On the Thermionic-Diffusion Theory of Minority Transport in Heterostructure Bipolar Transistors

Anatoly A. Grinberg and Serge Luryi, *Fellow, IEEE*

Abstract—**Theory of the minority-carrier transport in heterostructure bipolar transistors (HBT) is reconsidered with a particular emphasis on the difference between the cases of abrupt and graded emitter–base junctions and the role in the former case of the quasi-Fermi level discontinuity at the interface. Exact analytical formulas are derived for the current–voltage characteristics of a double-heterojunction HBT, valid for arbitrary levels of injection and base doping, including the degenerate case. The theory is applied to the static characterization of HBT which compares the forward and reverse dependences $I_C(V_{EB})$ and $I_E(V_{CB})$. It is shown that these characteristics coincide in the low-injection limit, if both the emitter–base and the collector–base diodes have ideality factors close to unity. The ratio of base currents in the reverse and forward modes of operation can be used to determine the abrupt emitter-base conduction band discontinuity and estimate the scattering length in the base.**

I. Introduction

As is well-known, the minority-carrier transport in the base of a bipolar transistor is adequately described by the drift-diffusion equation, provided the base width W is sufficiently large, $W >> l_{sc}$, where $l_{sc} \sim D/v_T$ is the characteristic scattering length in the base, D is the diffusion coefficient, and v_T the thermal velocity of minority carriers. For shorter bases, $W \leq l_{sc}$, the drift-diffusion equation breaks down, and a more refined Boltzmann-transport model is required. In the present work, we confine ourselves to the long-base limit. Even in this limit, the analysis of minority transport should carefully distinguish between the cases of a graded-gap (g-HBT) and an abrupt (a-HBT) emitter–base heterostructure. The g-HBT case is essentially similar to the homojunction transistor, except that forces acting on the electrons and holes must include energy-gap variations in addition to the electric fields [1]. A detailed theory, first presented by Marty *et al.* [2], and reproduced essentially without change in modern textbooks [3], [4], really applies only to the g-HBT case. However, the apparent generality of the approach [2] has often led to its uncritical use in the general case of HBT, even though since the publication of that paper, a number of authors [5]–[10] have pointed out that the correct mechanism of transport across an abrupt heterointerface is thermionic emission. One of the purposes of the present work is to develop a theory applicable to the case of a-HBT by following the method used in [2] and introducing the appropriate corrections along the way.

The basic relations of Marty *et al.* [2] in need of revision when applied to a-HBT are

$$\int_{C_1}^{C_2} \frac{J_n\,dz}{\mu n} = eV \tag{1}$$

$$np = n_i^2\, e^{eV/kT} \tag{2}$$

where J_n is the electron current, V the applied voltage, $\mu = eD/kT$ the electron mobility, and n, p, n_i are, respectively, the electron, hole, and intrinsic concentrations in the base near the emitter junction. The integral in (1) is taken between the emitter contact C_1 and an auxiliary contact C_2 deep in the base.

Both of these relations implicitly rely on the continuity of the quasi-Fermi level (imref) E_{Fn} for electrons at the emitter-base junction, an assumption which breaks down for a-HBT. For a continuous E_{Fn}, (1) can be derived by integrating the drift-diffusion expression

$$\frac{J_n}{\mu n} = \nabla E_{Fn} \tag{3}$$

from "contact to contact." However, (3) is not valid in the vicinity of an abrupt interface, where, as is well-known from the theory of Schottky diodes, the imref E_{Fn} suffers a discontinuity. Therefore, the integral (1) should be understood as a sum of integrals $\int_{C_1}^{-\epsilon}$ and $\int_{+\epsilon}^{C_2}$, excluding a small region $\pm\epsilon$ from the interface. In that small region, the transport is governed by thermionic emission instead of (3) and a correct model of the imref variation must include a finite drop $\delta E_{Fn} \equiv E_{Fn}(+\epsilon) - E_{Fn}(-\epsilon)$. The contribution of δE_{Fn} drops out in the proper integral and therefore in the right-hand side of (1) one should have eV replaced by $eV - \delta E_{Fn}$. The value of δE_{Fn} may depend on the current density, the base width W, and the recombination rate for electrons in the base.

As will be shown in Section II, the correct expressions[1] replacing (1) and (2) are of the form

$$\int_{C_1}^{C_2} \frac{J_n\,dz}{\mu n} = eV - \Delta + kT \ln\left(\frac{n_{(+)}}{n_{(-)}}\frac{N_C^{(E)}}{N_C^{(B)}}\right) \tag{4}$$

$$n_{(+)}p_{(+)}\frac{N_C^{(B)}}{N_C^{(E)}}\frac{n_{(-)}}{n_{(+)}}\, e^{\Delta/kT} = n_i^2\, e^{eV/kT} \tag{5}$$

---

[1] Equation (5) is valid only for nondegenerate $p_{(+)}$, just as its homojunction analog (2). The more general relation, valid for arbitrary hole concentrations in the base, is (19).

Manuscript received August 17, 1992; revised December 1, 1992. The review of this paper was arranged by Associate Editor M. Shur.

The authors are with AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 9207642.

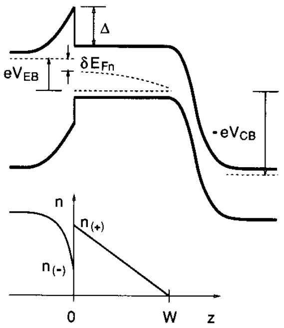


Fig. 1. Band diagram of an abrupt-junction heterostructure bipolar transistor under bias and a schematic electron concentration profile.

where Δ is the conduction-band discontinuity, cf. Fig. 1, $n_{(+)}$ and $p_{(+)}$ are, respectively, the electron and hole concentrations in the base near the emitter junction, $n_{(-)}$ is the electron concentration on the emitter side of the abrupt boundary, and $N_C^{(E)}$ and $N_C^{(B)}$ are the conduction band densities of states in the emitter and base, respectively, in the vicinity of the discontinuity. The assumption of a continuous electronic imref ($\delta E_{Fn} = 0$) is equivalent to

$$\frac{n_{(-)}}{N_{(+)}} = \frac{N_C^{(E)}}{N_C^{(B)}} \times e^{-\Delta/kT}. \tag{6}$$

Equations (4) and (5) reduce to (1) and (2), respectively, if (6) is assumed.

As will be discussed in Section III, the actual value of the ratio $n_{(-)}/n_{(+)}$ can be determined from the boundary condition on the electronic flux

$$-\frac{J_n}{e} = v_R N_C^{(B)} \left(\frac{n_{(-)}}{N_C^{(E)}} - \frac{n_{(+)}}{N_C^{(B)}} e^{-\Delta/kT} \right) \tag{7}$$

where v_R is the Richardson velocity, $v_R = \sqrt{kT/2\pi m}$, and m is the effective mass of electrons in the base.[2] The density-of-states factors enter because of the kinetics of electron transmission across an effective mass discontinuity [11]. Apart from these factors, the boundary conditions (7) have been first used by Grinberg *et al.* [5] in a theory of minority transport, applicable to both g-HBT and a-HBT cases, and subsequently by other authors [6]–[10]. For practically important a-HBT designs [9], the key difference from g-HBT arises from the fact that the emitter current becomes independent of the base width when $\Delta/kT >> 1$. This leads to significant modifications of the Early effect both in the static limit and at high frequencies [12].

In the present work, we shall be only concerned with the static limit. Section IV is devoted to the derivation of general analytic expressions for the current–voltage characteristics of a double-heterojunction HBT with one junction abrupt and the other continuous. The presented theory is valid for arbitrary levels of injection and includes the case of a degenerate base doping. In Section V we consider a commonly used [3], [4] characterization of the HBT comparing the forward and the reverse dependences $I_C(V_{BE})$ and $I_E(V_{BC})$. We show that in the low injection limit these characteristics coincide, quite irrespectively of the value of Δ. Contrary to assertions commonly encountered in the literature, it does not matter whether the current is limited by the potential barrier or by the base transport, so long as the ideality factors of both the emitter–base and the collector–base diodes are close to unity. Comparison of the common-emitter (respectively, common-collector) current gains β in the forward and reverse configurations allows an accurate determination of the conduction-band discontinuity Δ and at the same time gives a good estimate of the ratio between l_{sc} and W. Our conclusions will be summarized in Section VI.

II. General Relations

The drift-diffusion equation for electrons in a graded-gap semiconductor is of the form (3) with the imref E_{Fn} given by

$$E_{Fn} = E_C + kT \ln (n/N_C) \tag{8}$$

where the coordinate dependence of the conduction band edge $E_C = -e(\phi + \chi)$ results from both the electrostatic potential ϕ and the variable affinity χ. Let us rewrite (3) in full

$$\frac{J_n}{e\mu n} = -\frac{d(\phi + \chi)}{dz} + \frac{kT}{e}\frac{d}{dz} \ln \left(\frac{n}{N_C} \right). \tag{9}$$

Consider an abrupt N_p junction, Fig. 2, where $e(\chi_2 - \chi_1) = \Delta$. As discussed in the Introduction, integrating (9) over any region including the band discontinuity (taken at $z = 0$), for example, between the two contacts C_1 ($z < 0$) and C_2 ($z > 0$), we must exclude a special region of infinitesimal thickness near the origin

$$\int_{C_1}^{C_2} \frac{J_n\,dz}{e\mu n} = \int_{C_1}^{-\epsilon} \frac{J_n\,dz}{e\mu n} + \int_{+\epsilon}^{C_2} \frac{J_n\,dz}{e\mu n}. \tag{10}$$

Substituting (9) into (10), we obtain (4) or, equivalently,

$$\int_{C_1}^{C_2} \frac{J_n\,dz}{\mu n} = eV - \delta E_{Fn} \tag{11}$$

where the imref jump at the abrupt interface is given by

$$\delta E_{Fn} \equiv \Delta - kT \ln \left(\frac{n_{(+)}}{n_{(-)}} \frac{N_C^{(E)}}{N_C^{(B)}} \right) = kT \ln \left(1 - \frac{J_n e^{\Delta/kT}}{en_{(+)} v_R} \right). \tag{12}$$

In the second equation in the right-hand side of (12) we have used the boundary condition (7). It should be noted that if, in addition to the discontinuity, there is some affinity grading in the emitter and/or base regions, so that the quantities $N_C^{(E)}$ and $N_C^{(B)}$ themselves have a coordinate dependence, then in (12) one must take

$$N_C^{(E)} \equiv N_C^{(E)}(-\epsilon) \qquad N_C^{(B)} \equiv N_C^{(B)}(+\epsilon). \tag{13}$$

---

[2] It is assumed here that the electron effective mass in the emitter is higher than that in the base. In a thermionic emission process between materials of different effective mass, the effective Richardson constant corresponds to the lower mass [11].

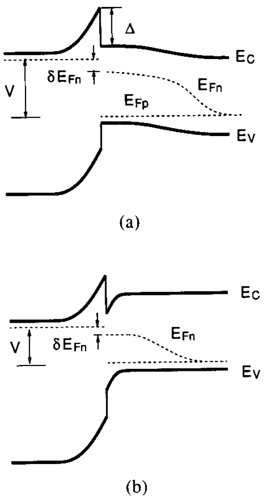


Fig. 2. Schematic band diagram of a forward-biased n-p heterojunction. (a) Case of $N_D \ll N_A$ so that base depletion is negligible. The band bending in the base illustrates the high-injection conditions. (b) Case of non-negligible base depletion ($N_D \lesssim N_A$).

Let us now discuss the derivation of (5). This equation enables one to include the effects of high injection. Just like (2) for a homogeneous p-n junction, (5) is valid only to the extent that the variation of E_{Fn} on the n side of the junction and the variation of the hole imref E_{Fp} on the p side are both negligible. A qualitative difference between these two equations lies in the fact that while (2) establishes a connection between the electron and the hole concentrations at the same point, its a-HBT analog, (5) connects the concentrations $n_{(-)}$ and $p_{(+)}$ evaluated at *different* points.

At a high level of injection, the concentration of holes in the quasi-neutral part of the base increases to maintain neutrality

$$p = p_0 + n - n_0 \tag{14}$$

where p_0 and $n_0 = n_i^2/p_0$ are the equilibrium hole and electron concentrations, respectively, and n_i is the intrinsic carrier concentration in the base

$$n_i^2 \equiv N_C^{(B)} N_V^{(B)} e^{-E_G^{(B)}/kT}. \tag{15}$$

The increased p implies an additional band bending in the base, cf. Fig. 2(a), such that the valence band edge moves upward with respect to E_{Fp}, and consequently, at a given forward bias, the barrier for electron injection increases. Assuming that at the top of the barrier the electron concentration $n_{(-)}$ is nondegenerate (which is almost always a good approximation for a-HBT), we have

$$\begin{aligned} n_{(-)} &= N_C^{(E)} e^{-[E_C(-\epsilon) - E_{Fn}]/kT} \\ &= N_C^{(E)} e^{-[E_G^{(B)} + \Delta]/kT} e^{eV/kT} e^{[E_{Fp} - E_V(+\epsilon)]/kT} \end{aligned} \tag{16}$$

where $E_V = E_C - E_G$ is the valence band edge. In deriving (16) we have used $eV = E_{Fn} - E_{Fp}$ and the obvious relations

$$\begin{aligned} [E_C(-\epsilon) - E_{Fn}] + eV + [E_{Fp} - E_V(-\epsilon)] &\\ = E_G(-\epsilon) &\\ = E_G(+\epsilon) + \Delta + [E_V(+\epsilon) - E_V(-\epsilon)]. & \end{aligned}$$

In general, the factor $\eta \equiv e^{[E_{Fp} - E_V(+\epsilon)]/kT}$ in (16) is related to $p_{(+)}$ by the Fermi transform $p_{(+)} = N_V^{(B)} \mathfrak{F}_{1/2}(\eta)$, where

$$\mathfrak{F}_{1/2}(\eta) \equiv \frac{2}{\pi^{1/2}} \int_0^\infty \frac{x^{1/2}\, dx}{1 + \eta e^x}, \qquad \eta = \mathfrak{F}_{1/2}^{-1}\left(\frac{p_{(+)}}{N_V^{(B)}}\right) \tag{17}$$

and $\mathfrak{F}_{1/2}^{-1}$ is the inverse transformation to $\mathfrak{F}_{1/2}$. Using (15) and (17), we can bring (16) in the following general form:

$$n_{(-)} = \frac{N_C^{(E)} n_i^2}{N_C^{(B)} N_V^{(B)}} e^{-\Delta/kT} e^{eV/kT} \mathfrak{F}_{1/2}^{-1}\left(\frac{p_{(+)}}{N_V^{(B)}}\right). \tag{18}$$

Equation (18) represents a generalization of (5) for degenerate distribution of majority carriers in the base. Using (12), we can eliminate $n_{(-)}$ from (18) to obtain

$$n_{(+)} e^{\delta E_{Fn}/kT} = \frac{n_i^2}{N_V^{(B)}} e^{eV/kT} \mathfrak{F}_{1/2}^{-1}\left(\frac{p_{(+)}}{N_V^{(B)}}\right). \tag{19}$$

For a nondegenerate hole concentration, we have $\mathfrak{F}(\eta) \approx \eta^{-1}$ and (19) becomes

$$n_{(+)} p_{(+)} e^{\delta E_{Fn}/kT} = n_i^2 e^{eV/kT} \tag{20}$$

which is equivalent to (5). For a strongly degenerate case, $\eta \ll 1$, often encountered in HBT, the Fermi integral and its inverse transform are asymptotically given by

$$\mathfrak{F}_{1/2}(\eta) \approx \frac{4[-\ln \eta]^{3/2}}{3\pi^{1/2}}$$

$$\mathfrak{F}_{1/2}^{-1}\left(\frac{p_{(+)}}{N_V^{(B)}}\right) = \exp\left[-\left(\frac{3\pi^{1/2} p_{(+)}}{4N_V^{(B)}}\right)^{2/3}\right]$$

and the relation between $n_{(+)}$ and $p_{(+)}$ assumes the form

$$n_{(+)}\, e^{\delta E_{Fn}/kT} = \frac{n_i^2}{N_V^{(B)}} e^{eV/kT} e^{-(3\pi^{1/2} p_{(+)}/4N_V^{(B)})^{2/3}}. \tag{21}$$

Equation (19) is the main result of this section. Consider first the case illustrated in Fig. 2(a), when the heterointerface coincides with the doping junction and the base doping N_A is much higher than the emitter doping N_D, so that the entire drop of the electrostatic potential occurs on the emitter side of the boundary. In this case, the values of $n_{(+)}$ and $p_{(+)}$ are practically coincident with the boundary concentrations $n(0)$ and $p(0)$ of the quasi-neutral base region and can be used as the boundary condition for the drift-diffusion equation in the base. Equation (19), combined with the neutrality condition (14), thus provides an expression for the boundary concentration $n_{(+)}$ in terms of the emitter-base bias V_{EB}.

In a more general case, illustrated in Fig. 2(b), there is an additional potential drop on the base side of the heterointerface. Tracing the above derivation, it is easy to see that (16), (18), and (19) remain valid in this situation. Of course, $n_{(+)}$ no longer equals $n(0)$ and $p_{(+)} \neq p(0)$. However, we can relate these concentrations by using the well-known fact that the imref variation in the depletion region of a forward-biased p-n junction is negligible[3] (outside the small $\pm\epsilon$ layer at the heterointerface). So long as both E_{Fn} and E_{Fp} remain constant in the depletion layer on the base side of the junction, (19) retains its form for n and p at any point in that layer. Therefore, we have in general

$$n(0)\, e^{\delta E_{Fn}/kT} = \frac{n_i^2}{N_V^{(B)}}\, e^{eV/kT}\, \mathfrak{F}_{1/2}^{-1}\left(\frac{p(0)}{N_V^{(B)}}\right). \tag{22}$$

If δE_{Fn} is known in terms of the current density, then (22) together with a similar relation at the base-collector interface determines the current-voltage characteristics of the transistor, cf. Section IV.

III. Evaluation of the Imref Discontinuity at an Abrupt Heterointerface

Let us first evaluate δE_{Fn} neglecting recombination in the base. Assume a sufficiently high forward emitter-base bias and a reverse (or zero) collector-base bias. In the simplest case of the emitter-base junction as in Fig. 2(a), we have to a good accuracy, $J_n \approx -eDn_{(+)}/W$ and from (12) we find

$$\begin{aligned} \delta E_{Fn} &= kT \ln\left(1 + \frac{De^{\Delta/kT}}{Wv_R}\right) \\ &\approx kT \ln\left(1 + \frac{4 l_{sc} e^{\Delta/kT}}{3W}\right) \end{aligned} \tag{23}$$

where the approximate relation in the second line corresponds to the simplest model of collisions in the base when the scattering length l_{sc} is independent of the electron energy; in this model one has [13]

$$D = \frac{4 v_R l_{sc}}{3}. \tag{24}$$

Our very use of the concept of a quasi-Fermi level for electrons in the base relies on the validity of the diffusive transport model and hence the assumption $l_{sc} \ll W$ must hold. It is then easy to see from (23) that δE_{Fn} is negligible for low values of the discontinuity, $\Delta \lesssim kT$. On the other hand, for a practical a-HBT with $\Delta \gg kT$ the imref discontinuity cannot be neglected. Boltzmann transport studies [13] show that the diffusion approximation is quite reasonable already for $W \gtrsim 4l_{sc}$ and nearly perfect for $W \gtrsim 10l_{sc}$. Modern transistors are rarely designed with a thicker base. Taking $W/l_{sc} \sim \Delta/kT \sim 10$, we find $\delta E_{Fn} \sim 8kT$, a tangible contribution indeed.

In the more general situation, corresponding to Fig. 2(b), $n(0) = n_{(+)} e^{-e\delta\phi/kT}$ where $\delta\phi$ is the electrostatic potential drop on the base side of the junction

$$\delta\phi \equiv \frac{(\phi_0 - V)\epsilon_E N_D}{\epsilon_E N_D + \epsilon_B N_A}$$

ϕ_0 is the built-in potential, and ϵ_E, ϵ_B are the permittivities of the emitter and the base, respectively. Substituting these expressions in (12) with $J_n = -eDn(0)/W$ we find

$$\delta E_{Fn} = kT \ln\left(1 + \frac{De^{(\Delta - e\delta\phi)/kT}}{Wv_R}\right). \tag{25a}$$

To include the effect of recombination, neglected above, one has to solve the diffusion equation in the base with a recombination term. The result, derived in the next section, is

$$e^{\delta E_{Fn}/kT} = 1 + \frac{De^{(\Delta - e\delta\phi)/kT}}{L_D v_R \tanh (W/L_D)} \tag{25b}$$

where L_D is the diffusion length of electrons in the base. Equation (25b) reduces to (25a) when the diffusion length is not too short, $L_D \gg W$.

It may appear that (25) are inaccurate because δE_{Fn} does not vanish when $J_n \to 0$. However, the correction to (25), which takes care of this discrepancy, occurs at nearly infinitesimal currents, corresponding to the regime when $n_{(+)}$ approaches the equilibrium concentration n_0 in the base. As soon as the forward emitter bias is sufficiently large, $V_{EB} \gg kT$, the imref discontinuity is accurately given by (25) and in the case (25a) it is practically independent of the current.

As an aside, let us briefly digress from transistors and stress that the existence of an imref discontinuity is a general feature of forward injection "over a cliff." Consider, for example, a double-heterostructure laser whose active narrow-gap layer W is sufficiently thin that the injected carrier concentration n can be regarded as uniform. In this case, the current is given by enW/τ_e and the imref discontinuity by

$$\delta E_{Fn} = kT \ln\left(1 + \frac{We^{\Delta/kT}}{\tau_e v_R}\right)$$

where τ_e is the electron lifetime in the active layer. A similar expression holds for holes at the other heterointerface.

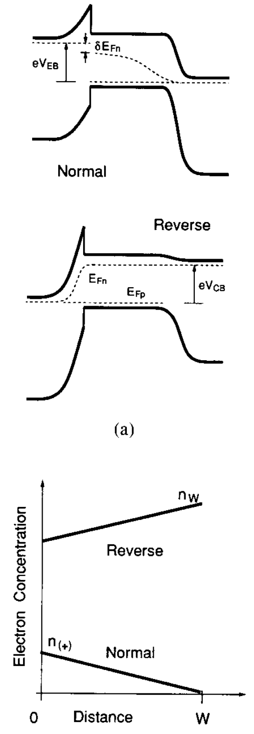
IV. Analytic Form of the Static Characteristics of an HBT

Consider a double-heterostructure transistor in which the emitter-base junction is abrupt while the collector-base junction is graded, Fig. 3. (The double heterostructure assumption is made for the sake of simplicity, as it allows us to neglect the hole component of the current.) The base bandgap and doping are assumed uniform and we shall further assume $N_A \gg N_D$. The diffusion equation in the base is of the form

$$\frac{\partial^2 n}{\partial z^2} = \frac{n - n_0}{L_D^2} \tag{26}$$

---

[3] We have already used this condition in writing down (16), where we assumed constancy of E_{Fn} on the emitter side and of E_{Fp} on the base side of the interface.



(a)

(b)

Fig. 3. Double-heterostructure transistor in the normal and reverse mode: (a) band diagram and (b) schematic electron concentration profile.

and subject to the boundary conditions $n(0) = n_{(+)}$ and $n(W) = n_W$, it has the following solution:

$$n - n_0 = \frac{[n_{(+)} - n_0]\sinh[\lambda_D(W - z)]}{\sinh(\lambda_D W)} + \frac{[n_W - n_0]\sinh(\lambda_D z)}{\sinh(\lambda_D W)} \tag{27}$$

where we have denoted $\lambda_D = 1/L_D$. Expressed in terms of the boundary concentrations, the emitter (J_E) and the collector (J_C) currents are

$$\frac{J_E}{eD} = -\frac{\lambda_D[n_{(+)} - n_0]}{\tanh(\lambda_D W)} + \frac{\lambda_D[n_W - n_0]}{\sinh(\lambda_D W)} \tag{28a}$$

$$\frac{J_C}{eD} = -\frac{\lambda_D[n_{(+)} - n_0]}{\sinh(\lambda_D W)} + \frac{\lambda_D[n_W - n_0]}{\tanh(\lambda_D W)}. \tag{28b}$$

In order to relate the boundary conditions $n_{(+)}$ and n_W to the applied voltages, we use (19) which is a general relation valid for abrupt as well as graded heterojunctions and homojunctions. For the latter two cases, of course, $\delta E_{Fn} = 0$. We shall assume that this applies to the base-collector junction[4] so that the corresponding imref factor $\exp(\delta E_{Fn}/kT)$ equals unity

$$n_W = \frac{n_i^2}{N_V^{(B)}} e^{eV_{CB}/kT}\,\mathfrak{F}_{1/2}^{-1}\left(\frac{p_W}{N_V^{(B)}}\right), \qquad p_W - n_W = p_0 - n_0. \tag{29}$$

At the base-emitter junction we have

$$n_{(+)}e^{\delta E_{Fn}/kT} = \frac{n_i^2}{N_V^{(B)}} e^{eV_{EB}/kT}\,\mathfrak{F}_{1/2}^{-1}\left(\frac{p_{(+)}}{N_V^{(B)}}\right),$$

$$p_{(+)} - n_{(+)} = p_0 - n_0 \tag{30}$$

where the imref factor[5] is given by (12) or, taking into account that $J_n(0) = J_E$ by

$$e^{\delta E_{Fn}/kT} = 1 - \frac{J_E e^{\Delta/kT}}{en_{(+)}v_R}. \tag{31}$$

Because of the nonlinearity introduced by the inverse Fermi transform, it is impossible to explicitly solve (29) and (30) for $n_{(+)}$ and n_W. Nevertheless, a close-form analytic expression of the transistor characteristics can be obtained in the general case. Inverting (28), we can express $n_{(+)}$ and n_W in terms of the currents

$$n_{(+)} - n_0 = -\frac{J_E\cosh(\lambda_D W) - J_C}{eD\lambda_D\sinh(\lambda_D W)} \tag{32a}$$

$$n_W - n_0 = -\frac{J_E - J_C\cosh(\lambda_D W)}{eD\lambda_D\sinh(\lambda_D W)}. \tag{32b}$$

Substituting (32) into (29)–(31), we obtain analytic expressions of the form

$$V_{EB} = V_{EB}(J_E, J_C) \qquad V_{CB} = V_{CB}(J_E, J_C) \tag{33}$$

for the current-voltage characteristics of a double-heterojunction bipolar transistor. These formulas, too cumbersome to put down explicitly, are valid for arbitrary injection levels and include the case of a degenerate base doping.

For low injection levels, we can replace p_W and $p_{(+)}$ in (29) and (30) by the equilibrium hole concentration p_0 and use the identity

$$\frac{n_i^2}{N_V^{(B)}}\,\mathfrak{F}_{1/2}^{-1}\left(\frac{p_0}{N_V^{(B)}}\right) \equiv n_0. \tag{34}$$

In this case it becomes possible to write down the characteristics in the "direct" form

$$\frac{J_E}{en_0D} = -\frac{\lambda_D\xi_D}{(1 + \xi_D)}\frac{e^{eV_{EB}/kT} - 1}{\tanh(\lambda_D W)} + \frac{\lambda_D\xi_D}{(1 + \xi_D)}\frac{e^{eV_{CB}/kT} - 1}{\sinh(\lambda_D W)}; \tag{35a}$$

---

[4] The case when the base-collector junction is also abrupt is treated without any additional complications.

[5] For a sufficiently strong forward emitter-base bias and a zero or reverse collector-base bias, both n_0 and n_W can be neglected in (28). In this case (28a) gives $n_{(+)} = -(J_E/eD)\tanh(\lambda_D W)/\lambda_D$, whence we obtain (25a) of the preceding section. Equation (25b) is obtained in a similar way.

$$\frac{J_C}{en_0D} = -\frac{\lambda_D\xi_D}{(1+\xi_D)}\frac{(e^{eV_{EB}/kT}-1)}{\sinh(\lambda_D W)} + \frac{\lambda_D[\tanh^2(\lambda_D W)+\xi_D]}{(1+\xi_D)}\frac{(e^{eV_{CB}/kT}-1)}{\tanh(\lambda_D W)} \tag{35b}$$

where

$$\xi_D \equiv \frac{v_R e^{-\Delta/kT}\tanh(\lambda_D W)}{D\lambda_D}. \tag{36}$$

Equations (35) have been obtained earlier [5] for nondegenerate base doping levels; we see that they retain the same form in the degenerate case.

For high-injection levels, the transistor characteristics in the "direct form"

$$J_E = J_E(V_{EB}, V_{CB}) \qquad J_C = J_C(V_{EB}, V_{CB}) \tag{37}$$

can be determined only for nondegenerate hole concentrations in the base. In this case, the substitution

$$\mathfrak{F}_{1/2}^{-1}\left(\frac{p}{N_V^{(B)}}\right) = \frac{N_V^{(B)}}{p} \tag{38}$$

simplifies (29) and (30) so that they can be solved explicitly for n_W and $n_{(+)}$

$$n_W - n_0 = -\frac{p_0}{2} + \left(\frac{p_0^2}{4} + n_i^2(e^{eV_{CB}/kT}-1)\right)^{1/2} \tag{39a}$$

$$n_{(+)} - n_0 = -\frac{1}{2}\left(p_0 - \frac{n_W - n_0}{(1+\xi_D)\cosh(\lambda_D W)}\right) + \left[\frac{1}{4}\left(p_0 + \frac{n_W - n_0}{(1+\xi_D)\cosh(\lambda_D W)}\right)^2 + \frac{n_i^2\xi_D(e^{eV_{EB}/kT}-1)}{1+\xi_D}\right]^{1/2}. \tag{39b}$$

Equations (39), when substituted in (28), yield the characteristics in the form (37).

V. Normal and Reverse-Bias Configurations

One of the common [3], [4] static characterizations of an HBT consists in the comparison of the collector currents $J_C^{(N)}$ and $J_E^{(R)}$ in the "normal" and "reverse" mode

$$J_C^{(N)} \equiv J_C(V_{EB} = V, V_{CB} = 0) \tag{40a}$$

$$J_E^{(R)} \equiv J_E(V_{EB} = 0, V_{CB} = V). \tag{40b}$$

Fig. 4 shows these characteristics, calculated from (28) and (39), valid for a nondegenerate base doping. We see that

$$J_C^{(N)}(V) = J_E^{(R)}(V) \tag{41}$$

except in the high injection limit. The fact that $J_C^{(N)}$ and $J_E^{(R)}$ should coincide at low injection levels, is evident from the exact analytic expressions (35), which remain valid at degenerate doping levels. Examining these expressions, we see that the validity of (41) is independent of the height of the heterojunction spike Δ. This means that the assertion commonly encountered in the literature that deviations from (41) are indicative of the electron flow being limited by the emitter-base junction, as opposed to the base transport, is incorrect. At high values of Δ (which is the case illustrated in Fig. 4) the electron flow is clearly limited by the emitter barrier.

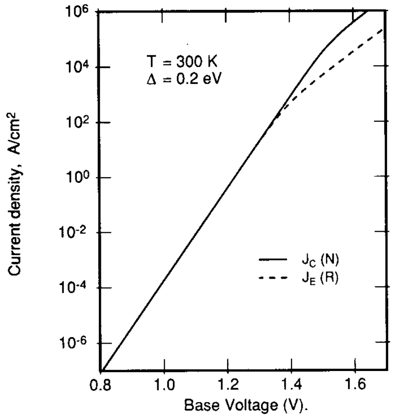


Fig. 4. Room-temperature dependences $J_C^{(N)}(V)$ [J_C as a function of V_{BE}] and $J_E^{(R)}(V)$ [J_E as a function of V_{CB}], calculated from (28) and (39) for the normal and reverse modes of operation of a double-heterojunction transistor with a graded base-collector junction and abrupt base-emitter junction. The following parameters were used: $\Delta = 0.2$ eV, $E_G^{(B)} = 1.424$ eV, $p_0 = 10^{18}$ cm^{-3}, $D = 20$ cm^2/s, $\lambda_D = 5 \times 10^{-5}$ cm, $W = 3 \times 10^{-5}$ cm, and $v_R = 10^7$ cm/s.

Experimentally, (41) is sometimes violated. Precisely why this happens is not well understood. One obvious reason may be related to different *ideality factors* for the forward and reverse diodes. For example, if the base doping N_A is not too heavy, then an appreciable fraction of the applied forward bias V_{BE} drops in the base, leading to an ideality factor

$$1 + \frac{\epsilon_E N_D}{\epsilon_B N_A}$$

as discussed already by Anderson [14] in his early paper on heterojunction diodes.[6] However, the different ideality factors imply different slopes of the characteristics $J_C^{(N)}(V)$ and $J_E^{(R)}(V)$ on the semi-logarithmic scale, whereas experimentally [4], [15] these characteristics are sometimes parallel over many orders of magnitude—displaced along the voltage axis by as much as 55 mV [15], with $J_E^{(R)} > J_C^{(N)}$ at the same bias. This discrepancy is hard to understand. Small deviations from (41) can result from an effective lowering of the barrier Δ by the junction electric field—for example, due to thermally assisted tunneling [16]. Our estimates show that in an AlGaAs/GaAs heterojunction with $N_D = 10^{18}$ cm^{-3} the effective Δ may be lowered by about 60 meV at $V_{BE} = 0$. However, similar lowering occurs also at moderate forward biases and the difference at low injection levels can hardly exceed 20 meV. Moreover, such effects are necessarily bias-dependent and must be accompanied by an ideality factor greater

---

[6] This factor readily follows from (25).

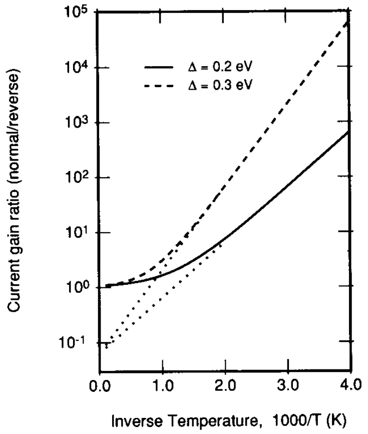


Fig. 5. Calculated [from (43)] temperature dependences of the base current ratio $J_B^{(R)}/J_B^{(N)}$ of a double-heterostructure transistor with parameters similar to those in Fig. 4, except Δ which is varied for two examples. The temperature dependence of D is assumed in accordance with (24) with an energy-independent scattering length l_{sc} = 140 Å.

than unity. In light of our results it should be really surprising if a device with *both* ideality factors close to unity would show a significant deviation from (41).

Next, consider the base currents $J_B^{(N)}$ and $J_B^{(R)}$ in the normal and reverse configurations, respectively. Let us restrict ourselves to low injection levels, where the common-emitter (common-collector) current gains $\beta^{(N)}(\beta^{(R)})$ are easily found from (35)

$$\beta^{(N)} \equiv \frac{J_C^{(N)}}{J_B^{(N)}} = \frac{1}{2\sinh^2(\lambda_D W/2)} \tag{42a}$$

$$\beta^{(R)} \equiv \frac{J_E^{(R)}}{J_B^{(R)}} = \frac{1}{2\sinh^2(\lambda_D W/2)} \cdot \frac{1}{1 + (D\lambda_D/v_R)e^{\Delta/kT}\coth(\lambda_D W/2)}. \tag{42b}$$

In light of (41) we then have

$$\frac{J_B^{(R)}}{J_B^{(N)}} = \frac{\beta^{(N)}}{\beta^{(R)}} = 1 + \frac{D\lambda_D e^{\Delta/kT}}{v_R \tanh(\lambda_D W/2)}. \tag{43}$$

For an a-HBT with $\Delta \gg kT$, (43) clearly shows that $J_B^{(R)} \gg J_B^{(N)}$. The physical origin of this is obvious: the concentration of minority carriers in the base in the reverse mode is much higher. For the same reason, the onset of high-injection condition occurs at lower current (Fig. 4).

The base current ratio is a strong function of the band discontinuity Δ. Equation (43) suggests that studies of the temperature dependence of this ratio can be used for a *measurement* of Δ. Fig. 5 shows an Arrhenius plot of $J_B^{(R)}/J_B^{(N)}$, calculated for two values of Δ. In this calculation we have assumed that the diffusivity D varies with the temperature in accordance with (24) where the scattering length l_{sc} is energy-independent. Besides the value of Δ, determined by the slope in the linear region, the intercept with the ordinate axis of the linear region, extrapolated to high temperatures, permits a determination of the prefactor in (43). For $\lambda_D W \ll 1$ (the case shown in Fig. 4) the intercept determines the ratio $2D/v_R W$ which in our model equals $8l_{sc}/3W$. These quantities are of interest to device physicists.

VI. Conclusions

Detailed analysis of the difference between phenomenological models of graded-gap and abrupt-heterojunction bipolar transistors has been carried out. It is shown that the approach of Marty *et al.* [2] can be modified without a loss of generality, so as to take into account the quasi-Fermi level discontinuity at an abrupt emitter-base interface. This permits a unified treatment of both abrupt-junction and graded-gap HBT's, as well as homojunction transistors. The model is valid for both high and low levels of injection, including the case of degenerate distribution of majority carriers in the base. Explicit analytical formulas are derived for the current-voltage characteristics.

The model is applied to the static characteristics $[I_C(V_{EB})]_{V_{CB}=0}$ in the normal mode of operation, and $[I_E(V_{CB})]_{V_{EB}=0}$ in the reverse mode. It is shown that these characteristics coincide in the low-injection limit, diverging only at high levels of injection, provided both the emitter-base and the collector-base junctions have near-unity ideality factors. It is also shown that measurements of the temperature dependence of the ratio of base currents in the reverse and forward modes can be used to determine the emitter-base conduction band discontinuity and estimate the scattering length in the base.

References

[1] H. Kroemer, "Heterostructure bipolar transistors and integrated circuits," *Proc. IEEE*, vol. 70, pp. 13-25, 1982.

[2] A. Marty, G. E. Rey, and J. P. Bailbe, "Electrical behavior of an Npn GaAlAs/GaAs heterojunction transistor," *Solid-State Electron.*, vol. 22, pp. 549-557, 1979.

[3] S. Tiwari, *Compound Semiconductor Device Physics.* San Diego, CA: Academic Press, 1992, ch. 7.

[4] P. M. Asbeck, "Bipolar transistors," in *High-Speed Semiconductor Devices*, S. M. Sze, Ed. New York: Wiley-Interscience, 1990, ch. 6, pp. 358-366.

[5] A. A. Grinberg, M. S. Shur, R. J. Fischer, and H. Morkoç, "An investigation of the effect of graded layers and tunneling on the performance of AlGaAs/GaAs heterojunction bipolar transistors," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 1758-1765, 1984.

[6] S.-C. Lee and H.-H. Lin, "Transport theory of the double heterojunction bipolar transistor based on current balancing concept," *J. Appl. Phys.*, vol. 59, pp. 1688-1695, 1986.

[7] M. S. Lundstrom, "An Ebers-Moll model for the heterostructure bipolar transistor," *Solid-State Electron.*, vol. 29, pp. 1173-1179, 1986.

[8] B. R. Ryum and I. M. Abdel-Motaleb, "A Gummel-Poon model for abrupt and graded heterojunction bipolar transistors (HBTs)," *Solid-State Electron.*, vol. 33, pp. 869-880, 1990.

[9] A. F. J. Levi, B. Jalali, R. N. Nottenburg, and A. Y. Cho, "Vertical scaling in heterojunction bipolar transistors with nonequilibrium base transport," *Appl. Phys. Lett.*, vol. 60, pp. 460-462, 1992.

[10] C. D. Parikh and F. A. Lindholm, "A new charge-control model for single- and double-heterojunction bipolar transistors," *IEEE Trans. Electron Devices*, vol. 39, pp. 1303-1311, 1992.

[11] A. A. Grinberg, "Thermionic emission in heterosystems with different effective electronic masses," *Phys. Rev. B*, pp. 7256-7258, 1986.

[12] A. A. Grinberg and S. Luryi, "Dynamic Early effect in heterojunction bipolar transistors," to be published.

[13] A. A. Grinberg and S. Luryi, "Diffusion in a short base," *Solid-State Electron.*, vol. 35, pp. 1299-1310, 1992.

[14] R. L. Anderson, "Experiments on Ge-GaAs heterojunctions," *Solid-State Electron.*, vol. 5, pp. 341-351, 1962.

[15] S. Tiwari, S. L. Wright, and A. W. Kleinsasser, "Transport and related properties of (Ga,Al)As/GaAs double heterostructure bipolar junction transistors," *IEEE Trans. Electron Devices*, vol. ED-34, pp. 185-197, 1984.

[16] E. H. Rhoderick, *Metal-Semiconductor Contacts.* Oxford, UK: Clarendon Press, 1980.

Anatoly A. Grinberg, photograph and biography not available at the time of publication.

Serge Luryi (M'81–SM'85–F'89) received the M.Sc. and Ph.D. degrees in theoretical physics from the University of Toronto, Toronto, Ont., Canada, in 1975 and 1978, respectively.

Since 1980, he has been with AT&T Bell Laboratories, Murray Hill, NJ, where he is currently a Distinguished Member of Technical Staff in Compound Semiconductor Device Research Laboratory. His main research interests are in the physics of exploratory semiconductor devices.