on Behaviour of Homojunction GaAs Bipolar Transistors," *Electron. Lett.* **20,** 258 (1984).
32. M. E. Klausmeir-Brown, M. S. Lundstrom, and M. R. Melloch, "Effects of Heavy Impurity Doping on AlGaAs/GaAs Bipolar Transistors," *IEEE Trans. Electron Devices* **36,** 2146 (1989).
33. J. W. Slotboom and H. C. de Graaf, "Measurements of Bandgap Narrowing in Si Bipolar Transistors," *Solid-State Electron.* **19,** 857 (1976).
34. S. Tiwari, "A New Effect in Heterostructure Bipolar Transistors," *IEEE Electron Device Lett.* **9,** 142 (1988).
35. S. Tiwari and D. J. Frank, "Analysis of the Operation of GaAlAs/GaAs HBTs," *IEEE Trans. Electron Devices* **36,** 2105 (1989).
36. R. W. Keyes, "Trends and Limits in Bipolar Transistor for Logic," presented at the IEEE Bipolar and Circuits Technology Meeting, Minneapolis, MN, 1988.
37. R. L. Pritchard, *Electrical Characteristics of Transistors*, McGraw-Hill Book Co., Inc., New York, 1967.
38. *Advanced Statistical Analysis Program (ASTAP), Program Reference Manual*, Order No. SH20-1118-0; available through IBM branch offices.

Received July 18, 1989; accepted for publication March 13, 1990

Sandip Tiwari *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.*
Dr. Tiwari received his education in electrical engineering from the Indian Institute of Technology at Kanpur (B. Tech, 1976), the Rensselaer Polytechnic Institute (M.S., 1977), and Cornell University (Ph.D., 1980). His work, interests, and contributions have been primarily related to semiconductor device physics and technology.

Steven L. Wright *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.*
Dr. Wright received his B.S. and M.S. degrees from the University of Colorado in 1975 and 1978, respectively, and his Ph.D. in electrical engineering from the University of California at Santa Barbara in 1982. His Master's thesis pertained to the liquid phase epitaxial growth of (In,Ga)(As,P). His doctorate focused on the MBE growth of polar on nonpolar heterojunctions with device-quality interfaces. The latter included the successful growth of GaP on Si, and the fabrication of a heterojunction bipolar transistor, employing GaP as a wide-gap emitter. Since 1982 Dr. Wright has been a Research Staff Member at the Thomas J. Watson Research Center, where his research activities have included work on MBE materials growth and III–V heterostructure devices. His primary research interests concern the impact of heterostructure material properties on high-speed device performance, and *in situ* measurement techniques during MBE growth.

David J. Frank *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.*
Dr. Frank received his B.S. degree from the California Institute of Technology, Pasadena, in 1977, and his Ph.D. degree in physics from Harvard University in 1983. Since then he has worked at the Thomas J. Watson Research Center, initially as a Postdoctoral Fellow in the Physical Sciences Department, studying nonequilibrium superconductivity, and currently as a Research Staff Member in the Logic, Memory and Packaging Department, modeling III–V devices. His interests include superconductor and semiconductor device physics, modeling and measurement, circuit design, and percolation in two-dimensional systems. Dr. Frank is a member of the American Physical Society.