Heterostructure Bipolar Transistors and Integrated Circuits

HERBERT KROEMER, FELLOW, IEEE

Invited Paper

Abstract—Two new epitaxial technologies have emerged in recent years (molecular beam epitaxy (MBE) and metal-organic chemical vapor deposition (MOCVD)), which offer the promise of making highly advanced heterostructures routinely available. While many kinds of devices will benefit, the principal and first beneficiary will be bipolar transistors. The underlying central principle is the use of energy gap variations beside electric fields to control the forces acting on electrons and holes, separately and independently of each other. The resulting greater design freedom permits a re-optimization of doping levels and geometries, leading to higher speed devices. Microwave transistors with maximum oscillation frequencies above 100 GHz and digital switching transistors with switching times below 10 ps should become available. An inverted transistor structure with a smaller collector on top and a larger emitter on the bottom becomes possible, with speed advantages over the common "emitter-up" design. Double-heterostructure (DH) transistors with both wide-gap emitters and collectors offer additional advantages. They exhibit better performance under saturated operation. Their emitters and collectors may be interchanged by simply changing biasing conditions, greatly simplifying the architecture of bipolar IC's. Examples of heterostructure implementations of I^2L and ECL are discussed. The present overwhelming dominance of the compound semiconductor device field by FET's is likely to come to an end, with bipolar devices assuming an at least equal role, and very likely a leading one.

"*What is claimed is:*

1) . . .
2) A device as set forth in claim 1 in which one of the separated zones is of a semiconductive material having a wider energy gap than that of the material in the other zones."

Claim 2 of U.S. Patent 2 569 347 to W. Shockley,
Filed 26 June 1948,
Issued 25 September 1951,
Expired 24 September 1968.

I. Introduction

THIS IS A PAPER about an idea whose time has come: A bipolar transistor with a wide-gap emitter. As the introductory quote shows, the idea is as old as the transistor itself. The great potential advantages of such a design over the conventional homostructure design have long been recognized [1]–[3], but until the early 70's, no technology existed to build practically useful transistors of this kind, even though numerous attempts had been made [3], [4]. The situation started to change with the emergence of liquid-phase epitaxy (LPE) as a technology for III/V-compound semiconductor heterostructures, and in recent years reports on increasingly impressive true three-terminal heterostructure bipolar transistors (HBT's) have appeared at an increasing rate [5]–[14]. In addition, there is also a rapidly growing literature on *two*-terminal *photo*transistors with wide-gap emitters [15]. Many of the phototransistors employ InP emitters with a lattice-matched (Ga, In) (P,As) base.

Since the mid-70's, two additional very promising heterostructure technologies have appeared: molecular beam epitaxy (MBE) [16] and metal-organic chemical vapor deposition (MOCVD) [17]. Impressive results on MOCVD-grown (Al,Ga)As-GaAs phototransistors have already been published [18]; HBT's grown by MBE have also been achieved [19].

Because of the pre-eminence of silicon in current IC technology, there exists a strong incentive to incorporate wide-gap emitters into Si transistors, in a way compatible with existing Si technology. A possible approach—and the most successful one so far—has been the use of heavily doped "semi-insulating polycrystalline" silicon (SIPOS) as emitter [20], utilizing the wider energy gap of "polycrystalline" (really: amorphous) Si compared to crystalline Si. An alternate approach has been the use of gallium phosphide, which has a room-temperature lattice constant within 0.3 percent of that of Si, grown on Si either by CVD [21] or by MBE [22]. But the results reported for the GaP-Si combination have so far been disappointing.

Finally, the first reports have recently appeared, in which HBT's have been integrated on the same chip with other devices, such as double-heterostructure (DH) lasers [23] or LED's [24].

In view of these recent developments it appears that Shockley's vision is about to become a reality. In fact, one of the purposes of this paper is to show that the possibilities for HBT's go far beyond simply replacing a homojunction emitter by a heterojunction emitter.

To appreciate these possibilities, it is useful first to view the wide-gap emitter as a simple example of a more general *central design principle* of heterostructure devices; it is discussed in Section II of this paper. Discussions of future device possibilities must be based on technological premises; they are discussed in Section III. In Section IV and V the concept and the high-speed benefits of the wide-gap emitter are reviewed, including some recent conceptual developments that do not appear to have been widely appreciated. Section VI discusses the promising concept of an inverted transistor design, in which the collector is made smaller than the emitter and placed on the surface of the structure, similar to I^2L, but using a heterostructure design applicable to all transistors. In Section VII the idea of a single-heterostructure transistor with a wide-gap emitter is generalized to DH transistors with both wide-gap emitters and wide-gap collectors. Such a design appears to

Manuscript received June 30, 1981; revised August 31, 1981. This work was supported in part by the Army Research Office and by the Office of Naval Research.

The author is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

0018-9219/82/0100-0013$00.75

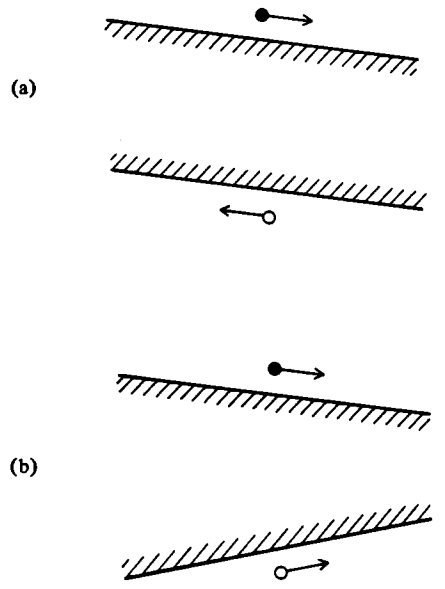

Fig. 1. Forces on electrons and holes. In a uniform-gap semiconductor (top) the two forces are equal and opposite to each other, and equal to the electrostatic force $\pm q\vec{E}$. In a graded-gap structure, the forces in electrons and holes may be in the same direction.

offer surprisingly large advantages for both microwave and digital devices, and especially for digital IC's. As examples of potential IC advantages, heterostructure modifications of both I^2L and ECL architecture are discussed. Finally, Section VIII offers some speculations on the question of FET's-versus-bipolars, and related questions.

In line with the character of this Special Issue (integrated) digital HBT's are emphasized over (discrete) microwave devices, but not to the point of exclusion of the latter. It would be artificial to attempt a complete separation: Not only was much of the past development of HBT's oriented towards discrete microwave devices, but several of the newer concepts originating in a digital context would improve microwave transistors as well.

II. The Central Design Principle of Heterostructure Devices

If one looks for a general principle underlying most heterostructure devices, one is led to the following considerations. If one ignores magnetic effects, the forces acting on the electrons and holes in a semiconductor are equal (except for a sign in the case of electrons) to the slopes of the edge of the band in which the carriers reside (Fig. 1). In ideal homostructures the energy gap is constant; hence the slopes of the two band edges are equal, and the forces acting on electrons and holes are necessarily equal in magnitude and opposite in sign. In fact they are equal to the ordinary electrostatic force $\pm q\vec{E}$ on a charge of magnitude $\pm q$ in an electric field $\vec{E}$. In a heterostructure, the energy gap may vary; hence the two band edge slopes and with it the magnitudes of the two forces need not be the same, nor need they be in any simple way related to the electrostatic force exerted by a field $\vec{E}$. In fact, the two slopes may have opposite signs (Fig. 1), implying forces on electrons and holes that act in the same direction, despite their opposite charges.

In effect, heterostructures utilize energy gap variations in addition to electric fields as forces acting on electrons and holes, to control their distribution and flow. This is what I would like to call the *Central Design Principle* of heterostructure devices. It is a very powerful principle, and one of the purposes of this paper is to give examples that show just how powerful it is.

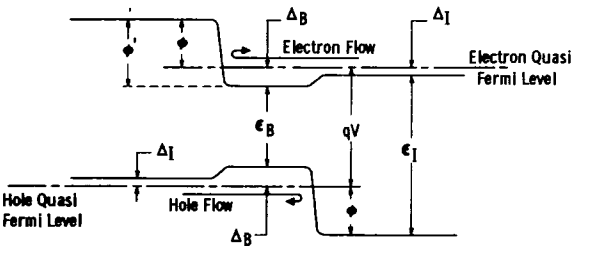

Fig. 2. Energy band diagram of a DH laser, showing the confinement forces driving both electrons and holes towards the active layer, on both sides of the latter. (From [25].)

Although by no means restricted to bipolar devices, the principle is especially powerful when, as in a bipolar transistor, the distribution and flow of *both* electrons and holes must be controlled. By a judicious combination of energy gap variations and electric fields it then becomes possible, within wide limits, to control the forces acting on electrons and holes, *separately and independently of each other*, a design freedom not achievable in homostructures.

The central design principle plays a role in almost all heterostructure devices, and it serves both to unify the ideas underlying different such devices, and as guidance in the development of new device concepts. No device demonstrates the central design principle better than the oldest and so far most important heterostructure device, the DH laser. This point is illustrated in Fig. 2, which shows the energy band structure of the device under lasing conditions, as anticipated (with only slight exaggeration) in the paper in which this device was first proposed [25], and from which Fig. 2 is taken. The drawing shows band edge slopes corresponding to forces that drive *both* electrons and holes towards the inside of the active layer, at *both* edges of the latter. This is the principal reason why the DH laser works, although it is not the only reason. The difference in refractive indices between the inner and outer semiconductors also plays an important role. Such a participation of additional concepts is not uncommon in other heterostructure devices either.

III. The Technological Premise

Throughout its history, heterostructure device design has chronically suffered from a technology bottleneck. Even LPE, whatever its merits as a superb laboratory technology, has outside the laboratory been largely limited to devices, such as injection lasers for fiberoptics use, which could simply not be built without heterostructures, but which were needed sufficiently urgently to put up with the limitations of LPE technology. Already for the "ordinary" three-terminal transistor (i.e., excepting phototransistors), the necessary high-performance combination of LPE and lithography was never developed to the point that the resulting heterostructures would reach the speed capability of state-of-the-art Si bipolars, much less reach their own theoretical potential exceeding that of Si.

As a result of the emergence of two new epitaxial technologies in the last few years, the heterostructure technology bottleneck is rapidly disappearing, to the point that the

incorporation of heterostructures into most compound semiconductor devices will probably be one of the dominant themes of compound semiconductor technology during the remainder of the present decade.

The two new technologies are MBE [16] and MOCVD [17]. Although differing in many ways, for the purposes of this paper the commonalities of the two technologies are more important than their differences, and there is no need to enter here into the debate as to which of the two technologies will eventually be best for doing what.

Both technologies are capable of growing epitaxial layers with high crystalline perfection and purity, comparable to state-of-the-art results with LPE and halide-CVD. Highly controlled doping levels up to 10^{19} impurities per cm^3 and more can be achieved, and highly controlled changes in doping level are possible during growth without interrupting the latter, and with at most a minor adjustment in growth parameters. The doping may be changed either gradually or abruptly. Because of the comparatively low growth temperatures (especially for MBE), diffusion effects during growth are weak, and with certain dopants much more abrupt doping steps can be achieved than with any other technique, not only when doping is "turned on," but also when it is "turned off."

Most important in our context of heterostructures, it is possible in both technologies to change from one III/V semiconductor to a different (lattice-matched) III/V semiconductor with greater ease than in any other technique. In both techniques, a change in semiconductor and hence in energy gap is not significantly harder to achieve than a change in doping level! In particular, the change can again be accomplished during growth without interruption, either gradually or abruptly and, if abruptly, over extremely short distances.

Finally, in both techniques the growth rates and hence the layer thicknesses can be very precisely controlled. Because the growth rates themselves are low (or can be made low), extremely thin layers can be achieved, to the point that effects due to the finite quantum-mechanical wavelengths of the electrons can be readily generated. It is in the context of the study of such quantum effects that both techniques have demonstrated their so far highest capability level. With both MOCVD and MBE, GaAs-(Al,Ga)As structures with over 100 epitaxial layers have been built [26], [27], and essentially arbitrary numbers appear possible. With MOCVD, layer thicknesses below 50 Å have been achieved, with MBE, below 10 Å. In either case, the capability far exceeds anything needed in the foreseeable future for transistor-like devices.

So far, these are laboratory results, mostly on GaAs-(Al,Ga)As structures. But it is the consensus of those working on the two technologies that much of this performance can be carried over into a production environment, with high yields and at an acceptable cost. Acceptable here means a cost low enough that it will not deter the use of the new technologies in most of those high-performance applications that need the performance potential of heterostructure devices.

An extension of both technologies to lattice-matched III/V-compound heterosystems beyond GaAs-(Al,Ga)As is an all but foregone conclusion, including GaAs-(Ga,In)P, InP-(Ga,In)(P,As), and InAs-(Al,Ga)Sb.

In view of these developments, the following scenario for the III/V-compound heterostructure technology of the 1990's is likely. Epitaxial technologies will be routinely available in which both the doping and the energy gap can be varied almost at will, over distances significantly below 100 Å, and covering a large fraction of their physically possible ranges, by what is essentially a software-controlled operation within a given growth run. The cost of the technology will be sufficiently low to encourage the development of high-performance devices that utilize this capability. The cost will be essentially a fixed cost per growth run, depending on the overall tolerance level but hardly at all on the number of layers and what they contain, similar to the cost of optical lithography, which has largely a fixed cost per masking step, almost independent of what is on the mask (at a given tolerance level). In particular, there will be only a negligible cost increment associated with using a heterojunction over using a homojunction (or no junction at all), and hence there will be only a negligible economic incentive *not* to use a heterojunction.

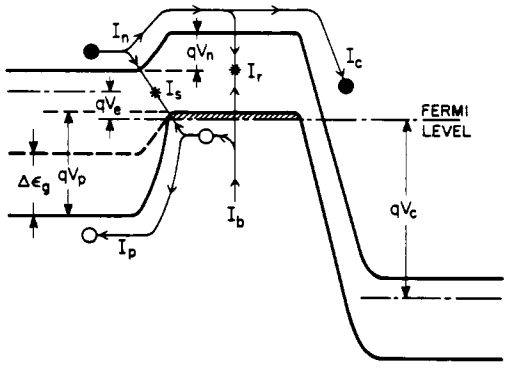

Fig. 3. Band diagram of an n-p-n transistor with a wide-gap emitter, showing the various current components, and the hole-repelling effect of the additional energy gap in the emitter.

What *will* be expensive, just as with masking, are multiple growth runs, in which the growth is interrupted and the wafer removed from the growth system for intermediate processing, with the growth to be resumed afterwards. Hence there will be a strong incentive to accomplish the desired device structure with the minimum number of growth runs, no matter how complicated the individual run might become.

The above scenario is the technological premise of the remainder of this paper. Although presented here in the context of bipolar transistors and IC's, this scenario, as well as the central design principle of Section II, obviously go far beyond these specific devices. Together, the two concepts might form the starting point for a fascinating speculation about the future of semiconductor devices beyond simple bipolar structures. However, such a discussion would go beyond the scope of this paper as well as of this Special Issue.

IV. The Wide-Gap Emitter

A. Basic Theory

The basic theory behind a wide-gap emitter is simple [1]. Consider the energy band structure of an n-p-n transistor, as in Fig. 3. In drawing the band edges as smooth monotonic curves we are implicitly assuming that the emitter junction has been graded sufficiently to obliterate any band edge discontinuities or even any nonmonotonic variations of the conduction band edge. We will return to this point later. There are the following injection-related dc currents flowing in such a transistor:

a) A current I_n of electrons injected from the emitter into the base;
b) A current I_p of holes injected from the base into the emitter;
c) A current I_s due to electron–hole recombination within

the forward biased emitter-base space charge layer.

d) A small part of I_r of the electron injection current I_n is lost due to bulk recombination.

The current contribution I_n is the principal current on which the device operation depends; the contributions I_p, I_s, and I_r are strictly nuisance currents, as are the capacitive currents (not shown in Fig. 3) that accompany any voltage changes. We have neglected any currents created by electron-hole pair generation in the collector depletion layer or the collector body.

Expressed in terms of these physical current contributions, the net currents at the three terminals are:

$$\text{Emitter current:} \quad I_e = I_n + I_p + I_s \tag{1a}$$

$$\text{Collector current:} \quad I_c = I_n - I_r \tag{1b}$$

$$\text{Base current:} \quad I_b = I_p + I_r + I_s. \tag{1c}$$

A figure of merit for such a transistor is the ratio

$$\beta = \frac{I_c}{I_b} = \frac{I_n - I_r}{I_p + I_r + I_s} < \frac{I_n}{I_p} \equiv \beta_{\max}. \tag{2}$$

Here, $\beta_{\max}$ is the highest possible value of β, in the limit of negligible recombination currents. It is the improvement of $\beta_{\max}$ to which the wide-gap emitter idea addresses itself.

To estimate $\beta_{\max}$ we assume that emitter and base are uniformly doped with the doping levels N_e and P_b. We denote with qV_n and qV_p the (not necessarily equal) heights of the potential energy barriers for electrons and holes, between emitter and base. We may then write the electron and hole injection current densities in the form

$$J_n = N_e v_{nb} \exp(-qV_n/kT) \tag{3a}$$

$$J_p = P_b v_{pe} \exp(-qV_p/kT). \tag{3b}$$

Here v_{nb} and v_{pe} are the mean speeds, due to the combined effects of drift and diffusion, of electrons at the emitter-end of the base, and of holes at the base-end of the emitter.

In writing (3a, b) with simple Boltzmann factors, we have implicitly assumed that both emitter and base are nondegenerate. In a homojunction transistor the emitter might be degenerate; in a heterojunction transistor the base might be degenerate, as is in fact assumed in Fig. 3. This requires small corrections either in (3a) for the homojunction case, or (3b) for the heterojunction case, which we neglect here for simplicity. We have also neglected correction factors allowing for the differences in the effective densities of states of the semiconductors.

We are interested here only in the ratio of the two currents. If the energy gap of the emitter is larger than that of the base by $\Delta\epsilon_g$, we have

$$q(V_p - V_n) = \Delta\epsilon_g \tag{4}$$

and we obtain

$$\frac{I_n}{I_p} = \beta_{\max} = \frac{N_e}{P_b}\frac{v_{nb}}{v_{pe}} \exp(\Delta\epsilon_g/kT). \tag{5}$$

For a good transistor, a value $\beta_{\max} \gtrsim 100$ is desirable.

Of the three factors in (5), the ratio v_{nb}/v_{pe} is least subject to manipulation. As a rule

$$5 < v_{nb}/v_{pe} < 50. \tag{6}$$

To obtain $\beta_{\max} \gtrsim 100$ it is therefore necessary that either

$$N_e \gg P_b \tag{7}$$

or that $\Delta\epsilon_g$ is at least a few-times kT.

Energy gap differences that are many-times kT are readily obtainable. As a result, very high values of I_n/I_p can be achieved *almost regardless of the doping ratio*. This does not mean that arbitrarily high β's can be obtained. It simply means that the hole injection current I_p becomes a negligible part of the base current compared to the two recombination currents: $I_b \cong I_s + I_r$. To have a useful transistor, we must still have $I_r \ll I_n$. If we approximate I_e by I_n, we obtain

$$\beta = \frac{I_n}{I_r + I_s}. \tag{8}$$

Based on the evidence from high-β HBT's that have been reported ($\beta \geqslant 10^3$),[1] the emitter-base hetero-interface can be made sufficiently defect-free to keep the interface recombination current I_s below $10^{-3} I_n$, at least at sufficiently high current levels I_n. At the same time, the base doping in a properly designed heterostructure transistor will be very high, and hence the minority carrier lifetime correspondingly low, to the point that the bulk recombination current I_r, rather than the interface recombination current I_s will dominate, in contrast to the situation in many homojunction transistors. We therefore neglect I_s beside I_r.

The bulk recombination current density may be written

$$J_r = \gamma n_e(0) w_b/\tau. \tag{9}$$

Here $n_e(0)$ is the injected electron concentration at the emitter end of the base, w_b is the base width, and τ the average electron lifetime in the base. The factor γ is a factor between 0.5 and 1.0, indicating by how much the average electron concentration differs from the electron concentration at the emitter end. If we insert (3a) and (9) into (8), and neglect I_s, we obtain

$$\beta \cong \frac{1}{\gamma}\frac{v_{nb}\tau}{w_b}. \tag{10}$$

This depends on the base doping only through the effect of the base doping on the lifetime. For heavy base doping levels the lifetimes may be short indeed.[2] Nevertheless, even for very short lifetimes, high β's should be achievable in transistors with a sufficiently thin base region, which is the case of dominant interest in any event. As an example, assume $w_b \cong 1000$ Å $= 10^{-5}$ cm. In such a transistor the electron velocity is likely to approach values close to bulk limited drift velocities $v_{nb} \cong 10^7$ cm $\cdot$ s^{-1}. Even for a lifetime as short as 10^{-9} s, this would lead to $\beta \cong 10^3$, a value that should satisfy even the most stringent demands. Evidently, no serious problems from reduced minority carrier lifetimes arise unless the latter drop to the vicinity of 10^{-10} s or lower, at least not for plausible base widths not exceeding 1000 Å.

Much of the remainder of this paper will deal with the tradeoffs made possible when high β-values can be obtained without a high emitter-to-base doping ratio. Before turning to these tradeoffs, it is instructive to return to (5) and to apply it to

[1] See, e.g., [7], [8], [9], [14], [18]. Even higher values have been found in some phototransistors. See [15] for further references.

[2] For GaAs, injection laser experience suggests lifetimes between 10^{-10} and 10^{-9} s for degenerate doping levels, slightly longer for nondegenerate doping.

energy gap variations in the conventional silicon transistor. The energy gap of Si, like that of the other semiconductors, is not strictly constant, but decreases slightly at the high doping levels that are desirable in the emitters of a homojunction transistor. As a result, a Si transistor is not strictly a uniform-gap transistor; it is itself a heterojunction transistor, but with a small yet highly undesirable *negative* value of $\Delta\epsilon_g$. The best available data, taken on actual transistor structures [28], indicate a gap shrinkage beginning at a doping level $N_d \sim 10^{17}$ cm^{-3}, and reducing the gap approximately logarithmically with doping level, reaching a gap shrinkage between 75 and 80 meV ($>3kT$) at $N_d \sim 10^{19}$ cm^{-3}. According to (5), an emitter gap shrinkage of $3kT$ reduces the ratio I_n/I_p by a factor $e^{-3} \sim 1/20$. The overall effect at this doping level is the same as if the emitter were doped only to 5×10^{17} cm^{-3}, without gap shrinkage. To obtain β-values larger than the ratio v_{nb}/v_{pe} (<50), the base region must be even less heavily doped than this value, which is far below what is metallurgically possible, and far below what would be desirable in the interest of almost all other performance characteristics, especially base resistance. Increasing the emitter doping beyond 10^{19} cm^{-3} improves β only very slowly, roughly proportionally to $N_e^{0.33}$. By pushing everything to the limit, state-of-the-art microwave transistors with P_b-values (averaged over the base region) of about 1×10^{18} cm^{-3} have been achieved [29]. But this is still far below what would be desirable.

Evidently, the conventional Si bipolar transistor behaves far less well than the naive uniform-gap textbook model would predict. In fact, the energy gap shrinkage and its consequences represent one of the dominant performance limitations of the device.

B. Graded Versus Abrupt Emitter Junctions

In Fig. 3, and in the discussion accompanying it, we had assumed that the emitter/base junction is compositionally graded, so as to yield smoothly and monotonically varying band edges. Such graded transistors are easily achieved, but unless the appropriate measures are taken to do so, the modern epitaxial technologies tend to produce abrupt transistors in which band edge discontinuities are present. As a rule, the conduction band on the wider gap side lies energetically above that on the narrower gap side. Applied to the wide-gap emitter in a transistor, this leads to the "spike-and-notch" energy band diagram shown in Fig. 4(a). Because the emitter-to-base doping ratio in an HBT tends to be low, most of the electrostatic potential drop will occur on the less heavily doped emitter side, and the potential spike will project above the conduction band in the neutral portion of the base, leading to a potential barrier of net height $\Delta\epsilon_B$. Such a barrier has both advantages and disadvantages, and a brief discussion is in order.

Consider first the potential notch accompanying the barrier on the base side. Such a notch will collect injected electrons, and therefore enhance recombination losses, a highly undesirable effect. Because of the low emitter-to-base doping ratio expected in an HBT, the notch will be quite shallow, with a depth given approximately by

$$\Delta\epsilon_N = (P_e/N_b)\, q V_n \tag{11}$$

which will typically be of the order 5 meV $\ll kT$. Nevertheless, because of the danger of interface recombination defects, it would be desirable to eliminate the notch altogether, and perhaps even replace it by a slightly repulsive potential, as

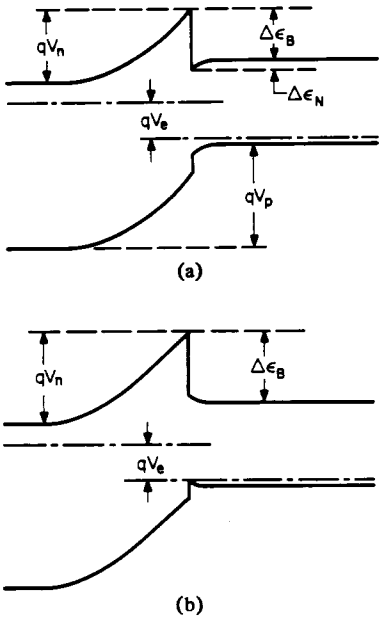

Fig. 4. Band structure of an abrupt wide-gap emitter, showing the spike barrier and the accompanying electron trapping notch (a) in the conduction band structure. The notch can be removed (b) by the incorporation of a planar acceptor doping sheet into the heterojunction.

shown in Fig. 4(b). This is easily accomplished by incorporating a very thin sheet with a very high acceptor concentration right at the interface. Typical required sheet doping concentrations will be of the order 10^{11} acceptors per *square* centimeter. The feasibility of such "planar doping" sheets has been demonstrated [30], at least with MBE, and there is little doubt that it can be accomplished by MOCVD as well.

As to the barrier itself, one minor drawback of its existence is the accompanying increase of the order $\Delta\epsilon_B/q$, in required emitter voltage to yield a given current density. More severe is the (related) drawback that the potential barrier $\Delta\epsilon_B$ drastically reduces the ratio J_n/J_p, from the value in (5), by a factor exp $(-\Delta\epsilon_B/kT)$. Instead of (4a), we now have

$$q(V_p - V_n) = \Delta\epsilon_g - \Delta\epsilon_B \cong \Delta\epsilon_V. \tag{4b}$$

The last equality results if the notch depth is small compared to kT, in which case $\Delta\epsilon_B = \Delta\epsilon_C$. Here $\Delta\epsilon_C$ and $\Delta\epsilon_V$ represent the conduction and valence band discontinuities. Instead of (5a), we obtain

$$\frac{I_n}{I_p} = \beta_{\max} = \frac{N_e}{P_b}\,\frac{v_{nb}}{v_{pe}} \exp\,(\Delta\epsilon_V/kT). \tag{5b}$$

If the valence band discontinuity is sufficiently large, a major improvement remains. Unfortunately, in the system of largest current interest, the (Al,Ga)As/GaAs system, the valence band discontinuity is quite small, $\Delta\epsilon_V = 0.15\ \Delta\epsilon_g$ [31], and the reduction of the spike by grading is probably essential. A detailed discussion of the detrimental effects of the spike is found in a paper by Marty *et al.* [10].

The above drawbacks of the extra potential barrier accompanying an abrupt emitter/base junction are partially compensated by the fact that such a barrier would inject the electrons into the base region with a substantial kinetic energy, and hence with a very high velocity ($\sim 10^8$ cm/s). Because of the directional dependence of the polar optical phonon scattering that is the dominant scattering process in III/V-compounds,

several collisions are required before the electrons have lost their high forward velocity. The result should be a highly efficient and very fast near-ballistic electron transport through the base. Such ballistic transport effects have been of great interest recently, and although their discussion has been largely in an FET context [32], [33], much of this discussion applies as well (or even more) to bipolar transistors with an emitter junction barrier that represents, in effect, a ballistic launching ramp.

Exactly what the balance between drawbacks and benefits will be for the abrupt emitter/base junction versus the graded one, remains to be seen. But it appears likely that ballistic effects will find their way into future transistors specifically designed around them.

An extreme case of high-energy electron injection into the base was discussed some time ago by Kroemer [34], in the form of a so-called *Auger transistor*. If the conduction band discontinuity $\Delta\epsilon_C$ becomes larger than the energy gap in the base, the electron injection may lead to Auger multiplication of electrons, and hence to a transistor with true current amplification in a grounded-base configuration $\alpha > 1$. Such a transistor might be of interest for power switching applications at very high microwave speeds. It remains to be seen what will come of this idea.

V. Speed Tradeoffs

A. The Emitter Capacitance Tradeoff

High beta-values above, say, 100 are of limited interest by themselves, except perhaps in phototransistors. The principal benefit of a wide-gap emitter is therefore not the ability to achieve high β-values, but the freedom to change doping levels in emitter and base without significant constraints by injection efficiency consideration, and thereby to re-optimize the transistor at a higher performance level.

We start our discussion with the choice of emitter doping. A wide-gap emitter permits a drop in emitter doping by several orders of magnitude without a deterioration of β, a prediction [1] that has been confirmed experimentally in almost all HBT's built. Now it is well known that the junction capacitance of a highly unsymmetrically doped p-n junction depends only on the doping level of the less heavily doped side. Suppose the base doping is initially kept fixed. If the emitter doping is now dropped below the base doping, the emitter capacitance of the transistor then depends principally on the emitter doping and drops with a decrease of the latter, roughly as

$$C_e \propto N_e^{1/2}. \tag{12}$$

Evidently, by dropping the emitter doping sufficiently far below the (initial) base doping a large reduction in emitter capacitance can be obtained [1], and this reduction remains if the base doping is subsequently increased. The result is an improvement in speed, but this effect is usually small, because the emitter capacitance is only one of several capacitances. The true significance of the reduction of the capacitance per unit area lies in two different facts. First, it permits an increase in the capacitive emitter area in the inverted transistor design discussed later, without increase in total emitter capacitance. Second, in HBT's for small-signal microwave amplification, a reduction in emitter capacitance will reduce the noise significantly [35].

Obviously, the doping in the emitter cannot be lowered arbitrarily far. Even if achievable crystal purities permitted it, the emitter series resistance would eventually become excessive, at least for a thick emitter body. However, under the technological scenario envisaged earlier, the weakly doped part of the emitter can always be kept very thin (say, a few-times 10^{-5} cm) to permit a drop in emitter capacitance per unit area by at least a factor 10 before emitter series resistance effects become serious.

A minor advantage of reduced emitter doping, mentioned by Milnes and Feucht [3], might be that the resulting emitters would have a significant reverse breakdown voltage. It is not clear how much of an advantage this would be.

B. The Base Resistance Tradeoff: Microwave Transistors

The most important single change made possible by a wide-gap emitter is a drastic increase in base doping, limited only by technological constraints and by the need to keep the minority lifetime in the base significantly above 10^{-10} s. The principal benefit is a major reduction in base resistance, which, in turn, increases the speed significantly [2]. A second benefit is a major improvement in overall transistor performance at high current densities [1], [3]–[5], including specifically an improvement in the speed-versus-power tradeoffs of microwave transistors.

Because we are principally interested in low-power speed aspects, we concentrate here on the effect of base resistance reduction. This effect is somewhat different in microwave transistors and in switching transistors.

For microwave transistors, Ladd and Feucht [2] have given a very detailed analysis, using the maximum oscillation frequency f_{max} as the figure of merit. It may be written in the form

$$f_{max} = \tfrac{1}{2}(f_t f_c)^{1/2} \tag{13}$$

where f_t has its familiar meaning as the frequency at which the current gain is reduced to unity, and f_c is the frequency equivalent of the RC time constant of the combination base resistance–collector capacitance,

$$f_c = 1/(2\pi R_b C_c). \tag{14}$$

Evidently, a reduction in R_b causes an increase in f_c and with it a smaller increase in f_{max}.

Ladd and Feucht's work was done in the late 60's and they give numerical values only for the "best" system known at the time, a GaAs emitter on a Ge base, of a construction previously demonstrated by Jadus and Feucht [36]. Because of severe limitations inherent in the then-available technology, the *external* base resistance (between the emitter edge and the base contact) could not be significantly decreased, and as a result, Ladd and Feucht concluded that only a negligible improvement in frequency could be achieved with the then-existing technology. If, however, the external base resistance problem could be solved, maximum oscillation frequencies f_{max} around 100 GHz would be achievable. Similarly high values can be predicted for other heterosystems such as (Al,Ga)As-on-GaAs or GaP-on-Si [37], [38]. There is little point in quoting more exact values, becaue the predictions depend noticeably on both technological and operating parameters whose choice would be applications-dependent. To pursue these matters in

detail would lead us too far away from our principal interest in digital switching transistors.

C. The Base Resistance Tradeoff: Digital Switching Transistors

The quantity of interest in digital switching transistors is not the maximum frequency of oscillation but the (somewhat vaguely defined) switching time. Although one would expect that any structural measures that improve the maximum oscillation frequency will also improve the switching speed, there is no simple one-to-one relationship between the two. The modes of operation are just too different. For example, in microwave transistors a high output power is usually of interest, while in highly integrated digital switching transistors the opposite is the case.

A comparison is further complicated by the fact that switching time depends on the circuit, and no standard measure for switching time, comparable to the frequencies f_t and f_{max} for oscillatory operation, has been agreed upon. Probably the best measure of switching time applicable to HBT's is the estimate by Dumke, Woodall, and Rideout (DWR) [5], who estimate the switching time as

$$\tau_s = \frac{5}{2} R_b C_c + \frac{R_b}{R_L} \tau_b + (3C_c + C_L) R_L. \qquad (15)$$

Here R_b is the base resistance, C_c the collector capacitance, and τ_b the base transit time, while R_L and C_L are load resistance and capacitance of the circuit. The result (15) is based on Ashar's analysis [39] of a two-transistor circuit, modified by Dumke. Dumke's modification simply consists of the following [40]. The load resistance must be large enough to develop a potential change equal to the necessary emitter swing ΔV on the next stage. Therefore, $R_L = \Delta V/I = R_E$, where I is the current that is switched to. Making the appropriate substitutions in Ashar's expression yields (15). Dumke *et al.* apply (15) to estimate the switching time of a hypothetical (Al,Ga)As-on-GaAs transistor with the following parameters. Base width: 1200 Å; base doping: 3×10^{18} cm^{-3}; base and emitter stripe widths: 2.5 μm, separated by 0.5-μm gaps; collector doping: 3×10^{16} cm^{-3}; load resistance: 50 Ω; load capacitance: negligible compared to collector capacitance. These values lead to the following values for the three terms in (15): 8.3 ps, 1.4 ps, and 8.3 ps, combining into an overall switching time of ~18 ps. The authors state that this is "roughly a factor of 5 or 8 faster than that which might be realized from the current post alloy diffused Ge or double diffused Si technologies respectively." Today, nearly 10 years later, post-alloy diffused Ge technology is all but forgotten (it never made it into IC's), and much of the then-predicted advantage over Si remains.

Just as in the case of Ladd and Feucht's estimate of f_{max}, much of the improvement is due to the reduction in base resistance that is associated with the high base doping possible in an HBT. In fact, two of the three terms in (15) depend linearly on R_b rather than with the square root as does f_{max}. This means that as long as those terms dominate τ_s, a reduction of R_b is even more effective in a digital switching transistor than in a microwave transistor. Only after the base resistance reduction has been carried so far that the $R_L C_L$ term dominates, does a further reduction in R_b lead to no further benefit. The hypothetical device analyzed by DWR lies at the borderline between the two regimes.

The specific numerical values quoted above should be viewed as approximations. To obtain an expression as simple as (15), Ashar and Dumke had to make numerous simplifications, just as the expression (13) for f_{max} is based on gross simplifications. The importance of the Ashar–Dumke result (15) is that it indicates the relative significance of the most important transistor parameters. A more detailed analysis is certainly needed, in particular, one that investigates the extent to which the various approximations made in deriving (15) remain applicable in HBT's that have been drastically modified from conventional design.

The assumption of different structural transistor parameters would, of course, have led to different values of τ_b. But the values assumed by DWR were quite reasonable in 1972; they are easily within the range of today's technology, and hence conservative. Further reductions in τ_b to below 10 ps appear readily achievable.

One possibility for improvement is to strive for a lower load resistance than the ad hoc value of 50 Ω assumed by DWR. One sees readily from (15) that the switching time goes through a minimum for

$$R_L = [R_b \tau_b/(3C_c + C_L)]^{1/2} \qquad (16)$$

for which (15) reduces the

$$\tau_s = \frac{5}{2} R_b C_c + 2[(3C_c + C_L) R_b \tau_b]^{1/2}. \qquad (17)$$

For the structural values assumed in DWR one would need $R_L \cong 21\ \Omega$, which would yield $\tau_s \cong 15$ ps. The improvement is not large, and the low load resistance might not be easy to achieve [40]. A much larger improvement would result from a reduction of the collector capacitance, obtained by inverting the transistor. This possibility will be discussed later.

D. The External Base Resistance Problem

In their detailed analysis of the (microwave) performance potential of HBT's, Ladd and Feucht go to great lengths to discuss the special problem posed by the highly detrimental external portion of the base resistance. Because their considerations also apply to digital switching transistors, and because they appear not to have been fully appreciated by subsequent workers on heterostructure bipolar transistors [41], it appears proper to re-emphasize the problem raised by Ladd and Feucht here, and to offer a remedy.

In all real transistors only part of the base resistance lies underneath the emitter, part lies between the edge of the emitter and the base contact. Usually, the outer region of the base is appreciably thicker than the inner region, and the near-surface portion of the outer base is more heavily doped than the remainder (Fig. 5(a)). This design minimizes the outer base resistance. If one wishes to obtain the postulated advantages of a wide-gap emitter, it is essential that the outer base resistance is not permitted to dominate the overall base resistance. This is easier said than done. For example, suppose that technological changes associated with the change from a (diffused or implanted) homojunction emitter to a heterojunction emitter, forced a change in geometry from that in Fig. 5(a) to that in Fig. 5(b) with a thin outer base. This is in fact the geometry used in the HBT's reported in the literature, except for the transistors reported by Ankri *et al.* [11], [14] and by Katz *et al.* [23]. Even though the doping level in the

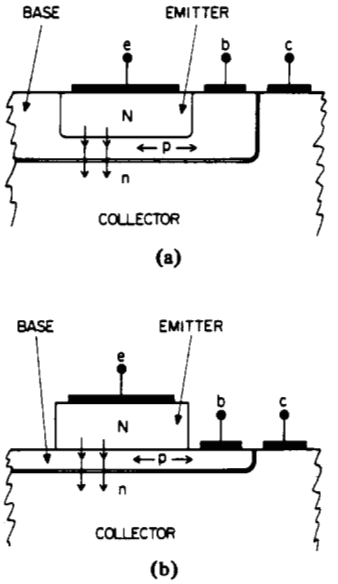

Fig. 5. In homojunction transistors of current technology (a), the base region is usually much thicker and more heavily doped outside the emitter than between the emitter and collector, reducing the external base resistance. This desirable feature would be lost in heterostructure transistors with the emitter island design shown in (b). To appreciate this point fully, recall that in actual structures the horizontal dimensions greatly exceed the vertical ones. In this drawing (and in Fig. 6) the vertical dimensions have been greatly exaggerated relative to the horizontal ones.

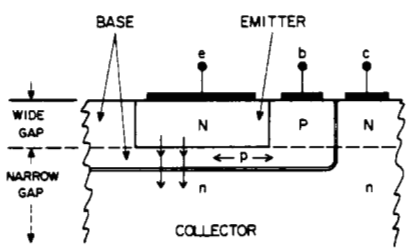

Fig. 6. Desirable emitter structure in which the p-n junction does not follow the planar hetero-interface, but is pulled up towards the surface.

outer base may have been increased, the beneficial effect of this change would be at least partially compensated by the reduction in thickness of the outer base. In unfavorable cases the outer base resistance might even have increased. Ladd and Feucht fully recognized the importance of this problem. They wrote "... *it is clear that the advantages of the low base resistance of the heterojunction devices will only be exploited if suitable geometries can be developed.*"

It is now important to realize that the wide-gap emitter configuration contains a built-in design possibility to keep the outer base resistance low [37], [38], [41], [42]. The design is shown in Fig. 6. Rather than constructing the wide-gap emitter as an island riding by itself on the top of a uniformly thin narrow-gap base layer, the wide-gap semiconductor may be extended beyond the emitter edge, forming part of the outer base region, with the emitter-base p-n junction pulled away from the heteroboundary and towards the surface. Such a configuration should be easily achievable by first growing the top wide-gap layer with the same relatively low n-type doping as the emitter, and then converting the region outside the emitter to heavy p-type doping by diffusion or ion implantation.

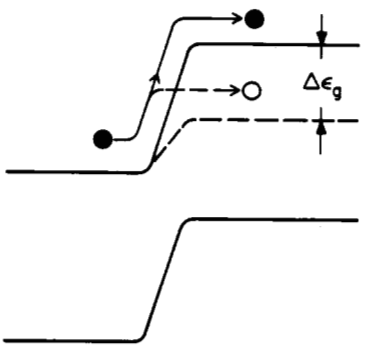

Fig. 7. Blocking of injection of electrons into the wide-gap portion of the base region in Fig. 6, due to the extra repulsive force generated by the wider energy gap.

In such a design the portion of the emitter that lies within the wide-gap region carries only a negligible current, compared to the wide–narrow portion. The reason for this is illustrated in Fig. 7. For injection into the wide-gap p-region, the electrons would have to climb a barrier that is higher by the energy gap difference $\Delta\epsilon_g$. But this reduces the injection current density by the same factor $\exp(-\Delta\epsilon_g/kT)$ that also reduces the hole injection into the wide-gap emitter.

This possibility does not appear to have been as widely recognized as it deserves; it *has* been used in the devices reported by Ankri *et al.* [11], [14], and by Katz *et al.* [23]. In both cases diffusion was used to convert the wide-gap portion of the base region to p-type.

VI. The "Inverted" Transistor

Since the first days of the alloy transistor, bipolar transistors have been built with a larger collector than emitter area, in the interest of efficient charge collection. In planar technology, the two junctions are necessarily of different area. The need for efficient charge collection then enforces the familiar configuration with the collector at the bottom and the emitter at the top. The exception to this rule is, of course, integrated injection logic (I^2L), where other considerations override this rule—at a price. I will say more about I^2L below. But apart from the I^2L exception, the "emitter-up rule" is so pervasive that it has become hard to imagine that a useful transistor could be built with the inverse order.

Now we have just seen that with a wide-gap emitter the emitter junction can be designed in such a way that part of the emitter-base junction does not inject carriers. Evidently, with such a design the need for efficient carrier collection can be met even with an emitter larger than the collector, IF those portions of the emitter-base junction that are not immediately opposite to a part of the collector-base junction are inactivated by pulling them onto the high-gap side of the hetero-interface. Once this is done, the transistor might just as well be "flipped," with the emitter on the substrate side and the collector on top, as shown in Fig. 8. The inverted configuration has several advantages, to the point that it might very well turn out the "canonical" configuration of future heterostructure bipolar transistor design [43].

The principal (but not the only) advantage of the inverted transistor is that it permits the use of a significantly smaller collector area, with an appropriately smaller collector capacitance. The consequences for the high-speed performance are obvious. Modern high speed transistors, both digital and (inter-

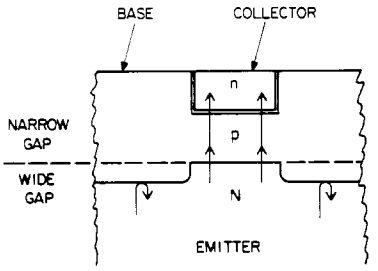

Fig. 8. Inverted "collector-up" transistor structure in which the emitter has a larger area than the collector, but the external portions of the emitter do not contribute to the injection, because there the p-n junction has been pulled into the wide-gap portion of the structure.

digitated) microwave transistors, typically have a collector area close to three-times the active (emitter) area. Inverting the structure thus permits a reduction of the collector capacitance by close to a factor of 3. For example, in the hypothetical switching transistor analyzed by Dumke *et al.* [5], the emitter area was 3.4-times the emitter area. If, in that device, one reduces the collector area by a factor $\frac{1}{3}$ and leaves all other quantities unchanged, the two dominant terms in (15) are reduced by the same factor, and the switching time is reduced from ~18 ps to ~7 ps. Similar improvements would occur in microwave power transistors.

However, some care is in order: Because now the total emitter area is larger than the active area, the emitter junction capacitance will increase, at least compared to a heterostructure transistor of conventional emitter-up configuration. But, as we saw earlier, the emitter junction capacitance per unit area of a heterojunction transistor can in any event be made significantly less than for a homojunction transistor. Hence, compared to the latter, a net reduction in emitter capacitance may result even in the face of a larger (inactive) emitter area.

A second advantage of the inverted configuration is the possibility of a major reduction of the large lead inductance in series with the emitter that is present in the conventional emitter-up configuration. Again an improvement in high-frequency properties will result.

A third advantage of an inverted transistor configuration, for digital switching transistors, will emerge later.

Technologically, the inverted structure should be achievable in essentially the same way as the pulled-up emitter junction: By first growing the top layer lightly n-type doped throughout, and then converting the region outside the collector to heavy p-doping by diffusion or ion implantation. Obviously, the collector layer must be chosen thick enough to support the intended collector bias voltage. Converting part of the surface inside the collector region to n^+ might be desirable.

VII. DH Transistors

A. Introduction: The Wide-Gap Collector

A reading of Shockley's patent quoted at the beginning of this paper leaves no doubt that the "*one. . . zone . . . having a wider energy gap than. . . the other zones*" is the emitter of the transistor. The question was soon raised whether there might also be advantages to a wide-gap collector [1]; but only the trivial and insignificant advantage of a reduction in the reverse-biased collector saturation was recognized.

This assessment must be revised in the light of the anticipated technological scenario discussed in Section III of this paper,

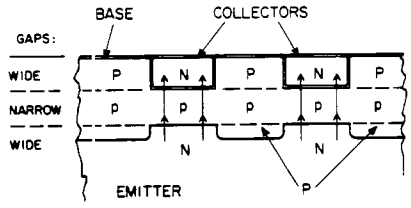

Fig. 9. A DH implementation of I^2L, combining wide-gap collectors with noninjecting emitter regions between the collectors.

and particularly in the light of the increased interest in highly integrated digital switching transistors. It appears that there are in fact several excellent reasons urging a wide-gap collector design, to the point that DH transistors with a wide-gap collector might very well be the rule rather than the exception for future bipolar transistor designs.

I give in this Section three examples that illustrate advantages to be gained by such a design. They fall into three groups:

a) Suppression of hole injection from base into collector in digital switching transistors under conditions of saturation;
b) Emitter/collector interchangeability in IC's;
c) Separate optimization of base and collector, especially in microwave power transistors.

The presentation does not attempt to give a complete and systematic critical evaluation of all aspects of DH transistor design. Its purpose is to initiate a discussion, not to end it.

B. Suppression of Hole Injection into the Collector under Saturated Conditions

In many digital logic families the collectors of the transistors are forward-biased during part of the logic cycle. If the base region is more heavily doped than the collector, as would normally be desirable, a copious injection of holes from the base into the collector takes place, which increases dissipation and slows down the switching speed. In a heterostructure technology, this highly deleterious phenomenon is easily suppressed the same way hole injection into the emitter is suppressed: By making the collector a wide-gap collector [40]. Such a design is an attractive alternative to the Schottky clamp in Schottky-TTL. Just as the wide-gap emitter, the wide-gap collector should be fairly lightly doped, in the interest of a low collector capacitance, and the base should remain heavily doped, in the interest of low base resistance. This choice of relative doping levels remains both possible and desirable in the inverted I^2L configuration, rather than calling for a weakly doped base to suppress collector injection, with its high base resistance penalty. In fact, in a recent paper [42], Kroemer has proposed a DH implementation of I^2L, which combines this idea with the idea of a selectively injecting emitter, discussed earlier. The structure is shown in Fig. 9. It avoids both the electron injection into those portions of the base where such injection is undesirable because of the absence of a collector opposite to the emitter, and the injection of holes into either collector or emitter. Even electrons spilling over at the edge of the active portion of the base region would not be able to penetrate into the upper part of the inactive portion of the base, because they would be repelled by the heterobarrier in the conduction band at the p–P interface. Because of the essentially complete suppression of parasitic charge storage, combined with greatly reduced *RC*-time constant effects due to the reduced base resistance, such an implementation of I^2L can be expected to have a much higher speed than the notoriously

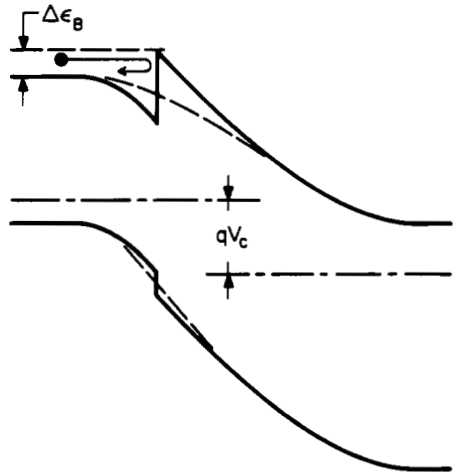

Fig. 10. Electron blocking action for low reverse bias at an abrupt p-n heterojunction collector. The blocking action can be prevented by grading the heterojunction, as indicated by the broken line.

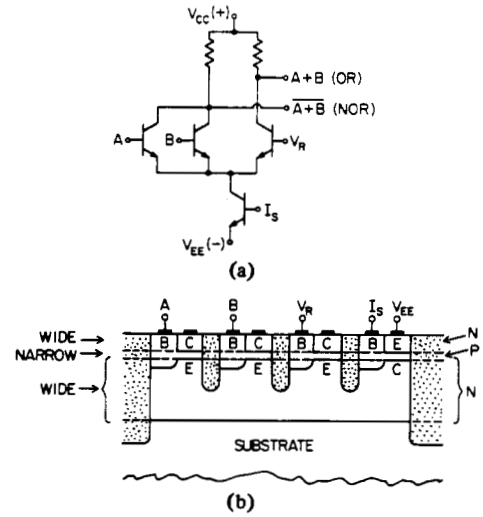

Fig. 11. Input stage of a DH implementation of ECL. The four transistors shown are implemented by three inverted and one noninverted transistor of identical structure, differing only in biasing. The dotted regions are isolation regions, prepared by proton bombardment or equivalent techniques.

slow homostructure implementations of I^2L, without increasing the highly desirable low dissipation levels of I^2L. Unfortunately, no quantitative estimates of the expected performance improvement have so far been published, but the possible improvements appear to be large.[3]

The referenced paper [43] also shows that the pnp horizontal transistor that serves as a current source in I^2L is easily incorporated into a DH design. It emerges as a rather peculiar structure that is basically a homostructure transistor with heterostructure sidewalls, which confine the current and improve the performance of the device.

There is one important restriction in the use of wide-gap collectors, which must not be overlooked. It is important that the free collection of electrons by the reverse-biased collector not be impeded by any heterobarrier due to a conduction band discontinuity (Fig. 10). Such barriers are easily eliminated by grading the heterostructure [44], [45].

C. Emitter/Collector Interchangeability

The advantages of a DH design for bipolar transistors are not restricted to the suppression of hole injection into the collector in saturating logic. A different advantage lies in the possibility of designing transistors in which the role of emitter and collector can be interchanged by simply changing the biasing conditions, while retaining the advantages of a wide-gap emitter regardless of which of the two terminal n-regions is used as the emitter. To achieve this freedom, the transistor need not be geometrically symmetrical: In the inverted structure shown earlier in Fig. 8, in which the active portion of the lower p-n junction covered the same area as the upper p-n junction; either the upper junction or the lower junction could be used as the emitter. While this might be no more than a mildly esoteric advantage in a discrete transistor, it offers a major new option in the architecture of digital IC's, be they of the saturating or nonsaturating variety: The DH design makes it possible, within a common three-layer n-p-n epitaxial layer structure, to integrate high-performance wide-gap emitter transistors having the conventional emitter-up configuration, with similar transistors having the I^2L-like inverted emitter-down configuration discussed previously.

The full power of this new option can probably not be appreciated without an example. The input stage of emitter-coupled logic (ECL), a nonsaturating logic family, serves admirably. Fig. 11(a) gives the basic circuit diagram of the parts of interest here. The top three transistors serve as a differential switch that compares the voltage levels of two logic signals A and B with a reference voltage V_R. The bottom transistor serves essentially as a constant-current source. (In some simpler versions of ECL it is replaced by a resistor.)

Evidently, the configuration calls for tying together the emitters of three transistors with the collector of a fourth. In a DH design, this integration is achieved easily, without sacrificing a high transistor performance, by implementing the top three transistors as inverted transistors, and the current supply transistor as a conventional emitter-up transistor, as shown in Fig. 11(b). The emitters of the three top transistors and the collector of the bottom transistor come together in a buried n-layer on top of the substrate. All four transistors are structurally identical; they differ merely in their biasing. Those readers who are familiar with ECL and its notorious integration difficulties will undoubtedly recognize the great integration advantages offered by what I would like to call HECL, for Heterostructure ECL.

A complete discussion of various other heterostructure modifications of ECL is intended for another place; the purpose of the present discussion was merely to demonstrate the central idea of the interchangeability of emitter and collector in a DH IC design.

D. Separate Optimization of Base and Collector

Except for the interrelated needs of a high mobility and a high saturated drift velocity for the electrons, the semiconductor properties desired for the base of a transistor are quite different from those for the collector and for the base/collector depletion layer. This is especially true in microwave power transistors. Evidently the different needs of base and collector regions can, at least in principle, be optimized best by selecting different materials in the two regions, that is, by a heterostruc-

[3] I have been informed by an anonymous reviewer that K. T. Alavi, in an unpublished M.S. thesis (M.I.T., 1980) has estimated that "over a 10-fold improvement in speed-power product can be anticipated." I did not have access to this work.

ture collector. In practice, this tends to mean a semiconductor with a wider energy gap in the collector and in the base/collector layer, compared to the base region.

Again, an example is called for to illustrate this idea. Consider the question as to the semiconductor combination offering the highest speed in a room-temperature microwave power transistor. One can argue that the fastest possible such transistor would be a GaAs-Ge-GaAs transistor [46] –IF such a transistor could in fact be built, which is by no means certain.

The reason for the choice of Ge as the ideal semiconductor for the base region is its high hole mobility, unexcelled by any other group-IV or III/V-compound semiconductor. Also, Ge is easily doped very heavily p-type. Taken together, the two properties assure a much lower base resistance than any other known useable semiconductor.

Admittedly, Ge has a lower electron mobility than several III/V compounds one might consider. But in a microwave power transistor with its necessarily fairly thick collector depletion layer (in the interest of a high breakdown voltage and a low collector capacitance) the transit time through the base is only a minor speed limitation compared to that through the collector depletion layer. Hence the beneficial effects of the high hole mobility in a Ge base layer are much larger than the detrimental effects of the lower electron mobility compared to, say, GaAs. On the other hand, Ge is hardly a desirable semiconductor for the collector and the base/collector depletion layer: Apart from a somewhat low saturated electron drift velocity ($v_s \cong 5 \times 10^6$ cm/s) and a high dielectric constant ($\epsilon \cong 16$), its low energy gap would lead to a low breakdown field and high thermally generated currents. Here a wider gap semiconductor is needed. Lattice-matching considerations suggest GaAs, which would be near-ideal in any event. One might be inclined to argue that the narrow gap of Ge also rules Ge out as a base region material of acceptably low thermal current generation rate. However, this is not the case: In a practical GaAs-Ge-GaAs transistor the Ge base region would be so thin and so heavily p-type doped that the thermal generation of electrons in the base would not contribute an unacceptably high collector saturating current.

Unfortunately, it is not at all clear whether or not GaAs-Ge-GaAs transistors with an acceptably low density of interface defects can be grown. Our own work at UCSB with the MBE growth of GaAs on Ge, and GaP on Si, has shown that the defect-free growth of a polar semiconductor such as GaAs on a nonpolar substrate such as Ge faces a number of quite fundamental difficulties, which have so far not been surmounted, and which may, in fact, be insurmountable [47].

However, none of the experimental uncertainties affect the principal point of our discussion here: The desirability of different semiconductors for base and collector, implying a heterostructure collector, is likely to be the rule rather than the exception in the technology of the future.

VIII. Some Speculations About the Future of Compound Semiconductor Devices

A. Bipolar Transistors versus FET's

If one ignores injection lasers and other optoelectronic devices, today's compound semiconductor device world is a pure FET world with essentially no bipolar inhabitants. A paper that predicts what amounts to a bipolar revolution in this FET world cannot simply ignore FET's. This is true even more once one realizes that the same technologies that promise to revolutionize bipolar transistors will also improve FET's [48]. In fact, very active and successful research into heterostructure FET's is already under way. However, on balance, heterostructures can be expected to benefit bipolar devices much more than they benefit FET's, and if so, this will naturally tend to shift the balance between the devices much more towards bipolars than past developments might suggest. There are several reasons for these expectations:

a) As was pointed out already in Section II, the Central Design Principle permits one to control the flow of electrons and holes separately and independently of each other. This makes heterostructures a very major advantage in bipolar devices (including lasers) in which there are in fact both kinds of carriers present. It does little for an FET, although a related benefit is obtained in FET's through the concept of modulation doping [49].

b) Every device has a dimension in the direction of current flow that controls the speed of the device. In FET's (other than VMOS) the current flow is parallel to the surface, and the critical control dimension is established by fine-line lithography. In a bipolar transistor, the speed-determining part of the current path is perpendicular to the surface (and to the epilayers), and to the first order, speed is governed by the layer thicknesses. Because vertical layer thicknesses can be easily made much smaller than horizontal lithography dimensions, there is, for given horizontal dimensions, an inherently higher speed potential in bipolar structures than in FET's. The two qualifiers "to the first order" and "for given horizontal dimensions" are important, though: Small horizontal dimensions are still needed to minimize speed-limiting *second-order* effects caused by horizontal resistive voltage drops in the thin base layers. These second-order effects are actually reduced in HBT's, due to the much higher base doping levels, and they are not as severe as the first-order limiting effects of the horizontal dimensions in FET's. But in any event, there is nothing in bipolar technologies that would require or even suggest the use of larger horizontal dimensions than in FET's. The same fine-line lithography technologies that are used for FET's, can and will be used for bipolar devices. The capability offered by the new epitaxial technologies is an *additional* capability, not an alternate.

c) Once sufficiently small dimensions have been achieved, "ballistic" effects become important [32], [33], and they are in fact extensively studied, so far predominantly in an FET context. On the whole, ballistic effects improve device performance by minimizing electron scattering. To obtain this benefit, two conditions must be satisfied. First, the electrons must be accelerated very quickly [32]. The most effective way to do this is by launching the electrons with a high kinetic energy from the conduction band discontinuity in a heterostructure, as discussed earlier. This is much more effective than acceleration by an ordinary nonuniform electric field, the rate of nonuniformity of which is limited by Debye-length considerations. Second, the path along which ballistic effects are to be utilized, must be short, at most a few thousand Angstrom units long. Evidently, both the abrupt launching and the short current paths call for a current flow perpendicular to the epitaxial layers rather than parallel to them, once again favoring the geometry of bipolar designs.

d) All digital switching transistors have a critical bias voltage (often called turn-on voltage), in the vicinity of which the switching action takes place. For high-performance digital IC's, especially VLSI circuits, it is important that this critical voltage be as reproducible as possible, not only across the chip in a single VLSI circuit, but also from wafer to wafer. This repro-

ducibility is easier to achieve in bipolar transistors than in FET's. In bipolar transistors the turn-on voltage is almost fixed for a fixed energy gap of the semiconductor in the base region. It depends logarithmically on the base doping and, apart from temperature, on hardly anything else. Hence it is easy to keep stable. One might say with little exaggeration that it is close to being a natural constant. The turn-on voltage in an FET is, by contrast, purely "man-made," depending at least linearly on both the electron concentration in the channel and the channel thicknesses. To achieve reproducible turn-on voltages, at least two separate quantities must be controlled tightly. Considering that processing differences tend to be very important in IC technology, this particular difference between bipolars and FET's might well turn out to be as important as the more fundamental differences, strongly favoring bipolars [50].

The above arguments suggest strongly that bipolar devices will play a much larger role in the future that they have in the past, eventually assuming a leading role ahead of FET's. Exactly where the border between the two technologies will be, is something too hazardous to predict.

B. A Change in Technological Philosophy?

We have witnessed, since about 1964, a steady growth in III/V-compound semiconductor devices, principally GaAs devices. The driving force behind this development has been the high performance of such devices, not attainable with mainstream Si devices. If we ignore once again lasers and other optoelectronic devices, and restrict ourselves to purely electronic amplifying and switching devices, high performance has been largely synonymous with high speed, made possible by the high electron mobility of GaAs, and by the availability of semi-insulating GaAs as a substrate. However, not even the most ardent advocate of GaAs ever claimed that GaAs was used because it had an attractive technology. We used GaAs despite its technology, not because of it, and the threat was never far away that Si devices, with their much simpler and more highly developed technology, would catch up with GaAs performance, the fundamental advantages of GaAs notwithstanding.

It is exactly this imbalance between fundamental promise and technological weakness that is being removed by the new epitaxial technologies. If the technological scenario postulated in Section III of this paper is even remotely correct, it means nothing less but that the great future strength of III/V-compounds lies precisely in their new technology, which permits an unprecedented complexity and diversity in epitaxial structures, going far beyond anything available in Si technology! This new technological strength is thus emerging as more important than the older fundamental strengths of high mobilities and semi-insulating substrates. It is a remarkable reversal of priorities indeed.

None of this means even remotely that III/V compounds will replace Si. They will not do so any more than aluminum, magnesium, and titanium replaced steel. The analogy of Si to steel is due to M. Lepselter, who called Si technology "the new steel" [51], to bring out the similarity in the role of Si in the new industrial revolution of our own days, to the role of steel in the industrial revolution of the early-19th century. I would like to carry this excellent analogy a bit further. Just as the *structural* metallurgy of the 19th century found it necessary eventually to go beyond steel, to aluminum, magnesium, titanium, and others taking their place beside steel, so the *electronic* metallurgy of our own age is going beyond Si, to the III/V-compounds and probably further, to take their own place beside Si.

We continue to build locomotives, ships, and automobiles from steel, but if it is airplanes and spacecraft we want, we need the other metals besides. And, of course, it took us a while to go from locomotives to spacecraft. The analogy to semiconductors is too obvious to require elaboration; only the time scale will be compressed.

All along the way from steel to titanium there were those who argued that the next step, while perhaps possible, was one for which no foreseeable need existed: All foreseeable needs of man could presumably be met by improvements of the technologies already in hand. Well, this too has not changed.

ACKNOWLEDGMENT

The work of this paper has greatly benefitted from uncounted discussions, over several years, with numerous individuals. Foremost amongst them were R. C. Eden, D. G. Chen, and S. I. Long, all (at the time) at the Rockwell Electronics Research Center. Others at Rockwell to whom I am indebted for discussions are J. S. Harris, R. Zucca, D. L. Miller, and P. Asbeck. I am grateful to W. P. Dumke (IBM) for a copy of his unpublished work on the switching time of bipolar transistors, which clarified many questions I had about that difficult problem. The final version of the paper benefitted from intense discussions with Prof. H. Beneking (Aachen) and from comments made by two anonymous reviewers.

REFERENCES

[1] H. Kroemer, "Theory of a wide-gap emitter for transistors," *Proc. IRE*, vol. 45, no. 11, pp. 1535–1537, Nov. 1957.

[2] G. O. Ladd and D. L. Feucht, "Performance potential of high-frequency heterojunction transistors," *IEEE Trans. Electron Devices*, vol. 17, pp. 413–420, May 1970.

[3] For a review see A. G. Milnes and D. L. Feucht, *Heterojunctions and Metal-Semiconductor Junctions*. New York: Academic, 1972. (See especially ch. 3.)

[4] For additional references see B. L. Sharma and R. K. Purohit, *Semiconductor Heterojunctions*. Elmsford, NY: Pergamon, 1974. (See especially sect. 7.6.)

[5] W. P. Dumke, J. M. Woodall, and V. L. Rideout, "GaAs-GaAlAs heterojunction transistor for high frequency operation," *Solid-State Electron.*, vol. 15, no. 12, pp. 1339–1334, Dec. 1972.

[6] a) M. Konagai and K. Takahashi, "Formation of GaAs-(GaAl)As heterojunction transistors by liquid phase epitaxy," *Elect. Eng. Japan*, vol. 94, no. 4, 1974;
b) ——, "(GaAl)As-GaAs heterojunction transistors with high injection efficiency," *J. Appl. Phys.*, vol. 46, no. 5, pp. 2120–2124, May 1975.

[7] B. W. Clark, H.G.B. Hicks, I.G.A. Davies, and J. S. Heeks, "A (GaAl)As-GaAs heterojunction structure for studying the role of cathode contacts on transferred electron devices," *Gallium Arsenide and Related Compounds 1974* (Deauville), Inst. Phys. Conf. Ser., vol. 24, 1975, pp. 373–375.

[8] M. Konagai, K. Katsukawa, and K. Takahashi, "(GaAl)As/GaAs heterojunction phototransistors with high current gain," *J. Appl. Phys.*, vol. 48, no. 10, pp. 4389–4394, Oct. 1977.

[9] P. W. Ross, H.G.B. Hicks, J. Froom, L. G. Davies, F. J. Probert, and J. E. Carroll, "Heterojunction transistors with enhanced gain," *Electron Eng.*, vol. 49, no. 589, pp. 36–38, Mar. 1977.

[10] A. Marty, G. Rey, and J. P. Bailbe, "Electrical behavior of an n-p-n GaAlAs/GaAs heterojunction transistor," *Solid-State Electron.*, vol. 22, no. 6, pp. 549–557, June 1979.

[11] D. Ankri and A. Scavennec, "Design and evaluation of a planar GaAlAs-GaAs bipolar transistor," *Electron. Lett.*, vol. 16, no. 1, pp. 41–47, Jan. 1980.

[12] J-P. Bailbe, A. Marty, P. H. Hiep, and G. E. Rey, "Design and fabrication of high-speed GaAlAs/GaAs heterojunction transistors," *IEEE Trans. Electron Devices*, vol. ED-27, pp. 1160–1164, June 1980.

[13] H. Beneking and L. M. Su, "GaAlAs/GaAs heterojunction microwave bipolar transistor," *Electron. Lett.*, vol. 17, no. 8, pp. 301–302, Apr. 1981.

[14] D. Ankri, A. Scavennec, C. Besombes, C. Courbet, F. Heliot, and J. Riou, "High frequency low current GaAlAs-GaAs bipolar transistor," presented at Dev. Res. Conf., Santa Barbara, June 1981, unpublished.

[15] For an up-to-date account, containing essentially complete earlier references, see three of the most recent papers on the subject:
a) M. Tobe, Y. Amemiya, S. Sakai, and M. Umeno, "High-sensitivity InGaAsP/InP phototransistors," *Appl. Phys. Lett.*, vol. 37, no. 1, pp. 73-75, July 1980;
b) M. N. Svilans, N. Grote, and H. Benking, "Sensitive GaAlAs/GaAs wide-gap emitter phototransistors for high current applications," *IEEE Electron Devices Lett.*, vol. ED-11, pp. 247-249, Dec. 1980;
c) J. C. Campbell, A. G. Dentai, C. A. Burrus, Jr., and J. F. Ferguson, "InP/InGaAs heterojunction phototransistors," *IEEE J. Quant. Electron.*, vol. QE-17, pp. 264-269, Feb. 1981.

[16] For two reviews see:
a) A. Y. Cho and J. R. Arthur, "Molecular beam epitaxy," *Prog. Solid State Chem.*, vol. 10, pt. 3, pp. 157-191, 1975;
b) K. Ploog, "Molecular beam epitaxy of III-V compounds," in *Crystals: Growth, Properties, and Applications*, H. C. Freyhardt, Ed. New York: Springer-Verlag, 1980, vol. 3, pp. 73-162.

[17] For a review with complete references to earlier work, see: R.D. Dupuis, L. A. Moudy, and P. D. Dapkus "Preparation and properties of $Ga_{1-x}Al_xAs$-GaAs heterojunctions grown by metal-organic chemical vapor deposition," *Gallium Arsenide and Related Compounds 1978* (St. Louis), Inst. Phys. Conf. Ser., vol. 45, pp. 1-9, 1979.

[18] R. A. Milano, T. H. Windhorn, E. R. Anderson, G. E. Stillman, R. D. Dupuis, and P. D. Dapkus, "$Al_{0.5}Ga_{0.5}As$-GaAs heterojunction phototransistors grown by metalorganic chemical vapor deposition," *Appl. Phys. Lett.*, vol. 39, no. 9, pp. 562-564, May 1979.

[19] D. L. Miller, personal communication.

[20] a) T. Matsushita, N. Oh-uchi, H. Hayashi, and H. Yamoto, "A silicon heterojunction transistor," *Appl. Phys. Lett.*, vol. 35, no. 7, pp. 549-550, Oct. 1979;
b) N. Oh-uchi, H. Hayashi, H. Yamoto, and T. Matsushita, "A new silicon heterojunction transistor using the doped SIPOS," *IEDM Dig.*, pp. 522-524, Dec. 1979;
c) T. Matsushita, H. Hayashi, N. Oh-uchi, and H. Yamamoto, "A SIPOS-Si heterojunction transistor," *Japan. J. Appl. Phys.*, vol. 20, suppl. 20-1, pp. 75-81, Jan. 1981 (Proc. 12th Conf. Solid-State Devices, Tokyo, Aug. 1980).

[21] T. Katoda and M. Kishi, "Heteroepitaxial growth of gallium phosphide on silicon," *J. Electron Mat.*, vol. 9, no. 4, pp. 783-796, Apr. 1980.

[22] S. L. Wright and H. Kroemer, to be published.

[23] J. Katz, N. Bar-Chaim, P. C. Chen, S. Margalit, I. Ury, D. Wilt, M. Yust, and A. Yariv, "A monolithic integration of GaAs/GaAlAs bipolar transistor and heterostructure laser," *Appl. Phys. Lett.*, vol. 37, no. 2, pp. 211-213, July 1980.

[24] H. Beneking, N. Grote, and M. N. Svilans, "Monolithic GaAlAs/GaAs infrared-to-visible wavelength converter with optical power amplification," *IEEE Trans. Electron Devices*, vol. ED-28, pp. 404-407, Apr. 1981.

[25] H. Kroemer, "A proposed class of heterojunction injection lasers," *Proc. IEEE*, vol. 51, pp. 1782-1783, Dec. 1963.

[26] See, e.g., J. J. Coleman, P. D. Dapkus, N. Holonyak, Jr., and W. D. Laidig, "Device-quality epitaxial AlAs by metalorganic-chemical vapor deposition," *Appl. Phys. Lett.*, vol. 38, no. 11, pp. 894-896, June 1981. This paper quotes only structures containing about 80 layers; much larger numbers have been achieved in unpublished work (personal communication).

[27] For two recent reviews see:
a) L. L. Chang and L. Esaki, "Semiconductor superlattices by MBE and their characterization," *Prog. Cryst. Growth Charact.*, vol. 2, no. 1, pp. 3-12, 1979;
b) A. C. Gossard, "Molecular beam epitaxy of superlattices in thin films," in *Thin Films: Preparation and Properties*, K. N. Tu and R. Rosenberg, Eds. New York: Academic, to be published.

[28] J. S. Slotboom and H. C. de Graaf, "Measurement of bandgap narrowing in Si bipolar transistors," *Solid-State Electron.*, vol. 19, no. 10, pp. 857-862, Oct. 1976.

[29] See, e.g.,
a) J. A. Archer, "Design and performance of small-signal microwave transistors," *Solid-State Electron.*, vol. 15, no. 3, pp. 249-258, Mar. 1972.
b) J. M. Gladstone, P. T. Chen, P. Wang, and S. Kakihana, "Computer aided design and fabrication of an X-band oscillator transistor," Int. Electron Devices Meeting (IEDM) 1973, *IEDM Dig.*, pp. 384-386, Dec. 1973.
c) T. W. Sigmon, "Characteristics of high performance microwave transistors fabricated by ion implantation," Int. Electron Devices Meeting (IEDM) 1973, *IEDM Dig.*, pp. 387-389, Dec. 1973.

[30] C.E.C. Wood, G. Metze, J. Berry, and L. F. Eastman, "Complex free-carrier profile synthesis by atomic-plane doping of MBE GaAs," *J. Appl. Phys.*, vol. 51, no. 1, pp. 383-387, Jan. 1980.

[31] R. Dingle, "Confined carrier quantum states in ultrathin semiconductor heterostructures," *Festkörperprobleme/Advances in Solid State Physics*, vol. 15, pp. 21-48, 1975.

[32] For a review, see H. Kroemer, "Hot electron relaxation effects in devices," *Solid-State Electron.*, vol. 21, no. 1, pp. 61-67, Jan. 1978.

[33] M. S. Shur and L. F. Eastman, "Ballistic and near ballistic transport in GaAs," *IEEE Electron Devices Lett.*, vol. EDL-1, pp. 147-148, Aug. 1980.

[34] H. Kroemer, "Heterojunction device concepts," U.S. Air Force Tech. Rep. AFAL-TR-65-243, Oct. 1965, unpublished. A published description is found in Milnes and Feucht [3], pp. 28-29.

[35] R. E. Yeats, personal communications.

[36] D. K. Jadus and D. L. Feucht, "The realization of a GaAs-Ge wide band gap emitter transistor," *IEEE Trans. Electron Devices*, vol. 16, pp. 102-107, Jan. 1969.

[37] H. Kroemer, Dev. Res. Conf. 1978, Santa Barbara; see *IEEE Trans. Electron Devices*, vol. ED-25, p. 1339, Nov. 1978.

[38] —, *Bull. Amer. Phys. Soc.*, vol. 24, p. 230, Mar. 1979.

[39] K. G. Ashar, "The method of estimating delay in switching circuits and the figure of merit of a switching transistor," *IEEE Trans. Electron Devices*, vol. ED-11, pp. 497-506, Nov. 1964.

[40] W. P. Dumke, personal communication, unpublished.

[41] The concept at issue here *has* been widely discussed in the DH laser literature. See, e.g., W. Susaki, H. Namizaki, H. Kan, and A. Ito, "A new geometry double-heterostructure injection laser for room-temperature continuous operation: Junction-stripe-geometry DH lasers," *J. Appl. Phys.*, vol. 44, no. 6, pp. 2893-2894, June 1973.

[42] H. Kroemer, "Heterostructures for everything—device principle of the 1980's?," *Japan. J. Appl. Phys.*, vol. 20, suppl. 20-1, pp. 9-13, Jan. 1981 (Proc. 12th Conf. Solid-State Devices, Tokyo, Aug. 1980).

[43] Inverted transistors (with a Schottky collector) with a wide-gap emitter, but without the idea of inactivating the "uncovered" part of the emitter area, have already been reported: H. Beneking, N. Grote, W. Roth, L. M. Su, and M. N. Svilans, "Realization of a bipolar GaAs/GaAlAs Schottky-collector transistor," *Gallium Arsenide and Related Compounds 1980* (Vienna), Inst. Phys. Conf. Ser., vol. 56, pp. 385-392, 1981.

[44] W. G. Oldham and A. G. Milnes, "n-n semiconductor heterojunctions," *Solid-State Electron.*, vol. 6, no. 2, pp. 121-132, Mar./Apr. 1963.

[45] D. T. Cheung, S. Y. Chiang, and G. L. Pearson, "A simplified model for graded-gap heterojunctions," *Solid-State Electron.*, vol. 18, no. 3, pp. 263-266, Mar. 1975.

[46] Some of the ideas on this subject were independently developed by Dr. Daniel G. Chen, to whom I owe several detailed discussions on this subject.

[47] For a discussion of those problems, see H. Kroemer, K. J. Polasko, and S. C. Wright, "On the (110) orientation as the preferred orientation for the molecular beam epitaxial growth of GaAs on Ge, GaP on Si, and similar zincblende-on-diamond systems," *Appl. Phys. Lett.*, vol. 36, no. 9, pp. 763-765, May 1980.—Unfortunately, even the switch to the (110) orientation has not solved the problems satisfactorily.

[48] See, e.g., D. Boccon-Gibod, J-P. André, P. Baudet, and J-P. Hallais, "The use of GaAs-(Ga, Al)As heterostructures for FET devices," *IEEE Trans. Electron Devices*, vol. ED-27, pp. 1141-1147, June 1980.

[49] See, e.g., S. Judaprawira, W. I. Wang, P. C. Chao, C.E.C. Wood, D. W. Woodard, and L. F. Eastman, "Modulation-doped MBE $GaAs/n\text{-}Al_xGa_{1-x}As$ MESFETs," *IEEE Electron Devices Lett.*, vol. EDL-2, pp. 14-15, Jan. 1981.

[50] The importance of this advantage of bipolar devices was first pointed out to me by Dr. R. C. Eden.

[51] M. Lepselter, "Integrated circuits—the new steel," Int. Electron Dev. Meeting (IEDM) 1974, *IEDM Dig.*, Dec. 1974.