A Theoretical Toolbox for the Simulation and Design of HBTs Constructed in the $Al_xGa_{1-x}As$ and $Si_{1-x}Ge_x$ Material Systems

by

Shawn Searles, P.Eng.

B.Sc.E.E., The University of Manitoba, 1987
M.Eng., Carleton University, 1989

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF ELECTRICAL ENGINEERING

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

July 12, 1995

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Electrical Engineering

The University of British Columbia
Vancouver, Canada

Date July 12, 1995

Abstract

A theoretical toolbox for the simulation of Heterojunction Bipolar Transistors (HBTs), including the effects of tunneling, recombination, and the optimum non-linear base profile (for the minimisation of the base transit time), is developed. The models developed are applicable to a general material system, and are analytic. Extensions specifically required by the complex $Si_{1-x}Ge_x$ material system are also developed. The optimum (to minimise base transit time) base doping is found to be non-exponential, and the optimum base bandgap grading is not linear. A general transport model for HBTs, including recombination processes, is developed that accounts for the complex nature of charge transport throughout the entire device. Unique methods for optimising HBT metrics, which cannot be employed for Bipolar Junction Transistors (BJTs), are also presented. A description of charge transport within the emitter-base Space-Charge Region (SCR), which accounts for tunneling and is not beholden to the usual drift-diffusion analysis, is developed. The implications of having different electron effective masses in the two sides of the heterojunction, leading to what is termed a mass boundary, is fully explored. It is found that the tunneling of electrons within the emitter-base SCR leads to a non-Maxwellian minority-particle ensemble distribution entering the neutral base. Finally, transport within SiGe HBTs is considered, with all of the relevant material models presented and multi-band transport models developed. This treatment leads to a variety of interesting conclusions regarding the operation of present-day SiGe HBTs and possible future designs.

Table of Contents

List of Tables

List of Figures

Acknowledgment

I would like to thank first of all, Professor Dave L. Pulfrey, my Ph.D. supervisor. I returned from industry to obtain my Ph.D. because I was interested in performing research that probed into the complex theories of solid-state device operation. Thanks to Dr. Pulfrey and his forthcoming guidance, I was able to navigate a steady course through the often turbulent waters of academic research, and attain the research goals I had planned to explore. Dr. Pulfrey provided constant encouragement to my work, offered valuable assistance, and provided me a learning experience that I know will serve me for the rest of my life. Dr. Pulfrey, however, went even further in his contributions during the time I worked on my Ph.D. He allowed and encouraged me to pursue other life interests, so that I can proudly say that my Ph.D. research was indeed a time that touched and enriched all aspects of my life. So to you Dr. Pulfrey I can only offer in return my simple but sincerest thanks.

I would also like to thank Professor Mike Jackson who provided me with many ideas throughout my Ph.D. research. Without the presence of Dr. Jackson, my Ph.D. research would not have been as interesting nor as fulfilling as is has been. I would also like to thank Professor Tom Tiedje, who offered me an excellent course in solid-state quantum mechanics; without which I could not have performed the Ph.D. research that I have done. Dr. Tiedje, you have challenged me and as a result, provided me a fundamental base from which I will solve many questions yet to come. I would also like to thank Dr. Jackson, Dr. Tiedje, Professor Nick Jaeger, Professor Matt Yedlin, Professor Jeff Young, and Professor Fred Lindholm, whose presence on my examining committees helped to ensure that my final thesis was the best it could possibly be.

Finally, I would like to thank Barbara Ippen, my wife to be on July 29th, for being a willing partner in my Ph.D. research efforts. Your caring presence has provided me a reference point that I could always count upon, no matter how hard things became during the course of my research goals.

CHAPTER 1

Introduction

The main objective of the Ph.D. research being presented in this thesis is the creation of models that will foster a deeper understanding regarding the physics surrounding a Heterojunction Bipolar Transistor (HBT). To this end, physically based models for the transport of charge within an HBT will be developed. These physics-based models will allow for the simulation of present-day HBT structures and novel structures for the future. By clearly identifying the relevant mechanisms by which charge transport takes place within the HBT, an optimum design for the device that incorporates the various compromises between competing device metrics (such as β, f_T, and R_B) can be obtained. A further goal is to reduce all of the models developed within this thesis to tractable, analytic forms. By obtaining analytic models for charge transport within the HBT, circuit level models that predict device performance can be developed in step with the emergence of HBT-based Integrated Circuit (IC) processes. Finally, the models that are developed within this thesis are in general free of any details specific to a single material system. However, given the importance of the $Al_xGa_{1-x}As$ and $Si_{1-x}Ge_x$ material systems, these two systems will be extensively studied and will serve as the chosen material systems for all examples presented.

The concept behind the HBT has been around since the time of Shockley [1]. Further, over 30 years ago, Kroemer developed much of the fundamental physics regarding the operation of the HBT [2]. However, it has not been until the last five years that industry has had the capability to manufacture HBTs with suitable yields to be commercially viable [3-5]. Also, the material research is still continuing and has a long way to go before HBT processes achieve the maturity of technologies such as CMOS. Furthermore, with experimental results becoming more prolific, and with rapidly diminishing device dimensions, we are finding that much of the physics laid down for modelling the HBT is inadequate for describing present-day devices [6-9].

With the increasing maturity of processes for the production of HBTs, comes an increase in the need for models that predict device operation. It is now possible to manufacture HBTs with active basewidths approaching 100 Å [10-12] and with features that change over distances of less than 10 Å [13-14]. As device dimensions approach the atomic lattice spacing of the crystal, the applicability of models based upon classical continuous fields becomes questionable [15]. There is already general agreement that one must consider higher order moments beyond the drift and diffusion terms in the Boltzmann Transport Equation (BTE) in order to model deep submicron devices [16-17]. The BTE is based upon classical physical models that in general do not incorporate

quantum mechanical (QM) phenomena. It has been recognised that the correct modelling of tunneling, a QM effect, is of paramount importance to the correct prediction of HBT operation [18-21]. Thus, models of HBTs that incorporate QM phenomena are becoming increasingly important in order to maintain accurate simulation of the HBT.

The general relationship between the terminal currents and voltages of an HBT can still be predicted today by models designed for Bipolar Junction Transistors (BJTs) [22]. However, it is not always clear why we can continue to apply BJT models to HBT operation when these BJT models were developed without consideration of the physical processes that govern transport within an HBT. Presumably, the BJT model has enough degrees of freedom so that it can be manipulated to cover HBT operation. For example, one of the most common discrepancies found when using BJT models for HBT simulation is that the injection indices (ideality factors) for the collector and base terminal currents do not correspond to what is theoretically predicted for BJT operation [23]. Thus, in order to accurately predict HBT operation, and to further develop HBT processes so as to advance device operation, one needs to understand such things as why the collector and base injection indices differ between an HBT and a BJT [24,25].

The $Si_{1-x}Ge_x$ material system has many unique physical considerations that other systems, such as the $Al_xGa_{1-x}As$ material system, do not have to contend with. The unique attributes of the $Si_{1-x}Ge_x$ material system are mostly due the effects of strain. Due to the large lattice mismatch between Si and Ge, $Si_{1-x}Ge_x$ films grown on top of $Si_{1-y}Ge_y$ substrates (where $x \neq y$) have a large degree of strain present within them if non-relaxed crystals with low defect density are to be manufactured. The presence of strain breaks the cubic symmetry of the crystal and changes the bulk electrical properties [26-28] of the film. By varying the Ge alloy content and the strain imparted to the SiGe film, it is possible to tailor both the bandgap and the offsets in the conduction and valence bands. Therefore, models specific to the $Si_{1-x}Ge_x$ material system must be developed in order to understand charge transport within the complex band structure that develops.

Finally, the reason for focussing on the $Si_{1-x}Ge_x$ and $Al_xGa_{1-x}As$ material systems stems from the maturity of AlGaAs devices, and the massive installed base of Si-based IC technologies that would easily admit SiGe devices. From a manufacturing standpoint the $Al_xGa_{1-x}As$ material system offers no redeeming features when compared to Si, save one - the lack of strain. Obviously, the key to the operation of an HBT is the formation of heterojunctions between two materials

characterised by different bandgaps. The $Al_xGa_{1-x}As$ material system has essentially a fixed lattice constant over the entire range of Al mole fraction x. For this reason, the $Al_xGa_{1-x}As$ material system is lattice matched and will admit an arbitrary heterojunction between $Al_xGa_{1-x}As$ and $Al_yGa_{1-y}As$ without developing a strain within one of the films. This lack of strain within the $Al_xGa_{1-x}As$ material system helps to ensure a defect-free heterointerface that greatly facilitates the manufacture of HBTs. For this reason, most commercially available HBTs are based in the $Al_xGa_{1-x}As$ material system [29]. However, most solid-state devices are Si based [30]. With the advancement of low-temperature Chemical Vapour Deposition (CVD) processing [31], the formation of high-quality commensurately strained $Si_{1-x}Ge_x$ films is becoming commercially viable. Therefore, given the manufacturing advantages of Si, it is expected that SiGe HBTs will shortly surpass AlGaAs HBTs as the most prolific commercially available HBT [32-37].

1.1 Modelling Details

Research has been conducted into the injection of electrons from the emitter into the base of $Al_xGa_{1-x}As$ npn HBTs [18,24,25]. The research has centred around abrupt HBTs where the heterojunction between the wide-energy-gap emitter and the narrow-energy-gap base is abrupt. In an abrupt $Al_xGa_{1-x}As$ HBT one finds the formation of a Conduction-Band Spike (CBS) between the emitter and the base (see Fig. 3.1). This spike, due ostensibly to differences in the electron affinity of the materials used for the formation of the emitter and the base, results in a large impediment to the flow of electrons from the emitter into the base. In fact, if the CBS were not taken into account when modelling the HBT, the collector current would be overestimated by over three orders of magnitude at room temperature (see Fig. 1.1). However, the modelling of charge transport through the CBS cannot be based upon simple thermionic injection alone. Since the width of the CBS is typically less than 100Å near the top of the spike, the occurrence of a tunneling current cannot be neglected. Finally, it will be shown that transport through the CBS can often be the limiting factor for the overall transport of charge within the HBT (*i.e.*, the determination of the collector current I_C). This occurrence of current-limited flow outside of the neutral base region will be studied and exploited for device optimisation. Therefore, the modelling of the relevant physical phenomena surrounding charge transport through the CBS, including tunneling and conservation of transverse momentum across the heterojunction in a diagonal mass tensor, will be investigated.

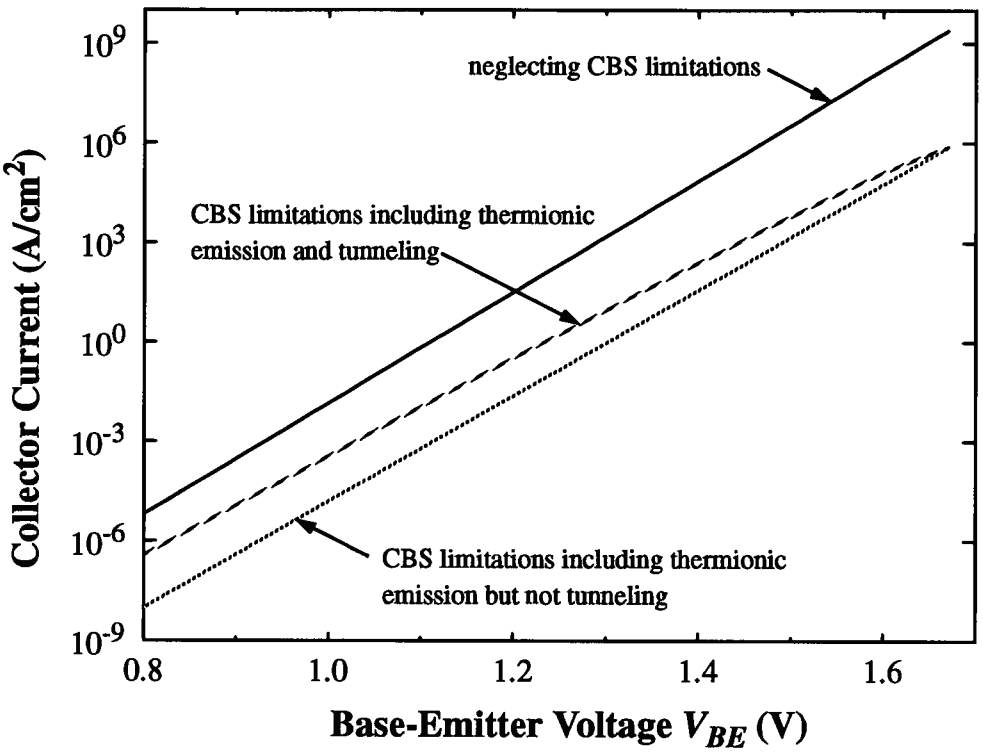


Fig. 1.1. Collector current for an abrupt AlGaAs HBT with 30% Al content in the emitter. The emitter doping is $5\times10^{17}cm^{-3}$, and the base doping is $1\times10^{19}cm^{-3}$ (see Section 4.5 for the complete device details). The top curve, where CBS limitations have been neglected, is arrived at by assuming Shockley boundary conditions and considering only neutral base transport.

The possibility of regions other than the neutral base controlling I_C is intriguing. However, from a modelling perspective, the immediate consequence of a multi-regional system controlling I_C is the question of how to join these various regions together to form one cohesive transport model. Furthermore, the possibility exists that under multi-regional control of I_C, older models, such as those for the neutral base [38], which assume that only the specific region being studied controls I_C, may not longer be valid. It will be shown in Chapter 2 that there is a very simple prescription for joining up all of the multi-regional transport models into a complete transport model for the determination of I_C. It will be further demonstrated in Chapter 6 that it is possible for two spatially separate regions to control I_C simultaneously by having essentially identical net-charge-transport capacity through both regions; the ramification of this is the inseparability of the two regions.

With the general model of Chapter 2 providing the overall method to link the various physical regions of the HBT together, then the problem of modelling charge transport within the entire HBT is effectively decoupled into a set of models; one model for each relevant region. To this end, Chapter 3 investigates and develops models for the various regions of the HBT, including the si-

multaneous optimisation of the base bandgap and doping profile (provisions are also made for the inclusion of bandgap narrowing due to heavy doping effects) for the minimisation of the base-transit time τ_B. Finally, the modelling of recombination events, which lead to the formation of the base current I_B, is developed in Chapter 5 with the specific attributes of a heterojunction included. These various regional models essentially form a toolbox for the study of charge transport within the HBT, with the general transport model of Chapter 2 forming the blueprint for the ultimate operation of the device.

The modelling efforts presented in this thesis regarding charge transport through the EB SCR are rigorous in that no appeal has been made to drift-diffusion analysis based upon phenomenological mobility models (*i.e.,* mobility models with an electric field dependency). Instead, models that include the quantum mechanics of charge transport, which have no appeal to said phenomenological mobility models, are analytically solved for. However, the neutral base charge transport models are based upon drift-diffusion analysis. The reason for resorting to simpler drift-diffusion analysis for the neutral base is its been found that the neutral base often does not represent the bottleneck to charge transport and thus does not dictate control over I_C ([25] and Fig. 1.1). Nevertheless, as the neutral base thickness approaches and becomes smaller than the mean free path, then a majority of the electrons will traverse the base without thermalising [39,40]. These un-thermalised, or hot, or ballistic electrons do not follow exactly the simple models of drift-diffusion contained within the BTE [16,41]. Instead, a general solution to the BTE is necessitated.

In present-day HBTs, and even in some of the emerging high performance BJTs, the understanding of hot electrons can be essential to the accurate modelling of the device's terminal characteristics [9,14]. The problem with general BTE solvers, such as Monte Carlo simulation, is that some important QM effects cannot be modelled. The BTE is based upon local potentials and therefore cannot include some QM effects, such as tunneling, which are inherently non-local. As was discussed and shown in Fig. 1.1, the failure to include tunneling results in a gross error regarding the transport of charge through the HBT. Section 4.6 will address the issue of merging classical BTE solvers with the models developed in Chapter 4 for charge transport through the CBS. Specifically, Section 4.6 will show that tunneling produces a considerable distortion to the minority-particle ensemble distribution entering the neutral base (deviations that are far from

Maxwellian or even hemi-Maxwellian). Finally, it should be noted that the use of drift-diffusion models in the neutral base will not produce gross errors like the failure to include tunneling through the CBS. Instead, drift-diffusion models can be employed in the neutral base, but with corrections that essentially amount to a 20 to 40% change to the diffusion coefficient D_n [42,43]. Even more importantly, if the neutral base does not control I_C, then in terms of D.C. calculations, no error will occur if these ballistic corrections to D_n are neglected; however, it terms of A.C. calculations, such as for τ_B, there would be an error.

The final modelling effort of this thesis pertains directly to the design and simulation of SiGe HBTs. As has been alluded to, the effect of strain on the electrical characteristics of $Si_{1-x}Ge_x$ films is dramatic. Chapter 6 reviews the various material models necessary for the description and study of the electrical characteristics of strained $Si_{1-x}Ge_x$. Specifically, once a review of the literature regarding the $Si_{1-x}Ge_x$ material models is presented, a comparison to experimental results is performed, and the most consistent set of material constants selected. The final result is a complete set of models for the calculation of the bandgap including conduction and valence band offsets. Furthermore, strained $Si_{1-x}Ge_x$ results in a two-band system both for the conduction and the valence band. Chapter 6 uses the $Si_{1-x}Ge_x$ material models and derives the necessary multi-band charge-transport models that are required to simulate SiGe HBTs. In fact, it is found that there is a substantial error incurred by replacing the two-band system with a single effective band. Finally, the charge-transport models are applied to the study of present-day as well as future SiGe HBT designs with some surprising results regarding operating voltages and critical layer thicknesses.

1.2 Thesis Organisation

This thesis is organised into five main chapters. Chapter 2 presents a general model for the HBT that is highly abstract in nature. The main tenet of the general model in Chapter 2 is that it can contain any number of physical regions to model the HBT, including sources and sinks within each region. Chapter 2 also introduces a method of optimisation through what is termed current-limited flow. Chapter 3 builds upon the ideas of Chapter 2 by considering specific examples of device optimisation that can be performed within an HBT but not a BJT. The main development in Chapter 3 is the solution for the optimum base bandgap and doping profile. Surprisingly, the optimum doping profile is not exponential, and the optimum base bandgap is not linear. Chapter 4

moves on to develop the necessary models for charge transport within the emitter-base SCR. Specifically, models for the tunneling of electrons through the CBS, including the effect of a spatially non-uniform effective mass, are developed. Finally, Chapter 4 goes on to show the effect of tunneling on the emerging minority-carrier ensemble distribution entering the neutral base. Chapter 5 rounds out the ideas presented in Chapter 2 by developing the necessary models for the recombination of minority carriers within the emitter-base SCR and the neutral base. Chapter 5 concludes by using the model of Chapter 2 to bring together the various regional models of Chapters 3 through 5 for the simulation of an AlGaAs HBT. Chapter 6 builds upon the models of Chapters 4 and 5 for the simulation of SiGe HBTs. Models that include the effects of strain on the conduction and valence bands in the $Si_{1-x}Ge_x$ material system are presented. Multi-band charge transport models, which include the material models of the $Si_{1-x}Ge_x$ material system, are then developed. Finally, Chapter 6 brings all of the models developed within the chapter together for the study of numerous present-day and future SiGe HBT designs.

CHAPTER 2

A Multi-Regional Model for HBTs Leading to Optimisation by Current-Limited Flow

Since the invention of the Bipolar Junction Transistor (BJT) in 1948 by Brattain, Bardeen and Shockley [44], continuous improvements have been made to its operation and reliability. Nowadays, BJTs are part of nearly every manufactured product sold within the world. This continuous development of the design and manufacture of the BJT shows no sign of ending nor any abating in the pace at which improvements are made. The question then, is what direction or directions will the course of BJT development take in the future?

The latest innovation in the evolution of the BJT has been termed Bandgap Engineering by Capasso [45]. By altering the actual semiconductor within the active portion of the BJT, generally by forming some sort of alloy, the shape of the bandgap can be altered to provide another force to govern the motion of electrons with the device. This idea, however, is not a new one. Shockley alluded to the use of Bandgap Engineering in his BJT patent of 1948 [1], and Kroemer first proposed the idea of using a wide-bandgap semiconductor for the emitter and a narrow-bandgap semiconductor for the base in 1957 [2]. This junction between two semiconductors with dissimilar bandgaps is a heterojunction, and leads to the creation of a Hetero-junction Bipolar Transistor (HBT). What makes the HBT of specific interest today, is that in 1957 it was not possible to manufacture HBTs due to the infancy of the art of semiconductor manufacture. It has only been in the late 1980's and the 1990's that commercially available HBTs have become feasible. Therefore, now is the time to fully explore the possibilities afforded by Bandgap Engineering to the continued development of the BJT.

2.1 Bandgap Engineering

The force acting upon an electron/hole within a semiconductor is the sum of the electric field due to any spatially varying charge, and the field of a spatially varying conduction/valence band (E_c/E_v) [7]. The electric field due to the spatially non-uniform charge is the standard force responsible for drift and it changes with applied bias. However, the effect of the field due to the variation of E_c/E_v is present from the construction of the device and is therefore ostensibly independent of the bias conditions (much the same as the electric field that is generated in the neutral base due to a spatially varying doping is independent of bias). It is this manufactured driving force, due to the spatial change in the bandgap and the band alignments, that gives rise to Bandgap Engineering. It is possible to effect such a rapid change in E_c/E_v that the affects of the standard

electric field are negligible and unimportant. One can therefore expect to create HBTs with markedly different terminal characteristics than those possible with BJTs. Finally, and most importantly, the terminal characteristics of HBTs can have a completely different dependence upon the physical construction of the device when compared to BJTs.

The final objective of Bandgap Engineering can be broken down into two distinct groups: techniques that provide for a slow change in E_c/E_v such that the overall electric field is modified (such as adding a gradient to E_c/E_v that aids in the transport of charge through the base) but is not overwhelmed by the engineered field; or techniques that afford extremely rapid or abrupt changes in E_c/E_v, so much so that electron/hole transport no longer depends upon the electric field due to the space-charge but is governed completely by the engineered bandgap.

The first group of Bandgap Engineering techniques was applied to the newly emerging HBT in the form of an additional adding field in the base and the collector, in order to afford a more rapid transit of the electron/hole through the device [2,46,47]. Shortly thereafter, the second group of Bandgap Engineering techniques resulted in the idea of placing an abrupt downwards change in E_c to provide a sudden increase in the kinetic energy to the electron as it entered the base (ballistic injection; see Fig. 2.1) [12,14,48]. The aiding field in the base produced results that were expected; the ballistic launcher, however, did not. In the end, it was the abrupt Bandgap Engineering technique that provided the most unique results in HBTs when compared to BJTs. Thus, abrupt Bandgap Engineering may be the more promising road to follow in seeking to continue the evolution of BJTs.

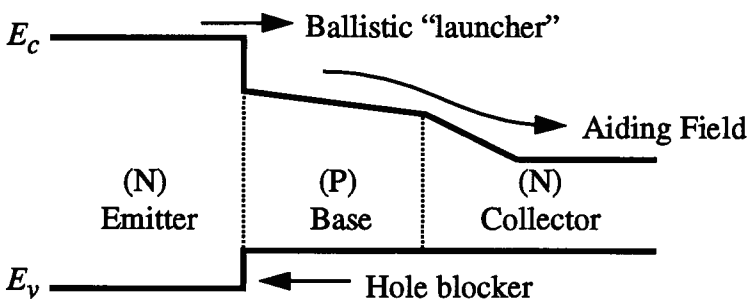


Fig. 2.1. The abrupt change of E_c in the emitter-base junction "launches" electrons into the base with a large kinetic energy. The gradual negative slope of E_c in the base and the collector helps to speed the electron through these regions. Finally, the abrupt change in E_v at the emitter-base junction suppresses hole back-injection into the emitter.

2.2 Regional Decoupling and Current-Limited Flow

Within any region of a solid-state device, charge flow or transport results in a spatial variation to the quasi-Fermi energy E_f. When the variation in the conduction and valence band is small (small, as defined by Berz [49], is a change of less than kT over one mean-free path λ), then one can speak of a continuous spatial change in E_f, and arrive at the standard drift-diffusion transport equations. However, when the change in the conduction or valence bands is not small, as occurs in abrupt Bandgap Engineering, then E_f does not vary in a continuous fashion but instead changes abruptly as well [50,7]. This abrupt change in E_f is due to a departure from conditions of quasi-equilibrium, where the transported current through the region is large in comparison to the equilibrium charge flows that result from the drift and diffusion of carriers [50,18].

To see the effect of a departure from quasi-equilibrium upon E_f, examine the effects due to an abrupt change in E_c, as shown in Fig. 2.1. Fig. 2.2 shows what the abrupt emitter-base heterojunction would look like, including the effect of the potential energy variation due to the Space-Charge-Region (SCR). The transport flux F is then given by the forward directed flux F_f minus the backward directed flux F_r. The forward and reverse directed fluxes are [18,20,21]:

$$F_f = q\upsilon n^{0^-} \quad \text{and} \quad F_r = q\upsilon n^{0^+} = q\upsilon n^{0^-} e^{-\frac{\Delta E_{fn}}{kT}} \tag{2.1}$$

which produces

$$F = F_f - F_r = q\upsilon n^{0^-}\left(1 - e^{-\frac{\Delta E_{fn}}{kT}}\right) = F_f\left(1 - e^{-\frac{\Delta E_{fn}}{kT}}\right), \tag{2.2}$$

where n^{0^-} is the electron concentration immediately to the left of the heterojunction, n^{0^+} is the electron concentration immediately to the right of the heterojunction that is capable of surmounting the barrier ΔE_c, υ is the ensemble average velocity of the flux (which can include tunneling), and ΔE_{fn} is the abrupt change in the electron quasi-Fermi-energy E_{fn}. The reason for the appearance of the term ΔE_{fn} in eqns (2.1) and (2.2) is due to the need for n^{0^+} to surmount the barrier ΔE_c. Therefore, the abrupt change in E_c generates the abrupt change in E_{fn}.

Eqn (2.2) clearly shows that as F goes towards zero, then so does ΔE_{fn}. In fact, if the conditions $F \ll F_f$ and $F \ll F_r$ are satisfied, then $\Delta E_{fn} \approx 0$. This is exactly what is meant by quasi-equilibrium; as long as the total transport current merely perturbs the equilibrium fluxes, the result will be a vanishingly small ΔE_{fn}. Conversely, if the transport current is not small compared to F_f and

F_r, then ΔE_{fn} will become substantial. Finally, in the limit of a large ΔE_{fn} (more than a few kT), F_r becomes very small compared to F_f, and $F \approx F_f$. Thus, it is not possible for the demanded transport current to exceed the available forward directed flux.

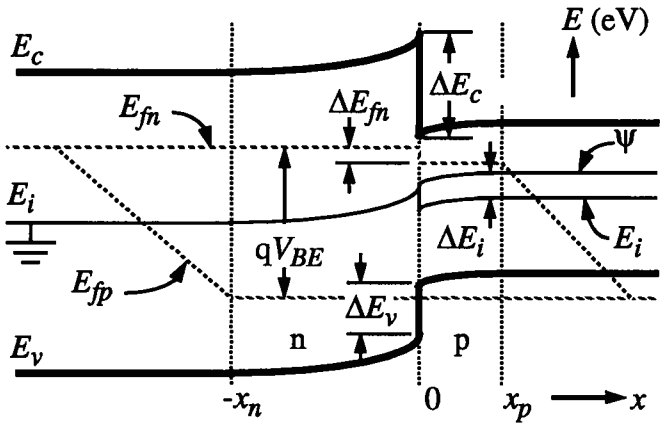


Fig. 2.2. Band diagram of the emitter-base junction showing the effect of the abrupt heterojunction on E_{fn} under an applied forward bias. ψ is the solution to the Poisson equation and is therefore continuous; however, the midgap energy E_i need not be.

The condition of $F \approx F_f$ is termed current-limited flow, and is a manifestation whereby quasi-equilibrium is grossly violated. The region in which current limiting has occurred responds by generating as large a ΔE_{fn} as necessary such as to reduce the demanded F to be no more than F_f. Obviously, the transport current through the entire device will be governed by the region in which current limiting has occurred. Furthermore, the physical construction of the region limiting the transport current will dictate the dependence of F_f, and thus F, on the applied bias. Therefore, abrupt Bandgap Engineering techniques can in principle generate regions which will govern the total transport current irrespective of any other physical portion of the device.

To examine the effects of current limiting by a region, consider the hypothetical structure shown in Fig. 2.3. Fig. 2.3 shows three different but adjoining regions with a total applied bias of V across them all. Charge is transported from Region 1 to Region 2 and finally through Region 3. Let the transport current be composed of electrons, although the same argument and solution results if holes are considered instead. To further generalise this picture consider a sink, or recombination process, existing in both Regions 2 and 3. Then, by the need to conserve particle flow, the electron flow must be continuous across the two boundaries separating the three regions. This

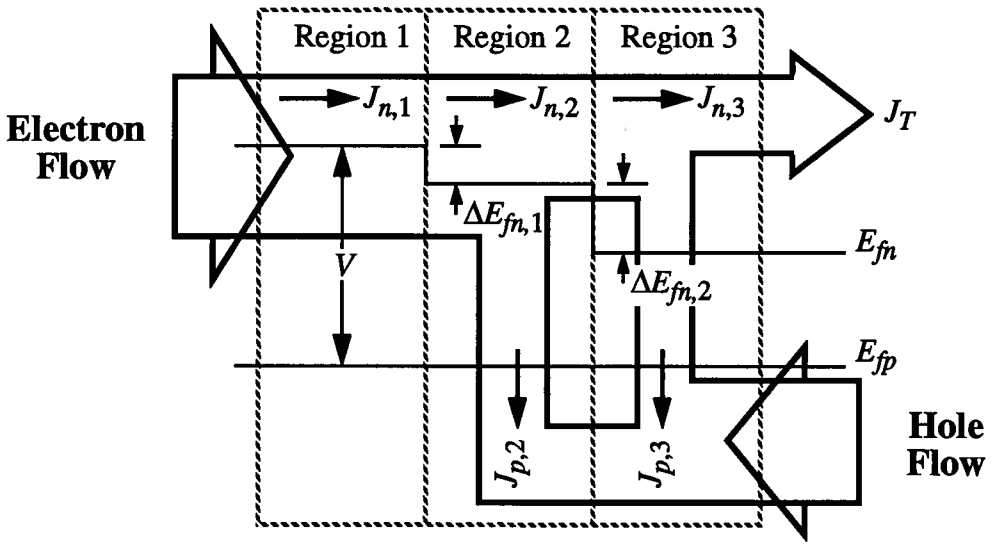


Fig. 2.3. Hypothetical HBT structure showing three physical regions that govern current transport. The applied bias is V, with a drop of $\Delta E_{fn,1}$ and $\Delta E_{fn,2}$ at the region boundaries. There are recombination processes in Regions 2 and 3 that generate currents $J_{p,2}$ and $J_{p,3}$ respectively. Conservation of current forces $J_{n,1} = J_{n,2} + J_{p,2}$, and $J_{n,2} = J_{n,3} + J_{p,3}$. Note: E_{fp} is assumed to be a constant.

procedure has been referred to as current balancing [51,52], but is generalised here to also allow for sinks (and with a simple extension, sources as well). Thus, the sink causes the electron and hole currents emanating from the region to couple together as the total electron flux entering the region must be conserved [24]. Now, the driving force in Region 1 is the full applied bias of V. However, at the boundaries, one needs to consider a drop of $\Delta E_{fn,x}$ (where $x = 1$ or 2) through the region. Thus, the driving force in Region 2 is not V but $V - \Delta E_{fn,1}$. Likewise, at the second boundary, another drop in the electron quasi-Fermi energy of $\Delta E_{fn,2}$ occurs, resulting in a driving force of $V - \Delta E_{fn,1} - \Delta E_{fn,2}$ in Region 3. Using the form given in eqn (2.2) for the transport current:

$$
\begin{aligned}
J_{n,1} &= J^0_{n,1}(V)\left(1 - e^{-\frac{\Delta E_{fn,1}}{kT}}\right), \\
J_{n,2} &= J^0_{n,2}(V - \Delta E_{fn,1})\left(1 - e^{-\frac{\Delta E_{fn,2}}{kT}}\right), \\
J_{n,3} &= J^0_{n,3}(V - \Delta E_{fn,1} - \Delta E_{fn,2}), \\
J_{p,2} &= J^0_{p,2}(V - \Delta E_{fn,1})\left(1 - e^{-\frac{\Delta E_{fn,2}}{kT}}\right), \\
J_{p,3} &= J^0_{p,3}(V - \Delta E_{fn,1} - \Delta E_{fn,2}).
\end{aligned}
\tag{2.3}
$$

It is important to realise that the hole currents $J_{p,x}$ represent electrons that have recombined; hence their direction of flow as presented in Fig. 2.3 and their connection with $\Delta E_{fn,x}$.

If the $J^0(V - \Delta E_{fn,x})$ functions can be expressed as $J^0(V)\exp(-\Delta E_{fn,x}/kT)$, then, equating $J_{n,2}$ with $J_{n,3} + J_{p,3}$ gives:

$$J^0_{n,2}(V)\left(1 - e^{-\frac{\Delta E_{fn,2}}{kT}}\right) = [J^0_{n,3}(V) + J^0_{p,3}(V)]\, e^{-\frac{\Delta E_{fn,2}}{kT}},$$

which produces, after dropping the explicit dependence upon V,

$$e^{-\frac{\Delta E_{fn,2}}{kT}} = \frac{J^0_{n,2}}{J^0_{n,2} + J^0_{n,3} + J^0_{p,3}}. \tag{2.4}$$

Then, equating $J_{n,1}$ with $J_{n,2} + J_{p,2}$ gives:

$$J^0_{n,1}(V)\left(1 - e^{-\frac{\Delta E_{fn,1}}{kT}}\right) = [J^0_{n,2}(V) + J^0_{p,2}(V)]\, e^{-\frac{\Delta E_{fn,1}}{kT}}\left(1 - e^{-\frac{\Delta E_{fn,2}}{kT}}\right).$$

Using eqn (2.4) in the above, and once again dropping the explicit dependence upon V, produces:

$$e^{-\frac{\Delta E_{fn,1}}{kT}} = \frac{J^0_{n,1}(J^0_{n,2} + J^0_{n,3} + J^0_{p,3})}{(J^0_{n,2} + J^0_{p,2})(J^0_{n,3} + J^0_{p,3}) + J^0_{n,1}(J^0_{n,2} + J^0_{n,3} + J^0_{p,3})}. \tag{2.5}$$

The final transport current J_T exiting the device is simply equal to $J_{n,3}$. Substituting eqn (2.4) and (2.5) into $J_{n,3}$ given in eqn (2.3) produces:

$$J_T(V) = J_{n,3}(V) = \frac{1}{\frac{\gamma_2\gamma_3}{J^0_{n,1}(V)} + \frac{\gamma_3}{J^0_{n,2}(V)} + \frac{1}{J^0_{n,3}(V)}}, \tag{2.6}$$

where

$$\gamma_2 = \frac{J^0_{n,2} + J^0_{p,2}}{J^0_{n,2}} \qquad \text{and} \qquad \gamma_3 = \frac{J^0_{n,3} + J^0_{p,3}}{J^0_{n,3}}.$$

Eqn (2.6) provides a very simple form for the ultimate transport current J_T emanating from the device, and extends eqn (34) in [52]. It includes all of the recombination effects of Regions 2 and 3, while allowing for a completely general relationship between the applied bias and the forward directed flux F_f (where the $J^0(V)$ functions are F_f). The only stipulation placed upon the use of eqn (2.6) is that $J^0(V - \Delta E_{fn,x}) = J^0(V)\exp(-\Delta E_{fn,x}/kT)$ (as will be seen in later chapters, where eqn (2.6) is applied, this is exactly the functional form that results). Therefore, to determine the trans-

port current that results from coupling three regions together, it is sufficient to calculate the forward directed fluxes through each region in isolation, and then use these results directly in eqn (2.6).

It is a very simple mathematical problem to generalise eqn (2.6) to a system of N regions. To do this, simply treat Regions 1, 2 and 3 as a single super-region, with the transport current given by eqn (2.6) used to define $J_{n,1}$; Regions 4 and 5 then become Regions 2 and 3 in the analysis leading up to eqn (2.6). Finally, a recursive application of the above procedure gives:

$$J_T(V) = \left(\sum_{i=1}^{N} \frac{1}{J^0_{n,i}(V)} \prod_{j=i+1}^{N+1} \gamma_j(V) \right)^{-1}, \tag{2.7}$$

where $\gamma_{N+1} \equiv 1$, and $\gamma_j = \dfrac{J^0_{n,j} + J^0_{p,j}}{J^0_{n,j}}$.

Eqn (2.7) is the general formula for the calculation of transport current through any multi-regional HBT (of which a BJT is a subset). The ramifications of eqn (2.7) are striking and generally lead to current-limited flow within a single region. An examination of eqn (2.7) begins with the γ_j functions, which are termed the recombination loss; γ_j, therefore, represents the additional current that must exist in order to satisfy the recombination events within Region j. Then, the transported current through each successive region is not $J^0_{n,j}$ but $J^0_{n,j}/\gamma_j$. Now, the form of eqn (2.7) is exactly the same as that used for the calculation of a connected series of conductors. This immediately leads to the picture of a series of pipes through which a current J_T must pass (see Fig. 2.4).

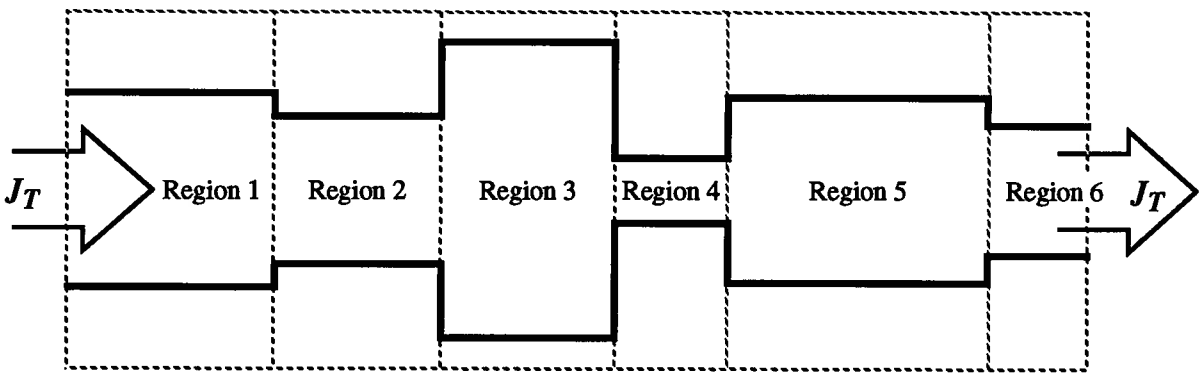


Fig. 2.4. The flow J_T that results from a series connection of six pipes (the flow entering only equals the flow leaving (= J_T) when there is no recombination in any of the regions). Obviously, the pipe in Region 4 is the most restrictive and J_T will accordingly be governed mostly by this region alone.

Looking at eqn (2.7) and letting $J^0_{n,j} \ll J^0_{n,k}$, where $j \neq k$ and k can range over all N, then Region j will be responsible for the current-limited flow of J_T and produce:

$$J_T(V) \approx J^0_{n,j}(V) \prod_{k=j+1}^{N} \alpha_k(V), \tag{2.8}$$

where α_j (= $1/\gamma_j$) is the transport efficiency of Region j and expresses the fraction of the transport current that is lost to recombination within the region. Eqn (2.8) is exactly the form expected from the arguments presented in Fig. 2.4. For if Region j is responsible for the current-limited flow, J_T would equal $J^0_{n,j}$ in the absence of recombination. However, each subsequent region downstream will lose α_k electrons to recombination. Therefore, the current $J^0_{n,j}$ will be diminished by α_k in each region encountered, leaving a final current of J_T exiting the device. This immediately leads to eqn (2.8). Thus, in a device with say six regions, if Region 3 produces the limiting flow, then $J_T = J^0_{n,3}\alpha_4\alpha_5\alpha_6$.

Finally, looking once again at eqn (2.8), recombination events upstream of Region j play no part in the ultimate current J_T. This is no surprise since all of the regions upstream of Region j can supply the demanded current within Region j. However, every region from 1 to N contributes to the recombination current, and must be included in the calculation of the total hole current J_P. Adding all of the recombination events together gives:

$$J_P(V) = \sum_{i=1}^{N} J^0_{p,i} = \sum_{i=1}^{N} J_T(\gamma_i - 1) \prod_{j=i+1}^{N+1} \gamma_j = J_T\left[\sum_{i=1}^{N} \gamma_i \prod_{j=i+1}^{N+1} \gamma_j - \sum_{i=1}^{N} \prod_{j=i+1}^{N+1} \gamma_j\right].$$

Then, after bringing γ_i into the multiplication and letting $i = i' - 1$ in the second term:

$$J_P(V) = J_T\left[\sum_{i=1}^{N}\prod_{j=i}^{N+1}\gamma_j - \sum_{i'=2}^{N+1}\prod_{j=i'}^{N+1}\gamma_j\right] = J_T\left[\prod_{j=1}^{N+1}\gamma_j + \sum_{i=2}^{N}\prod_{j=i}^{N+1}\gamma_j - \sum_{i'=2}^{N}\prod_{j=i'}^{N+1}\gamma_j - \gamma_{N+1}\right].$$

Finally, since $\gamma_{N+1} \equiv 1$ from eqn (2.7), and i' is a dummy variable, the above reduces to:

$$J_P(V) = J_T(V)\left[\prod_{j=1}^{N}\gamma_j(V) - 1\right]. \tag{2.9}$$

Eqn (2.9) provides for the total hole current generated within the device. Combining both eqns (2.7) and (2.9), the total electron and hole current entering and leaving the device is known. As will almost always be the case, one region alone will dictate the transport current and lead to current-limited flow. Then, eqn (2.7) can be replaced by its approximate form, eqn (2.8), to yield after substitution into eqn (2.9):

$$J_P(V) = J^0_{n,j}(V)\left[\prod_{k=1}^{j}\gamma_k(V) - \prod_{k=j+1}^{N+1}\alpha_k(V)\right]. \qquad (2.10)$$

The results of this section are models for the total electron and hole currents entering and leaving an HBT. These models are free of essentially any restrictions upon their functional form, and can therefore be applied to a wide variety of physical processes. Furthermore, the form of the models presented is based upon a simple, modular approach, that is easy to apply to any device. The important ramification is that one region alone will tend to determine the overall transport through the entire device; creating a situation of current-limited flow. The key to achieving a situation of current-limited flow is the existence of a substantial ΔE_{fn} in one region. Finally, abrupt Bandgap Engineering techniques provide the capacity to create a situation of current-limited flow in any region of the device. In the next section and chapters to come, the concept of current-limited flow will be exploited in the optimisation and modelling of HBTs. In the end, eqns (2.7) and (2.9) (or their approximate forms, eqns (2.8) and (2.10) respectively), will be used to bring together all of the models for each of the relevant regions of an HBT.

2.3 Optimisation Through Current-Limited Flow

The main conceptual result of the last section was that one region, or physical process, will tend to dictate the transport current through the entire device. This section examines how to intentionally design a specific region, through Bandgap Engineering techniques, to result in current-limited flow; thereby allowing for a decoupling of J_T from the physical transport processes in all other regions of the device. Finally, once J_T is decoupled from a specific region, by ensuring that transport through the region is much larger than the demanded J_T, then one is free to optimise that specific region without affecting J_T.

Fig. 2.5 shows the transport current that would result from a hypothetical three region device. For case (a), Region 3 controls J_T under low bias and Region 2 controls under high bias; while Region 1 plays no part at all. In case (b), the transport current in Region 1 has been lowered so that Region 1, and neither Regions 2 or 3, controls J_T under all bias conditions. This demonstrates, in principle, the feasibility of engineering a specific region to be the source of current-limited flow, and thereby link J_T to the physical process in that region alone.

In order to see how optimisation can occur by engineering a specific region to be the source of current-limited flow, one begins by identifying the need for decoupling. Imagine there are two specific metrics, say Early voltage (V_A) and collector resistance (R_C), that are to be optimised. If these two metrics are connected to one parameter, in this case collector doping, and the two metrics do not both move towards their optimum value with either an increase or decrease in the one parameter, then only a compromise and not a true optimum can be reached. In the example given, V_A is to be maximised and R_C minimised. However, increased collector doping decreases both V_A and R_C, forcing a compromise between the two metrics to be made. If it were possible to decouple either of these metrics from the one parameter, then it would be possible (in terms of this one parameter only) to optimise both metrics. Therefore, decoupling the metrics from their common competing parameter is the key to removing the compromise and achieving a true optimum.

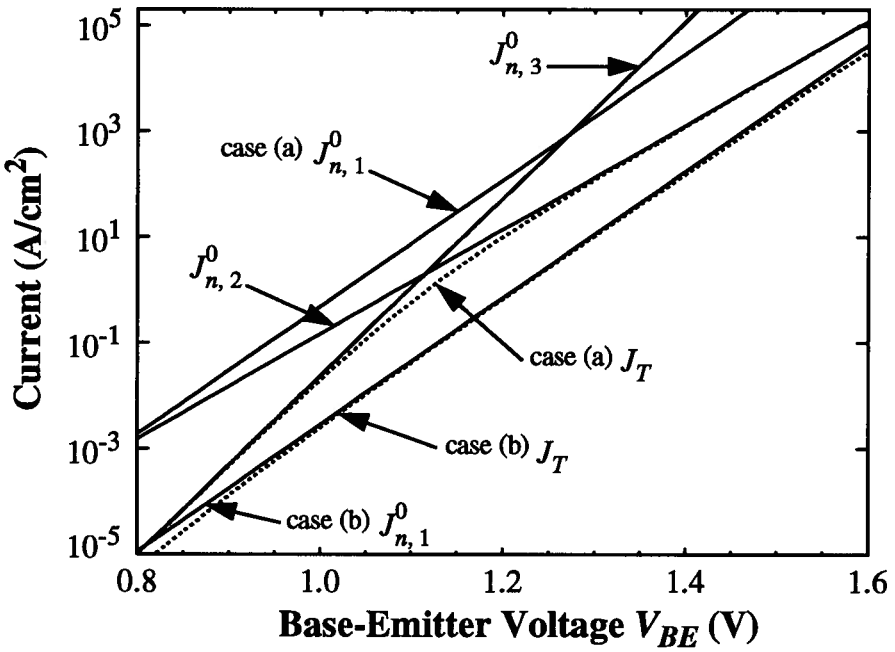


Fig. 2.5. J_T for a three-region HBT in the absence of recombination. The solid lines represent the maximum regional currents J^0, while the dashed lines are J_T. For case (a), Region 1 is never the limiting region; while for case (b), Region 1 is the source of current-limited flow.

At the heart of decoupling is the separation of the transport current from the physical process that is to be optimised. For if the transport current is not affected, or at least not in a detrimental fashion, then one is free to optimise the desired metric. Current-limited flow provides the necessary tool to decouple J_T from all regions, and therefore all physical transport processes, save

one. Continuing on with the example of simultaneously optimising V_A and R_C, if J_T were decoupled from the construction of the base and collector, say by making the emitter-base SCR the source of current-limited flow, then V_A would no longer depend upon the collector doping; enabling the optimisation of R_C without affecting V_A. With base-width modulation no longer an issue, in terms of the collector current and therefore V_A, it would be possible to increase the intrinsic collector doping adjacent to the base and thereby reduce R_C. A further optimisation, in terms of the base-collector capacitance C_{BC}, could also be had by placing a low-doped collector region within the CB SCR (say at 10^{16}cm^{-3} for 2000Å) in order to set C_{BC}, followed immediately by a highly doped extrinsic collector to reduce R_C. Optimisation of competing metrics is thus achieved by first identifying the coupling parameter; then, one other region that does not contain the coupling parameter is constructed (generally through abrupt Bandgap Engineering techniques) to be the source of current-limited flow in order to provide for the control of J_T (*i.e.*, the collector current).

This chapter has provided a logical course to decouple otherwise competing metrics so they may be simultaneously optimised. The tool for decoupling the competing metrics being the creation of current-limited flow outside of the region or regions to be optimised. It is possible to achieve current-limited flow in any given region by resorting to abrupt Bandgap Engineering techniques. Thus, abrupt Bandgap Engineering provides the necessary tool to further optimise BJTs. Finally, all the models for the various regions of the HBT are neatly brought together through eqns (2.7) and (2.9) (or their approximate forms eqns (2.8) and (2.10)) for the calculation of the total electron and hole currents entering and leaving the device.

CHAPTER 3

Base Layer Decoupling and Optimisation

Traditionally, the base region, or more specifically the neutral base region, has determined the overall performance of the BJT. As such, the physical construction of the base is of paramount importance to the function of the BJT. At issue with the base is the fact that there are basically two degrees of freedom within the base; namely the base doping profile $N_{AB}(x)$ and the neutral base width W_B (in an HBT a third parameter, namely the bandgap in the base $E_g(x)$, is also available). Against these two (or three) independent parameters lie numerous device metrics that are to be optimised. Obviously, with more metrics than independently controllable parameters, it is impossible to simultaneously optimise all of the metrics. Thus, an inherent compromise is forced to exist between many of the metrics, which leads to an unnecessary limit to the peak performance of the BJT.

Chapter 2 dealt with the effects of abrupt Bandgap Engineering techniques upon the transport current within an HBT. It was found that through abrupt Bandgap Engineering, it was possible to construct a specific region in such a fashion that the transport current J_T depended on this region alone; thereby decoupling J_T from all other regions of the device. Once J_T has been decoupled from all other regions of the device, save one, the task of independently optimising each region becomes trivial.

The possibility of decoupling J_T from the physical construction of the base promises to eliminate the interdependence that the base-controlled metrics have upon each other. Once the base metrics are free of each other then one can finally consider a truly optimised BJT and thus, achieve a significant improvement to the peak performance of the BJT. Parameters such as the intrinsic base sheet-resistance $R_{B\square}$, base-emitter capacitance C_{BE}, injection index γ (not to be confused with the γ in Section 2.2 which is the recombination loss), Early voltage V_A, base transit time τ_B, and the base-collector capacitance C_{BC} could then be simultaneously optimised. The key to the optimisation of these base metrics rests simply on the decoupling of J_T from the base by constructing one other region of the device in such a manner that it results in current-limited flow.

This chapter takes the abstract concept of optimisation through current-limited flow and applies it to the base region. The methods used to achieve the simultaneous optimisation of the base region metrics follow directly from the prescriptions of Chapter 2. Specifically, the base metrics $R_{B\square}$, C_{BE}, γ, V_A, and τ_B are considered for optimisation. Finally, once the optimum models for each of these metrics within the base region have been derived, they are linked together for the calculation of the total electron and hole currents by the methods derived in Chapter 2.

3.1 Independent Optimisation Of $R_{B\square}$, C_{BE}, And γ

In the design of any transistor, the sufficient design criteria is to provide for a gain that is greater than one. However, it is generally desirable to design a gain that is much larger than one. In the case of a BJT, this translates into maximising the current gain β (= collector current I_C divided by the base current I_B). In current-day BJTs, the manufactured materials are so pure, that for the most part, the recombination of minority carriers being transported through the neutral base represents only a small fraction of the total I_B [53]. Therefore, β will depend on the injection efficiency γ of the Emitter-Base (EB) junction. For an npn BJT the EB γ is given by:

$$\gamma = \frac{J_{n,B}}{J_{n,B} + J_{p,E}}, \tag{3.1}$$

where $J_{n,B}$ is the electron transport current through the base, and $J_{p,E}$ is the hole current injected into the emitter (also known as hole back-injection). Using eqn (3.1), in the absence of neutral-base recombination, the gain is:

$$\beta = \frac{\gamma}{1-\gamma} = \frac{J_{n,B}}{J_{p,E}}. \tag{3.2}$$

Thus, β is maximised as γ is driven towards 1; meaning that $J_{p,E}$ is driven towards zero and/or $J_{n,B}$ is made as large as possible.

In an npn BJT, $J_{n,B}$ is inversely proportional to the base Gummel number $G_{\#B}$ [54-56] given by:

$$G_{\#B} = \int_{W_B} \frac{p(x)}{D_n(x)} dx, \tag{3.3}$$

where D_n is the electron minority carrier diffusion coefficient, W_B is the neutral base width, and p is the base majority hole concentration (= base doping N_{AB} except under high-level injection [56]). Furthermore, for a transparent emitter (an emitter where there is little hole recombination), $J_{p,E}$ is inversely proportional to the emitter Gummel number $G_{\#E}$ [54-56] given by:

$$G_{\#E} = \int_{W_E} \frac{n(x)}{D_p(x)} dx, \tag{3.4}$$

where D_p is the hole minority carrier diffusion coefficient, W_E is the neutral emitter width, and n is the emitter majority electron concentration (= emitter doping N_{DE} except under high-level injection). Thus, β is proportional to $G_{\#E}/G_{\#B}$. Now, the intrinsic base sheet-resistance $R_{B\square}$ is also inversely proportional to $G_{\#B}$ [54]. However, unlike the case for β, where it is desirable to reach a maximum, $R_{B\square}$ is to be minimised in order to improve the high-frequency operation of the BJT.

Since $R_{B\square}$ and β are both tied to the parameter $G_{\#B}$, and increasing $G_{\#B}$ optimises $R_{B\square}$ while de-optimising β, we realise these two metrics are competing and therefore cannot be simultaneously optimised (at least in terms of the parameter $G_{\#B}$).

As was discussed in Section 2.3, the key to optimising two otherwise competing metrics is to identify their common parameter (in this case $G_{\#B}$) and remove its dependency from one of the two metrics. Continuing on with the case of optimising $R_{B\square}$ and β, it would appear possible to increase $G_{\#B}$ and thereby minimise $R_{B\square}$, while also increasing $G_{\#E}$ and thereby maximise β. However, $G_{\#E}$ in a BJT cannot be increased because N_{DE} is either at or very near its maximum physical limit ($\approx 10^{21}$ cm^{-3}). Thus, without resorting to Bandgap Engineering techniques, the only available parameter is $G_{\#B}$, meaning that a compromise has to be made between $R_{B\square}$ and β. This was the motivation for the first HBT; to decouple β from its sole dependence upon $G_{\#B}$.

Looking at Fig. 3.1(a), the band-diagram for a BJT shows that it is just as easy for an electron to enter the base as it is for a hole to enter the emitter (the two carriers see exactly the same potential barrier of $V_{bi} - V_{BE}$). Therefore, the ratio of $J_{n,B}$ to $J_{p,E}$ (= β) will be proportional to the ratio of the available number of electrons in the emitter to the available number of holes in the base ($= N_{DE}/N_{AB} \sim G_{\#E}/G_{\#B}$). Now, if it were possible to alter the bandgap of the EB junction so that the holes had to surmount a larger barrier than the electrons, then $J_{p,E}$ would be significantly reduced and β increased (see Fig. 3.1(b)). Finally, if Bandgap Engineering were employed to achieve an initial 1000-fold increase in β (by reducing $J_{p,E}$ through a Bandgap Engineered ΔE_v), then $G_{\#B}$ could be increased 32-fold, thereby reducing $R_{B\square}$ 32-fold, while still leaving a net 32-fold increase in β. Thus, by creating a heterojunction at the EB metallurgical junction, it is possible to reduce $J_{p,E}$ without increasing $G_{\#E}$. Then, the gains provided by a reduced $J_{p,E}$ are shared between an increase in β and a decrease in $R_{B\square}$.

The methods just described for the simultaneous optimisation of $R_{B\square}$ and β demonstrate the potential gains of abrupt Bandgap Engineering. However, the techniques described above did not follow the exact prescription given in Section 2.3, and thus maintain a coupling between $R_{B\square}$ and β. Instead of decoupling β from $G_{\#B}$, another degree of freedom was added to $G_{\#E}$; namely the abrupt change of ΔE_v in the valence band at the EB junction. The dependence of β upon $G_{\#B}$ still exists, but $J_{p,E}$ and thus β, by the addition of a heterojunction within the EB SCR, now has another dependence of $\exp(-\Delta E_v/kT)$ [2,46,47] through the intrinsic carrier concentration in the emitter

$n_{i,E}$. However, since β still depends upon $G_{\#B}$, any change in $G_{\#B}$ due to bias (such as the Early effect [57], Kirk effect [58] or high level injection [56,59]), will still affect and generally degrade β. The reduction of $J_{p,E}$ through simple abrupt Bandgap Engineering is thus seen as a good first step, but falls short of the optimum case where β is decoupled from $G_{\#B}$ altogether.

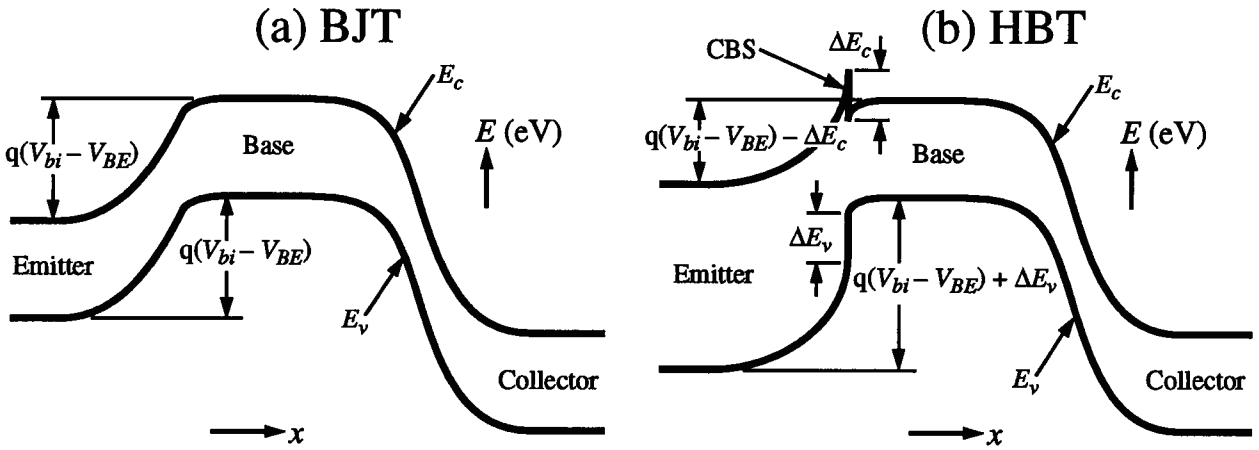


Fig. 3.1. (a): Band diagram of a homojunction BJT. Clearly, the potential barrier seen by a hole trying to go from the base to the emitter is the same barrier seen by an electron trying to go from the emitter to the base. (b): Band diagram of an HBT. Through abrupt Bandgap Engineering, the barrier seen by a hole trying to enter the emitter is a least ΔE_v larger than the barrier seen by an electron trying to enter the base. Also note the formation of the Conduction-Band Spike (CBS).

To fully decouple β from $G_{\#B}$ one looks at the spike in E_c at the EB junction shown in Fig. 3.1(b). This Conduction-Band Spike (CBS) occurs in HBTs where the base is made of GaAs and the emitter is made of $Al_xGa_{1-x}As$ [25]. The barrier to electrons entering the base lies somewhere between $q(V_{bi} - V_{BE})$ and $q(V_{bi} - V_{BE}) - \Delta E_c$, depending on the amount of tunneling through the CBS. In general, it is found that the drop ΔE_c is sufficient to cause the CBS to be the region of current-limited flow (this will be fully discussed in Chapter 4). Thus, J_T (= $J_{n,B}$ in the absence of significant neutral-base recombination) will be governed by the physical process of transport through the CBS, and not by the transport through the neutral base. Furthermore, transport through the CBS has little dependence upon $G_{\#B}$ (as long as the base doping is much larger than the emitter doping). Therefore, J_T and thus $J_{n,B}$ are decoupled from $G_{\#B}$ through the condition of current-limited flow at the CBS.

The condition of current-limited flow in the region of the CBS follows exactly the prescriptions of Section 2.3. $J_{n,B}$ has now been decoupled from $G_{\#B}$, meaning that processes connected to

$G_{\#B}$ such as the Early effect, Kirk effect, and high-level injection, which degraded the collector current of BJTs, are no longer an issue for the abrupt HBT (the term abrupt refers to the abrupt change of ΔE_c and ΔE_v at the EB junction). With the collector current decoupled from $G_{\#B}$, $R_{B\square}$ can be minimised by increasing $G_{\#B}$ through an increase in N_{AB}, while leaving β and therefore γ unaffected.

Before leaving this section to discuss the further optimisation of the base and collector, it should be noted that the EB junction capacitance C_{BE} can also be minimised due to the condition of current-limited flow at the CBS. The high-frequency performance of a BJT improves as C_{BE} decreases. Most notably f_T (the frequency at which β, under the conditions of an A.C. short circuit between emitter and collector, has dropped to unity) increases as C_{BE} is reduced. Since C_{BE} is given by:

$$C_{BE} = \sqrt{\frac{qN_{DE}N_{rat}\varepsilon}{2(V_{bi}-V_{BE})}} \quad \text{where} \quad N_{rat} = \frac{N_{AB}}{N_{AB}+N_{DE}}, \tag{3.5}$$

then N_{DE} and N_{rat} need to be minimised in order to reduce C_{BE}. In a BJT, the need to maximise β forces $N_{DE} \gg N_{AB}$, meaning that N_{AB} is reduced in order to reduce C_{BE}. Thus, C_{BE} is connected to $R_{B\square}$ as well, and leads to another condition where only a compromise and not a true optimum can be reached. CBS-limited flow in an abrupt HBT decouples β from N_{AB}, so that $R_{B\square}$ can be optimised by increasing N_{AB}. Finally, C_{BE} is reduced in an abrupt HBT through the reduction of N_{DE} (for HBTs, $N_{AB} \gg N_{DE}$ so that $N_{rat} \approx 1$). The only limit to the reduction in N_{DE} being the point at which a significant intrinsic emitter resistance R_E begins to occur (see Fig. 3.2).

This section has presented the methods to simultaneously optimise $R_{B\square}$, C_{BE}, and γ. Optimisation of these metrics begins by decoupling from γ and C_{BE} the dependence upon N_{AB}. This decoupling is afforded by the creation of current-limited flow at the CBS. With γ and C_{BE} decoupled from N_{AB}, $R_{B\square}$ is optimised by increasing N_{AB}. Then, C_{BE} is optimised by reducing N_{DE}. Finally, the optimisation of γ depends first of all upon $J_{p,E}$ (which depends heavily on SCR recombination [24]) and secondly upon $J_{n,B}$ (which is governed by the flow of J_T through the CBS); EB SCR recombination, which accounts for most of $J_{p,E}$, is covered in Chapter 5, while the current within the CBS is covered in Chapter 4. The optimisation afforded by the abrupt HBT in comparison to the BJT is stunning, as none of the methods discussed in this section would have been applicable to a BJT because the gain of the transistor would have been reduced below unity.

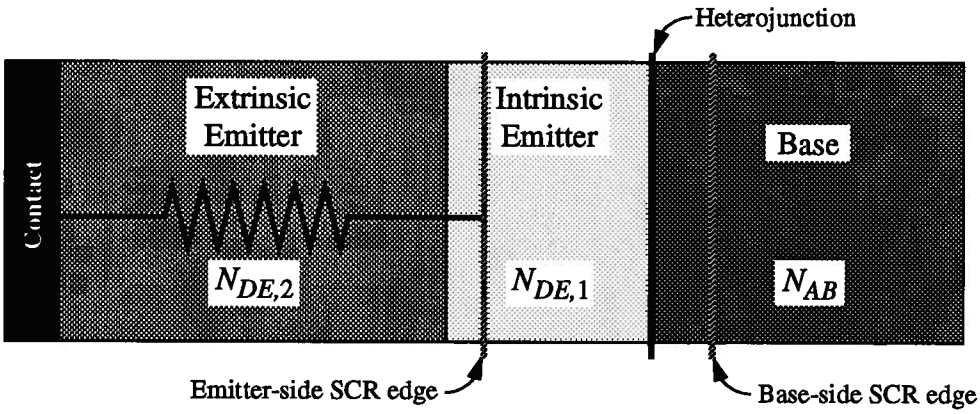


Fig. 3.2. The resistance of the intrinsic emitter will become considerable if $N_{DE,1}$ is reduced without bound. To minimise this parasitic resistance, the width of the intrinsic emitter is only made large enough to contain the emitter extent of the EB SCR. Then, a highly doped $N_{DE,2}$ extrinsic emitter is placed as a cap layer on top of the device, where the eventual contact layer is formed.

3.2 Reducing τ_B by Decoupling the Base from I_C

Chapter 2 discussed the merits of Bandgap Engineering, where the natural evolutionary path of the BJT produces the HBT. Two Bandgap Engineering techniques were considered: techniques that created abrupt changes in E_c/E_v leading to the creation of current-limited flow; and techniques that created gradual changes in E_c/E_v that produced additional aiding fields for the transport of charge. Then, Section 3.1 focussed upon the benefits of current-limited flow produced by an abrupt change of ΔE_c within the EB SCR. This section carries on with the benefits to be derived from current-limited flow, but delves into the second group of Bandgap Engineering techniques - namely the creation of fields in the base to aid in the transport of charge through the region.

A major component of the total transit time for a BJT or HBT is still the neutral-base transit time τ_B. In the absence of any spatial variation to the bandgap or N_{AB}, then under low level injection conditions, with the neutral base width W_B larger than a few mean-free paths λ, the base transit time is given by the standard equation:

$$\tau_B = \frac{W_B^2}{2D_n}. \tag{3.6}$$

τ_B can be reduced from the value given in eqn (3.6), without reducing W_B, by introducing an aiding field in the base (as is shown in Fig. 2.1). BJTs where an aiding field has been placed in the base are termed drift-base transistors [60]. This aiding field implies, for an npn BJT, a downwards

slope to E_c in the neutral base. Before the creation of HBTs, a negative slope in E_c could only be achieved by varying N_{AB} from a high value near the emitter-side of the neutral base, to a low value near the collector-side of the neutral base [60]. This non-uniform $N_{AB}(x)$ would indeed reduce τ_B but at the expense of having a low base doping nearest the collector; leading to a reduced magnitude of the Early voltage. Therefore, the drift-base transistor had a rather limited range of optimisation as the aiding field was coupled in a compromising fashion to the Early voltage. Add to this the fact that the optimum $N_{AB}(x)$ was an un-manufacturable exponential, then the optimum drift-base transistor was a good idea that was generally beyond the manufacturing capabilities of the day.

Enter Bandgap Engineering once again. The issue with the drift-base transistor was the low base doping near the collector. By using a graded bandgap in the base (where the bandgap is large near the emitter-side of the neutral base and small near the collector-side of the neutral base), an aiding field can be created without the need to vary N_{AB} [38]. Thus, by using Bandgap Engineering techniques to create a gradual down-slope to E_c in the base, τ_B can be reduced without lowering N_{AB} and compromising the Early voltage. Kroemer calculated τ_B for a non-uniform bandgap E_g across the neutral base, and found [38]:

$$\tau_B = \int_0^{W_B} \frac{n_i^2(x)}{p(x)} \int_x^{W_B} \frac{p(z)}{D_n(z)\, n_i^2(z)} dz dx, \tag{3.7}$$

where n_i is the intrinsic carrier concentration. The derivation in [38] which leads to eqn (3.7) is based upon Shockley boundary conditions. However, it is a simple extension to show that eqn (3.7) is actually quite general, and is applicable to cases where a ΔE_{fn} is present. Finally, if a linear grading of the bandgap in the base is used, such that $n_i^2(x) = n_i^2(x=0)\exp(qFx/kT)$, eqn (3.7) gives:

$$\tau_B = \frac{W_B^2}{2D_n}\left[e^{-\frac{\Delta E_g}{kT}} + \frac{\Delta E_g}{kT} - 1\right] 2\left(\frac{kT}{\Delta E_g}\right)^2, \tag{3.8}$$

where $F = \Delta E_g/(qW_B)$, and ΔE_g represents the difference between the bandgap at the emitter-side of the neutral base and the bandgap at the collector-side of the neutral base. As an example, if D_n = 30cm^{2}s^{-1}, W_B = 1000Å, and ΔE_g = 3kT, then using eqn (3.6) τ_B = 1.67ps, while using eqn (3.8) τ_B = 0.76ps, a 2.2-fold reduction in τ_B through the addition of a graded bandgap in the base.

The reduction of τ_B through a graded-base transistor is very attractive. When coupled to the fact that the Early voltage is not compromised, Bandgap Engineering in the base appears to hold nothing but gains. The only requirement of a graded-base transistor is the need to create a graded alloy in the base in order to provide for the downwards slope in E_c. In the case of $Al_xGa_{1-x}As$/GaAs HBTs, the bandgap is increased with an increase in the Al mole fraction x; while in $Si_{1-x}Ge_x$ HBTs, the bandgap is decreased with an increase in the Ge mole fraction x. Now, in AlGaAs HBTs the Al mole fraction must remain below a maximum of $x = 0.45$, for this is the point at which the material changes from a direct to an indirect bandgap [61]. In a similar fashion, $Si_{1-x}Ge_x$/Si HBTs have an upper limit of $\Delta x < 0.2$ due to the effects of strain (this is discussed fully in Chapter 6). Thus, an "alloy budget" exists in the HBT, meaning that a decision must be made in the allocation of alloy mole fraction among the various regions of the HBT. Therefore, a compromise must be made in the amount of Bandgap Engineering allocated to the formation of the graded-base versus all the other bandgap-engineered regions of the device.

Since the heterojunction in the EB SCR provides the most important gains in terms of optimising the metrics of the device (namely decoupling γ from $G_{\#B}$), part of the total alloy budget must be allocated to its formation. In the case of AlGaAs HBTs, fully 66% of the maximum total alloy budget ($\Delta x = 0.3$ of a maximum 0.45) is spent in the formation of the EB heterojunction (In reality, $\Delta x < 0.45$ is a maximum upper limit that is generally reduced to 0.30 for practical applications. With this reduced alloy budget, the EB heterojunction would consume the entire budget). In SiGe HBTs, virtually the entire alloy budget of $\Delta x < 0.2$ is spent in the formation of the EB heterojunction. Therefore, irrespective of the material system used to form the HBT, little if any of the alloy budget remains for the Engineered Bandgap in the base once the EB heterojunction has been formed. This means there is little room to reduce τ_B through a manipulation of the bandgap within the base.

The reduction of τ_B is a desirable goal, even in the face of very real practical limitations. Bandgap Engineering in the base may not play a significant role due to the restricted alloy budget; but drift-base transistors, based upon a non-uniform $N_{AB}(x)$, might become plausible by the creation of an abrupt EB heterojunction. The reasons for abandoning drift-base transistors were: it was not possible to manufacture the steep doping profile in the base required to generate the aiding field; and the low base doping near the collector-side of the neutral base resulted in an intoler-

ably low Early voltage. The first problem, namely the manufacture of the highly non-uniform $N_{AB}(x)$, is no longer an issue with advanced MBE and MOCVD growth techniques. The second problem, a decrease to the Early voltage, is solved by decoupling the collector current I_C from the base, so that modulations to $G_{\#B}$ from changes to V_{CB} no longer matter, provided punch-through is avoided, of course. Following, once again, the prescriptions of Section 2.3, I_C is decoupled from $G_{\#B}$ by creating a situation of current-limited flow at the CBS formed by the EB heterojunction; thereby linking I_C to the physical transport mechanisms associated with the CBS instead of the neutral base region. With the two old problems associated with using a non-uniform $N_{AB}(x)$ for the reduction of τ_B solved, the optimum $N_{AB}(x)$ for the reduction of τ_B is investigated.

3.3 Optimum Base Doping Profile to Minimise τ_B

Bandgap Engineering in the base is not really being considered in this section; however, it can be included in the optimisation without any changes in the arguments to follow (this includes effects due to a manufactured change in the bandgap and changes to the bandgap due to heavy doping effects). Starting with eqn (3.7), then after substituting $p = N_{AB}$, τ_B becomes:

$$\tau_B = \int_0^{W_B} \frac{n_i^2(x)}{N_{AB}(x)} \int_x^{W_B} \frac{N_{AB}(z)}{D_n(z)\, n_i^2(z)} \mathrm{d}z\mathrm{d}x. \tag{3.9}$$

If D_n is taken as some average constant, then eqn (3.9) is simplified even further to become:

$$\tau_B = \frac{1}{D_n} \int_0^{W_B} \frac{n_i^2(x)}{N_{AB}(x)} \int_x^{W_B} \frac{N_{AB}(z)}{n_i^2(z)} \mathrm{d}z\mathrm{d}x. \tag{3.10}$$

Eqn (3.10) provides the functional form of τ_B to be minimised. Using the calculus of variations, and searching for the weak variations in $N_{AB}(x)/n_i^2(x)$, then the Euler-Lagrange characteristic equation that minimises eqn (3.10) is:

$$\frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x} = \mathrm{C}, \tag{3.11}$$

where

$$y(x) = \int_x^{W_B} \frac{N_{AB}(z)}{n_i^2(z)} \mathrm{d}z, \tag{3.12}$$

and C is an arbitrary constant. The solution of eqn (3.11) is straightforward and yields:

$$y(x) = -A_1 e^{A_2 x} \tag{3.13}$$

where A_1 and A_2 are arbitrary constants. The beauty about eqn (3.13) is it solves for both $N_{AB}(x)$ and $n_i^2(x)$ simultaneously. The next section will deal with non-uniform bandgap effects, so taking for now that $n_i(x)$ is constant, then differentiating eqn (3.12) and substituting in eqn (3.13) gives:

$$N_{AB}(x) = A_1 A_2 e^{A_2 x} = a e^{bx} \tag{3.14}$$

Eqn (3.14) gives the standard exponential solution [60] for the doping profile in the base that leads to a minimum in τ_B.

Within the confines of weak variations, a possible minimum could occur by admitting a piece-wise solution for $N_{AB}(x)$ composed of N sections whose form within each section is given by eqn (3.14). The conditions of continuity at any break-point joining two regions being [62]:

$$\frac{\partial F}{\partial y'} \text{ and } F - y'\left(\frac{\partial F}{\partial y'}\right) \text{ be continuous,} \tag{3.15}$$

where F is the integrand that is to be made stationary, and primes denote differentiation with respect to the dependent variable x. In the case being considered, $F = y/y'$. Then, using the exponential solution for $y(x)$ in eqn (3.13), and applying the second continuity condition of eqn (3.15) produces:

$$F - y'\left(\frac{\partial F}{\partial y'}\right) = \frac{2y}{y'} = \frac{2}{A_2},$$

which must be continuous at the break-point x_0 joining the two regions. If we let Region 1 join with Region 2, where the solution in Region 1 is $A_{1,1} e^{A_{2,1} x}$ and the solution in Region 2 is $A_{1,2} e^{A_{2,2} x}$, then the above equation requires that $A_{2,1} = A_{2,2} = A_2$. Applying the first continuity condition of eqn (3.15) at the point $x = x_0$ produces:

$$\frac{\partial F}{\partial y'} = \frac{y}{y'^2} = \frac{1}{A_{1,1} A_2^2 e^{A_2 x_0}} = \frac{1}{A_{1,2} A_2^2 e^{A_2 x_0}} \Rightarrow A_{1,1} = A_{1,2} = A_1.$$

Thus, a piece-wise connection of exponentials is not admitted as a stationary solution for $N_{AB}(x)$. However, if the last equation is rewritten as $A_{1,1} A_2^2 e^{A_2 x_0} = A_{1,2} A_2^2 e^{A_2 x_0}$, then as $A_2 \to 0$ no restriction is placed on the values admitted for $A_{1,1}$ and $A_{1,2}$. This admitted solution for $y(x)$ is also a piece-wise and discontinuous set of constant solutions. As such, this solution for $y(x)$ tends towards a strong variation and care must be exercised in the absolute applicability of the weak variational principles used to obtain this result. With that cautionary note in mind, if the form of $A_{1,1}$ and $A_{1,2}$ are carefully chosen to be a_1/A_2 and a_2/A_2 respectively, then as $A_2 \to 0$, $N_{AB}(x)$ also becomes a piece-wise and discontinuous set of constant solutions.

The weak variational principles used to find the $N_{AB}(x)$ that renders τ_B stationary are constructed in a such a manner that only $y(x)$ be defined at the end points of the integration. Since $N_{AB}(x)$ is given by y', there is no simple way to specify the doping values at the end points of the integration (namely the emitter and collector edges to the neutral base). Further examination of eqn (3.14) shows that there are no bounds to the value of the constant b in the exponent of the exponential defining $N_{AB}(x)$. In fact, by letting $b \rightarrow -\infty$, an infinitely large aiding field can be created in the base and τ_B will be reduced to zero. To see this, eqn (3.14) is used in (3.10) to give:

$$\tau_B = \tau_{B0} \frac{2\,(e^b - b - 1)}{b^2}, \tag{3.16}$$

where τ_{B0} is the τ_B given by eqn (3.6). Clearly, as $b \rightarrow -\infty$, $\tau_B \rightarrow 0$. As a check, as $b \rightarrow 0$, τ_B given by eqn (3.16) goes to τ_{B0}. Thus, no matter what N_{AB} is forced to be at the emitter-side of the neutral base, $N_{AB}(x)$ can be made to decrease at a rate such that τ_B is ostensibly reduced to zero. Therefore, the variational principles used to deduce that the optimum $N_{AB}(x)$ is a pure exponential are based upon an unrestricted doping at the collector-side of the neutral base

It is not reasonable to allow the doping at the collector-side of the neutral base to become arbitrarily small, even in the presence of current-limited flow at the CBS. For even though I_C is decoupled from $G_{\#B}$, $R_{B\square}$ still depends on $G_{\#B}$ and would become unreasonably large as $b \rightarrow -\infty$. Eqn (3.10) is revisited, but this time τ_B is made stationary subject to boundary conditions upon $N_{AB}(x)$ at the emitter- and collector-sides of the neutral base. Since there appears to be no simple way of including these boundary conditions into the variational principles, a numerical minimisation was constructed [63]. The results of numerical attempts to render τ_B stationary, subject to the boundary conditions placed upon $N_{AB}(x)$, produced a form that suggests $N_{AB}(x)$ be exponential in the middle of the base but have two constant regions attached on the ends (see Fig. 3.3). This result seems plausible in light of the variational analysis performed so far, where a constant was admitted as a solution to $N_{AB}(x)$. Even more convincing, the form being suggested from the numerical analysis is not a piece-wise connected set of exponentials (which was rejected as a possible stationary solution from the variational analysis), but is a piece-wise connection involving constant regions of doping, as is admissible from the variational analysis. In any event, it is clear that the boundary conditions placed upon $N_{AB}(x)$ cause the exponential solution from simple variational analysis to become non-stationary.

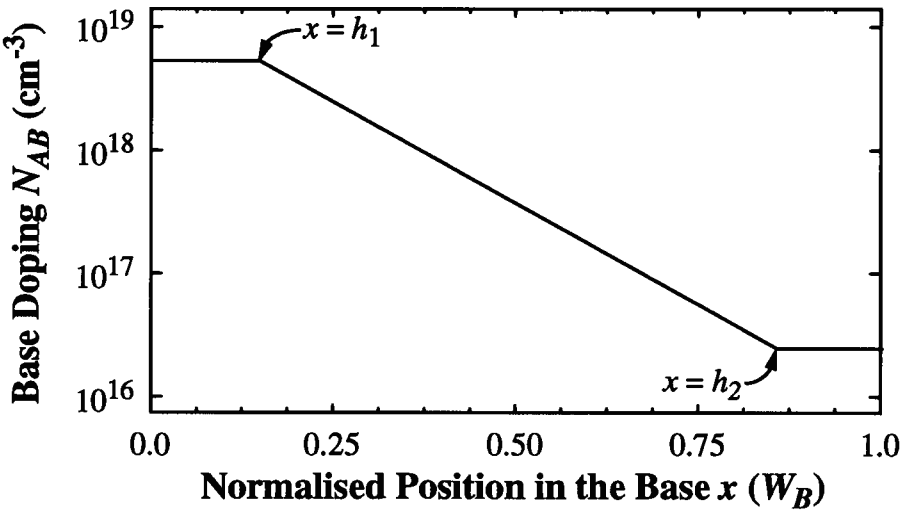


Fig. 3.3. Optimum doping profile $N_{AB}(x)$ obtaining by numerically minimising eqn (3.10) with the boundary conditions $N_{AB}(x=0) = 5\times10^{18}\text{cm}^{-3}$ and $N_{AB}(x=W_B) = 2\times10^{16}\text{cm}^{-3}$.

Using the form for $N_{AB}(x)$ suggested from the numerical work, namely exponentials separated by regions of constant doping, analytic methods were employed to find the break points between the exponentials and the constant regions that minimised τ_B. Using the form of $N_{AB}(x)$ given in Fig. 3.4, then finding the break-point h that minimises τ_B given by eqn (3.10) produces, after considerable algebraic manipulation with the symbolic mathematics tool MACSYMA (see Appendix A):

$$h = \frac{(U\ln U + 1 - U)\ln U}{(U\ln U + 2)\ln U + 2(1-U)} \quad \text{and} \quad \tau_B = \tau_{B0}\frac{U[U(2\ln U - 3) + 4] - 1}{U[(U\ln U + 2)\ln U + 2(1-U)]} \tag{3.17}$$

where τ_{B0} is still the τ_B given by eqn (3.6), h is normalised to the neutral base width W_B (and therefore ranges from 0 at the emitter-side to 1 at the collector-side of the neutral base), and U is the doping ratio given by $N_{AB}(x=0)/N_{AB}(x=W_B)$. The interesting thing to note about eqn (3.17) is that it depends only on the relative doping ratio U. Further, the exact same solution results (save $h \rightarrow 1 - h$) if $N_{AB}(x)$ is changed, in a symmetrical fashion to that shown in Fig. 3.4, so that the constant region occurs first followed by the exponential region. Eqn (3.17) represents the solution of the simplest form of $N_{AB}(x)$ suggested from the numerical analysis.

The process described above is repeated again, but this time with the optimum form (shown in Fig. 3.3) obtained from numerical analysis. Again, substituting this form of $N_{AB}(x)$ into eqn (3.10) and minimising τ_B produces, after considerable algebraic manipulation with the symbolic mathematics tool MACSYMA (see Appendix B):

$$h_1 \equiv 1 - h_2 = \frac{1}{\ln U + 2} \quad \text{and} \quad \tau_B = \tau_{B0}\frac{2}{\ln U + 2} = \tau_{B0} 2h_1 , \tag{3.18}$$

where, h_1 and h_2 are normalised to the neutral base width W_B. Eqn (3.18) shows the beauty of the symmetric form used for $N_{AB}(x)$; namely that the length of each of the constant regions is the same, and the exponential region is perfectly centred within the base. It is very simple to prove that τ_B given by eqn (3.18) is always smaller then that given by eqn (3.17). Therefore, the form of $N_{AB}(x)$ given in Fig. 3.3 produces a smaller τ_B then the form given in Fig. 3.4.

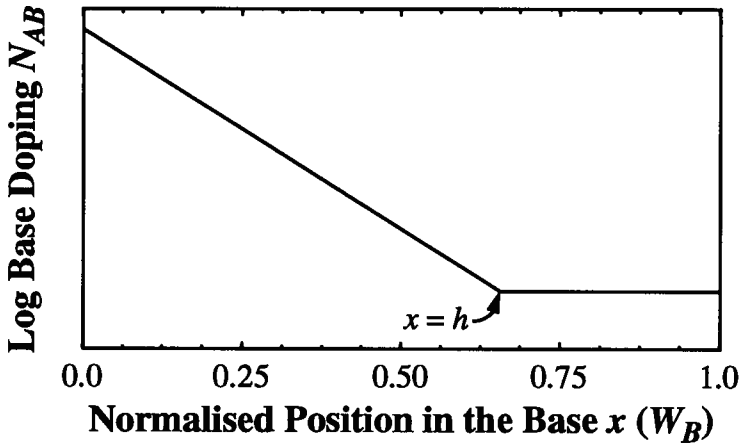


Fig. 3.4. The first trial function for $N_{AB}(x)$ inspired by the form suggested by Fig. 3.3.

The process is continued by constructing more complex forms based upon an extension to $N_{AB}(x)$ given in Fig. 3.3. When eqn (3.10) is minimised using the $N_{AB}(x)$ given by the form shown in Fig. 3.5(a), it is possible to find a stationary result where $h_1 \neq 0$ ($h_1 = 0$ would give $N_{AB}(x)$ as shown in Fig. 3.3). Even though $N_{AB}(x)$ given by Fig. 3.5(a) renders τ_B stationary, when compared to the result obtained from eqn (3.18), it does not produce the absolute minimum value for τ_B. In fact, taking one final progression to using the $N_{AB}(x)$ as shown in Fig. 3.5(b), a stationary result is again obtained, but it is larger still than the case shown in Fig. 3.5(a) and therefore does not produce the absolute minimum value for τ_B. Therefore, eqn (3.18), with $N_{AB}(x)$ as shown in Fig. 3.3, produces the absolute minimum in τ_B subject to the boundary conditions for the doping at the emitter- and collector-sides of the neutral base. The most notable thing about the optimum form of $N_{AB}(x)$, as shown in Fig. 3.3, is that it is not the pure exponential the device community has been lead to believe is the optimum. This result answers the problem posed in [64,65], where the authors used third order perturbation theory to show that an exponential was indeed stationary but it did not produce the absolute minimum for τ_B.

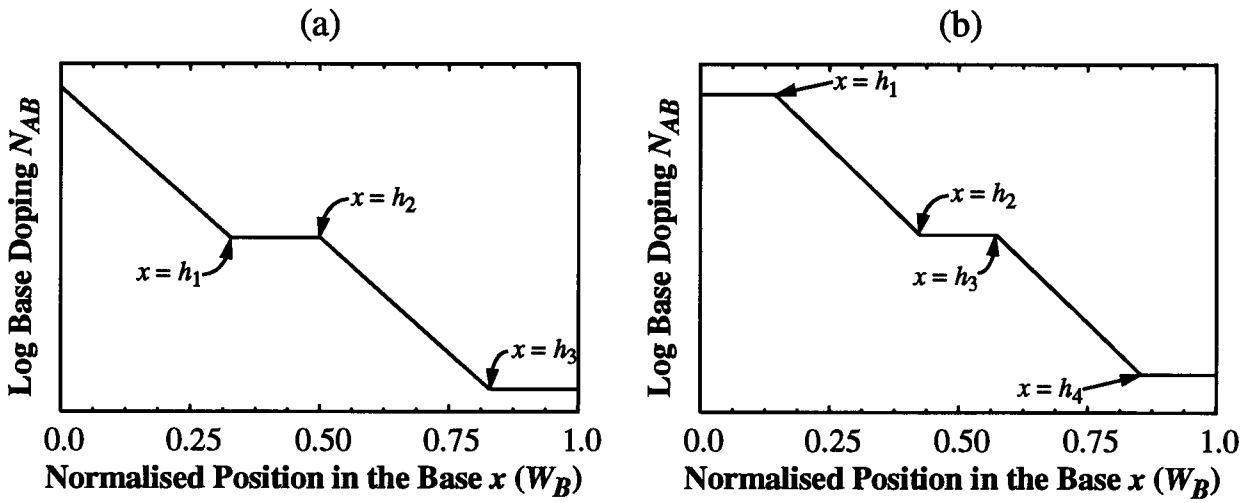


Fig. 3.5. (a): the second trial function for $N_{AB}(x)$, which is an extension of the form shown by Fig. 3.3; (b): the final trial function for $N_{AB}(x)$.

As a final consideration, it is instructive to use the $N_{AB}(x)$ suggested by the analysis surrounding eqn (3.15). In the proof that showed $N_{AB}(x)$ could not be constructed of piece-wise continuous exponentials, it was found that $N_{AB}(x)$ could be constructed of piece-wise discontinuous constants. In the simplest case, if $N_{AB}(x)$ is constructed as shown in Fig. 3.6, then it is straight forward to show that τ_B is minimised when:

$$h = \frac{1}{2} \qquad \text{and} \qquad \tau_B = \tau_{B0}\frac{U+1}{2U} \ . \tag{3.19}$$

Eqn (3.19) shows that a very simple jump discontinuity, or step, in the base doping profile at exactly the half-way point in the neutral base, can reduce the base transit time by a factor of two when compared to the uniform base case (τ_{B0}). In fact, for any $U \geq 10$, the full two-fold reduction in τ_B is achieved. Still, for all relevant U, τ_B given by the step-doping case of eqn (3.19) is larger than that achieved by the optimum-doping case of eqn (3.18). However, the step-doping case shows that even a very simple change to the base doping profile can produce a significant reduction in the transit time through the neutral base. As for the technological objection that a perfect step-doping profile is impossible to create, any deviations from a step, say due to diffusion of dopant during the thermal-cycle of the manufacturing process, will only tend to drive $N_{AB}(x)$ towards the optimum profile and reduce τ_B even further: this result is obvious as a spreading of the step-discontinuity increases the spatial extent of the aiding field and thereby decreases the transit time. Therefore, the step-doping profile, although not as beneficial as the optimum doping profile, still provides for a significant reduction of τ_B, but with very little complexity in terms of manufacturing.

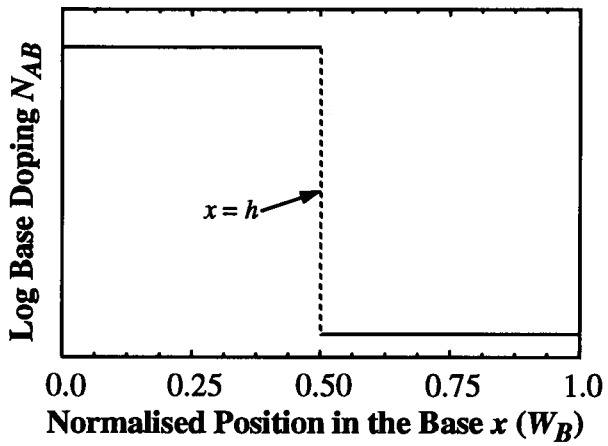


Fig. 3.6. Step-doping profile for $N_{AB}(x)$.

Comparing τ_B given by the optimum $N_{AB}(x)$ (eqn (3.18)), to the ramped $N_{AB}(x)$ (eqn (3.17)), then to the step-doping case (eqn (3.19)), and finally to the pure exponential case (eqn (3.16), with $b = -\ln U$), shows some interesting results (see Fig. 3.7). In all four cases as $U \rightarrow 1$, $\tau_B \rightarrow \tau_{B0}$: this is required and acts as a check to the validity of the four models. As was stated before, for the entire useful range of U (*i.e.*, > 1), τ_B is minimised by the optimum doping profile leading to eqn (3.18). However, for the range $1 \leq U \leq 7.389 = e^2$, τ_B from the step-doping profile is **smaller** than that from the pure exponential profile. Thus, not only have we found out that the pure exponential is not the optimum, we have also found that for small doping ratios the step-doping profile is better than the exponential. An examination of Table 3.1 shows that as U becomes large, the pure exponential case and the ramped case both approach the optimum case for the minimisation of τ_B. This result shows that the optimum-doping case initially starts out looking much like the step-doping case, then as U increases, slowly transforms itself into the pure exponential case. Finally, for U = 300, the optimum-doing case has $h_1 \equiv 1 - h_2 = 0.13$ and τ_B is only 10% less when compared to the pure-exponential case; however, the optimum-doping case has a 49% larger Gummel number and thus a 49% smaller $R_{B\square}$ when compared to the pure-exponential case. Clearly, the pure exponential case is not the optimum doping profile to use, either in terms of minimising τ_B or $R_{B\square}$. Therefore, the optimum-doping case shown in Fig. 3.3 and governed by eqn (3.18) is the best base-doping profile to use in order to minimise τ_B with the smallest impact on $R_{B\square}$.

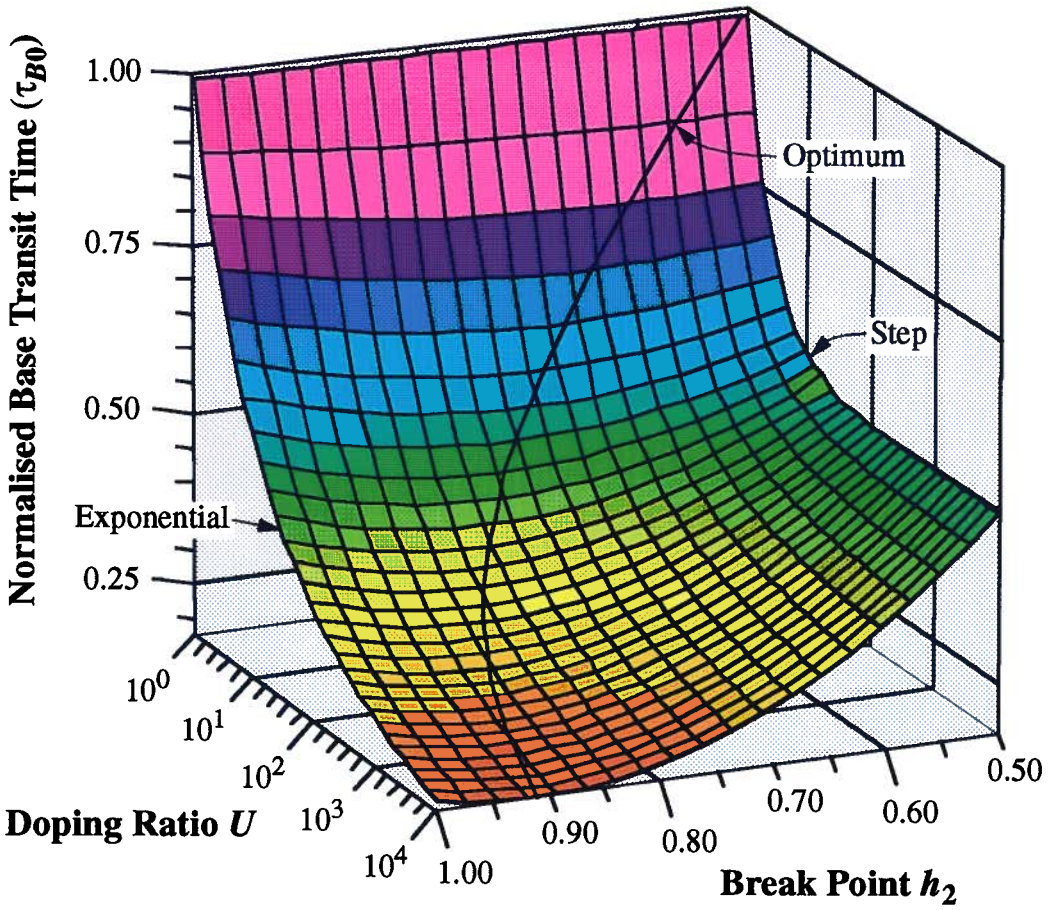


Fig. 3.7. τ_B using $N_{AB}(x)$ from Fig. 3.3, where $h_1 \equiv 1 - h_2$ but h_2 is varied as a parameter instead of being given by eqn (3.18). $h_2 = 0.5$ corresponds to the step-doping case, while $h_2 = 1$ corresponds to the pure exponential case. Finally, the line drawn on the surface is the τ_B that results from the optimum-doping case given by eqn (3.18).

Table 3.1: τ_B for the four doping cases: Optimum, Ramp, Step, and Exponential as given by eqns (3.18), (3.17), (3.19), and (3.16) respectively. NOTE: all values are given in units of τ_{B0}.

U	Optimum	Ramp	Step	Exponential
3	0.65	0.69	0.67	0.72
$7.389 = e^2$	0.50	0.54	0.57	0.57
10	0.46	0.50	0.55	0.53
30	0.37	0.40	0.52	0.42
100	0.30	0.32	0.51	0.34
300	0.26	0.27	0.50	0.29

There are two major restrictions placed on the use of the optimum-doping case for the minimisation of τ_B. These two restrictions are: that the aiding field produced by the non-uniform $N_{AB}(x)$ be small enough to neglect high field effects; and the variation in $D_n(x)$ be small enough to ignore. The first requirement is not terribly restrictive, for even with a base width of 1000Å, and U = 100, the aiding field is 1.7×10^4V/cm which is acceptable for heavy-doped Si and would be at the edge where high field effects begin to occur in heavy-doped GaAs. However, the second requirement that $D_n(x)$ be ostensibly constant over the entire base width is much harder to accept; for even though the base region of an HBT is very heavily doped, $D_n(x)$ would still have a significant variation with U in the range of 7 to 30. The issue of a non-uniform $D_n(x)$, as well as variations in $n_i(x)$ due to Bandgap Engineering and heavy doping, are considered in the next section. In any event, τ_B will always be reduced by using a monotonically decreasing (from emitter towards the collector) non-uniform $N_{AB}(x)$. Therefore, if exact values and not general trends are required, then the optimum-doping case presented in this section must be applied with caution if there is considerable variation in either $D_n(x)$ or $n_i(x)$ across the base.

This section has provided for the optimum $N_{AB}(x)$, given a set of boundary condition to the neutral base, in order to minimise τ_B. It was found that the optimum $N_{AB}(x)$ only depends on the relative doping ratio U, and not on the absolute doping given by the boundary conditions. Furthermore, the optimum $N_{AB}(x)$ is not the pure exponential that the device community has thought was the case, but is an augmented exponential as shown in Fig. 3.3. They key to applying the results of this section hinge on the decoupling of I_C from $G_{\#B}$ afforded by the creation of current-limited flow at the CBS. Therefore, only by creating an abrupt HBT[1] can the drift-base BJT be manufactured without a significant reduction to the Early voltage.

3.4 The Effect of a Non-Uniform n_i and D_n on the Optimum τ_B

Section 3.3 derived the optimum base doping profile for the minimisation of τ_B. It was found that if the base doping was fixed at the emitter- and collector-sides of the neutral base, then the optimum $N_{AB}(x)$ was an augmented exponential shown in Fig. 3.3 and governed by eqn (3.18).

1. It is possible to decouple I_C from $G_{\#B}$ by using a varying bandgap in the base [66]. However, as was discussed in Section 3.2, the alloy budget generally prohibits any significant Bandgap Engineering in the base if an EB heterojunction is to be formed in order to control β. Therefore, the technique of current-limited flow is the only practical method to decouple I_C from $G_{\#B}$.

Due the arguments presented in Section 3.2, Section 3.3 found the optimum $N_{AB}(x)$ without regard to the optimum $n_i(x)$. However, due to the heavy base doping that is characteristic of HBTs, bandgap narrowing will certainly cause variations to $n_i(x)$ when a non-uniform $N_{AB}(x)$ is present. This section will consider the joint optimisation of $N_{AB}(x)$ and $n_i(x)$ in terms of minimising τ_B. Also, the effects of a non-uniform $D_n(x)$ will be discussed.

τ_B is given in full by eqn (3.9). If the variation of $D_n(x)$ with respect to $N_{AB}(x)$ is for the moment ignored, then eqn (3.10) results. Section 3.3 finds the functions $y(x)$ that render eqn (3.10) stationary and then finds the one $y(x)$ that minimises τ_B. $y(x)$ is given by eqn (3.12), which produces after differentiation with respect to x:

$$y'(x) = -\frac{N_{AB}(x)}{n_i^2(x)}.$$

Using eqn (3.11), which is the O.D.E. that renders $y(x)$ stationary, in the above equation yields:

$$\frac{N_{AB}(x)}{n_i^2(x)} = -Cy(x). \tag{3.20}$$

At this point Section 3.3 lets $n_i(x)$ be a constant, which can then be absorbed into the arbitrary constant C, to yield eqn (3.14). However, one could just as easily let $N_{AB}(x)$ be a constant and solve for $n_i^2(x)$. If this is done, then all of the results of Section 3.3 are still applicable to the optimisation of $n_i^2(x)$; for the stationary function $y(x)$ has no dependence on either $N_{AB}(x)$ or $n_i(x)$. This immediately results in the optimum $n_i^2(x)$ being given by the reciprocal to $N_{AB}(x)$ shown in Fig. 3.3, with eqn (3.18) governing the placement of h_1 and h_2 and solving for τ_B. The only change is that U is now given by the ratio $n_i^2(x=W_B)/n_i^2(x=0)$ (the endpoints have been interchanged to keep $U > 1$). If the variation in the effective density of states for E_c and E_v is ignored, then $n_i^2(x) = n_i^2(x=0)\exp(-\Delta E_g(x)/\mathrm{kT})$, where $\Delta E_g(x)$ is now defined as the difference in E_g at x relative to E_g at the emitter-side of the neutral base. Since the optimum $n_i^2(x)$ is given by the reciprocal to $N_{AB}(x)$ shown in Fig. 3.3, and given that Fig. 3.3 is a log plot, then $\Delta E_g(x)$ looks exactly like Fig. 3.3 but it would be linear and not log (see Fig. 3.8). Therefore, just like in the optimum doping case, the optimum bandgap-graded-base HBT is not purely linear, but is the augmented ramp shown in Fig. 3.8.

There is no reason to consider a pure optimisation of either $n_i^2(x)$ or $N_{AB}(x)$. Eqn (3.20) solves for the simultaneous optimisation of both $n_i^2(x)$ and $N_{AB}(x)$. Thus, part of the aiding field can be created by a non-uniform $N_{AB}(x)$, and the rest of the aiding field can be created by a Band-

gap Engineered $n_i^2(x)$. This realisation allows the burden of generating an aiding field to be shared between two physically different parameters. By using both $n_i^2(x)$ and $N_{AB}(x)$, far less of the alloy budget needs to be used in order to generate $n_i^2(x)$, and a smaller decrease in $N_{AB}(x)$ will necessarily have a smaller impact on $G_{\#B}$ and $R_{B\square}$. As an example, let $\tau_B = 0.5\tau_{B0}$. This requires that $U = 7.389 = e^2$, where

$$U = \left.\frac{N_{AB}(x)}{n_i^2(x)}\right|_{x=0} \cdot \left.\frac{n_i^2(x)}{N_{AB}(x)}\right|_{x=W_B} . \tag{3.21}$$

Letting both $N_{AB}(x)$ and $n_i^2(x)$ share equally in generating the aiding field gives $U_{N_{AB}}$ (which is the U for eqn (3.18)) equal to $U_{n_i^2}$ (which is the U for $n_i^2(x)$ shown in Fig. 3.8) which is equal to $\sqrt{7.389} = e$. Thus, the doping in the base as well as $n_i^2(x)$ change by only 2.7-fold, meaning that ΔE_g is only 1kT.

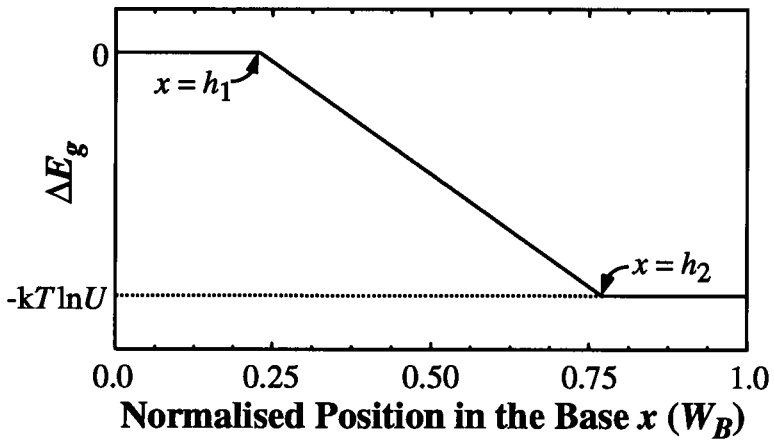


Fig. 3.8. Optimum bandgap in the base to minimise τ_B. The bandgap at the emitter-side of the neutral base (x=0) is the reference point. $U = n_i^2(x=W_B)/n_i^2(x=0)$, where h_1, h_2 and τ_B are given by eqn (3.18).

So far this section has only presented the case where $N_{AB}(x)$ and $n_i^2(x)$ are treated independently of each other. This will not be the case when N_{AB} is large enough to cause bandgap narrowing that couples $n_i^2(x)$ to $N_{AB}(x)$. Since HBTs are characterised by their very high base doping, bandgap narrowing effects need to be considered. Fortunately, the optimisation process that renders $y(x)$ stationary in eqn (3.20) does not depend upon the relationship between $N_{AB}(x)$ and $n_i^2(x)$. Indeed, using eqns (3.21) and (3.18), the optimum $y(x)$ has exactly the same form as the optimum $N_{AB}(x)$ shown in Fig. 3.3 (see Fig. 3.9). Therefore, with the optimum $y(x)$ shown in Fig. 3.9, eqn (3.20) is used to solve for $N_{AB}(x)$ where $n_i^2(x) = n_i^2(\Delta E_g(x), N_{AB}(x))$.

In general, the dependence that n_i^2 has with respect to N_{AB} will be too complex to allow for a closed-form analytic solution. In this case a possible solution process is to use an iterative approach where $y(x)$ is first solved for using eqns (3.21) and (3.18). A trial function $N_{AB}^T(x)$ for the actual $N_{AB}(x)$ is constructed by using h_1 and h_2 from $y(x)$, and forcing $N_{AB}^T(x)$ to take the form of Fig. 3.3 (while obeying the original doping boundary conditions). Finally, using eqn (3.20), a new $N_{AB}(x)$ is solved for using $n_i^2(\Delta E_g(x), N_{AB}^T(x))$ and $y(x)$. This process can be repeated until little change is observed in $N_{AB}(x)$. In the event that the convergence of this iterative method is too slow, then higher-order numerical methods such as Newton-Raphson iteration could be used instead. Thus, it is a simple matter to include bandgap narrowing into the optimum base profile for the minimisation of τ_B, for the stationary function $y(x)$ that defines both $N_{AB}(x)$ and $n_i^2(x)$ is independent of both these functions.

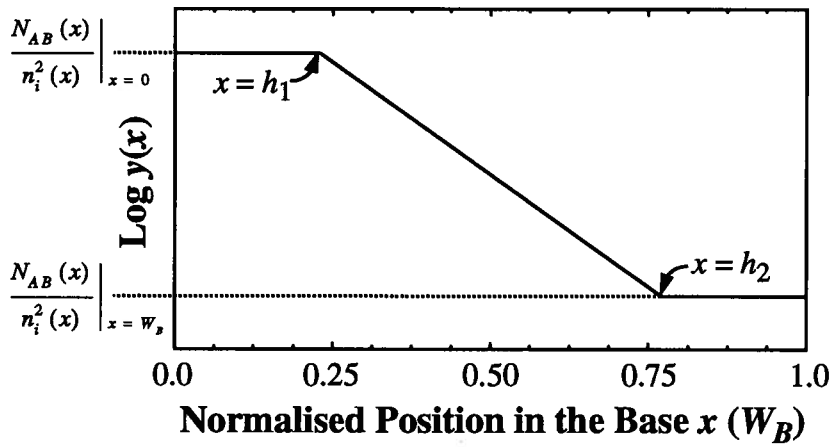


Fig. 3.9. The optimum stationary function $y(x)$ that minimises τ_B. The break points h_1 and h_2, as well as the transit time τ_B are given by eqn (3.18) with U defined in eqn (3.21). $N_{AB}(x)$ and $n_i^2(x)$ are solved for using $y(x)$ in eqn (3.20) along with C = -1. $y(x)$, as shown here, does not depend on the functional form of either n_i^2 or N_{AB}, but only on the boundary condition U.

The last issue to tackle is the effect of a non-constant $D_n(x)$ on the optimum profile found thus far. Strictly, to accomplish this minimisation, one must apply the methods of variational calculus to eqn (3.9) directly; which leads to an O.D.E. that is not soluble in terms of any know transcendental functions. The effect of a non-uniform $D_n(x)$ is investigated numerically in [63] for large U, and the result is a solution that has elements of the stationary functions presented in this chapter, but as a whole cannot be construed as the same. However, current day BJTs (and HBTs) are such that τ_B is an important but not dominant part of the total transit time (in the area of 30%).

Therefore, more than a 2-fold reduction in τ_B is really not warranted as the point of diminishing returns would be surpassed. From the results presented earlier in this section, τ_B can be reduced 2-fold with only a 2.7-fold reduction in $N_{AB}(x)$ across the base when coupled with a ΔE_g of 1kT. With $N_{AB}(x)$ changing by only 2.7-fold, it is reasonable to assert that $D_n(x)$ is ostensibly constant. However, if larger changes to $N_{AB}(x)$ are pursued, then the results of this chapter will certainly reduce τ_B, but only a full numerical optimisation will provide the true minimum [63].

This section has found the optimum base profile for the minimisation of τ_B when both the doping and the bandgap have been constrained at the emitter- and collector-sides of the neutral base. The optimum base profile has the form of an augmented exponential shown in Fig. 3.9, not the long-established pure exponential [60] that has been mistakenly assumed. Further, the solution presented allows for the simultaneous optimisation of $N_{AB}(x)$ and $n_i^2(x)$, and can also include the effects of bandgap narrowing due to heavy doping. Perhaps the most interesting and startling result occurs by using both $N_{AB}(x)$ and $n_i^2(x)$ to generate the aiding field in the base, thereby reducing the overall variation in each parameter across the neutral base. Finally, all of the models and methods presented and discussed in this chapter have no particular material system in mind. Therefore, this chapter can be applied to an HBT build in any material system (such as AlGaAs or SiGe).

CHAPTER 4

Transport Through the EB SCR

In BJTs it is customary to apply the Shockley boundary condition at both edges to the EB SCR in order to determine the quasi-Fermi levels [67]. The Shockley boundary conditions are based upon the assumption that no matter what physical process is responsible for the movement of charge through the EB SCR, the total transport current J_T will be very small compared to the forward and reverse directed fluxes at any point within the region. This argument follows exactly the development of Section 2.1. Applying eqn (2.2) under the conditions described in this paragraph leads to $\Delta E_{fn} \approx 0$. In fact, the Shockley boundary conditions simply state that E_{fn} and E_{fp} are constant across the EB SCR. These boundary conditions allow for an enormous simplification because the exact details of the transport through the EB SCR no longer need to be understood or included in the final model for the device.

By their very nature, HBTs can generate spikes (such as the CBS in Fig. 3.1) in the conduction and valence bands that reduce the forward directed flux. If one of these spikes is large enough, then J_T could be constrained by the flux through this one feature alone. Fig. 3.1(b) shows the general band diagram for HBTs built in the $Al_xGa_{1-x}As$ material system, where there is an abrupt heterojunction between the emitter and the base. The very nature of the sign of ΔE_c, when coupled to the fact that the emitter doping is much smaller that the base doping, produces a feature in E_c called the CBS [25]. The CBS can easily force the electrons to take a path that requires an increase in energy of nearly 240meV. To increase the electron energy by 240meV, with respect to a homojunction, would reduce the available number of electrons, and therefore the forward directed flux, by four orders of magnitude at room temperature. A reduction by 10^4 in the forward directed flux will most certainly result in current-limited flow in the region containing the CBS. This will invalidate the quasi-equilibrium assumption of the Shockley boundary conditions. Thus, one must consider the limits imposed by the movement of charge through the CBS upon the transport current within the EB SCR.

The thermionic injection of electrons over the top of the CBS is not the only method of transport through the region. Due to the quantum mechanical nature of the electron, and the fact that the width of the CBS is typically of the same order as the de Broglie wavelength, the electron could tunnel through the CBS instead of trying to increase its energy in order to surmount the barrier. Since a reduction in the required energy to surmount the CBS leads to an exponential increase in the forward directed flux, tunneling and therefore the quantum mechanical nature of the

electron also needs to be considered when deriving the physical models for transport through the CBS. Failure to include this tunneling current will underestimate J_T by up to two orders of magnitude [25] (see also Fig. 4.8). Therefore, no matter how powerful a model is used (such as Monte Carlo modelling), if tunneling is not accounted for through the CBS, the terminal characteristics of the device will be greatly underestimated.

All of the previous chapters have relied on the existence of current-limited flow in one region of the device that is separated from both the base and the collector. Specifically, the region providing current-limited flow occurred at the EB heterojunction where the CBS is formed. Since the transport current through the device leads to I_C, and because current-limited flow at the CBS controls the transport current, then I_C is governed completely by the transport mechanisms of the CBS. Under the condition of CBS control, I_C has no dependence on the physical construction of either the base or the collector. By constructing the HBT in a fashion where the CBS controls I_C, a detailed understanding of the physics surrounding the CBS must be undertaken if one hopes to accurately predict the terminal characteristics of the device. This chapter investigates and derives models for the transport of charge through the region containing the CBS, including effects due to tunneling and a varying effective mass.

4.1 Formulation of Charge Transport at the CBS

The transport of charge through the region where the CBS is formed can be found by viewing the system as a set of forward and reverse directed fluxes (F_f and F_r respectively) entering the region from opposite sides (see Fig. 4.1). If there is no source or sink of carriers within the region considered, then just like eqn (2.2) $F = F_f(-x_n) - F_r(x_p)$, where F is the transport flux, x_n is the thickness of the SCR extending from the heterojunction into the emitter, and x_p is the thickness of the SCR extending from the heterojunction into the base. If at the points $-x_n$ and x_p it is acceptable to state that the system is fully thermalised, based upon a local Fermi energy E_f, then the carrier distribution with respect to total energy U is:

$$\mathrm{f}(U) = \frac{1}{1 + \mathrm{e}^{\frac{U-\mu}{kT}}}, \tag{4.1}$$

where f is the Fermi-Dirac distribution function and μ is the electrochemical potential (which is usually termed the Fermi energy E_f). Using eqn (4.1) and the quantum mechanics of crystalline

solids, the transport flux through the region containing the CBS in the x-direction can be written in the standard form [68-70]:

$$F = F_f - F_r = \frac{2q}{(2\pi)^3}\int_{R_f} d^3k\, f_1(U)\,(1 - f_2(U))\, W^+(U_x)\frac{1}{\hbar}\frac{\partial U}{\partial k_x} - \frac{2q}{(2\pi)^3}\int_{R_r} d^3k\, f_2(U)\,(1 - f_1(U))\, W^-(U_x)\frac{1}{\hbar}\frac{\partial U}{\partial k_x} \quad (4.2)$$

where $W^+(U_x)$ and $W^-(U_x)$ are the forward and reverse directed transmission probabilities respectively, f_1 is the Fermi-Dirac distribution at $-x_n$, f_2 is the Fermi-Dirac distribution at x_p, R_f is the valid energy range considering forward flux, R_r is the valid energy range considering reverse flux, U is total energy, U_x is the x-directed energy, and $\boldsymbol{k}$ is three dimensional k-space.

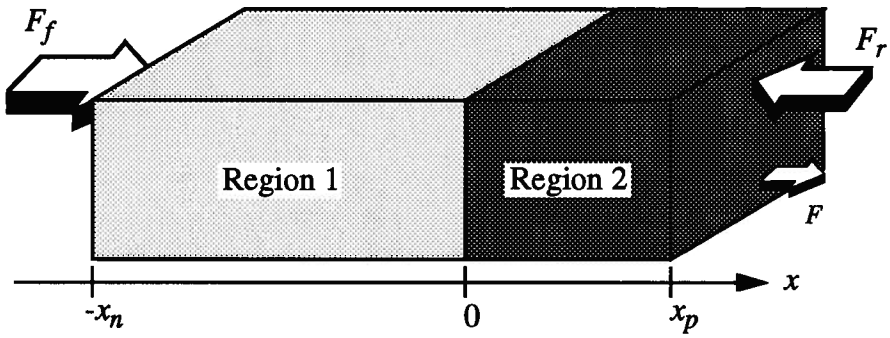


Fig. 4.1. Abstract model of current flux within the region containing the CBS. The EB heterojunction is centred at $x = 0$, with x_n being the excursion into the emitter (Region 1), and x_p being the excursion into the base (Region 2). There is a flux F_f entering the region at $x = -x_n$ and another flux F_r entering from $x = x_p$. The net transport flux F is equal to $F_f - F_r$ in the absence of any sinks or sources within the region.

The interpretation of eqn (4.2) is straight forward in that: there are $2(2\pi)^{-3}$ electron states per unit volume in k-space (including spin degeneracy); $f_1(1\text{-}f_2)$ (in the case of the forward directed flux) is the probability of an electron existing in Region 1 and being able to move to an empty state in Region 2; $W^+(U_x)$ is the probability of the electron moving from $-x_n$ to x_p with a forward directed energy of U_x; and $(1/\hbar)(\partial U/\partial k_x)$ is the group velocity of the electron [15]. As eqn (4.2) stands, the forward and reverse directed transmission probabilities are treated separately using $W^+(U_x)$ and $W^-(U_x)$ respectively. This allows for a non-reversible system to be studied, where electron collisions with the lattice (but not with other electrons) can be included. Strictly, if collisions are considered that change the total energy U of the electron, and not simply its direction in k-space, then the vacancy probability $1 - f_2(U)$ (in the case of the forward directed flux) will not depend on U, but will depend on the exit energy in Region 2. However, if any type of collision is

considered, then $W^{+}(U_x)$ and $W^{-}(U_x)$ will be of an extremely complex nature and would require a numerical calculation of eqn (4.2) (this could be accomplished by a Monte-Carlo simulator using non-local mathematics; however, no such simulator exists at this time). As a result, eqn (4.2) is simplified by considering collision-less or ballistic transport throughout the entire region, leading to $W^{+}(U_x) = W^{-}(U_x) = W(U_x)$. With the assumption of ballistic transport throughout the region from $-x_n$ to x_p, and converting from k to momentum p (= $\hbar k$), eqn (4.2) yields:

$$F = F_f - F_r = \frac{2q}{h^3}\int_{R_f} d^3p\, f_1(U)\,(1 - f_2(U))\,W(U_x)\frac{\partial U}{\partial p_x} - \frac{2q}{h^3}\int_{R_r} d^3p\, f_2(U)\,(1 - f_1(U))\,W(U_x)\frac{\partial U}{\partial p_x}. \tag{4.3}$$

Examining eqn (4.3) shows that if the regions of integration R_f and R_r were equal, then the two integrals could be reduced to one integral with an integrand of $(f_1 - f_2)W(\partial U/\partial p_x)$. One could then identify an F_f and F_r from this integrand (which strictly speaking is not the same as that defined in eqns (4.2) and (4.3), but for all practical situations is identical), giving:

$$F_f \equiv \frac{2q}{h^3}\int_{R_f} d^3p\, f_1(U)W(U_x)\frac{\partial U}{\partial p_x} \tag{4.4}$$

and

$$F_r \equiv \frac{2q}{h^3}\int_{R_r} d^3p\, f_2(U)W(U_x)\frac{\partial U}{\partial p_x}. \tag{4.5}$$

The key to the definitions of eqns (4.4) and (4.5) is the equivalence of R_f and R_r. The fact that this is indeed true is proven later on in Section 4.3 once the effects of a non-uniform effective mass have been brought into the picture.

The solution of F_f and F_r defined in eqns (4.4) and (4.5) begins by determining the transmission probability $W(U_x)$. Strictly, $W(U_x)$ must be calculated by solving the Schrödinger equation, based upon the potential profile encountered within the EB SCR. The solution of the Schrödinger equation, even for a potential obtained from the depletion approximation, is complex enough to require a numerical solution. Failure to obtain an analytic form for $W(U_x)$ would hide the rich interplay that exists between the final transport model for the CBS and the physical attributes such as doping concentration, temperature, effective mass, electron affinity, and bias conditions. An approximate but analytic form is thus sought for the solution of $W(U_x)$. To this end, one could ap-

peal to the asymptotic formalisms in the complex plane used by Landau and Lifshitz [71], or to the JWKB method [72], to obtain:

$$W(U_x) = \exp\left(\Re e\left\{\frac{2i}{\hbar}\int_{-x_n}^{x_p} p\,dx\right\}\right) = \exp\left(\Re e\left\{-\frac{2\sqrt{2m}}{\hbar}\int_{-x_n}^{x_p}\sqrt{V(x)-U_x}\,dx\right\}\right), \quad (4.6)$$

where $V(x)$ is the potential profile of the CBS, and only the real part of the exponent in eqn (4.6) is retained (*i.e.*, $U_x < V(x)$), such that particles with energies larger than the potential energy move without any quantum mechanical reflection. Eqn (4.6) presents a simple analytic solution for $W(U_x)$, where the particle mass m is in general not equal to the electron mass m_e, but to the more general effective mass m^* that is characteristic of semiconductors.

$W(U_x)$ is solved for using eqn (4.6) and a $V(x)$ obtained from the depletion approximation. Fig. 4.2 shows the CBS, which is an enlargement of Fig. 3.1(b). Since the depletion approximation results in a parabolic form for $V(x)$, then one can write:

$$V(x) = V_{pk}\left(1+\frac{x}{x_n}\right)^2 \quad \text{for } -x_n \le x \le 0, \quad (4.7)$$

where V_{pk} is the peak energy of the CBS, and the reference energy is at the bottom of the conduction band where $x = -x_n$. Eqn (4.7) is appropriate to the case where the heterojunction and the metallurgical junction are coincident. The domain $0 \le x \le x_p$ will be considered separately so that $W(U_x)$ may be separated into two functions; one for Region 1 ($W_{CBS}(U_x)$) and another for Region 2 ($W_N(U_x)$, where N stands for Notch), leading to:

$$W(U_x) = W_{CBS}(U_x)\,W_N(U_x). \quad (4.8)$$

Using eqns (4.6)-(4.8) with $W_N(U_x) = 1$ produces:

$$W_{CBS}(U_x) = W_{CBS}(U_x' V_{pk}) = \exp\left[\frac{x_n\sqrt{2mV_{pk}}}{\hbar}\left(\ln\left(\frac{\sqrt{1-U_x'}+1}{\sqrt{U_x'}}\right)U_x' - \sqrt{1-U_x'}\right)\right], \quad (4.9)$$

where U_x' is normalised energy in terms of V_{pk} (*i.e.*, $U_x' = U_x/V_{pk}$). Eqn (4.9) forms the basic kernel for the transmission probability, and it is written in a most general form where V_{pk} and x_n have not yet been defined in terms of the material parameters and applied bias for the EB SCR.

With $W(U_x)$ solved for using eqn (4.8) and (4.9) ($W_N(U_x)$ will be solved for when the regions of integration R_f and R_r are determined), F_f and F_r can be obtained once the energy dispersion relationship $U(p)$ has been set out. The following section will determine $U(p)$ and include the effects of a non-uniform effective mass m^* that generally occurs at an abrupt heterojunction.

Once $U(p)$ has been determined, the regions of integration R_f and R_r are set out so that F_f and F_r can be solved for using eqns (4.4) and (4.5) in the next section.

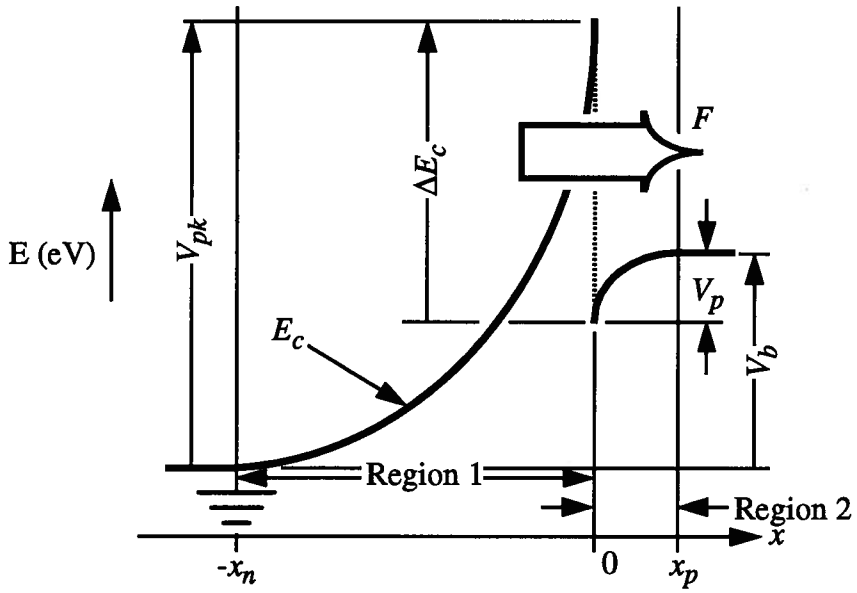


Fig. 4.2. Blow-up of the CBS from Fig. 3.1(b), showing the various energies and their reference.

4.2 Incorporation of Effective Mass Changes

In general, the two materials that form the abrupt heterojunction shown in Fig. 4.2 are characterised by a different effective mass m^*. This change in m^* can either enhance or diminish the flux F in transit through the CBS when compared to the case where m^* is uniform throughout the region. Failure to account for the change in m^* can result in a significant error. Worse yet, this error is not simply a multiplicative constant as is stated by Grinberg [51], but has a dependence on the applied bias. Therefore, in solving for F_f and F_r using eqns (4.4) and (4.5), the dispersion relationship $U(p)$ needs to be determined in concert with the effects of a non-uniform m^*.

Concentrating on eqn (4.4) for F_f (the exact same results will apply to eqn (4.5) for F_r), it is realised that the integration is being performed over p-space. As the entire integrand is dependent upon total energy U and x-directed energy U_x, it would be beneficial to cause a change of variables in the domain of integration from $\boldsymbol{p}$ to $\boldsymbol{U}$. To this end, the dispersion relationship will be taken as parabolic, but left as a diagonal mass tensor to yield:

$$\mathrm{U}(p) = U_x(p) + \mathrm{U}_\perp(p) = \frac{p_x^2}{2m_x} + \frac{1}{2}\left(\frac{p_y^2}{m_y} + \frac{p_z^2}{m_z}\right), \tag{4.10}$$

where m_x, m_y, and m_z are the effective masses for particles that have momenta of p_x, p_y, and p_z re-

spectively. As before, U_x is the x-directed energy, and now, $U_\perp$ is the transverse directed energy. It is also important to realise that eqn (4.10) implicitly places the energy reference at the band extrema. A further simplification can be achieved by a change from cartesian momentum coordinates to cylindrical momentum coordinates. Since we are considering devices that behave essentially as one-dimensional, symmetry dictates that the azimuth direction in the cylindrical system be chosen parallel to the x axis (see Fig. 4.3). This yields:

$$p_y = p_\perp \cos\Theta \quad \text{and} \quad p_z = p_\perp \sin\Theta, \tag{4.11}$$

where eqn (4.10) has that:

$$U_x = \frac{p_x^2}{2m_x} \quad \text{and} \quad U_\perp = \frac{p_y^2}{2m_y} + \frac{p_z^2}{2m_z}. \tag{4.12}$$

Eqns (4.10)-(4.12) together allow for the solution of F_f. The only approximation being made is that $U(p)$ can be adequately described within the parabolic approximation. However, the full mass tensor has been retained (albeit in diagonal form) so that anisotropic materials such as Si, SiGe and strained semiconductors can be modelled with the results to follow in this chapter.

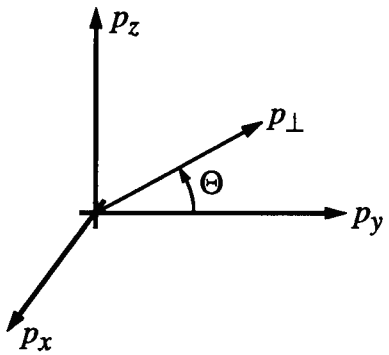


Fig. 4.3. Diagram showing the definitions of the cylindrical momentum space coordinates.

At present, the non-uniformity of m^* has not been included, but it has also not been precluded. Setting aside the issues of a spatially varying m^* for the moment, the integration over $\boldsymbol{p}$ is transformed to $\boldsymbol{U}$ by the Jacobian:

$$\mathrm{J}\left(\frac{p_x, p_y, p_z}{U_x, \Theta, U_\perp}\right) = \left|\begin{bmatrix} \frac{\partial p_x}{\partial U_x} & \frac{\partial p_x}{\partial \Theta} & \frac{\partial p_x}{\partial U_\perp} \\ \frac{\partial p_y}{\partial U_x} & \frac{\partial p_y}{\partial \Theta} & \frac{\partial p_y}{\partial U_\perp} \\ \frac{\partial p_z}{\partial U_x} & \frac{\partial p_z}{\partial \Theta} & \frac{\partial p_z}{\partial U_\perp} \end{bmatrix}\right|. \tag{4.13}$$

The solution of the Jacobian in eqn (4.13) rests on the definitions in eqns (4.11) and (4.12). Looking at eqn (4.12) for the definition of U_x shows a dependence upon the canonical coordinate p_x alone. From this realisation it immediately follows that:

$$\frac{\partial p_x}{\partial \Theta} = 0 \qquad \text{and} \qquad \frac{\partial p_x}{\partial U_\perp} = 0 . \tag{4.14}$$

Furthermore, eqn (4.12) also produces:

$$\frac{\partial p_x}{\partial U_x} = \frac{m_x}{p_x}. \tag{4.15}$$

So far, eqns (4.14) and (4.15) have quickly solved for the first row of the Jacobian in eqn (4.13).

Moving on to the second row and looking once again to eqn (4.12), but this time taking the definition for $U_\perp$ and performing partial implicit differentiation with respect to $U_\perp$, gives:

$$1 = \frac{p_y}{m_y}\frac{\partial p_y}{\partial U_\perp} + \frac{p_z}{m_z}\frac{\partial p_z}{\partial U_\perp} \Rightarrow \qquad \frac{\partial p_y}{\partial U_\perp} = \frac{m_y}{p_y} - \frac{m_y}{m_z}\frac{p_z}{p_y}\frac{\partial p_z}{\partial U_\perp} . \tag{4.16}$$

Then using eqn (4.11), which can be condensed and rewritten as $p_z^2 = p_y^2 \tan^2\Theta$, produces after implicit differentiation with respect to $U_\perp$:

$$\frac{\partial p_z}{\partial U_\perp} = \frac{p_y}{p_z}\tan^2\Theta\frac{\partial p_y}{\partial U_\perp}. \tag{4.17}$$

Finally, substituting eqn (4.17) into (4.16), yields:

$$\frac{\partial p_y}{\partial U_\perp} = \frac{1}{p_y}\,\frac{m_y m_z \cos^2\Theta}{m_z\cos^2\Theta + m_y\sin^2\Theta}. \tag{4.18}$$

Pressing on and using $p_z^2 = p_y^2 \tan^2\Theta$, but this time performing implicit differentiation with respect to Θ, gives after some algebraic manipulation:

$$\frac{\partial p_y}{\partial \Theta} = \frac{p_z}{p_y}\frac{\cos\Theta}{\sin^2\Theta}\left[\cos\Theta\frac{\partial p_z}{\partial \Theta} - \frac{p_z}{\sin\Theta}\right] \Rightarrow \qquad \frac{\partial p_y}{\partial \Theta} = \frac{1}{\sin\Theta}\left[\cos\Theta\frac{\partial p_z}{\partial \Theta} - \frac{p_z}{\sin\Theta}\right]. \tag{4.19}$$

Then, returning back to eqn (4.12) for $U_\perp$ and performing implicit differentiation with respect to Θ yields:

$$0 = \frac{p_y}{m_y}\frac{\partial p_y}{\partial \Theta} + \frac{p_z}{m_z}\frac{\partial p_z}{\partial \Theta} \Rightarrow \qquad \frac{\partial p_z}{\partial \Theta} = -\frac{m_z}{m_y}\frac{p_y}{p_z}\frac{\partial p_y}{\partial \Theta} . \tag{4.20}$$

Finally, substituting eqn (4.20) into (4.19), and using eqn (4.11) where $p_y = p_z\cos\Theta/\sin\Theta$, produces:

$$\frac{\partial p_y}{\partial \Theta} = -p_z\,\frac{m_y}{m_z\cos^2\Theta + m_y\sin^2\Theta}. \tag{4.21}$$

The second row of the Jacobian is then finished off by realising that $U_\perp$, as given in eqn (4.12), has no dependence upon U_x, which immediately produces:

$$\frac{\partial p_y}{\partial U_x} = 0. \tag{4.22}$$

Eqns (4.18), (4.21) and (4.22) provide the solution for the second row of the Jacobian in eqn (4.13).

Moving on to the third row of the Jacobian, and substituting eqn (4.18) into (4.17) yields:

$$\frac{\partial p_z}{\partial U_\perp} = \frac{1}{p_z} \frac{m_y m_z \sin^2\Theta}{m_z \cos^2\Theta + m_y \sin^2\Theta}. \tag{4.23}$$

Then, substituting eqn (4.21) into (4.20) produces:

$$\frac{\partial p_z}{\partial \Theta} = p_y \frac{m_z}{m_z \cos^2\Theta + m_y \sin^2\Theta}. \tag{4.24}$$

Finally, using exactly the same logic that lead to eqn (4.22), gives:

$$\frac{\partial p_z}{\partial U_x} = 0. \tag{4.25}$$

The Jacobian in eqn (4.13) is solved for by using eqns (4.14), (4.15), (4.18), (4.21)-(4.25) to yield:

$$\mathrm{J}\left(\frac{p_x, p_y, p_z}{U_x, \Theta, U_\perp}\right) = \left\| \begin{bmatrix} \frac{m_x}{p_x} & 0 & 0 \\ 0 & -\frac{p_z m_y}{m_z \cos^2\Theta + m_y \sin^2\Theta} & \frac{m_y m_z \cos^2\Theta / p_y}{m_z \cos^2\Theta + m_y \sin^2\Theta} \\ 0 & \frac{p_y m_z}{m_z \cos^2\Theta + m_y \sin^2\Theta} & \frac{m_y m_z \sin^2\Theta / p_z}{m_z \cos^2\Theta + m_y \sin^2\Theta} \end{bmatrix} \right\|. \tag{4.26}$$

Given the sparse nature of the matrix in eqn (4.26), the solution of the determinant quickly yields:

$$\mathrm{J}\left(\frac{p_x, p_y, p_z}{U_x, \Theta, U_\perp}\right) = \frac{m_x}{p_x} \frac{m_y m_z}{m_z \cos^2\Theta + m_y \sin^2\Theta}. \tag{4.27}$$

Eqn (4.27) is the Jacobian that allows the integral definitions in eqns (4.4) and (4.5) to be transformed from $\boldsymbol{p}$ to $\boldsymbol{U}$. As will be seen shortly, this greatly facilitates the development of the models for F_f and F_r.

Maintaining the focus upon eqn (4.4), as set out at the start of this section, and using eqns (4.27) and (4.10) to transform from $\boldsymbol{p}$ to $\boldsymbol{U}$, yields:

$$F_f = \frac{2q}{h^3}\int_{R_1} d^3U J\left(\frac{p_x, p_y, p_z}{U_x, \Theta, U_\perp}\right) f_1(U) W(U_x) \frac{\partial U}{\partial p_x}$$

$$= \frac{2q}{h^3}\int_{R_1} d\Theta dU_x dU_\perp \left(\frac{m_x}{p_x}\frac{m_y m_z}{m_z \cos^2\Theta + m_y \sin^2\Theta}\right) f_1(U_x + U_\perp) W(U_x) \frac{p_x}{m_x}$$

$$= \frac{2q}{h^3}\int_{R_1} d\Theta dU_x dU_\perp \frac{m_y m_z}{m_z \cos^2\Theta + m_y \sin^2\Theta} f_1(U_x + U_\perp)\, W_{CBS}(U_x)\, W_N(U_x) \qquad (4.28)$$

where $R_1 = R_f$ to reflect that F_f originates at $-x_n$ within Region 1. Eqn (4.28) is the full model for transport through the CBS. However, as was stated previously, the effect of a non-uniform m^* has not been included. It is instructive to pause at this point and determine, under simpler conditions, $W_N(U_x)$ and thus the region of integration R_1 before moving on to include the effect of a spatially varying m^*.

With m_x, m_y, and m_z as constants throughout the system, there is no coupling between Θ, $U_\perp$, and U_x, so that all canonical coordinates can be considered independently of each other (this is not the case when m^* is non-constant). Re-examining Fig. 4.2 shows that in the region $0 < x \le x_p$ the potential profile that generates $W_N(U_x)$ is of a strictly monotonically increasing nature (unlike the CBS within the domain $-x_n \le x \le 0$, which contains ΔE_c). Since we are considering a system in which their are no collisions that could either raise or lower the particle's total energy, the particle must emerge from the EB SCR with sufficient energy to enter into the neutral base with an energy that is above E_c; else one would be admitting particle transport within the forbidden bandgap. This fact allows for a considerable simplification to the definition of $W_N(U_x)$; namely:

$$W_N(U_x) = \begin{cases} 1 & \text{if} \quad U_x \ge V_b \\ 0 & \text{if} \quad U_x < V_b \end{cases} . \qquad (4.29)$$

Although strictly speaking eqn (4.29) is not the full form for $W_N(U_x)$, it captures the ultimate result since any particle that enters the neutral base within the forbidden bandgap (*i.e.*, $U_x < V_b$) will within short order be attenuated to the point where it no longer carries any current. Therefore, since we are only interested in calculating the transport current, the exact form for $W_N(U_x)$ is irrelevant, and eqn (4.29) suffices as it captures the essential feature of $W_N(U_x)$.

With $W_N(U_x)$ defined in eqn (4.29), that last task to accomplish before F_f can be solved for by eqn (4.28) is to determine R_1. Re-examining Fig. 4.1, it is obvious that for a particle to enter the

EB SCR at $x = -x_n$ and contribute to F_f, it must possess a positive x-directed momentum. With the energy reference shown in Fig. 4.2, a positive x-directed momentum translates into $p_x \geq 0$. Furthermore, examination of eqn (4.12) shows that $p_x \geq 0$ translates into $U_x \geq 0$ (This is for the case where $m_x > 0$ and therefore applies to electrons. To consider holes, it is best to use a negative hole energy instead of a negative hole mass so that all of the results in this chapter may be applied directly.). If the requirement that $U \equiv U_\perp + U_x \leq E$ be imposed (where E is the bandwidth for $U(p)$), then together with $U_x \geq 0$, and $U_x \geq V_b$ from eqn (4.29), then R_1 will be as shown in Fig. 4.4.

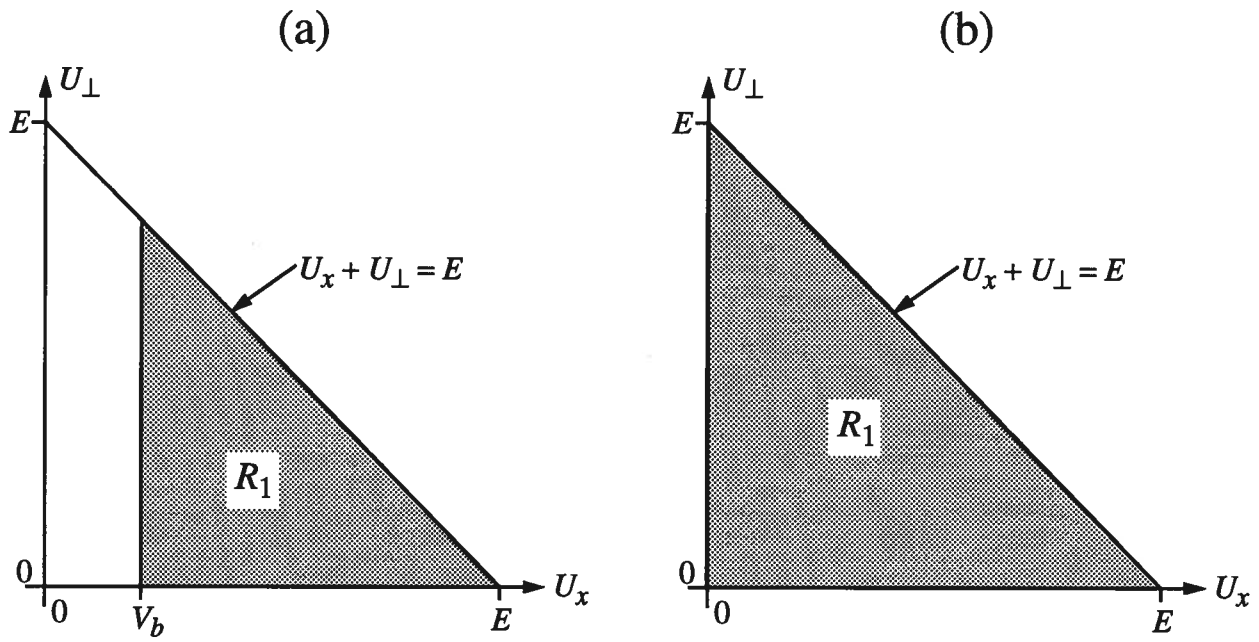


Fig. 4.4. Domain of integration R_1 for a uniform m^*. (a): case where the applied bias is such that $V_b \geq 0$; (b): case where the applied bias is such that $V_b \leq 0$. Note: Fig. 4.2 defines V_b.

F_f can now be solved for by using eqns (4.28), (4.29), (4.9), (4.1), and the region of integration R_1 as shown in Fig. 4.4. Since R_1 takes into account $W_N(U_x)$, then solving eqn (4.28) yields:

$$F_f = \frac{2q}{h^3}\int_0^{2\pi} d\Theta \frac{m_y m_z}{m_z\cos^2\Theta + m_y\sin^2\Theta}\int_{\max(V_b,0)}^{E} dU_x\, W_{CBS}(U_x)\int_0^{E-U_x} dU_\perp f_1(U_x+U_\perp). \qquad (4.30)$$

Examination of eqn (4.30) reveals that the integral over Θ has no dependence on the results of the second and third integrals. This allows the Θ integral to be performed independently to yield:

$$\int_0^{2\pi} d\Theta \frac{m_y m_z}{m_z\cos^2\Theta + m_y\sin^2\Theta} = 4\int_0^{\pi/2} d\Theta \frac{m_y m_z}{m_z\cos^2\Theta + m_y\sin^2\Theta} = 4\sqrt{m_y m_z}\tan^{-1}\left(\sqrt{\frac{m_y}{m_z}}\tan\Theta\right)\Bigg|_0^{\frac{\pi}{2}}$$

The above equation is evaluated by letting Θ approach $\pi/2$ from the left, giving:

$$\int_0^{2\pi} d\Theta \frac{m_y m_z}{m_z \cos^2\Theta + m_y \sin^2\Theta} = 2\pi\sqrt{m_y m_z}. \tag{4.31}$$

Eqn (4.31) solves for the anisotropic effective mass tensor and is evaluated in such a manner than all branch points of the inverse tangent are respected. Therefore, as long as one can assert that the second and third integrals of eqn (4.30) are indeed independent of Θ, then one can substitute eqn (4.31) into (4.30) to obtain:

$$F_f = \frac{4\pi q\sqrt{m_y m_z}}{h^3} \int_{\max(V_b,\,0)}^{E} dU_x \, W_{CBS}(U_x) \int_0^{E-U_x} dU_\perp \, f_1(U_x + U_\perp) \ . \tag{4.32}$$

Eqn (4.32), with the region of integration R_1 as shown in Fig. 4.4, gives us a flavour for the transport current through the CBS. The interpretation of eqn (4.32) yields: a thermalised ensemble of electrons at $x = -x_n$ (characterised by the distribution f_1 with an electrochemical potential μ of $E_{fn,1}$) is injected to the right, towards the CBS; each electron within the ensemble is characterised by a forward-directed energy U_x and a transverse directed energy $U_\perp$ which is random but evenly distributed in all directions; every electron then passes through the CBS with a probability of transmission given by W_{CBS} which is dependent upon U_x alone; the transverse directed portion of the electron's energy leads to a contribution given by the geometric mean of the two transverse effective masses; finally, only electrons that can enter the neutral base outside of the forbidden bandgap (*i.e.*, $U_x \geq V_b$), and are within the bandwidth E of the conduction band, are allowed to contribute to the transport current. Eqn (4.32) solves for F_f under the condition that the effective mass tensor is a constant throughout the CBS.

Returning back to eqn (4.28), the main thrust of this section is continued; namely the incorporation of a spatially varying m^* into the transport current. The inclusion of a non-constant m^* requires that the electron energy U ($\equiv U_\perp + U_x$) be generalised to:

$$U_1 = U_{x,1} + U_{\perp,1} \quad \text{and} \quad U_2 = U_{x,2} + U_{\perp,2}\,, \tag{4.33}$$

where energies with a subscript of 1 refer to transit within Region 1 (*i.e.*, $-x_n \leq x \leq 0$), while energies with a subscript of 2 refer to transit within Region 2 (*i.e.*, $0 < x \leq x_p$). The reason for the generalisation that leads to eqn (4.33) is that the spatial change in the effective mass tensor results in a mixing of the x-directed and transverse directed energies. Therefore, one cannot maintain a to-

tally separate view of U_x and $U_\perp$. Now, the energy reference continues to be located at $E_c(x=-x_n)$, so that using eqn (4.12) produces:

$$U_{x,1} = \frac{p_{x,1}^2}{2m_{x,1}} \quad \text{and} \quad U_{\perp,1} = \frac{p_{y,1}^2}{2m_{y,1}} + \frac{p_{z,1}^2}{2m_{z,1}}, \tag{4.34}$$

while

$$U_{x,2} = \frac{p_{x,2}^2}{2m_{x,2}} + V_b \quad \text{and} \quad U_{\perp,2} = \frac{p_{y,2}^2}{2m_{y,2}} + \frac{p_{z,2}^2}{2m_{z,2}}. \tag{4.35}$$

It is important to understand the exact meaning of eqns (4.33)-(4.35). To begin with, the energies U, U_x, and $U_\perp$ represent total energies within their respective regions. Band diagrams such as those shown in Fig. 4.2 do not show the total energy U, but instead show only U_x. In the event that the system possesses transverse symmetry, then the potential energy is $V(x,y,z) \equiv V(x)$. When there is transverse symmetry, it is possible cast the full three dimensional problem into two decoupled one dimensional problems whose solution only depends upon U_x or $U_\perp$ respectively. For this reason, $U_{x,2}$ is not simply given by the kinetic energy term containing $p_{x,2}$, it must also include the offset potential energy of V_b. Thus, eqn (4.35) gives the total energy U_2 located at $x = x_p$, while eqn (4.34) gives the total energy U_1 located at $x = -x_n$. The reason for defining the energies at $-x_n$ and x_p being that F_f is based upon particles injected to the right from $x = -x_n$, while F_r is based upon particles injected to the left from $x = x_p$. Furthermore, because the potential energy $V(x,y,z) \equiv V(x)$ does not vary in the transverse direction, $U_{\perp,1}$ and $U_{\perp,2}$ do not contain an offset potential energy term.

The cumbersome nature of the energy relations given by eqns (4.34) and (4.35) arise from the quantum mechanical nature of the problem. Looking back to eqn (4.3) shows the flux being calculated by an integration over p-space. Strictly speaking, quantum mechanics does not allow one to consider momentum and position simultaneously. Eqn (4.3) must be interpreted with care, because F_f is based upon a distribution in p-space located at $x = -x_n$, while F_r is based upon a distribution in p-space located at $x = x_p$. Essentially, due to the slow variation of $V(x)$ over the atomic dimensions, it is possible to cast the problem into quasi-classical form [15] where one can speak of distinct p-space distributions at largely separated positions in real space. Finally, because we transform $\boldsymbol{p}$ into $\boldsymbol{U}$, the same concerns for p-space apply to U-space as well.

Due to the translational invariance of the potential V(*x*,*y*,*z*) along the transverse direction, the transverse momentum $p_\perp$ commutes with the Hamiltonian of the system; leading to the conservation of $p_\perp$. Therefore, at the heterojunction separating Region 1 from Region 2 (*i.e.*, at $x = 0$), $p_{\perp,1} \equiv p_{\perp,2}$ (where eqn (4.11) has $p_\perp = \sqrt{p_{y,1}^2 + p_{z,1}^2}$) so that:

$$p_{y,1} = p_{y,2} = p_y \qquad \text{and} \qquad p_{z,1} = p_{z,2} = p_z. \tag{4.36}$$

Since the potential energy V(*x*,*y*,*z*) (≡ V(*x*)) does not vary in the transverse direction, then $p_{\perp,1} \equiv p_{\perp,2}$ cannot vary with x if collisions are prohibited. Using eqn (4.36) in eqns (4.34) and (4.35) shows that $U_{\perp,1}$ and $U_{\perp,2}$ must remain constants of the motion. Therefore, eqn (4.36) must hold equally well at any x within Regions 1 and 2, and more specifically at $-x_n$ and x_p where eqns (4.33)-(4.35) are defined.

Using eqn (4.36) in eqns (4.34) and (4.35) leads to:

$$U_{\perp,1} = \frac{p_y^2}{2m_{y,1}} + \frac{p_z^2}{2m_{z,1}} \quad \text{and} \quad U_{\perp,2} = \frac{p_y^2}{2m_{y,2}} + \frac{p_z^2}{2m_{z,2}}.$$

Applying eqn (4.11) to the above yields, after a little algebraic manipulation:

$$\frac{U_{\perp,1}}{U_{\perp,2}} = \frac{m_{y,2}m_{z,2}}{m_{y,1}m_{z,1}}\mathrm{R}(\Theta) \qquad \text{where} \qquad \mathrm{R}(\Theta) = \frac{m_{z,1}\cos^2\Theta + m_{y,1}\sin^2\Theta}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta}. \tag{4.37}$$

Examination of eqn (4.37) shows the necessary condition that if $m_{y,1} = m_{z,1} = m_{y,2} = m_{z,2}$, then $U_{\perp,1}/U_{\perp,2} = 1$. Eqn (4.37) represents the change in the transverse energy that must occur to conserve $p_\perp$ in the face of a spatially varying effective mass tensor. It is instructive at this point to reveal the full implications of eqn (4.37) upon the total energy within the system. Fig. 4.5 shows the effect of eqn (4.37) when $m_{y,1} = m_{z,1} = m_1$, and $m_{y,2} = m_{z,2} = m_2$. When $m_1 < m_2$, then $U_{\perp,1} > U_{\perp,2}$. As will be described in the next paragraph, total energy must be conserved throughout Regions 1 and 2. Thus, when $m_1 < m_2$, the positive difference $U_{\perp,1} - U_{\perp,2}$ is transferred into $U_{x,2}$ which leads to an enhancement in the forward directed flux. Conversely, when $m_1 > m_2$, then $U_{\perp,1} < U_{\perp,2}$. Thus, when $m_1 > m_2$, the negative difference $U_{\perp,1} - U_{\perp,2}$ is removed from $U_{x,2}$ which leads to an diminution in the forward directed flux.

Since eqn (4.3) is based upon a collision-less system within Regions 1 and 2, then the total energy must be conserved at the heterojunction separating Region 1 from Region 2 (*i.e.*, $x = 0$). Thus,

$$U_1 \equiv U_2.$$

Furthermore, since there are no collisions within the two regions, the above conservation require-

ment applies equally well at all x within Regions 1 and 2, and more specifically at $-x_n$ and x_p where eqns (4.33)-(4.35) are defined. Using the above equation in eqn (4.33) produces:

$$U_{x,1} + U_{\perp,1} \equiv U_{x,2} + U_{\perp,2}. \tag{4.38}$$

Eqns (4.38) and (4.36) are the conservation requirements imposed at the abrupt heterojunction separating Region 1 from Region 2. It is important to remember that most of the proceeding arguments are based upon the conservation of $p_\perp$. This conservation can only be asserted if the Hamiltonian of the entire system has translational symmetry along the transverse spatial dimension. If the heterojunction contains a corrugation or surface roughness, then one could not assert that $p_\perp$ is conserved. This would lead to a considerable increase in the complexity of the model that would necessarily require a detailed view of the device at the atomic level.

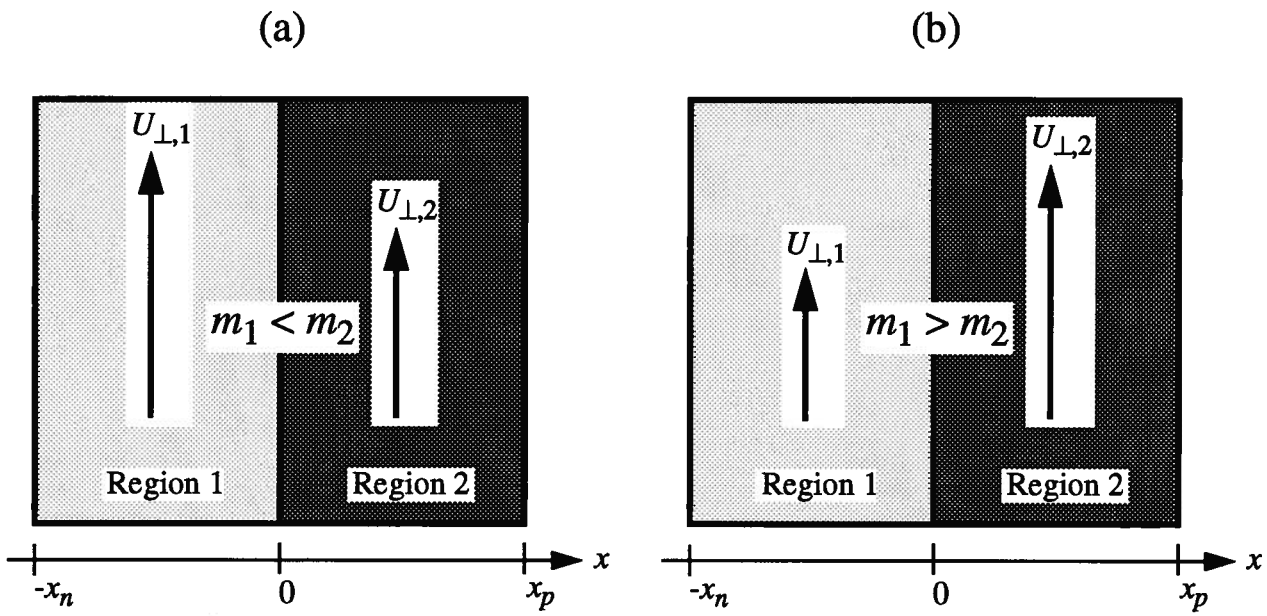


Fig. 4.5. The effect that conservation of $p_\perp$ has upon $U_{\perp,1}$ and $U_{\perp,2}$ when a mass boundary is placed at $x = 0$. Using $m_{y,1} = m_{z,1} = m_1$ and $m_{y,2} = m_{z,2} = m_2$ in eqn (4.37), then $U_{\perp,1}/U_{\perp,2} = m_2/m_1$. (a): when $m_1 < m_2$, energy is removed from $U_{\perp,1}$ and transferred to $U_{x,2}$ when moving from the left to the right; (b): when $m_1 > m_2$, energy is removed from $U_{x,2}$ and transferred to $U_{\perp,2}$ when moving from the left to the right.

With eqns (4.38), (4.37), (4.35) and (4.34), the effect of a spatially varying effective mass tensor can be completed. The abrupt change to the effective mass tensor, as described in Fig. 4.5, results in a mixing of $U_{x,1}$ and $U_{\perp,1}$ with $U_{x,2}$ and $U_{\perp,2}$ when passing through the mass barrier (*i.e.*, heterojunction) at $x = 0$. This mixing, along with the assumption that there are no collisions, results in a one-to-one mapping between energy state ($U_{x,1}$, $U_{\perp,1}$) in Region 1 and energy state ($U_{x,2}$, $U_{\perp,2}$) in Region 2. This mapping is solved for by substituting eqn (4.37) into (4.38), giving:

$$U_{x,2} = U_{x,1} + \gamma(\Theta) U_{\perp,1}, \tag{4.39}$$

and

$$U_{x,1} = U_{x,2} + \gamma'(\Theta) U_{\perp,2}, \tag{4.40}$$

and

$$U = U_{x,1} + U_{\perp,1} = U_{x,2} + U_{\perp,2}, \tag{4.41}$$

where

$$\gamma(\Theta) = 1 - \frac{m_{y,1} m_{z,1}}{m_{y,2} m_{z,2}} R^{-1}(\Theta) \quad \text{and} \quad \gamma'(\Theta) = 1 - \frac{m_{y,2} m_{z,2}}{m_{y,1} m_{z,1}} R(\Theta) = \frac{\gamma(\Theta)}{\gamma(\Theta) - 1}. \tag{4.42}$$

Finally, using this simplified form based on the function γ (the notation for γ was initially set forth by Christov [70,73], but has been extended here to include anisotropic effects), eqn (4.37) becomes:

$$\frac{U_{\perp,1}}{U_{\perp,2}} = \frac{1}{1-\gamma} = 1 - \gamma', \tag{4.43}$$

where the explicit dependence upon Θ has been dropped for simplification. Eqn (4.41) simply asserts the fact that a collision-less system is being considered, while eqns (4.39) and (4.40) represent the energy mapping that occurs when crossing the heterojunction at $x = 0$ from the left or from the right respectively.

Returning back to eqn (4.28) for the calculation of F_f, the integral is being performed over U-space located at $x = -x_n$ with a domain of integration R_1. Using the formalisms for passing through the heterojunction that were developed in eqns (4.39)-(4.43), it is important to realise that the transmission probability $W(U_x)$, as defined in eqn (4.8), must be extended to:

$$W(U_x) = W_{CBS}(U_{x,1}) W_N(U_{x,2}), \tag{4.44}$$

for W_{CBS} is defined in Region 1 and thus depends upon $U_{x,1}$, while W_N is defined in Region 2 and thus depends upon $U_{x,2}$. However, any function that depends upon total energy U (such as the Fermi-Dirac distribution function $f_x(U)$) remains unaffected by the mass barrier due to the conservation of total energy set out in eqn (4.41). Therefore, eqn (4.29) for W_N is rewritten as:

$$W_N(U_{x,2}) = \begin{cases} 1 & \text{if} \quad U_{x,2} \geq V_b \\ 0 & \text{if} \quad U_{x,2} < V_b \end{cases}. \tag{4.45}$$

The domain of integration R_1, which is used for p- or U-space integrations performed at $x = -x_n$, will be modified from what is shown in Fig. 4.4 by the non-uniform effective mass tensor. One still requires that for a particle to enter the EB SCR at $x = -x_n$ and contribute to F_f, it must possess $p_{x,1} \geq 0$; or in terms of energy, $U_{x,1} \geq 0$. And, the requirement that U (= $U_{x,1} + U_{\perp,1}$) $\leq E$ (where E is the bandwidth for $U(p)$) is still maintained. However, eqn (4.45) imposes the condi-

tion that $U_{x,2} \geq V_b$, which results in a coupling between $U_{x,1}$ and $U_{\perp,1}$ when eqn (4.39) is used to map from Region 2 into Region 1. Therefore, using the three boundary conditions set out in this paragraph, along with eqn (4.39), yields the following boundary for R_1:

$$\begin{aligned} U_{x,1} &\geq 0, \\ U_{x,1} + U_{\perp,1} &\leq \mathrm{E}, \\ U_{x,1} + \gamma U_{\perp,1} &\geq V_b. \end{aligned} \tag{4.46}$$

It is also possible to transform R_1 (which is applicable to an integration carried out at $x = -x_n$) into R_2 (which is applicable to an integration carried out at $x = x_p$) by substituting eqns (4.40) and (4.41) into (4.46) to produce the following boundary for R_2:

$$\begin{aligned} U_{x,2} + \gamma' U_{\perp,2} &\geq 0, \\ U_{x,2} + U_{\perp,2} &\leq \mathrm{E}, \\ U_{x,2} &\geq V_b. \end{aligned} \tag{4.47}$$

When the effective mass tensor is uniform, then eqns (4.42) and (4.37) produce $\gamma = \gamma' = 0$. Under these uniform conditions, then indeed eqn (4.46) produces the R_1 as shown in Fig. 4.4. However, when $\gamma \neq \gamma' \neq 0$, R_1 becomes distorted from that shown in Fig. 4.4. γ and γ' can take on any value in the range $-\infty \leq (\gamma,\ \gamma') \leq 1$. As was discussed in the examination of eqn (4.37) that lead to Fig. 4.5, two distinctly different domains occur for γ: firstly, when $m_1 < m_2$ where $0 < \gamma \leq 1$ (and $-\infty < \gamma' < 0$), and energy is transferred from $U_{\perp,1}$ into $U_{x,2}$ which leads to an enhancement in the forward directed flux; secondly, when $m_1 > m_2$ where $-\infty < \gamma < 0$ (and $0 < \gamma' \leq 1$), and energy is removed from $U_{x,2}$ and transferred into $U_{\perp,2}$ which leads to a reduction in the forward directed flux. Fig. 4.6 shows R_1 and R_2 for the case where $\gamma > 0$, while Fig. 4.7 shows R_1 and R_2 for the case where $\gamma < 0$. Examination of Fig. 4.6 shows a focussing of R_2 towards the direction of charge flow. This is due to the energy transfer into $U_{x,2}$ when passing through the heterojunction, leading to what is termed current enhancement. Conversely, examination of Fig. 4.7 shows a reflection in R_1 against the direction of charge flow. This is due to the energy removal from $U_{x,2}$ past the heterojunction, leading to what is termed current reflection. The current reflection occurs because ultimately, no carrier may enter the base within the forbidden bandgap (*i.e.*, $U_{x,2} < V_b$). As a result of Figs. 4.6 and 4.7, care must be exercised in applying the integration boundary R_1 (or R_2) to the solution of F_f in eqn (4.28).

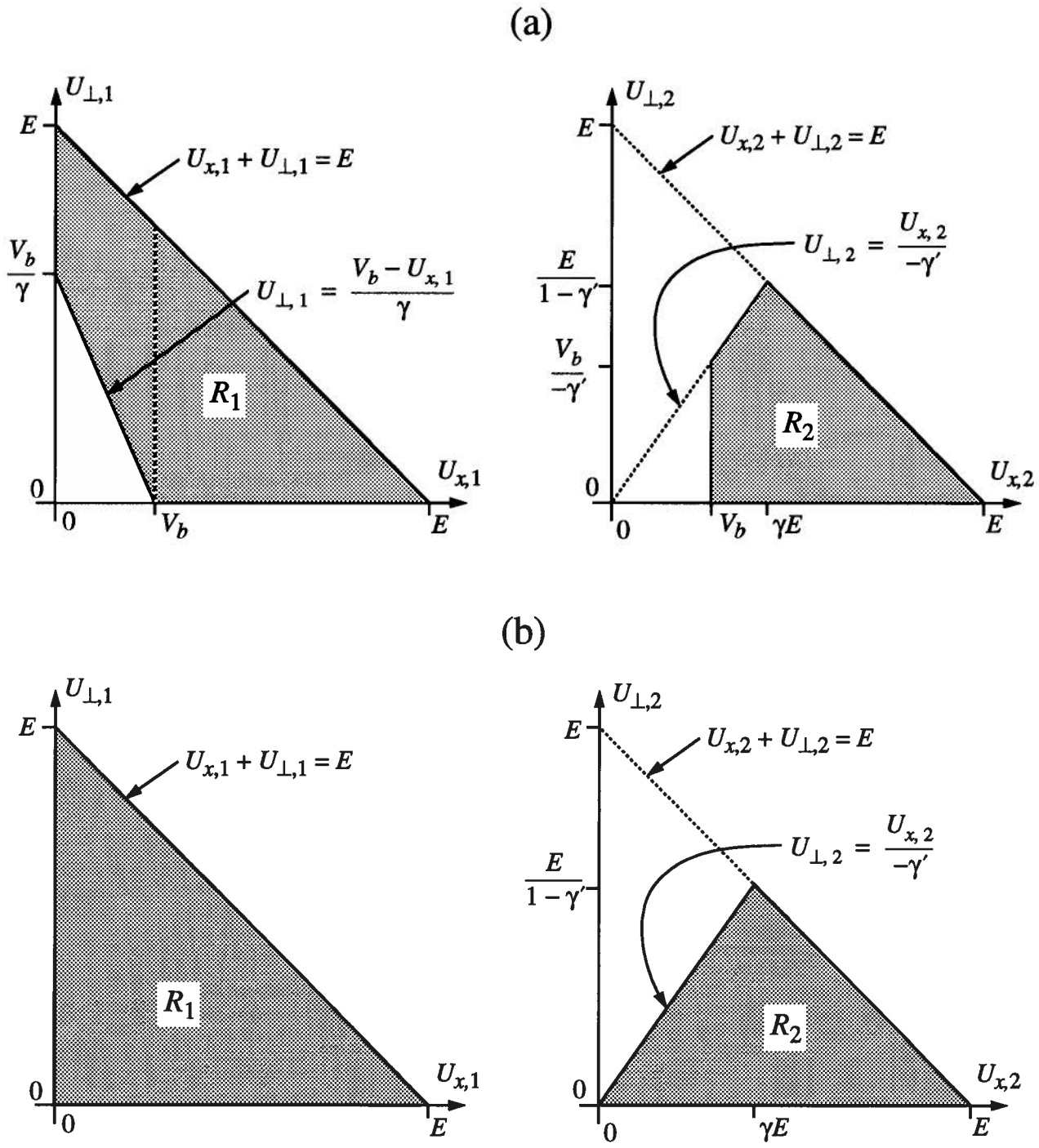


Fig. 4.6. Enhancement case where $m_1 < m_2$ (*i.e.*, $\gamma > 0$ and $\gamma' < 0$). Domains of integration R_1 and R_2 from eqns (4.46) and (4.47) for the calculation of F_f at $x = -x_n$ and $x = x_p$ respectively: (a) the applied bias is such that $V_b \geq 0$; (b) the applied bias is such that $V_b \leq 0$. Each domain of integration represents the ensemble of particles that contribute to F_f. Notice in R_2 how the transfer of energy from $U_{\perp,1}$ into $U_{x,2}$, due to the increasing m^* in the direction of charge flow, leads to a focussing of the particles towards the direction of charge flow.

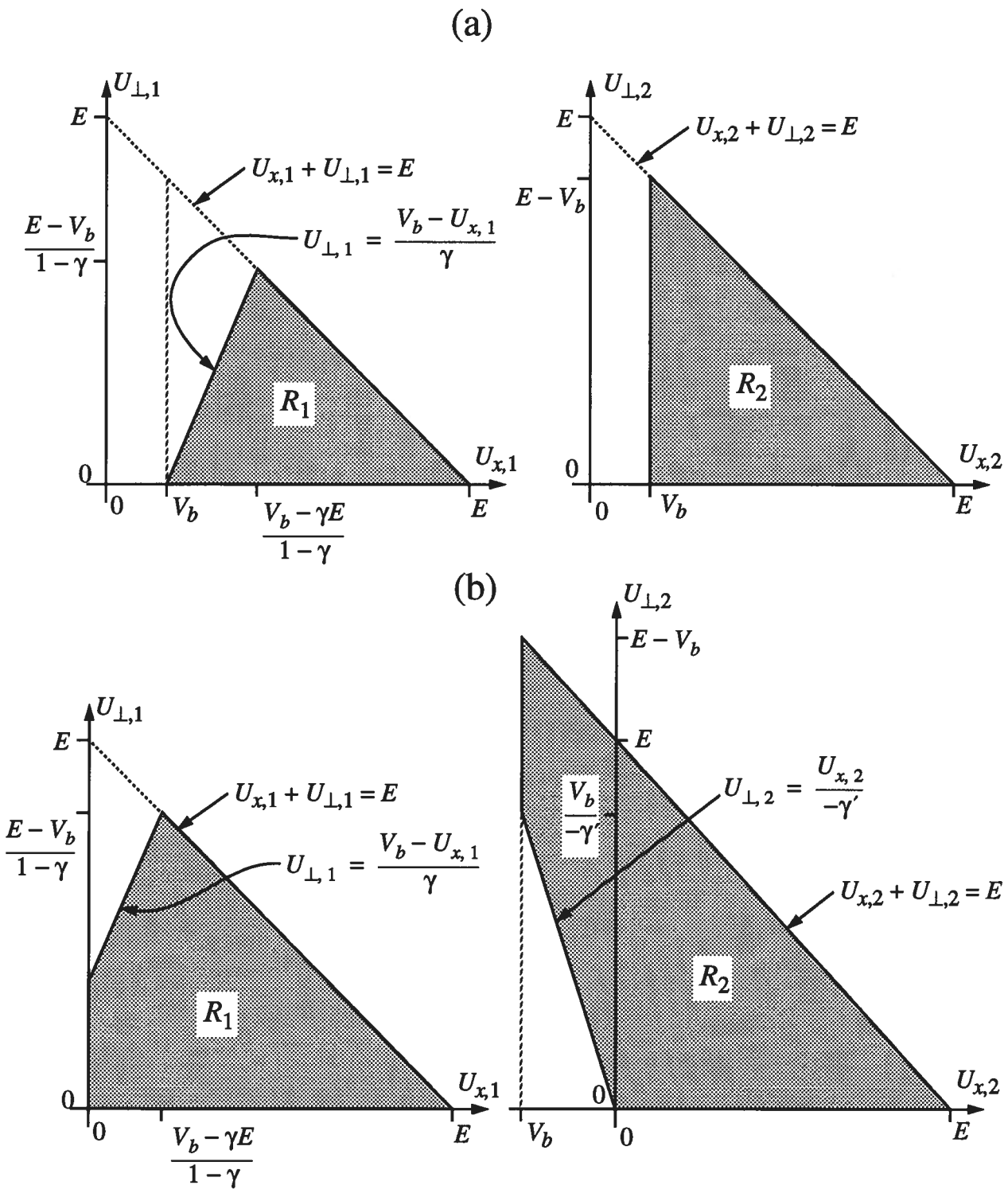


Fig. 4.7. Reflection case where $m_1 > m_2$ (*i.e.*, $\gamma < 0$ and $\gamma' > 0$). Domains of integration R_1 and R_2 from eqns (4.46) and (4.47) for the calculation of F_f at $x = -x_n$ and $x = x_p$ respectively: (a) the applied bias is such that $V_b \geq 0$; (b) the applied bias is such that $V_b \leq 0$. Each domain of integration represents the ensemble of particles that contribute to F_f. Notice in R_1 how the removal of energy from $U_{x,2}$ into $U_{\perp,2}$, due to the decreasing m^* in the direction of charge flow, leads to a reflection of the particles against the direction of charge flow. The reflection occurs because of the necessity for particles to enter the base outside of the forbidden bandgap (*i.e.*, $U_{x,2} \geq V_b$, or $p_{x,2} \geq 0$).

Before eqn (4.28) is recast to include changes to m^* (by including R_1 from Figs. 4.6 and 4.7) it is instructive to calculate the Jacobian that transforms integrations performed within Region 1 into those performed within Region 2. In other words, we wish to determine:

$$J\left(\frac{U_{x,1}, \Theta, U_{\perp,1}}{U_{x,2}, \Theta, U_{\perp,2}}\right) = J\left(\frac{U_{x,1}, U_{\perp,1}}{U_{x,2}, U_{\perp,2}}\right) = \left\| \begin{bmatrix} \frac{\partial U_{x,1}}{\partial U_{x,2}} & \frac{\partial U_{x,1}}{\partial U_{\perp,2}} \\ \frac{\partial U_{\perp,1}}{\partial U_{x,2}} & \frac{\partial U_{\perp,1}}{\partial U_{\perp,2}} \end{bmatrix} \right\|.$$

Using eqns (4.40) and (4.43) produces:

$$J\left(\frac{U_{x,1}, U_{\perp,1}}{U_{x,2}, U_{\perp,2}}\right) = \left\| \begin{bmatrix} 1 & 1 - \frac{m_{y,2}m_{z,2}}{m_{y,1}m_{z,1}}R(\Theta) \\ 0 & \frac{m_{y,2}m_{z,2}}{m_{y,1}m_{z,1}}R(\Theta) \end{bmatrix} \right\|,$$

where $R(\Theta)$ is defined in eqn (4.37). Using eqn (4.37) yields, after substitution into the above:

$$J\left(\frac{U_{x,1}, \Theta, U_{\perp,1}}{U_{x,2}, \Theta, U_{\perp,2}}\right) = J\left(\frac{U_{x,1}, U_{\perp,1}}{U_{x,2}, U_{\perp,2}}\right) = \frac{m_{y,2}m_{z,2}}{m_{y,1}m_{z,1}} \frac{m_{z,1}\cos^2\Theta + m_{y,1}\sin^2\Theta}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta}. \tag{4.48}$$

Finally, by combining the above Jacobian for a change in variables from Region 1 to Region 2 with the Jacobian given by eqn (4.27) for a change in variables from $\boldsymbol{p}$ to $\boldsymbol{U}$ (which in this case is subscripted to reflect calculations within Region 1), gives:

$$J\left(\frac{p_{x,1}, p_{y,1}, p_{z,1}}{U_{x,1}, \Theta, U_{\perp,1}}\right) J\left(\frac{U_{x,1}, U_{\perp,1}}{U_{x,2}, U_{\perp,2}}\right) = J\left(\frac{p_{x,1}, p_{y,1}, p_{z,1}}{U_{x,2}, \Theta, U_{\perp,2}}\right) = \frac{m_{x,1}}{p_{x,1}} \frac{m_{y,2}m_{z,2}}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta} \tag{4.49}$$

Examination of eqn (4.49) shows it to be almost identical to the Region 1 Jacobian in eqn (4.27), but with subscripts denoting Region 2 instead of Region 1. This is to be expected because the energy versus momentum relations in Regions 1 and 2 (eqns (4.34) and (4.35) respectively) differ only by a constant of V_b, which will not result in a deformation of the differential volume element. However, the term $m_{x,1}/p_{x,1}$ and not $m_{x,2}/p_{x,2}$ remains in eqn (4.49). The reason for this discrepancy from perfect symmetry lies in the fact that F_f is an ensemble of particles originating at $x = -x_n$. As such, it is the particle velocity at the point of origin that will dictate the current flux. Once the ensemble population is cast in phase space, then by Liouville's theorem [74], the flux is conserved at all other points in phase space and must equal the current at the point of origin. Therefore, the term $\partial U/\partial p_x$ in eqn (4.4) for F_f remains $\partial U/\partial p_{x,1}$ ($= p_{x,1}/m_{x,1}$) and not $\partial U/\partial p_{x,2}$.

The final transport model for F_f, including the effects of a non-uniform m^*, is presented. For the enhancement case (*i.e.*, $m_1 < m_2$ and $\gamma > 0$), then using eqn (4.28) with calculations based at $x = -x_n$ and R_1 defined in Fig. 4.6, produces that:

$$F_f = \frac{2q}{h^3}\int_0^{2\pi} d\Theta \frac{m_{y,1}m_{z,1}}{m_{z,1}\cos^2\Theta + m_{y,1}\sin^2\Theta}\left[\int_{\max(V_b,0)}^{E} dU_{x,1}\, W_{CBS}(U_{x,1}) \int_0^{E-U_{x,1}} dU_{\perp,1}\, f_1(U_{x,1}+U_{\perp,1})\right. \tag{4.50}$$

$$\left. + \int_0^{\max(V_b,0)} dU_{x,1}\, W_{CBS}(U_{x,1}) \int_{\frac{V_b-U_{x,1}}{\gamma}}^{E-U_{x,1}} dU_{\perp,1}\, f_1(U_{x,1}+U_{\perp,1})\right].$$

The term $W_N(U_{x,2})$ is equal to 1 within the domain R_1 and has been removed for clarity. However, if W_N does not have this simple form, then the full $W_N(U_{x,2}) = W_N(U_{x,1} + \gamma U_{\perp,1})$ must remain in eqn (4.50), where the coupling of the canonical variables forces it to remain nested within the third integral over $U_{\perp,1}$. If this is the case, it may be beneficial to calculate F_f at $x = x_p$. Using R_2 as defined in Fig. 4.6, along with eqns (4.48) and (4.28), produces:

$$F_f = \frac{2q}{h^3}\int_0^{2\pi} d\Theta \frac{m_{y,2}m_{z,2}}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta}\left[\int_{\gamma E}^{E} dU_{x,2}\, W_N(U_{x,2}) \int_0^{E-U_{x,2}} dU_{\perp,2}\, f_1(U_{x,2}+U_{\perp,2})\, W_{CBS}(U_{x,1})\right. \tag{4.51}$$

$$\left. + \int_{\max(V_b,0)}^{\gamma E} dU_{x,2}\, W_N(U_{x,2}) \int_0^{\frac{U_{x,2}}{-\gamma'}} dU_{\perp,2}\, f_1(U_{x,2}+U_{\perp,2}) W_{CBS}(U_{x,1})\right].$$

In this case W_N (= 1 within the domain R_2) has been left in to show its general inclusion for the calculation of F_f. Eqn (4.51) is useful in applications where W_N does not have a simple form. However, W_{CBS} remains nested within the third integral over $U_{\perp,2}$ and cannot be easily removed due to its dependence upon $U_{x,1}$, which by way of eqn (4.40) is equal to $U_{x,2} + \gamma' U_{\perp,2}$.

It should be noted that all of the fluxes considered within this chapter are electron fluxes. Thus, to calculate conventional current densities from these fluxes (such as F_f), one must multiply by "-1".

Finally, for the reflection case (*i.e.*, $m_1 > m_2$ and $\gamma < 0$), then using eqn (4.28) with calculations based at $x = -x_n$ and R_1 defined in Fig. 4.7, produces:

$$F_f = \frac{2q}{h^3}\int_0^{2\pi} d\Theta \frac{m_{y,1}m_{z,1}}{m_{z,1}\cos^2\Theta + m_{y,1}\sin^2\Theta}\left[\int_{\max(V_b,0)}^{E} dU_{x,1}\, W_{CBS}(U_{x,1}) \int_0^{E-U_{x,1}} dU_{\perp,1}\, f_1(U_{x,1}+U_{\perp,1})\right.$$
$$\left. - \int_{\max(V_b,0)}^{\frac{V_b-\gamma E}{1-\gamma}} dU_{x,1}\, W_{CBS}(U_{x,1}) \int_{\frac{V_b-U_{x,1}}{\gamma}}^{E-U_{x,1}} dU_{\perp,1}\, f_1(U_{x,1}+U_{\perp,1})\right]. \tag{4.52}$$

As was done with the enhancement case, the term $W_N(U_{x,2})$ (= 1 within the domain R_1) has been removed for clarity. However, as is true for the enhancement case, if W_N does not have this simple form, then the full $W_N(U_{x,2}) = W_N(U_{x,1} + \gamma U_{\perp,1})$ must remain in eqn (4.52), where the coupling of the canonical variables forces it to stay nested within the third integral over $U_{\perp,1}$. If this is the case, it may be beneficial to calculate F_f at $x = x_p$. Using R_2 as defined in Fig. 4.7 along with eqns (4.48) and (4.28), produces:

$$F_f = \frac{2q}{h^3}\int_0^{2\pi} d\Theta \frac{m_{y,2}m_{z,2}}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta}\left[\int_{\max(V_b,0)}^{E} dU_{x,2}\, W_N(U_{x,2}) \int_0^{E-U_{x,2}} dU_{\perp,2}\, f_1(U_{x,2}+U_{\perp,2})\, W_{CBS}(U_{x,1})\right.$$
$$\left. + \int_{\min(V_b,0)}^{0} dU_{x,2}\, W_N(U_{x,2}) \int_{\frac{U_{x,2}}{-\gamma'}}^{E-U_{x,2}} dU_{\perp,2}\, f_1(U_{x,2}+U_{\perp,2}) W_{CBS}(U_{x,1})\right]. \tag{4.53}$$

Again, as with the enhancement case, eqn (4.53) simplifies the problem of calculations involving a complex W_N, but at the expense of making calculations of W_{CBS} far more complex. Basically, if W_N has a simple form then use either eqn (4.50) or (4.52) for the calculation of F_f under enhancement or reflection respectively. On the other hand, if W_{CBS} has a simple form then use either eqn (4.51) or (4.53) for the calculation of F_f under enhancement or reflection respectively. Finally, if both W_N and W_{CBS} have a complex form then little can be done to reduce the complexity of the problem.

Eqns (4.50)-(4.53) present a rigorous model, that includes the effect of quantum mechanical tunneling, for the calculation of the forward flux entering a two region system with an abrupt mass- and hetero-junction in-between. These equations solve, for the first time, the transport current within a complex region while allowing for an anisotropic media. As such, these equations

represent a significant progression from the models derived by Stratton, Padovani, Christov, Crowell and Rideout [69,70,73,75-78]. The models presented here allow for all of the features found within HBT structures which were not accounted for by the aforementioned authors in their study of Schottky diodes. Furthermore, the models presented here overcome the problem encountered by Perlman and Feucht [79], who solved the same system but neglected tunneling. Due to the neglect of tunneling, the models in [79] have an un-physical discontinuous change when the mass boundary is placed coincidently with the potential boundary. It is important to be able to model transport through complex regions like the CBS, for in modern abrupt HBT structures this transport current is often what defines the ultimate terminal characteristics of the device. Finally, the models presented in this section have no bias toward, or any specific requirement on, any one material system. Therefore, the results of this section can be applied equally well to any material system.

In concluding this section it is important to mention some cautionary comments and shed some physical insight into eqns (4.50)-(4.53). First of all, examination of eqns (4.50) and (4.52) shows the first double integral over $U_{x,1}$ and $U_{\perp,1}$ to be identical in both equations and also equal to eqn (4.30) which is for a constant m^*. For this reason, this double integral is termed the standard forward flux $F_{f,\text{standard}}$ as this is the standard flux that would flow in the absence of the mass barrier. The last double integral in eqn (4.50) represents an additional flux that would normally have entered the base within the forbidden bandgap, but due to the mass boundary transferring energy from $U_{\perp,1}$ into $U_{x,2}$, it is raised up into E_c within the base to contribute to the total F_f. As such, this current is termed the enhancement forward flux $F_{f,\text{enhance}}$. Finally, the last double integral in eqn (4.52) represents a flux that would normally have entered the base within E_c, but due to the mass boundary removing energy from $U_{x,2}$, it is lowered into the forbidden bandgap within the base and is lost from the total F_f. As such this current is termed the reflected forward flux $F_{f,\text{reflect}}$. It is also important to remember that when solving eqns (4.50)-(4.53), γ and γ' have a dependence upon Θ in general. Therefore, unlike eqn (4.30) (and thus $F_{f,\text{standard}}$) where the Θ integration can be treated as an independent multiplier to yield eqn (4.31), the calculation of $F_{f,\text{enhance}}$ and $F_{f,\text{reflect}}$ will have $\gamma(\Theta)$ and $\gamma'(\Theta)$ nested within the integrand, making for a potentially stiff problem to solve due to the complex nature of the Θ integral.

4.3 Calculation of F_r and a Unified Model for F

The total transport flux F is equal to $F_f - F_r$ as is given by eqn (4.2). The models of the previous section, given in eqns (4.50)-(4.53), concentrate on the calculation of F_f. The reason for maintaining a focus upon F_f, while neglecting F_r, is that the two the fluxes are essentially identical, save for a change in the electrochemical potential within the distribution functions f_1 and f_2 used to determine F_f and F_r respectively. Furthermore, under the condition of current-limited-flow due to a given region, eqn (2.2) shows that it is F_f that defines the transport current through that region. However, as was discussed in Section 4.1, before one can assert that F_f and F_r share a dependency that is indicative of eqn (2.2), it is necessary to prove that the regions of integration for F_f and F_r provide for the form given in eqn (2.2). The calculation of F_r, and the ultimate proof that eqn (4.2) (and thus the transport flux through the CBS) has the form of eqn (2.2), begins by returning back to eqn (4.3).

Eqn (4.3) sets out the general models for F, F_f and F_r, but does not explicitly show the effect of a mass boundary. Included within eqn (4.3) is the requirement that tunneling, or any other conduction process for that matter, that moves electrons from one state to another depend upon the probability that the final state be unoccupied (= (1 − f) ≡ h). Using eqn (4.3) for F_f, eqn (4.44) for W, eqn (4.34) for $U_{x,1}$, and the Jacobian given by eqn (4.49) to move calculations to x_p, yields:

$$F_f = \frac{2q}{h^3}\int_{-\infty}^{\infty} dU^f_{x,2}\int_0^{\infty} dU^f_{\perp,2}\int_0^{2\pi} d\Theta \frac{m_{y,2}m_{z,2}}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta} f_1^f(U^f)\ h_2^f(U^f)\, W^f_{CBS}(U^f_{x,1})\, W^f_N(U^f_{x,2}) \tag{4.54}$$

where the superscript f refers to functions that have their energy reference located at the bottom of the conduction band at $x = -x_n$. To arrive at the infinite extent for the region of integration it is only necessary to extend the definitions of W_{CBS}, W_N, and f_1 to implicitly account for the fact that the flux density must be zero outside of the region R_2 defined by eqn (4.47) (*i.e.*, $W_{CBS}(U_{x,1}) \equiv 0$ when $U_{x,1} \le 0$, $W_N(U_{x,2}) \equiv 0$ when $U_{x,2} \le V_b$, and $f_1(U) \equiv 0$ when $U \ge E$). No loss to the generality of these function occurs as a result of this extension. Likewise for F_r, but using only the p to U Jacobian of eqn (4.27) in order to maintain the calculations at x_p, yields:

$$F_r = \frac{2q}{h^3}\int_{-\infty}^{\infty} dU^r_{x,1}\int_0^{\infty} dU^r_{\perp,1}\int_0^{2\pi} d\Theta \frac{m_{y,2}m_{z,2}}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta} f_2^r(U^r)\ h_1^r(U^r)\, W^r_{CBS}(U^r_{x,2})\, W^r_N(U^r_{x,1}) \tag{4.55}$$

where the superscript r refers to functions that have their energy reference located at the bottom of the conduction band at $x = x_p$. Note that in eqn (4.55) the subscripts referring to Regions 1 and 2 have been interchanged to reflect the reverse direction of flow for F_r in comparison to F_f. Therefore, both eqns (4.54) and (4.55) have been constructed so that the integration over U-space occurs at the point $x = x_p$. This will facilitate direct comparison between F_r and F_f.

The task that remains is to recast the r-superscripted functions of eqn (4.55) into the f-superscripted functions of eqn (4.54). The only difference that exists between the f- and r-functions is their energy reference. Since $E_c(x=x_p) - E_c(x=-x_n) = V_b$, and there is transverse symmetry, then using eqns (4.34) and (4.35):

$$U^r_{x,1} = U^f_{x,2} - V_b \quad \text{and} \quad U^r_{\perp,1} = U^f_{\perp,2} \qquad \Rightarrow U^r = U^r_{x,1} + U^r_{\perp,1} = U^f - V_b. \tag{4.56}$$

Finally, recasting eqns (4.39) and (4.40) into r and f form, gives:

$$U^f_{x,1} = U^f_{x,2} + \gamma' U^f_{\perp,2} \quad \text{and} \quad U^r_{x,2} = U^r_{x,1} + \gamma' U^r_{\perp,1}. \tag{4.57}$$

The reason γ' and not γ is used in the definition for $U^r_{x,2}$, is because Regions 1 and 2 are interchanged for the calculation of F_r. This regional interchange maintains consistency with Section 4.2 where the flux always originates in Region 1. With the interchange of Regions 1 and 2, all of the effective masses are also interchanged. Finally, observation of eqns (4.42) and (4.37) shows that interchanging the 1 and 2 subscripts maps γ into γ'. Since all of the functions used in eqns (4.54) and (4.55) are thermodynamically reversible (due to the fact the system is collision-less), then a general function $g^r(U)$ is the same as $g^f(U + V_b)$ (where U can be either r- or f-superscripted). Using this functional translation, along with eqns (4.57) and (4.56) gives:

$$\begin{aligned}
W^r_{CBS}(U^r_{x,2}) = W^f_{CBS}(U^r_{x,2} + V_b) &= W^f_{CBS}(U^r_{x,1} + \gamma' U^r_{\perp,1} + V_b) = W^f_{CBS}(U^f_{x,2} + \gamma' U^f_{\perp,2}) \\
\Rightarrow W^r_{CBS}(U^r_{x,2}) &= W^f_{CBS}(U^f_{x,1}), \\
W^r_N(U^r_{x,1}) &= W^f_N(U^r_{x,1} + V_b) = W^f_N(U^f_{x,2}), \\
f^r_2(U^r) &= f^f_2(U^r + V_b) = f^f_2(U^f), \\
f^r_1(U^r) &= f^f_1(U^r + V_b) = f^f_1(U^f).
\end{aligned}$$

The above equations recast the r-superscripted functions into the desired f-superscripted functions.

Using the above equations, along with the fact that the probability of hole occupancy h is equal to 1 – f, eqn (4.55) becomes:

$$F_r = \frac{2q}{h^3}\int_{-\infty}^{\infty} dU^r_{x,1}\int_0^{\infty} dU^r_{\perp,1}\int_0^{2\pi} d\Theta \frac{m_{y,2}m_{z,2}}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta}\, f_2^f(U^f)\; h_1^f(U^f)\, W^f_{CBS}(U^f_{x,1})\, W^f_N(U^f_{x,2}). \tag{4.58}$$

Then, the only thing left to do before a direct comparison between eqn (4.58) for F_r and eqn (4.54) for F_f can be made, is to determine the Jacobian that transforms $(U^r_{x,1}, U^r_{\perp,1})$ into $(U^f_{x,2}, U^f_{\perp,2})$. Examination of eqn (4.56) shows that the only difference between points in $(U^r_{x,1}, U^r_{\perp,1})$ space and points in $(U^f_{x,2}, U^f_{\perp,2})$ space is a constant V_b. Since the addition of a constant does not distort the differential volume element, the Jacobian is unity. This allows eqn (4.58), along with $U^r_{\perp,1} = U^f_{\perp,2}$, to immediately transform into:

$$F_r = \frac{2q}{h^3}\int_{-\infty}^{\infty} dU^f_{x,2}\int_0^{\infty} dU^f_{\perp,2}\int_0^{2\pi} d\Theta \frac{m_{y,2}m_{z,2}}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta}\, f_2^f(U^f)\; h_1^f(U^f)\, W^f_{CBS}(U^f_{x,1})\, W^f_N(U^f_{x,2}). \tag{4.59}$$

Comparison of eqn (4.59) for F_r and eqn (4.54) for F_f shows almost exactly the same functions; save the fact that F_r deals with transport from Region 2 to Region 1 (*i.e.*, $f_2^f(U^f)\; h_1^f(U^f)$), while F_f deals with transport from Region 1 to Region 2 (*i.e.*, $f_1^f(U^f)\; h_2^f(U^f)$). Therefore, the transport flux is:

$$F = F_f - F_r = \frac{2q}{h^3}\int_{-\infty}^{\infty} dU^f_{x,2}\int_0^{\infty} dU^f_{\perp,2}\int_0^{2\pi} d\Theta \frac{m_{y,2}m_{z,2}}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta}\cdot$$

$$[\, f_1^f(U^f)\; h_2^f(U^f) - f_2^f(U^f)\; h_1^f(U^f)\,]\, W^f_{CBS}(U^f_{x,1})\, W^f_N(U^f_{x,2})$$

$$= \frac{2q}{h^3}\int_{-\infty}^{\infty} dU^f_{x,2}\int_0^{\infty} dU^f_{\perp,2}\int_0^{2\pi} d\Theta \frac{m_{y,2}m_{z,2}}{m_{z,2}\cos^2\Theta + m_{y,2}\sin^2\Theta}\cdot$$

$$[\, f_1^f(U^f) - f_2^f(U^f)\,]\, W^f_{CBS}(U^f_{x,1})\, W^f_N(U^f_{x,2})\ . \tag{4.60}$$

The *f* superscripts have been included as a reminder that the energy reference is located at the bottom of the conduction band at $x = -x_n$.

Eqn (4.60) completes the proof that F_f and F_r share a dependency that is indicative of eqn (2.2). It also validates the modified definitions for F_f and F_r given by eqns (4.4) and (4.5) respectively. Eqn (4.60) is brought into exact agreement with eqn (2.2) when the f_1 and f_2 distribution functions of eqn (4.1) are given by the Boltzmann approximation, leading to:

$$f_1(U) = \frac{1}{1+e^{\frac{U-\mu_1}{kT}}} \approx e^{-\frac{U-\mu_1}{kT}}$$
$$f_2(U) = \frac{1}{1+e^{\frac{U-\mu_2}{kT}}} \approx e^{-\frac{U-\mu_2}{kT}} \tag{4.61}$$

Eqn (4.61), under the Boltzmann approximation, produces:

$$f_1^f(U^f) - f_2^f(U^f) = e^{-\frac{U}{kT}} e^{\frac{\mu_1}{kT}} \left(1 - e^{-\frac{\mu_1-\mu_2}{kT}}\right) = f_1^f(U^f)\left(1 - e^{-\frac{\Delta E_{fn}}{kT}}\right), \tag{4.62}$$

where $\Delta E_{fn} \equiv \mu_1 - \mu_2$. Since ΔE_{fn} is a constant with respect to the canonical variables defining the integration in eqn (4.60), then substituting eqns (4.62) and (4.54) into (4.60) gives:

$$F = F_f - F_r = F_f\left(1 - e^{-\frac{\Delta E_{fn}}{kT}}\right). \tag{4.63}$$

Thus, the transport flux through the CBS has exactly the same form as eqn (2.2). This will allow the models of this chapter to be used with the results of Chapter 2. Eqn (4.63) also justifies the methodology used within this chapter where F_f alone is calculated. Finally, examination of eqn (4.63) shows that it possesses two simple but fundamental requirements: as the driving force ΔE_{fn} increases, so does F increase; when the system is at equilibrium ($\Delta E_{fn} \equiv 0$), the transport flux vanishes.

4.4 Analytic CBS Transport Models

Section 4.2 presented the general models for the calculation of the transport flux F_f through a complex two region system with an abrupt mass barrier in-between. The models also allow for an anisotropic effective mass tensor m^*. This section will take the models of Section 4.2 (eqns (4.50)-(4.53)) and derive analytic solutions for the calculation of F through the CBS. By obtaining analytic models, and not simply resorting to numerical calculation, the rich interplay that exists between the physical attributes such as doping concentration, temperature, effective mass, electron affinity, and bias conditions, will be brought out for study in the final transport model of the CBS. The key component to all of the models presented in this chapter is the inclusion of the effects due to tunneling. Any model or simulator (such as the highly acclaimed Monte Carlo simu-

lator) that fails to account for the vast increase in transport current through the CBS due to tunneling, will be grossly inaccurate even if every conceivable scattering process and other driving force outside of tunneling is accounted for (see Fig. 4.8).

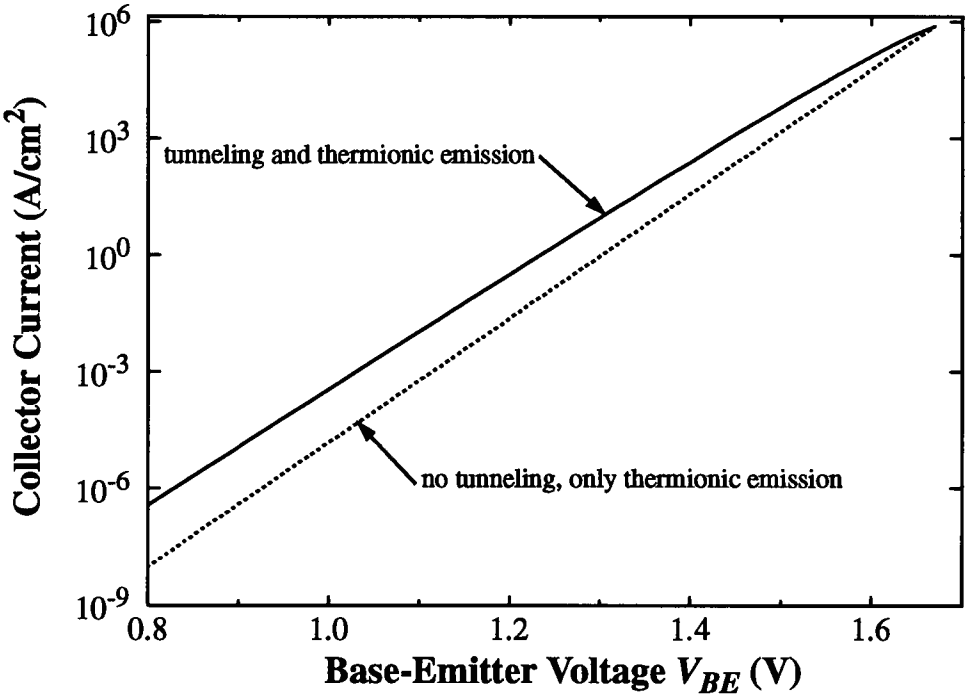


Fig. 4.8. Collector current for an abrupt AlGaAs HBT with 30% Al content in the emitter. The emitter doping is 5×10^{17}cm^{-3}, and the base doping is 1×10^{19}cm^{-3}. Notice the large error that results if the tunneling current through the CBS is not accounted for. Also, the tunneling current has a bias dependence that alters the current to voltage relationship from the form $\exp(qV_{BE}/kT)$ (which characterises the thermionic emission curve quite well) to $\exp(qV_{BE}/nkT)$, where $n > 1$.

4.4.1 Analytic Model for the Standard Flux $F_{f,s}$

With the result of eqn (4.63), the development returns to the main goal of this section; deriving analytic models for F_f from eqns (4.50)-(4.53). For the problems being considered, the form of W_N in eqn (4.45) suggests that eqn (4.50) be used for the enhancement case (*i.e.*, $m_1 < m_2$ and $\gamma > 0$), and eqn (4.52) be used for the reflection case (*i.e.*, $m_1 > m_2$ and $\gamma < 0$). As was discussed near the very end of Section 4.2, eqns (4.50) and (4.52) share a common term called $F_{f,\text{standard}}$ (or $F_{f,s}$ for short), plus a unique term for the enhancement case of $F_{f,\text{enhance}}$ (or $F_{f,e}$ for short), and a unique term for the reflection case of $F_{f,\text{reflect}}$ (or $F_{f,r}$ for short). These terms, using eqns (4.50) and (4.52) are:

$$F_{f,s} = \frac{2q}{h^3}\int_0^{2\pi} d\Theta \frac{m_{y,1}m_{z,1}}{m_{z,1}\cos^2\Theta + m_{y,1}\sin^2\Theta}\int_{\max(V_b,0)}^{E} dU_{x,1}\, W_{CBS}(U_{x,1})\int_0^{E-U_{x,1}} dU_{\perp,1}\, f_1(U_{x,1}+U_{\perp,1}), \quad (4.64)$$

$$F_{f,e} = \frac{2q}{h^3}\int_0^{2\pi} d\Theta \frac{m_{y,1}m_{z,1}}{m_{z,1}\cos^2\Theta + m_{y,1}\sin^2\Theta}\int_0^{\max(V_b,0)} dU_{x,1}\, W_{CBS}(U_{x,1})\int_{\frac{V_b-U_{x,1}}{\gamma}}^{E-U_{x,1}} dU_{\perp,1}\, f_1(U_{x,1}+U_{\perp,1}), \quad (4.65)$$

$$F_{f,r} = \frac{2q}{h^3}\int_0^{2\pi} d\Theta \frac{m_{y,1}m_{z,1}}{m_{z,1}\cos^2\Theta + m_{y,1}\sin^2\Theta}\int_{\max(V_b,0)}^{\frac{V_b-\gamma E}{1-\gamma}} dU_{x,1}\, W_{CBS}(U_{x,1})\int_{\frac{V_b-U_{x,1}}{\gamma}}^{E-U_{x,1}} dU_{\perp,1}\, f_1(U_{x,1}+U_{\perp,1}). \quad (4.66)$$

The derivation of the analytic models begins with $F_{f,s}$. $F_{f,s}$ is the most important term, and as it will turn out, the essential equation for the solution of $F_{f,r}$ as well.

The analytic solution of eqn (4.64) for $F_{f,s}$ begins by noting that the integrals over $U_{x,1}$ and $U_{\perp,1}$ contain no term with a dependence upon Θ. This allows the Θ integral to be performed independently, as in eqn (4.31), to yield the same result as eqn (4.32) but with $m_y = m_{y,1}$ and $m_z = m_{z,1}$. Essentially repeating eqn (4.32), but with a change to the dummy variables in eqns (4.64), yields after performing the integration over $U_\perp$ using the full Fermi-Dirac distribution:

$$F_{f,s} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}kT}{h^3}\int_{\max(V_b,0)}^{E} dU_x\, W_{CBS}(U_x)\ln\left(\frac{1+e^{-\frac{U_x-\mu_1}{kT}}}{1+e^{-\frac{E-\mu_1}{kT}}}\right).$$

The integrand above becomes vanishingly small (at an exponential rate) for large U_x, allowing for a simplification by letting $E \to \infty$ to produce:

$$F_{f,s} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}kT}{h^3}\int_{\max(V_b,0)}^{\infty} dU_x\, W_{CBS}(U_x)\ln\left(1+e^{-\frac{U_x-\mu_1}{kT}}\right).$$

In general, even if the emitter is degenerately doped, the energies U_x at which the above integrand produces significant contributions to $F_{f,s}$ occurs at energies where U_x is a few kT larger than μ_1. This allows what is essentially an assertion of the Boltzmann approximation that leads to eqn (4.61), so that:

$$F_{f,s} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}kT}{h^3} e^{\frac{\mu_1}{kT}} \int_{\max(V_b,0)}^{\infty} dU_x \, W_{CBS}(U_x) e^{-\frac{U_x}{kT}}. \tag{4.67}$$

Eqn (4.67) provides for the model of the standard flux, where the integrand multiplied by the leading constants is the standard flux density.

Eqn (4.67) is now solved for by substituting in W_{CBS} from eqn (4.9) and making a change of variables from absolute energy U_x to normalised energy U_x' (where $U_x' = U_x/V_{pk}$, and V_{pk} is the height of the CBS as defined in Fig. 4.2). Before performing these changes to eqn (4.67), the solution process is further facilitated by the following change of variables:

$$x = \frac{\sqrt{1-U_x'}+1}{\sqrt{U_x'}} \Rightarrow \qquad U_x' = \left(\frac{2x}{1+x^2}\right)^2 \quad \text{and} \quad \sqrt{1-U_x'} = \frac{x^2-1}{x^2+1}.$$

Letting

$$x = e^y \Rightarrow \qquad U_x' = \frac{1}{ch^2(y)} \quad \text{and} \quad \sqrt{1-U_x'} = th(y),$$

where ch(y) is the hyperbolic cosine of y, and th(y) is the hyperbolic tangent of y. Using the above equations, along with the normalised energies from the start of the paragraph, yields for $V_b < V_{pk}$:

$$F_{f,s} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}\,kT\,V_{pk}}{h^3} e^{\frac{\mu_1}{kT}} \left[\int_{\max(V_b',0)}^{1} dU_x' \, e^{\frac{x_n\sqrt{2m_{x,1}V_{pk}}}{\hbar}\left(\frac{y}{ch^2(y)} - th(y)\right) - \frac{V_{pk}}{kT\,ch^2(y)}} + \frac{kT}{V_{pk}} e^{-\frac{V_{pk}}{kT}} \right] \tag{4.68}$$

where all energies, including V_b', are in terms of normalised energy (*i.e.*, $V_b = V_b' V_{pk}$). The last term inside of the square brackets is the thermionic injection term where $W_{CBS} = 1$. In the event that $V_b' > 1$ (*i.e.*, $V_b > V_{pk}$), then the CBS is at an energy too low to effect the transport current and:

$$F_{f,s} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}\,(kT)^2}{h^3} e^{\frac{\mu_1}{kT}} e^{-\frac{V_b}{kT}}.$$

Up to this point, the parameters x_n (which is the n-side extent of the EB SCR) and V_{pk} (which is the n-side portion of the potential drop across the EB SCR) have been left as is without connection to the material parameters of the device (where the device is arbitrarily chosen as an npn HBT). However, using the depletion approximation gives [24,80]:

$$V_{pk} = qN_{rat}(V_{bi} - V_{BE}), \quad V_p = q(1 - N_{rat})(V_{bi} - V_{BE}), \Rightarrow \quad \frac{V_p}{V_{pk}} = \frac{\varepsilon_1 N_D}{\varepsilon_2 N_A}$$

$$V_b = q(V_{bi} - V_{BE}) - \Delta E_c, \qquad x_n = \sqrt{\frac{2\varepsilon_1 V_{pk}}{q^2 N_D}}, \; x_p = \sqrt{\frac{2\varepsilon_2 V_p}{q^2 N_A}} \Rightarrow \frac{x_p}{x_n} = \frac{N_D}{N_A} \tag{4.69}$$

$$\text{where } N_{rat} = \frac{\varepsilon_2 N_A}{\varepsilon_2 N_A + \varepsilon_1 N_D} \quad \text{and} \quad V_{bi} = \frac{kT}{q}\ln\left(\frac{N_A N_D}{n_{i,2}^2}\right) + \frac{\Delta E_c}{q}.$$

V_{bi} is the built-in potential of the junction, $n_{i,2}$ is the intrinsic carrier concentration in Region 2, N_D is the emitter doping, N_A is the base doping, ε_i is the permittivity of the respective region, and V_{BE} is the forward bias across the EB junction. The doping ratio N_{rat} differs slightly from that in eqn (3.5) due to a nonuniform ε. Concentrating on the case $V_b < V_{pk}$, then using eqn (4.69) within eqn (4.68), along with

$$y = U_p + r \qquad \text{where} \qquad U_p = \frac{\hbar}{2V_t}\sqrt{\frac{N_D}{\varepsilon_1 m_{x,1}}} \quad \text{gives:} \tag{4.70}$$

$$F_{f,s} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}\,kT\,V_{pk}}{h^3} e^{\frac{\mu_1}{kT}} \left[\int_{\max(V_b', 0)}^{1} dU_x' \, e^{\frac{N_{rat}(V_{bi}-V_{BE})}{U_p V_t}\left(\frac{r}{\mathrm{ch}^2(U_p+r)} - \mathrm{th}(U_p+r)\right)} + \frac{kT}{V_{pk}} e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}} \right] \tag{4.71}$$

where V_t is the thermal voltage kT/q, and $U_x' = \mathrm{ch}^{-2}(U_p + r)$. As will be shown shortly, eqn (4.71) can be solved in a tractable and analytic fashion. However, the integrand within eqn (4.71) is still the flux density, and is worthy of separate investigation. It is worthwhile to note that eqn (4.71), and the transform used to obtain it, follows that of Crowell and Rideout [78] used in the development of Schottky diodes. Furthermore, U_p is the V_t normalised version of E_{00} from [75].

The standard forward flux density $\Phi_{f,s}$ for a given energy U_x' is:

$$\Phi_{f,s}(U_x') = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}\,kT\,V_{pk}}{h^3} e^{\frac{\mu_1}{kT}} e^{\frac{N_{rat}(V_{bi}-V_{BE})}{U_p V_t}\left(\frac{r}{\mathrm{ch}^2(U_p+r)} - \mathrm{th}(U_p+r)\right)}, \tag{4.72}$$

where the energy U_x' ($= \mathrm{ch}^{-2}(U_p + r)$) is defined in terms of r. The energy at which the maximum $\Phi_{f,s}$ occurs can be found directly from eqn (4.72). In terms of the variable r, and given that exponentials are analytic functions, $\Phi_{f,s}$ will be at a maximum when the exponent containing r in eqn

(4.72) is at a maximum. To this end it is found that:

$$\frac{d}{dr}\left(\frac{r}{ch^2(U_p+r)} - th(U_p+r)\right) = -\frac{2r sh^3(U_p+r)}{ch^3(U_p+r)} \to 0 \Rightarrow \qquad r = 0, -U_p, \pm\infty. \tag{4.73}$$

sh(y) is the hyperbolic sine of y. Examination of the definition for U_x', in terms of r, shows that r has a range of $-U_p \le r < \infty$. Furthermore, when $r = -U_p$ then $U_x' = 1$, which corresponds to the top of the CBS, and when $r \to \infty$ then $U_x' = 0$ (it should be noted that $U_x' < 1$ deals with the tunneling of electrons through the CBS while $U_x' > 1$ deals with thermionic injection over the CBS). The solutions of $r = -U_p$ and $-\infty$ occur due to the mapping used to define U_x', in terms of r, and do not represent the absolute maximum that is being sought. Thus, the maximum $\Phi_{f,s}$ occurs when $r = 0$ and gives:

$$\Phi_{max} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}\,kT\,V_{pk}}{h^3} e^{\frac{\mu_1}{kT}} e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}\frac{th(U_p)}{U_p}} \qquad \text{at} \qquad U'_{max} = ch^{-2}(U_p). \tag{4.74}$$

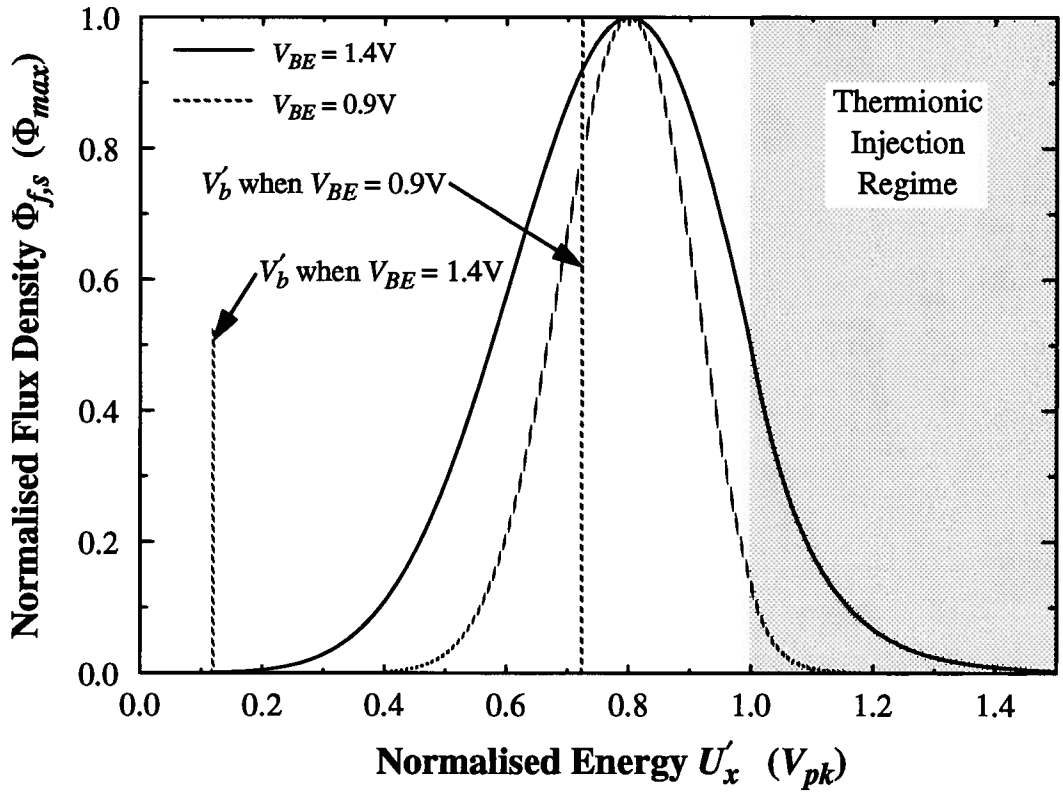


Fig. 4.9. Flux density $\Phi_{f,s}$, normalised to Φ_{max}, for an $Al_{0.3}Ga_{0.7}As/GaAs$ abrupt HBT at two different forward biases. The material parameters, the same as in Fig. 4.8, are: emitter doping N_D $5\times10^{17}cm^{-3}$; base doping N_A $1\times10^{19}cm^{-3}$; emitter permittivity ε_1 $11.9\varepsilon_0$; ΔE_c is 0.24 eV; $n_{i,2}$ is $2.25\times10^6cm^{-3}$; V_{bi} is 1.671 V; $m_{x,1}$ is $0.091m_0$; T is 300K. Note that energies $U_x' < V_b'$ would enter the base within the forbidden bandgap, and although displayed here are reflected in reality.

As was also found in [78], eqn (4.74) presents a surprising result that the energy U'_{max} at which the peak flux density Φ_{max} occurs is independent of the applied bias. Therefore, relative to the top of the CBS, Φ_{max} occurs at the same place regardless of the applied bias (see Fig. 4.9). Further consideration of U'_{max} reveals the following general traits: as U_p (from eqn (4.70)) increases from 0 towards infinity, U'_{max} moves from 1 towards zero, and tunneling becomes increasingly dominant over thermionic emission; as N_D increases, or ε_1 decreases, the width x_n of the CBS decreases and U'_{max} becomes smaller, showing that tunneling is increasing; as $m_{x,1}$ decreases the probability of tunneling should increase, as is confirmed by the associated reduction in U'_{max}; also, as temperature decreases, U'_{max} becomes smaller since it is easier for electrons to tunnel through the barrier then it is to obtain enough thermal energy to pass overtop of the CBS; finally, in the limit as $\hbar$ goes to zero, the system should evolve to a state that is purely describable by classical mechanics, and it is found that U'_{max} goes to 1, which indicates that there is indeed no tunneling. Therefore, the general traits of the flux density, as presented, follow physical expectations.

Returning to the solution of eqn (4.71), the integration over U'_x is converted into an integration over ζ. Using eqn (4.73), it is found that for:

$$\zeta = \frac{r}{\mathrm{ch}^2(U_p + r)} - \mathrm{th}(U_p + r), \qquad \frac{d\zeta dr}{dr dU'_x} = \left(-\frac{2r\,\mathrm{sh}(U_p + r)}{\mathrm{ch}^3(U_p + r)}\right)\left(-\frac{\mathrm{ch}^3(U_p + r)}{2\,\mathrm{sh}(U_p + r)}\right) = r,$$

and then eqn (4.71) becomes (under the condition that $V_b < V_{pk}$):

$$F_{f,s} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}\,kT\,V_{pk}}{h^3} e^{\frac{\mu_1}{kT}}\left[\int dr \frac{d\zeta}{dr}\frac{1}{r} e^{\frac{N_{rat}(V_{bi}-V_{BE})}{U_p V_t}\zeta} + \frac{kT}{V_{pk}} e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}}\right]. \qquad (4.75)$$

Eqn (4.75) has had the limits of integration from eqn (4.71) temporarily removed for clarity. At this point no approximations have been introduced into the solution. At issue with the solution of eqn (4.75) is that $r(\zeta)$ cannot be determined in closed form. If $\zeta(r)$ were invertible then eqn (4.75) could potentially be solved analytically. Observation of Fig. 4.9 shows that $\Phi_{f,s}$, the integrand of eqn (4.71), has the form of a Gaussian. Indeed, $\Phi_{f,s}$ is extremely symmetric and suggests that a Taylor series expansion about U'_{max} (*i.e.*, $r = 0$) for $\zeta(r)$ is a potentially good approximation. Performing a Taylor expansion of $\zeta(r)$ about $r = 0$ up to second order produces:

$$\zeta \approx -r^2\frac{\mathrm{sh}(U_p)}{\mathrm{ch}^3(U_p)} - \mathrm{th}(U_p) \Rightarrow \qquad \frac{d\zeta}{dr} \approx -2r\frac{\mathrm{sh}(U_p)}{\mathrm{ch}^3(U_p)}.$$

Finally, substituting the above approximate equation for $\zeta(r)$ back into the integral within eqn

(4.75) yields:

$$\int dr \frac{d\zeta}{dr}\frac{1}{r} e^{\frac{N_{rat}(V_{bi}-V_{BE})}{U_p V_t}\zeta} = -2\frac{\mathrm{sh}(U_p)}{\mathrm{ch}^3(U_p)} e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}\frac{\mathrm{th}(U_p)}{U_p}} \int dr e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{U_p V_t}\frac{\mathrm{sh}(U_p)}{\mathrm{ch}^3(U_p)} r^2}.$$

The above equation is simply the integration of a Gaussian, and results in an error-function solution. With the limits of integration from eqn (4.71) reintroduced, the solution of the above is:

$$\sqrt{\pi}\frac{\mathrm{sh}(U_p)}{\mathrm{ch}^3(U_p)}\sigma_r e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}\frac{\mathrm{th}(U_p)}{U_p}}\left[\mathrm{erf}\left(\frac{U_p}{\sigma_r}\right)+\mathrm{erf}\left(\frac{\left(\mathrm{ach}\left(\frac{1}{\sqrt{\max(V_b',0)}}\right)-U_p\right)}{\sigma_r}\right)\right], \quad (4.76)$$

where

$$\sigma_r = \sqrt{\frac{\mathrm{ch}^3(U_p)U_p\ \mathrm{k}T}{V_{pk}\ \mathrm{sh}(U_p)}}.$$

Eqn (4.76) solves for the integral in eqn (4.75) and produces the analytic model for $F_{f,s}$ that is sought after. The complexity of eqn (4.76) stems mainly from the evaluation of the boundary conditions. Fig. 4.9 shows $\Phi_{f,s}$ and the boundaries of integration. As long as the majority of $\Phi_{f,s}$ is contained within the two boundaries, then the error functions will both approach 1, and eqn (4.76) can be approximated by:

$$2\sqrt{\pi}\frac{\mathrm{sh}(U_p)}{\mathrm{ch}^3(U_p)}\sigma_r e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}\frac{\mathrm{th}(U_p)}{U_p}}. \quad (4.77)$$

Eqn (4.77) is the simplified model for the integral in eqn (4.75), but it still contains most of the important features regarding CBS transport. Thus, the final and approximate models for $F_{f,s}$ are found by substituting either eqn (4.76) or eqn (4.77) respectively, into the integral of eqn (4.75) to obtain (under the condition that $V_b < V_{pk}$) (see also eqn (4.92) for low temperature considerations):

$$F_{f,s} = F_{f,s0}\sqrt{\frac{\pi V_{pk}\mathrm{sh}(U_p)U_p V_t}{q\mathrm{ch}^3(U_p)}} e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}\frac{\mathrm{th}(U_p)}{U_p}}\left[\mathrm{erf}\left(\frac{U_p}{\sigma_r}\right)+\right. \quad (4.78)$$

$$\left.+\mathrm{erf}\left(\frac{\left(\mathrm{ach}\left(\frac{1}{\sqrt{\max(V_b',0)}}\right)-U_p\right)}{\sigma_r}\right)\right]+F_{f,s0}V_t e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}}$$

or approximately as

$$F_{f,s} \approx 2F_{f,s0}\sqrt{\frac{\pi V_{pk}\mathrm{sh}(U_p)U_pV_t}{\mathrm{qch}^3(U_p)}}\,\mathrm{e}^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}\frac{\mathrm{th}(U_p)}{U_p}} + F_{f,s0}V_t\,\mathrm{e}^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_t}},$$

where

$$F_{f,s0} = \frac{4\pi q^2\sqrt{m_{y,1}m_{z,1}}\,kT}{h^3}\mathrm{e}^{\frac{\mu_1}{kT}}.$$

Finally, under the condition where $V_b \geq V_{pk}$:

$$F_{f,s} = F_{f,s0}V_t\mathrm{e}^{-\frac{V_{bi}-V_{BE}-\Delta E_c/q}{V_t}}. \tag{4.79}$$

4.4.2 Analytic Model for the Enhancement Flux $F_{f,e}$

With the analytic model for $F_{f,s}$ presented in eqns (4.78)-(4.79), attention is focussed upon the solution of the enhancement term $F_{f,e}$. Examination of eqn (4.65) shows that the integration over $U_{\perp,1}$ has a lower limit that includes $\gamma(\Theta)$. Thus, unlike the solution for $F_{f,s}$, the Θ integration to calculate $F_{f,e}$ cannot be performed independently. Further, eqns (4.42) and (4.37) show that γ has a complex dependence upon Θ that would most likely cause the final integration over Θ, for the calculation of $F_{f,e}$, to become analytically intractable. To alleviate this complexity an approximation is made. So far, all of the models presented use a general mass tensor that is diagonal with respect to the direction of transport. This general mass tensor formulation is maintained, but the mass barrier will be confined to the study of an isotropic change in the transverse direction of the mass tensor. Thus, $m_{y,2} = a_m m_{y,1}$ and $m_{z,2} = a_m m_{z,1}$. With this approximation, then using eqns (4.42) and (4.37) it is found that:

$$\gamma(\Theta) = 1 - \frac{m_{y,1}m_{z,1}}{a_m^2 m_{y,1}m_{z,1}}\left(\frac{a_m m_{z,1}\cos^2\Theta + a_m m_{y,1}\sin^2\Theta}{m_{z,1}\cos^2\Theta + m_{y,1}\sin^2\Theta}\right) = 1 - \frac{1}{a_m}. \tag{4.80}$$

Eqn (4.80) reduces γ (and also γ') to a constant. With this simplification, the Θ integral in eqn (4.65) can be performed independently using eqn (4.31). Then, the development of $F_{f,e}$ will follow exactly the one for the calculation of $F_{f,s}$ but with a slight modification to the limits of integration. Therefore, using eqn (4.67), but with the limits of integration obtained from eqn (4.65), yields:

$$F_{f,e} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}kT}{h^3}\mathrm{e}^{\frac{\mu_1}{kT}}\mathrm{e}^{-\frac{V_b}{\gamma kT}}\int_0^{\max(V_b,0)} dU_x\, W_{CBS}(U_x)\mathrm{e}^{-\frac{U_x}{kT}\frac{\gamma-1}{\gamma}}.$$

Examination of the above equation shows that T inside of the integral can be redefined with:

$$T_{eff} = T\frac{\gamma}{\gamma - 1} = T\gamma' = T(1 - a_m) \quad \text{where} \quad a_m = \frac{m_{y,2}}{m_{y,1}} = \frac{m_{z,2}}{m_{z,1}}. \tag{4.81}$$

T_{eff} is then the effective temperature of the flux density. Under the enhancement case $\gamma > 0$ and thus $a_m > 1$, leading to $T_{eff} < 0$. With eqn (4.81) substituted into the equation preceding it, then:

$$F_{f,e} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}kT}{h^3} e^{\frac{\mu_1}{kT}} e^{-\frac{V_b}{\gamma kT}} \int_0^{\max(V_b, 0)} dU_x \, W_{CBS}(U_x) e^{-\frac{U_x}{kT_{eff}}}. \tag{4.82}$$

Eqn (4.82) is the same as eqn (4.67) except the limits of integration are slightly different. However, the effective temperature of the flux density is now negative. The effect of the negative temperature T_{eff} is to cause an increase to the electron distribution as one proceeds to higher energies. This leads to a condition of population inversion that is similar to what is found in lasers. The solution of eqn (4.82) does indeed follow the one presented for $F_{f,s}$, but the fact that $T_{eff} < 0$ must be accounted for. Population inversion, when combined with the fact that W_{CBS} also increases with increased energy, means that the peak flux density will no longer occur at an energy of U'_{max} given in eqn (4.74), but will instead occur at the upper energy boundary allowed into the problem.

The integral inside of eqn (4.71), although derived for the solution of $F_{f,s}$, will solve eqn (4.82) for $F_{f,e}$ when the limits of integration from eqn (4.82) are employed. However, it no longer makes sense to use an expansion that is centred about U'_{max}, as population inversion moves the peak flux density to an energy of $\max(V_b, 0)$. Eqn (4.82) is solved by returning to eqn (4.71) and introducing T_{eff} into all relevant equations to yield (under the condition that $V_b < V_{pk}$):

$$F_{f,e} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}kT}{h^3} e^{\frac{\mu_1}{kT}} e^{-\frac{V_b}{\gamma kT}} V_{pk} \int_0^{\max(V'_b, 0)} dU'_x \, e^{\frac{V_{pk}/q}{U_{p,eff}V_{t,eff}}\left(\frac{r}{ch^2(U_{p,eff}+r)} - th(U_{p,eff}+r)\right)} \tag{4.83}$$

where

$$U_{p,eff} = \frac{\hbar}{2V_{t,eff}}\sqrt{\frac{N_D}{\varepsilon_1 m_{x,1}}} \quad \text{and} \quad V_{t,eff} = \frac{kT_{eff}}{q}.$$

The primes on the energies still denote normalisation with respect to V_{pk}. Unlike the development that took eqn (4.71) into (4.75), eqn (4.83) is expanded about V'_b. Furthermore, the condition of population inversion causes the integrand in eqn (4.83) to become basically exponential in terms of U'_x. Remembering from eqn (4.71) that $U'_x = ch^{-2}(U_{p,eff} + r)$, and $\sqrt{1 - U'_x} = th(U_{p,eff} + r)$,

then a Taylor expansion about $\max(V_b, 0)$ for the exponent inside of the integral of eqn (4.83), up to and including linear terms, is:

$$\frac{V_{pk}/q}{U_{p,eff}V_{t,eff}}\left(\frac{r}{\mathrm{ch}^2(U_{p,eff}+r)}-\mathrm{th}(U_{p,eff}+r)\right)\approx\frac{V_{pk}/q}{U_{p,eff}V_{t,eff}}\left(r_b U_x' - \sqrt{1-\max(V_b',0)}\right), \quad (4.84)$$

where

$$r_b = \mathrm{ach}\left(\frac{1}{\sqrt{\max(V_b',0)}}\right) - U_{p,eff}.$$

The final model for $F_{f,e}$ is arrived at by substituting eqn (4.84) into eqn (4.83) and solving. The only concern when performing this integration is to ensure that $V_b < V_{pk}$. If $V_b > V_{pk}$, then the integral in eqn (4.83) is broken down into two integrals: one integral from 0 up to 1 (remember, normalised energies are being used so that $U_x' = 1$ corresponds to $U_x = V_{pk}$); and a second integral from 1 up to V_b' (over which $W_{CBS} = 1$). Finally for $V_b < V_{pk}$:

$$F_{f,e} = F_{f,s0}\frac{U_{p,eff}V_{t,eff}}{r_b}e^{-\frac{V_b}{\gamma kT}}e^{-\frac{V_{pk}\sqrt{1-\max(V_b',0)}}{qU_{p,eff}V_{t,eff}}}\left(e^{\frac{V_{pk}r_b\max(V_b',0)}{qU_{p,eff}V_{t,eff}}}-1\right), \quad (4.85)$$

while for $V_b \geq V_{pk}$:

$$F_{f,e} = F_{f,s0}V_{t,eff}e^{-\frac{V_b}{\gamma kT}}\left(1-e^{-\frac{V_b}{qV_{t,eff}}}\right). \quad (4.86)$$

As a final check on the validity of the model for $F_{f,e}$ (*i.e.*, eqns (4.85)-(4.86)), observation of eqn (4.65) and the region of integration in Fig. 4.6 shows that as $\gamma \to 0^+$, $F_{f,e} \to 0$. This occurs because when $\gamma = 0$ there is no mass barrier and $F = F_{f,s}$. Obviously, when $V_b \leq 0$, the upper limit of integration in eqn (4.82) is zero and the integral itself vanishes. For the case where $V_b > 0$, examination of eqn (4.82) shows that the terms containing γ are:

$$e^{\frac{U_x(1-\gamma)-V_b}{\gamma kT}}.$$

The enhancement case is being considered, where $0 < \gamma < 1$. Furthermore, since the limits of integration have it that $0 < U_x < V_b$, then $U_x(1-\gamma) - V_b < -\gamma U_x < 0$. Therefore, the terms that makeup the exponent of the above equation are always negative. Then, as γ approaches 0 from the positive side, the exponent goes to negative infinity and eqn (4.82) goes to zero. The exact same development occurs for eqns (4.85) and (4.86), so that the previous argument is applicable, and eqns (4.85)-(4.86) do indeed vanish as $\gamma \to 0^+$.

4.4.3 Analytic Model for the Reflection Flux $F_{f,r}$

With the analytic model for $F_{f,e}$ presented in eqns (4.85)-(4.86), attention is finally focussed upon the solution of the reflection term $F_{f,r}$. Eqn (4.66) is the general model for $F_{f,r}$, and it also contains γ within the $U_{\perp,1}$ as well as the $U_{x,1}$ integrations. Therefore, as was the case with the solution of $F_{f,e}$, the Θ integration to calculate $F_{f,r}$ cannot be performed independently. To simplify this problem, as was done with $F_{f,e}$, the mass barrier is assumed to consist of an isotropic change in the transverse direction of the mass tensor. This allows eqn (4.80) to be used in the solution of $F_{f,r}$. In fact, using the same basic steps from eqns (4.80) to (4.82) will also solve for $F_{f,r}$. The only change that occurs is to the upper limit of integration over $U_{x,1}$, which will approach infinity as $E \rightarrow \infty$ (this is because $\gamma < 0$ for the reflection case). The final result is:

$$F_{f,r} = \frac{4\pi q\sqrt{m_{y,1}m_{z,1}}kT}{h^3} e^{\frac{\mu_1}{kT}} e^{-\frac{V_b}{\gamma kT}} \int_{\max(V_b,\,0)}^{\infty} dU_x \, W_{CBS}(U_x) e^{-\frac{U_x}{kT_{eff}}}. \tag{4.87}$$

Eqn (4.87) is identical to eqn (4.82) save the limits of integration. This fact occurs because of the symmetry of the problem being considered. As was stated before, $F_{f,s}$ is the standard flux that would flow if there was no mass barrier at all. $F_{f,e}$ on the other hand, is the flux of carriers that would normally enter the base within the forbidden bandgap (*i.e.*, $U_{x,2} < V_b$), but due to the mass barrier, is raised up into the conduction band to contribute to the total flux; thus the integration is carried out from $0 < U_{x,1} < V_b$. Finally, $F_{f,r}$ is the flux of carriers that would normally enter the base within the conduction band (*i.e.*, $U_{x,2} > V_b$), but due to the mass barrier, is lowered down into the forbidden bandgap to become reflected and take away from the total flux; thus the integration is carried out from $V_b < U_{x,1} < \infty$. The form of the integral for $F_{f,e}$ and $F_{f,r}$ must be the same since the Jacobian transforms and the boundary conditions given in eqns (4.46)-(4.47) do not depend on the sign of γ.

Even though the models for $F_{f,e}$ and $F_{f,r}$ are ostensibly identical, their analytic solutions are not. This occurs because in $F_{f,r}$, T_{eff} is positive (the same as for $F_{f,s}$). In fact, eqn (4.87) is identical to eqn (4.67) for $F_{f,s}$, except the temperature of the flux density is no longer T but T_{eff} (there is also a constant multiplier of $\exp(-V_b/\gamma kT)$ that occurs in eqn (4.87) that is not present in the model for $F_{f,s}$). Examination of eqn (4.81) shows that when $\gamma < 0$ (as it is for the reflection case), then T_{eff} has a range of $0 < T_{eff} < T$; where $T_{eff} \rightarrow 0$ as $\gamma \rightarrow 0$, and $T_{eff} \rightarrow T$ as $\gamma \rightarrow -\infty$. Therefore, the flux

density in the reflection case is characterised by a temperature that is always less than the lattice temperature T, but unlike the enhancement case it remains positive under all conditions. Thus, the reflection case is identical to, and can be calculated by, the standard case but with a flux density characterised by T_{eff} instead of T (of course, the $\exp(-V_b/\gamma kT)$ term must also be included).

With T_{eff} instead of T used for the flux density in eqn (4.67), along with the $\exp(-V_b/\gamma kT)$ term, the final model for the reflection case becomes (under the condition that $V_b < V_{pk}$):

$$F_{f,r} = F_{f,r0}\sqrt{\frac{\pi V_{pk}\mathrm{sh}(U_{p,eff})U_{p,eff}V_{t,eff}}{q\mathrm{ch}^3(U_{p,eff})}}e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_{t,eff}}\frac{\mathrm{th}(U_{p,eff})}{U_{p,eff}}}\left[\mathrm{erf}\left(\frac{U_{p,eff}}{\sigma_{r,eff}}\right)+\right. \tag{4.88}$$

$$\left.+\mathrm{erf}\left(\frac{\left(\mathrm{ach}\left(\frac{1}{\sqrt{\max(V_b',0)}}\right)-U_{p,eff}\right)}{\sigma_{r,eff}}\right)\right]+F_{f,r0}V_{t,eff}e^{-\frac{N_{rat}(V_{bi}-V_{BE})}{V_{t,eff}}}$$

where

$$F_{f,r0} = F_{f,s0}e^{-\frac{V_b}{\gamma kT}} \quad \text{and} \quad \sigma_{r,eff} = \sqrt{\frac{\mathrm{ch}^3(U_{p,eff})U_{p,eff}V_{t,eff}}{(V_{pk}/q)\ \mathrm{sh}(U_{p,eff})}}.$$

Finally, when $V_b \geq V_{pk}$, then:

$$F_{f,r} = F_{f,r0}V_{t,eff}e^{-\frac{V_{bi}-V_{BE}-\Delta E_c/q}{V_{t,eff}}}. \tag{4.89}$$

Eqns (4.88)-(4.89) present the analytic model for $F_{f,r}$, which is basically the same as the model for $F_{f,s}$ but with the flux density characterised by T_{eff}. The only potential issue (as concerns error due to approximation) with eqn (4.88) (and eqn (4.78) as well) occurs at very low temperatures where tunneling is extremely large. Observation of Fig. 4.9 shows that for V_{BE} = 0.9V, the lower limit of integration is approaching the point at which the peak flux density occurs. However, when the temperature is reduced from 300K to 77K, then U'_{max} moves from 0.80 down to 0.086 (relative to V_{pk}), and the lower limit of integration ends up past the peak flux density. When the peak flux density occurs outside of the region of integration, error will begin to occur with the model because the model is based upon a Taylor expansion about U'_{max}. This potential error at low temperature is exacerbated in the calculation of $F_{f,r}$ because T_{eff} is even less than T (for an $Al_{0.3}Ga_{0.7}As$ to GaAs flux, $m_{x,1} = 0.092m_e$ and $m_{x,2} = 0.067m_e$, so that T_{eff} = 81K when T

= 300K). The solution to this problem is to perform the Taylor expansion of the integrand in eqn (4.87) for $F_{f,r}$ and eqn (4.67) for $F_{f,s}$ about the lower limit of integration; namely max(V_b, 0).

Fortunately, in the course of solving the enhancement case, the desired expansion about V_b has already been performed. Eqn (4.84) is the expansion about V_b up to and including linear terms. If the second order terms are included, then using the transform preceding eqn (4.75) gives:

$$\frac{r}{\mathrm{ch}^2(U_{p,eff}+r)} - \mathrm{th}(U_{p,eff}+r) \approx r_b U_x' - \sqrt{1-V_b'} - (U_x' - V_b')^2 \frac{1}{4V_b'\sqrt{1-V_b'}}, \qquad (4.90)$$

where the condition that $V_b' > 0$ is assured as this expansion is being used to solve the case where $V_b' > U_{max}'$. Substituting eqn (4.90) into eqn (4.83), but using the limits of integration set out in eqn (4.87), produces, after performing the integral over the Gaussian [81,#3.322.2]:

$$F_{f,r} = F_{f,r0} \frac{U_{p,eff} V_{t,eff} \sqrt{\pi}}{\sigma_{r,eff}} e^{\frac{V_{pk}/q}{U_{p,eff} V_{t,eff}} (r_b V_b' - \sqrt{1-V_b'})} e^{\frac{r_b^2}{\sigma_{r,eff}^2}} \left[1 - \mathrm{erf}(-\frac{r_b}{\sigma_{r,eff}}) \right], \qquad (4.91)$$

where $F_{f,r0}$ is defined in eqn (4.88), and $\sigma_{r,eff}$ is altered from its definition in eqn (4.88) to:

$$\sigma_{r,eff} = \sqrt{\frac{U_{p,eff} V_{t,eff}}{(V_{pk}/q)\, V_b' \sqrt{1-V_b'}}}.$$

Eqn (4.91) solves for $F_{f,r}$ when $U_{max}' < V_b' < 1$, and is used instead of eqn (4.88). Eqn (4.88) is used only when $V_b' < U_{max}'$ (which is generally the case except under very low temperatures, or if the heterojunction is such that ΔE_c is quite small).

In a similar fashion, eqn (4.78) for the calculation of $F_{f,s}$ is further restricted to $V_b' < U_{max}'$. Then, when $U_{max}' < V_b' < 1$ occurs, $F_{f,s}$ is given by (after a simple extension from eqn (4.91)):

$$F_{f,s} = F_{f,s0} \frac{U_p V_t \sqrt{\pi}}{\sigma_r} e^{\frac{V_{pk}/q}{U_p V_t} (r_b V_b' - \sqrt{1-V_b'})} e^{\frac{r_b^2}{\sigma_r^2}} \left[1 - \mathrm{erf}(-\frac{r_b}{\sigma_r}) \right], \qquad (4.92)$$

where, in this case only:

$$\sigma_r = \sqrt{\frac{U_p V_t}{(V_{pk}/q)\, V_b' \sqrt{1-V_b'}}} \quad \text{and} \quad r_b = \mathrm{ach}\left(\frac{1}{\sqrt{V_b'}}\right) - U_p.$$

Eqn (4.92), in concert with eqns (4.78)-(4.79) form the model for $F_{f,s}$ with an unrestricted placement of the base barrier potential V_b, and the ability to model very low temperatures. Likewise, eqns (4.88)-(4.89) and (4.91) form the complete model for $F_{f,r}$. Finally, without any further extensions, eqns (4.85)-(4.86) form the model for $F_{f,e}$.

Before leaving this section a cautionary note regarding the numerical calculation of eqns (4.91) and (4.92) is in order. As V_b' surpasses U'_{max} by more that 3 $\sigma_{r,eff}$ (or σ_r), then the term 1 – erf(x) (which is the complementary error function) rapidly approaches zero. One must ensure that the numerical code that generates erf(x) has the proper asymptotic form or else the result will be incorrectly forced to zero (*i.e.*, $1 - \text{erf}(x) \rightarrow e^{-x^2} / (x\sqrt{\pi})$). Analytically, as $\sigma_{r,eff}$ (or σ_r) $\rightarrow 0$, then by simply using the asymptotic form for 1 – erf(x), eqn (4.91) is seen to become eqn (4.85) for $F_{f,e}$, where the "–1" term in eqn (4.85) is dropped; this result is expected because under these conditions the linear Taylor expansion is sufficient.

4.5 The Effect of Emitter-Base SCR Control on I_C

The previous section presented the analytic models for the calculation of the forward flux F_f, and included the mass boundary effects. The only assumption made in the development of the models of the previous section was that the mass boundary be isotropic in terms of the transverse directed effective mass terms. In the event a material system is studied where this is not true, where such a system must posses an indirect bandgap because an anisotropic effective mass tensor is required, then the models of the previous section can be used, but the final Θ integration must be performed using the general models of eqns (4.50)-(4.53) given at the end of Section 4.2. This section will connect the models of the previous section together to simulate an abrupt HBT where the CBS is responsible for current-limited-flow. This will provide insight into the models and allow for the effect of the mass boundary to be fully explored.

Returning back to eqn (2.6) for a three-section device, the collector current density will be equal to J_T. Let the simulated device be governed by the CBS in Section 1 (where $J^0_{n,1} = F_{f,CBS}$), the neutral base in Section 2 ($J^0_{n,2} = F_{f,base}$), and the collector in Section 3 ($J^0_{n,3} = F_{f,coll}$). As long as the demanded currents in the base and collector greatly exceed what the CBS can provide (*i.e.*, $F_{f,base}$ and $F_{f,coll} \gg F_{f,CBS}$), then if no significant recombination occurs throughout the base and collector sections (*i.e.*, $\gamma_2 = \gamma_3 \approx 1$), eqn (2.6) produces:

$$J_C = J_T \approx F_{f,CBS} = \begin{cases} F_{f,s} & \text{if} \quad \gamma = 0 \\ F_{f,s} + F_{f,e} & \text{if} \quad \gamma > 0 \\ F_{f,s} - F_{f,r} & \text{if} \quad \gamma < 0 \end{cases} . \tag{4.93}$$

where the multiplication of the electron flux by "-1" is not required due to the definition of J_C.

It is very interesting to see that when the CBS is responsible for current-limited-flow, I_C will peer directly into the quantum mechanical nature of the CBS. Thus, the quantum mechanical effect of tunneling, including the effects of the mass barrier at the heterojunction itself, will be observable by simply measuring I_C.

The simulated HBT will be based essentially on the following AlGaAs/GaAs HBT at 300K: emitter is $Al_{0.3}Ga_{0.7}As$; base is GaAs; emitter doping N_D $5\times10^{17}cm^{-3}$; base doping N_A $1\times10^{19}cm^{-3}$; emitter permittivity ε_1 is $11.9\varepsilon_0$; base permittivity ε_2 is $12.9\varepsilon_0$; ΔE_c is 0.24eV; $n_{i,2}$ is $2.25\times10^{6}cm^{-3}$; m_1 is $0.092m_0$; m_2 is $0.067m_0$; $\rightarrow N_{rat}$ is 0.956; V_{bi} is 1.671V; $x_n(V_{BE}=0)$ is 649Å; U_p is 0.488; U'_{max} is 0.795; γ is -0.373; T_{eff} is 81.5K; $U_{p,eff}$ is 1.80; $U'_{max,eff}$ is 0.104; V_b is >0 when $V_{BE} < 1.431$V. Two other plausible devices are also considered for the reflection case; in order to make the comparisons direct, all parameters are identically maintained except m_2 is either lowered to $\frac{1}{2}$ of m_1 ($= 0.046m_0$), or to $\frac{1}{4}$ of m_1 ($= 0.023m_0$). The enhancement case typically does not occur for electrons, but most certainly occurs for holes. Using the reciprocal relations to the reflection case gives m_2: $0.126m_0$; $0.184m_0$; $0.368m_0$. Changes to the effective density of states due to the changing m_2 are not reflected into V_{bi} nor ΔE_c. Therefore, the simulations that are about to be presented are contrived in terms of a physical analogue but as such allow for the most direct observation and comparison, regarding CBS transport, that is possible.

Beginning with the reflection case, Fig. 4.10 plots $F_{f,s}$ as well as $F_{f,r}$, using the analytic models of the previous section for the three m_2 cases of: $0.067m_0$; $0.046m_0$; $0.023m_0$. At T equal to 300K as well as 200K, decreasing m_2 (and thus making γ a larger negative number) results in an increase to $F_{f,r}$. Physically, as m_2 decreases, the mass barrier will demand a larger transfer of energy from $U_{x,2}$ into $U_{\perp,2}$ in order to conserve transverse momentum (see Fig. 4.5); thus, a larger number of particles will be reflected as they will not possess a sufficient amount of $U_{x,2}$ energy to satisfy the momentum conservation requirements and enter the neutral base. Furthermore, as V_{BE} is increased, $F_{f,r}$ begins to decrease and then decrease quite rapidly. The physical cause for this is the interplay between the base potential V_b and the mass barrier. As was just stated, the mass barrier moves energy from $U_{x,2}$ into $U_{\perp,2}$. The point at which reflection occurs is when $U_{x,2} < V_b$. Obviously, as V_b is made smaller, more energy can be removed from $U_{x,2}$ without encountering reflection. Since V_b decreases as V_{BE} increases then $F_{f,r}$ must decrease, relative to $F_{f,s}$, as V_{BE} increases. The sudden decrease in $F_{f,r}$ for $V_{BE} > 1.4$V corresponds to the point at which V_b goes below the reference poten-

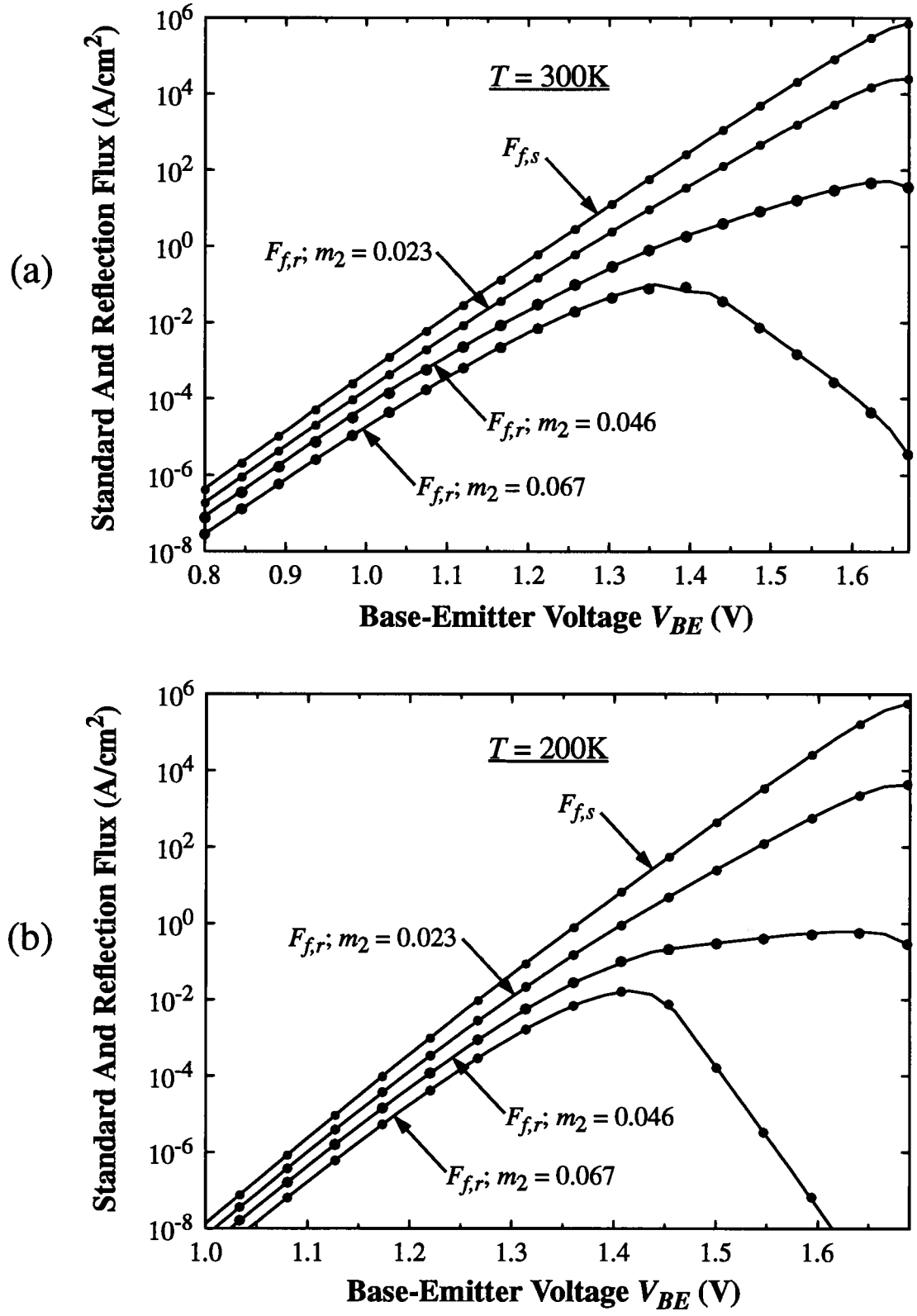


Fig. 4.10. Standard Flux $F_{f,s}$ and Reflection Flux $F_{f,r}$ for an HBT with the parameters given near the start of this section. The only parameter being varied is the base side effective mass m_2. The lines are obtained from the analytic models of eqns (4.78) (4.79) (4.92) for $F_{f,s}$ and eqns (4.88) (4.89) (4.91) for $F_{f,r}$ while the solid dots are from the numerical calculation of eqn (4.67) for $F_{f,s}$ and eqn (4.87) for $F_{f,r}$. (a) results for $T = 300$K. (b) results for $T = 200$K.

tial energy E_c in the neutral emitter (V_b is < 0 when V_{BE} is > 1.431 V). Since the neutral emitter generates the flux that impinges upon the CBS, very few particles will have $U_{x,2}$ reduced below zero by the mass barrier (unless the mass barrier is very strong due to a small m_2/m_1). Thus, once V_b decreases below zero, reflection will taper off quickly as there are essentially no more particles to reflect from the V_b barrier.

Looking now at Fig. 4.11(b), as T is reduced from 300K to 200K, there is an increase in $F_{f,r}$ relative to $F_{f,s}$ at low bias where $V_b > 0$. The physical explanation for this fact is more complex. First of all, any particle where $U_{\perp,1}$ is zero will be unaffected by the mass barrier because momentum conservation is guaranteed when $p_\perp$ is zero (see eqn (4.39)). This means that only particles where $-\gamma U_{\perp,1}$ is comparable to, or larger than, $U_{x,1}$ will be affected by the mass barrier. Now, to tunnel through the potential barrier requires that the particle obtain a sufficient $U_{x,1}$ in order to pass through the CBS (on average an energy of $U'_{max} V_{pk}$ is required). Any energy gained by $U_{\perp,1}$ will do nothing to improve the particle's chances of passing through the barrier; in fact it will only serve to lower the particle's availability because the occupancy decreases exponentially with any increase in total energy. Thus, the CBS preferentially picks out, from the random ensemble of particles impinging upon the barrier, those particles that possess a sufficiently high $U_{x,1}$ to pass through the barrier, while being blind to the amount of $U_{\perp,1}$ contained by each particle. Since U'_{max} decreases rapidly along with a decrease in T, $-\gamma U_{\perp,1}$ will become larger relative to $U_{x,1}$ as T decreases, and the mass barrier will cause a larger reflection flux.

Maintaining the focus upon Fig. 4.11, the effect of the mass boundary can be seen quite readily. In Fig. 4.11(a) the temperature is held constant and all three mass cases are presented. This clearly shows that as the mass barrier is strengthened by reducing m_2, the relative importance of $F_{f,r}$ rapidly increases. Perhaps even more importantly, the effect that $F_{f,r}$ has on the total flux F is bias dependent. This shows that the mass barrier cannot be described by a simple multiplicative constant as has been suggested in the literature [51,79,82]. Another important feature that is clearly brought out in both Figs. 4.11(a) and (b) is that for $V_{BE} > 1.43$ (which corresponds to $V_b < 0$), the effect of $F_{f,r}$ is negligible. As was discussed earlier, once $V_b < 0$ there will be few particles left that can reflect from the potential barrier in the base. However, as the mass barrier is significantly strengthened to the point where m_1 is four times larger than m_2, the mass barrier is able to reflect particles from V_b even when $V_b < 0$. These results clearly indicate that the position of V_b is very

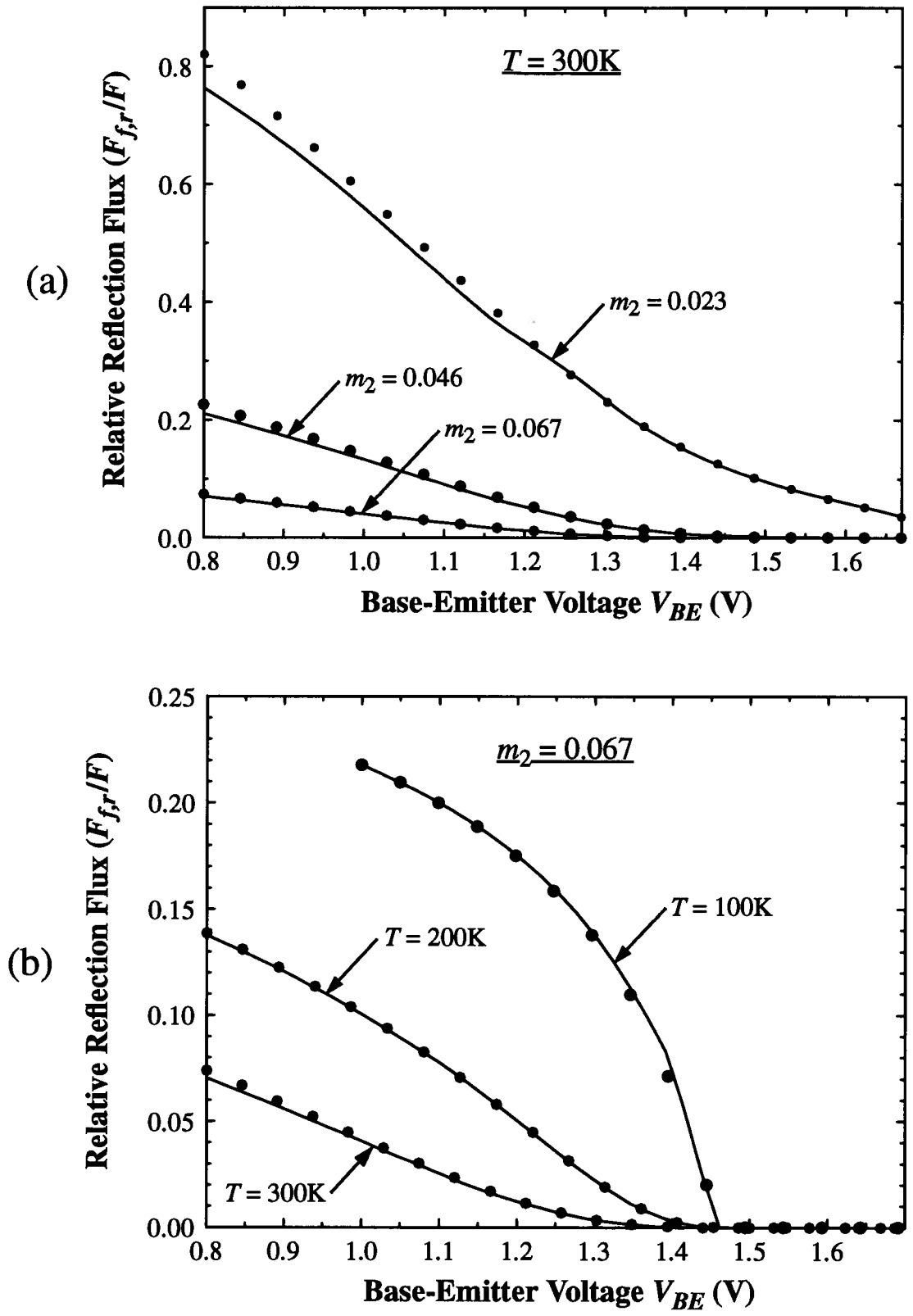


Fig. 4.11. Relative importance of $F_{f,r}$ to the total flux F ($= F_{f,s} - F_{f,r}$) for an HBT with the same parameters as Fig. 4.10. The lines are obtained from the analytic models, while the solid dots are from numerical calculation. (a) results for $T = 300$K. (b) results for $m_2 = 0.067$. Note: usable currents (*i.e.*, $> 10^{-8}$ Acm^{-2}) begin at $V_{BE} > 1.0$V for $T = 200$K, and $V_{BE} > 1.2$V for $T = 100$K.

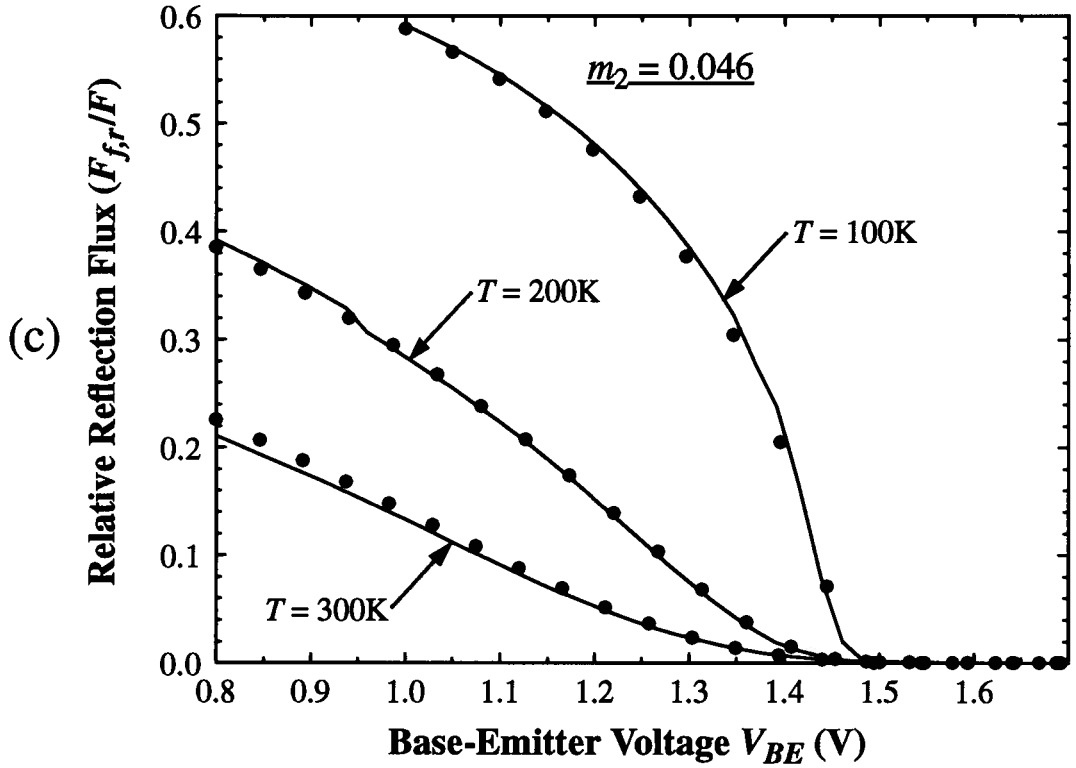


Fig. 4.11. Continuation of Fig. 4.11 from the previous page. (c) results for $m_2 = 0.046$.

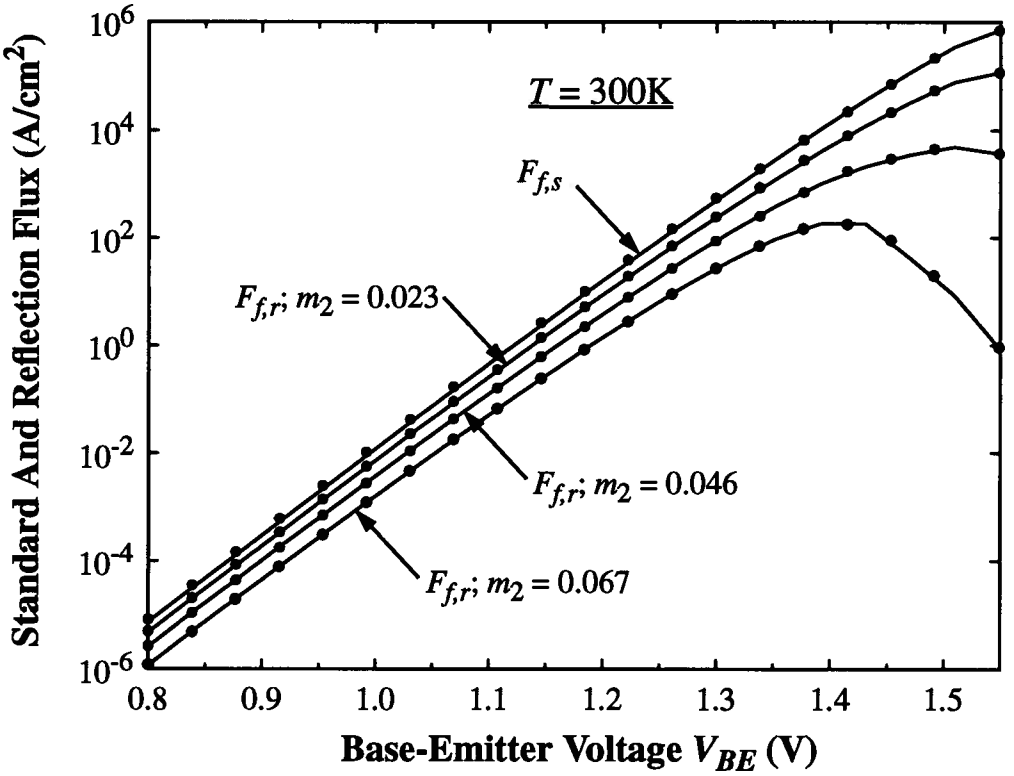


Fig. 4.12. Standard Flux $F_{f,s}$ and Reflection Flux $F_{f,r}$ for an HBT with the same parameters as Fig. 4.10, but with ΔE_c reduced from 0.24eV down to 0.12eV. Note how reducing ΔE_c increases the relative importance of the reflecting potential barrier V_b by lowering V_{bi} (see Fig. 4.10(a)).

important to the transport flux through the CBS. The conclusion is that during the design of the device it is beneficial to have a large ΔE_c so that V_b is lowered, and the mass boundary will have a reduced effect. Finally, examination of Fig. 4.11(b) and (c) clearly demonstrates that lowering the temperature increases the relative importance of $F_{f,r}$ in all cases. Obviously, the combination of lower temperatures and a stronger mass barriers produces the largest reflections.

The case of an $Al_{0.3}Ga_{0.7}As$/GaAs HBT produces the rather fortuitous result that V_b is below zero right around the bias at which the device would routinely be operated. There are other material systems (like SiGe) and devices (HBTs with a smaller emitter Al content) where this is not the case. In these systems ΔE_c is smaller so that V_b stands as a larger reflector. Fig. 4.12 shows what the effect of reducing ΔE_c from 0.24eV down to 0.12eV has on the transport flux. Under these conditions V_b remains unchanged but V_{bi} is reduced by 0.12V to 1.551V. Therefore, relatively speaking, the mass barrier has a larger effect, and the effect occurs over a larger bias range.

Reexamination of Figs. 4.10 and 4.11 show an excellent agreement between the analytic models of the previous section and the exact numerical calculation of eqns (4.67) and (4.87). These results clearly show that the approximations used to obtain the analytic models do not compromise the accuracy of the final results. This means that it is reasonable to look at the functional dependencies within these analytic models in order to obtain a deeper insight into the mechanisms by which transport occurs through the CBS. In the end, these analytic models will facilitate a full model for the HBT when other regions of the device (such as the neutral base, or the collector), are brought into the problem.

Attention is now moved from the reflection to the enhancement case. As was stated at the start of this section, three cases will be considered for the enhancement case. In order to make comparisons with the reflection case simple, only m_2 is varied and it is chosen to be the reciprocal to the three reflection cases; namely $0.126m_0$, $0.184m_0$, and $0.368m_0$. Fig. 4.13 is basically the same as Fig. 4.10 (except that γ is now positive under the case of enhancement), and plots $F_{f,s}$ as well as $F_{f,e}$. The same basic trends are observed for the enhancement case as were observed in the reflection case. In Fig. 4.13 at T equal to 300K as well as 200K, increasing m_2 (and thus increasing γ) results in an increase to $F_{f,e}$. Physically, as m_2 increases, the mass barrier will transfer more energy from $U_{\perp,1}$ into $U_{x,2}$ in order to conserve $p_\perp$ (see Fig. 4.5); thus, a larger number of particles will be moved from out of the base bandgap and into E_c to contribute to $F_{f,e}$. Furthermore, as

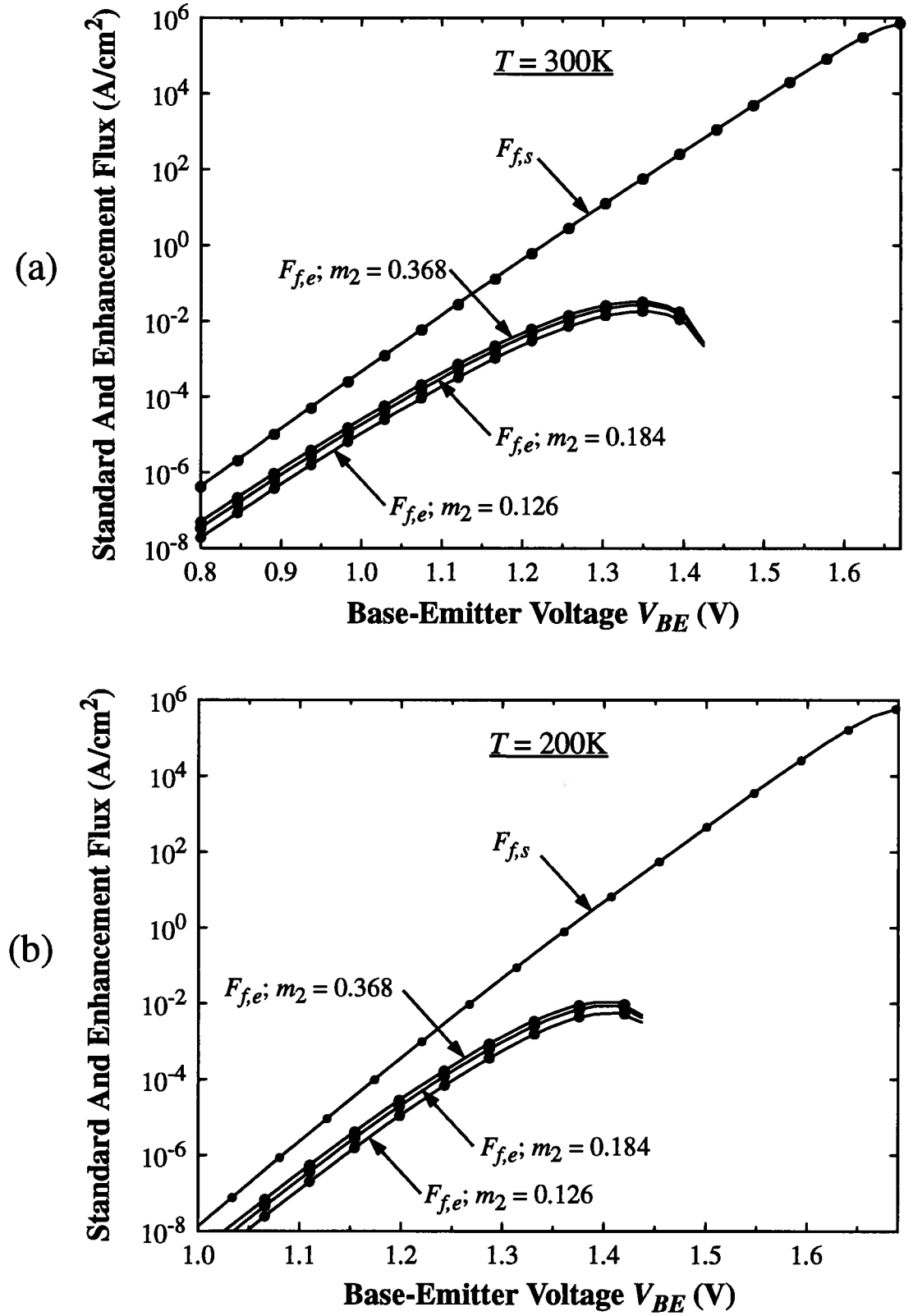


Fig. 4.13. Standard Flux $F_{f,s}$ and the Enhancement Flux $F_{f,e}$ for an HBT with the parameters given near the start of this section. The only parameter being varied is the base side effective mass m_2. The lines are obtained from the analytic models of eqns (4.78) (4.79) (4.92) for $F_{f,s}$ and eqns (4.85) (4.86) for $F_{f,e}$, while the solid dots are from the numerical calculation of eqn (4.67) for $F_{f,s}$ and eqn (4.82) for $F_{f,e}$. (a) results for $T = 300$K. (b) results for $T = 200$K.

V_{BE} is increased, $F_{f,e}$ begins to decrease and then decrease abruptly. The physical cause for this is exactly the same as for the reflection case. As V_{BE} increases V_b decreases so that fewer particles need to be helped over the barrier and $F_{f,e}$ decreases. In the event that $V_b < 0$, every particle that makes it through the CBS must enter the base, since the enhancing mass barrier can only raise $U_{x,2}$ and the minimum $U_{x,2}$ is zero. Thus, once V_b decreases below zero, $F_{f,e}$ must abruptly vanish.

Moving on to Fig. 4.14(b), as T is reduced from 300K to 200K, there is an increase in $F_{f,e}$ relative to $F_{f,s}$. The physical explanation for this fact is identical to the reflection case. Since U'_{max} decreases rapidly along with a decrease in T, the particles will emerge from the CBS with a smaller $U_{x,1}$. As such, an increased number of particles will be available below V_b. With more particles existing below V_b, the mass barrier may effect a larger transfer of particles from below to above the base barrier, and thus increase $F_{f,e}$ as T is reduced.

The most important difference to note between the enhancement and the reflection case is that a smaller increase occurs in $F_{f,e}$ when compared to $F_{f,r}$ for a similar increase in the strength of the mass barrier (which is affected by increasing or decreasing m_2 respectively). The reason for this arises purely because of the nature of enhancement and reflection. For the reflection case, as m_2 becomes arbitrarily small $\gamma \rightarrow -\infty$. With $\gamma \rightarrow -\infty$, every particle that hits the mass barrier will also have its $U_{x,2} \rightarrow -\infty$, leading to a total reflection of all carriers (examination of eqn (4.81) shows that as $\gamma \rightarrow -\infty$ then $T_{eff} \rightarrow T$ so that $F_{f,r} \rightarrow F_{f,s}$ and $F \rightarrow 0$). Thus, it is possible for the reflecting mass barrier to become so effective that the transport flux is reduced to zero. For the enhancement case, there is a fixed ensemble of carriers launched from the neutral emitter towards the CBS that attempts to enter into the base. Once the CBS has removed its portion of the ensemble, the enhancing mass barrier is left to increase $U_{x,2}$ by removing energy from $U_{\perp,1}$. At the limiting strength of the enhancing mass barrier (*i.e.*, $\gamma = 1$), the entire amount of $U_{\perp,1}$ is transferred into $U_{x,2}$ (see eqn (4.39)). Since the particles will have a one kT spread of energy in $U_{\perp,1}$, starting from $U_{\perp,1} = 0$, the enhancing barrier will rapidly reach a limit by which it can no longer increase $F_{f,e}$. Thus, the enhancing barrier will have a smaller effect on F than the reflecting barrier, and as such will not experience the same increase in $F_{f,e}$ due to an increase in m_2 that $F_{f,r}$ would realise for a similar decrease in m_2.

The differences just described between the reflecting and the enhancing case in the previous paragraph can also be understood from a graphical analysis of Figs. 4.6 and 4.7. For the enhance-

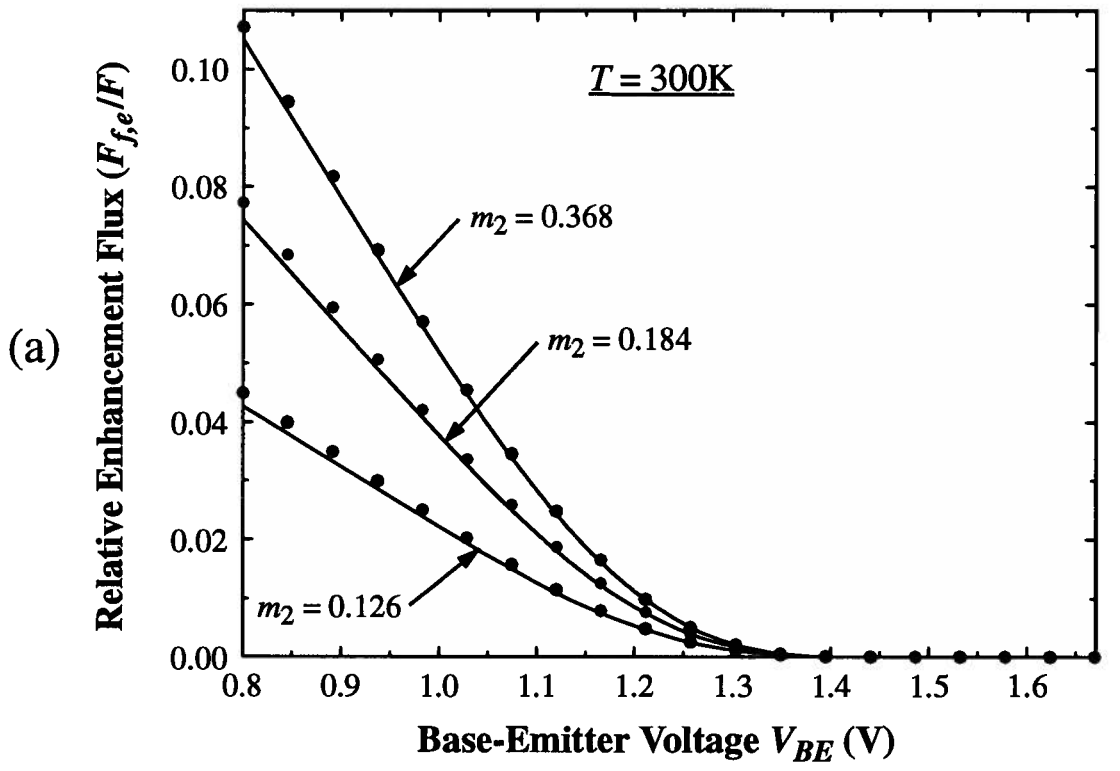
ment case, there is a limit of $\gamma = 1$. Looking at Fig. 4.6 for the integration in R_1, then obviously in the limit when $\gamma = 1$, R_1 will take on a fixed, non-vanishing shape with no possibility of an increase due to a change in the mass barrier. This leads to a maximum value for $F_{f,e}$ and thus F as well. For the reflection case of Fig. 4.7, there is a limit of $\gamma \rightarrow -\infty$. When $\gamma \rightarrow -\infty$, the region of integration R_1 will be reduced to zero, and likewise, so will F. This clearly shows that reflection can produce a far larger effect upon F than enhancement can.

Fig. 4.14 clearly demonstrates the effect of m_2, V_b and T upon $F_{f,e}$. Concentrating on Fig. 4.14(a), there is clearly an increase in $F_{f,e}$ as the strength of the mass barrier increases (*i.e.*, as m_2 increases). However, looking back to Fig. 4.11(a) confirms that the enhancing case does indeed produce less of an effect than the reflecting case. Examination of Fig. 4.14(a) and (b) also shows that once V_b is reduced below zero for $V_{BE} > \approx 1.43\,\text{V}$ (V_{bi} changes with T), $F_{f,e} = 0$ as there is no longer a base barrier to surmount. Finally, Fig. 4.14(b) shows that reducing T increases $F_{f,e}$ in much the same manner as for the reflecting case.

Reexamination of Figs. 4.13 and 4.14 show an excellent agreement between the analytic models of the previous section and the exact numerical calculation of eqns (4.67) and (4.82). These results clearly show that the approximations used to obtain the analytic models do not compromise the accuracy of the final answer.

It is important to keep in mind that under the condition where the CBS is responsible for current-limited-flow, then the results that have been displayed in this section are equal to J_C. Since for most abrupt HBTs the CBS is indeed responsible for limiting the current, then the modelling of CBS transport becomes of paramount importance to the understanding of the device. With the analytic models presented in Section 4.4, and the general models of Sections 4.2 and 4.3, transport through complex structures like the CBS is now fully developed.

Finally, it should be realised that the models of Sections 4.2 to 4.4 determine the transport of charge through the entire EB SCR, and not just the CBS. Eqns (4.50)-(4.53) take into account any quantum mechanical effects, including transport via standard Drift-Diffusion (DD), without the need to appeal to high-energy phenomenological mobility models. By treating transport as a system of collision-less particles that originate from a thermal distribution, the problem of carrier heating and cooling, which needs to be included in DD models [83-85], is ameliorated. Thus, velocity overshoot, including carrier cooling as the electron surmounts V_{bi}, is modelled throughout the entire EB SCR.



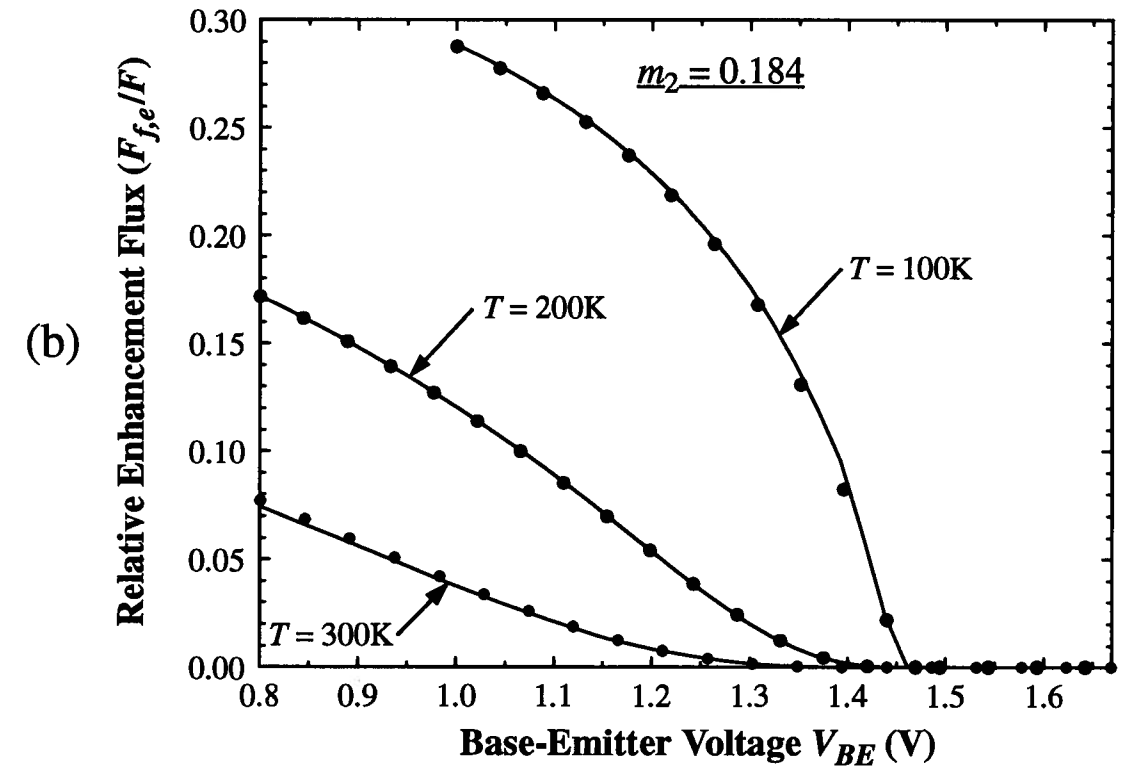


Fig. 4.14. Relative importance of $F_{f,e}$ to the total flux F ($= F_{f,s} + F_{f,e}$) for an HBT with the same parameters as Fig. 4.13. The lines are obtained from the analytic models, while the solid dots are from numerical calculation. (a) results for $T = 300$K. (b) results for $m_2 = 0.184$. Note: usable currents (*i.e.*, $> 10^{-8}$Acm^{-2}) begin at $V_{BE} > 1.0$V for $T = 200$K, and $V_{BE} > 1.2$V for $T = 100$K.

4.6 Deviations from Maxwellian Forms and Non-Ballistic Effects

This section will use the models of the previous sections in order to gain an understanding of the electron distribution that is injected into the neutral base from the emitter. With the neutral base width W_B being pressed below 1000Å, the truly ballistic device is being approached. In the regime where the electron in transit through the neutral base suffers only a few collisions, then one cannot appeal to classical solutions that depend upon a thermalised distribution (*i.e.*, drift-diffusion analysis), nor can one avoid the effect of collisions altogether and treat the ensemble ballistically throughout. In this in-between region, where collisions are important but do not dominate the transport characteristics, solution methods that solve the Boltzmann Transport Equation (BTE) must be used [42,43]. The issue with solving the BTE often hinges upon the shape of the particle ensemble distribution entering the neutral base. As there are less collisions within the base it becomes important to obtain the correct initial ensemble distribution. This section will provide a method to determine the correct ensemble distribution that enters the neutral base. Furthermore, the effect of collisions, or non-ballistic effects within the CBS will also be examined.

It has long been recognized that the particle ensemble distribution entering the neutral base of abrupt HBTs is not Maxwellian [14,39-41]. A Maxwellian distribution is characterised by a Boltzmann distribution in energy, with a parabolic relationship between momentum (or **k**) and energy. Therefore, the Maxwellian distribution appears as a Gaussian distribution in k-space centred at $\mathbf{k} = 0$ (see Fig. 4.15(a)). In the thermionic analysis of the EB heterojunction (*i.e.*, no tunneling is considered through the CBS), one would have a Maxwellian distribution near the top of the CBS (see Fig. 4.2 at $x = 0$). Then, because of the abrupt potential drop beyond the CBS when going towards the base, the Maxwellian distribution is pulled apart so that only the right-going half of the ensemble enters the base. This halved distribution is termed a hemi-Maxwellian (see Fig. 4.15(b)), and is identical to the full Maxwellian except that for $k_x < 0$ the distribution is zero (because the particles are only moving in the positive x-direction). Once the hemi-Maxwellian ensemble has entered the neutral base, and if there have been no collisions from $x = 0$ to $x = x_p$, the distribution will no longer peak at $\boldsymbol{k} = 0$ with an energy of 0, but will be shifted towards larger k_x with an increased energy of $\Delta E_c - V_p$ relative to E_c at $x = x_p$. This shifted hemi-Maxwellian is termed "hot" because it appears to look like a distribution that is characterised by a temperature which is higher that the lattice temperature T.

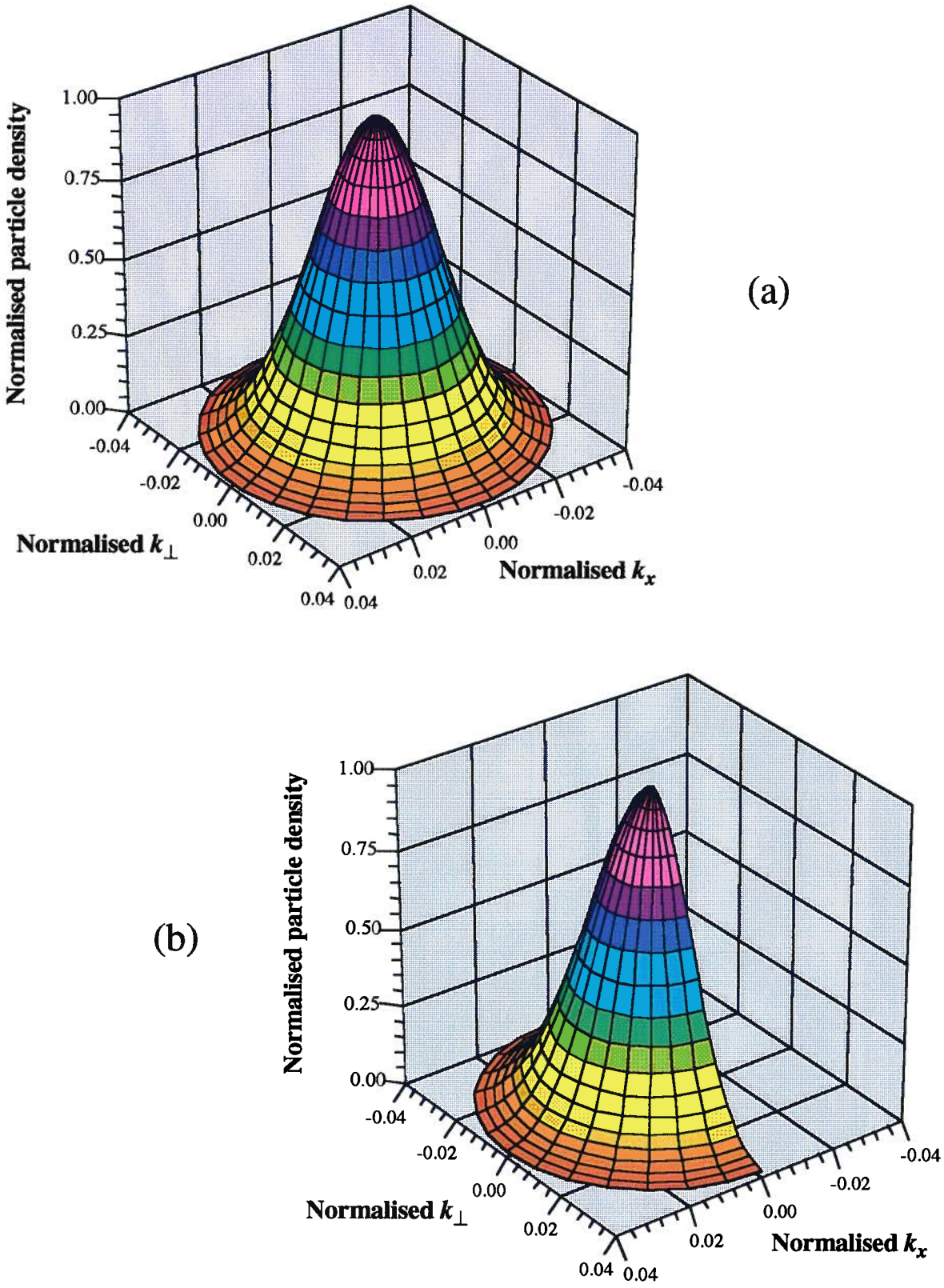


Fig. 4.15. Ensemble particle distributions assuming a purely thermalised thermionic injection from the peak of the CBS in Fig. 4.2. (a) the initial Maxwellian distribution at $x = 0$. (b) the hemi-Maxwellian distribution that is injected towards the neutral base (positive x-direction). **k** is normalised to the length of the GaAs reciprocal lattice vector using an effective mass of 0.067.

From the results of Fig. 4.15, and the arguments of the previous paragraph, the distribution entering the neutral base at $x = x_p$ is clearly not Maxwellian. However, in terms of being able to analyse the neutral base using drift-diffusion analysis, solutions based upon a hemi-Maxwellian distribution will differ from a full Maxwellian distribution by only a multiplicative constant. The issue of the hemi-Maxwellian being hot, however, will require that an energy-balancing scheme also be included by using the second moments of the BTE to arrive at hydro-dynamic drift-diffusion analysis [16,17]. Many researchers who have studied transport within the EB SCR, or the neutral base, have relied on the assumption that the worst-case deviation from a Maxwellian would be a shifted or hot hemi-Maxwellian. This assumption is shown to be false when a structure like the CBS of Fig. 4.2 is present within the EB SCR. In fact, the distribution function entering the neutral base is appreciably distorted from either a Maxwellian, hemi-Maxwellian, or hot hemi-Maxwellian. Furthermore, the distortion to the ensemble distribution has a considerable bias dependence.

Setting aside for the moment the issue of the mass barrier, which serves to distort the ensemble distribution even further, tunneling through the CBS results in a profound change in the shape of the ensemble distribution. As was discussed in the explanation of Fig. 4.10, tunneling through the CBS preferentially picks out from the random Maxwellian ensemble of particles impinging upon the barrier, those particles that possess a sufficiently high $U_{x,1}$ to pass through the barrier, while being blind to the amount of $U_{\perp,1}$ contained by each particle. Clearly, this will tend to focus the ensemble at $x = 0$ towards higher $U_{x,1}$ and destroy the circular symmetry that exists between k_x and $k_\perp$ shown in Fig. 4.15(b) for the hemi-Maxwellian distribution. Finally, in moving from $x = 0$ to $x = x_p$, a number of particles will be reflected by the neutral base potential V_b which will clip off the distribution (much like a hemi-Maxwellian is cut from a Maxwellian) and result in a potentially hot ensemble entering the neutral base.

Fig. 4.9 shows the ensemble distribution after an integration has occurred along the transverse direction. The result, which was formally proven in Section 4.4, is essentially a Gaussian distribution versus $U_{x,1}$. Since momentum **p** and wave vector **k** vary as the square root of $U_{x,1}$, the ensemble distribution plotted in Fig. 4.9 will give a very distorted, non-Gaussian (*i.e.*, non-Maxwellian) shape when plotted against $k_{x,1}$. Furthermore, V_b cuts the distribution off for particles where $U_{x,2}$ (= $U_{x,1} - V_b$ because there is no mass barrier) < V_b. This results in a form that is indicative of, but distinctly different from, a hot hemi-Maxwellian (see Fig. 4.16).

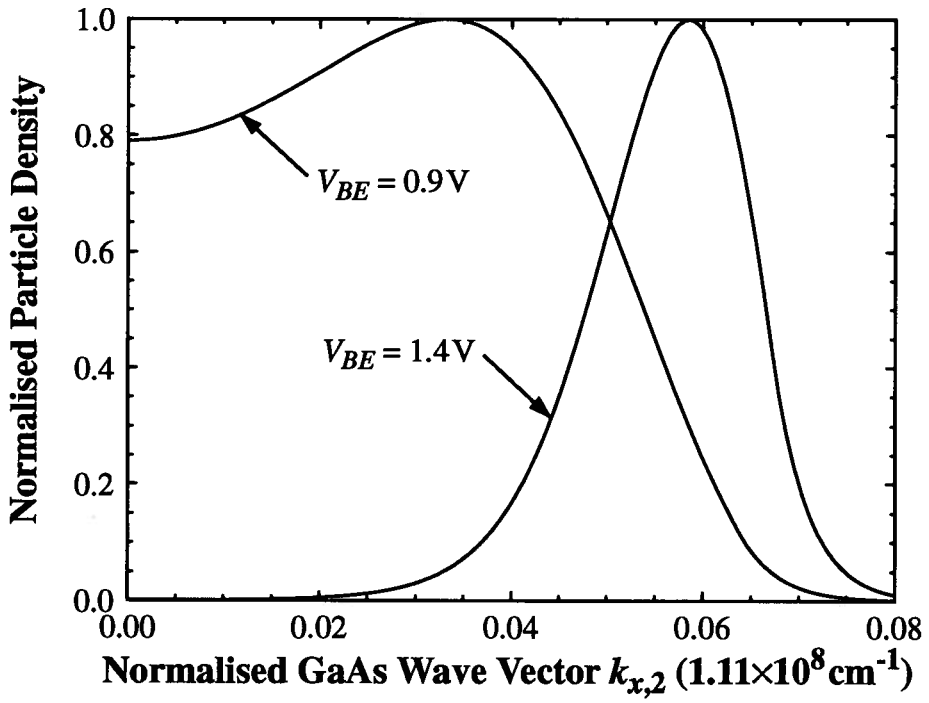


Fig. 4.16. Ensemble distribution versus wave vector $k_{x,2}$ entering the neutral base (*i.e.*, at $x = x_p$) (T = 300K). This is essentially a replot of Fig. 4.9 except when $U_{x,2} < V_b$ the distribution is cut-off and not displayed in order to see the effect of the reflecting base potential. Also, Fig. 4.9 is a plot of the ensemble approaching the CBS from $x = -x_n$. Finally, $k_{x,2}$ is normalised to the length of the reciprocal lattice vector (*i.e.*, $2\pi/a$ where a is the lattice constant).

Fig. 4.16 shows the distortion to the ensemble distribution along $k_{x,2}$. At low bias, where V_b is approaching U_{max}, the ensemble distribution is clipped very near the peak of the distribution, but, unlike a hemi-Maxwellian, not right at the peak. Further, the Gaussian form with respect to energy results in a very flat-topped and non-Gaussian form with respect to **k**. As the bias is increased, V_b recedes when compared to U_{max} so that the distribution no longer has a clipped form. This results in a hot distribution that is asymmetric and which looks quite different from a shifted Maxwellian. Fig. 4.16 clearly shows the non-Maxwellian nature of the ensemble distribution entering the neutral base. However, it does not show the distortion that occurs along $k_\perp$ ($k_\perp = k_{\perp,1} = k_{\perp,2}$ because of momentum conservation in eqn (4.36)). In order to see the full ensemble distribution entering the neutral base ($= W_{CBS}(U_{x,1})f_1(U_{x,1} + U_{\perp,1})$), a three dimensional plot versus $k_{x,2}$ and $k_{\perp,2}$, is displayed in Fig. 4.17. Observation of Fig. 4.17 clearly shows the non-Maxwellian or non-hemi-Maxwellian shape of the electron ensemble distribution entering the neutral base at $x = x_p$. Furthermore, Fig. 4.17 also demonstrates that it would be a gross approximation to assume

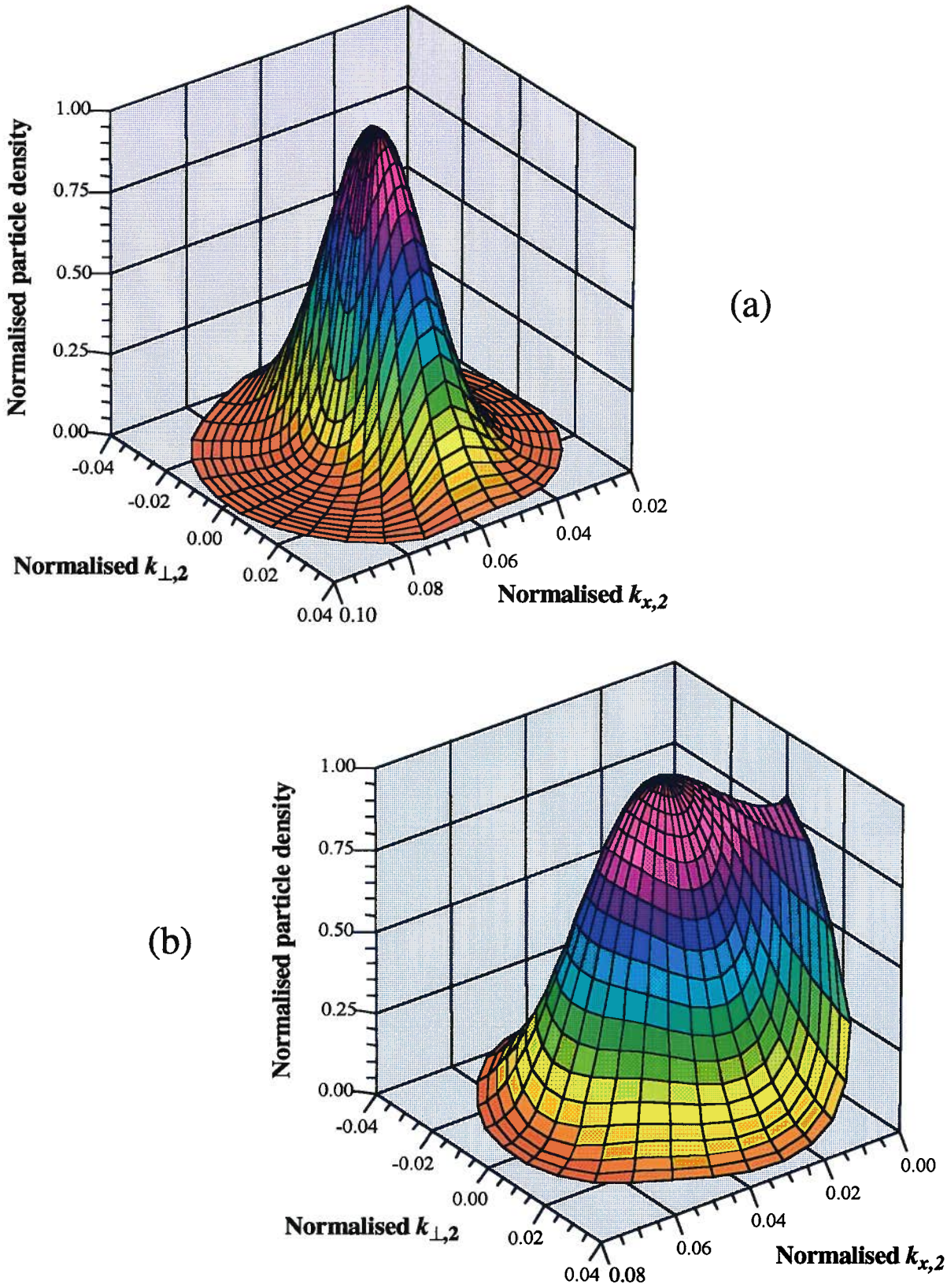


Fig. 4.17. Ensemble electron distribution entering the neutral base versus **k** (T=300K). The particle density is normalised to the peak of the distribution, and **k** is normalised to the length of the GaAs reciprocal lattice vector (= $1.11\times10^{8}\text{cm}^{-1}$). (a) V_{BE} = 1.4V. (b) V_{BE} = 0.9V. Comparing (a) and (b) to Fig. 4.15 shows that these distributions are neither Maxwellian nor hemi-Maxwellian.

that the shape of the ensemble distribution is invariant under a change in bias. These results clearly indicate that the assumption of a hot Maxwellian or hemi-Maxwellian entering the neutral base in an abrupt HBT is erroneous.

Figs. 4.16 and 4.17 have used the full HBT parameters that Section 4.5 has been based upon, except that the mass boundary has been neglected by setting $m_2 = m_1$. As was alluded to earlier in this section, the mass boundary will have the effect of further distorting the ensemble distribution. Fig. 4.18 plots the electron ensemble distribution entering the base under the condition where $m_2 = 0.023$ (*i.e.*, the reflecting case) to clearly observe the mass barrier effects. The effect of the reflecting mass barrier is to simultaneously pull the distribution towards lower $k_{x,2}$ and higher $k_{\perp,2}$. Looking at Fig. 4.18(a) and comparing to Fig. 4.17(a) clearly shows the extension in $k_{\perp,2}$; while careful observation of the constant $k_{x,2}$ line from the peak shows that the distribution is indeed being pulled and distorted towards lower $k_{x,2}$. Comparison of Figs. 4.18(b) and 4.17(b) clearly demonstrates the distortion due to the reflecting mass barrier upon the ensemble distribution. It is important to realise that, although the volume of the distribution is larger in Fig. 4.18 then in Fig. 4.17, there is an overall multiplicative factor of 0.25 (for this reflecting mass barrier) when computing the flux, leading to a net reduction in the total flux.

Fig. 4.19 plots the electron ensemble entering the neutral base with an enhancing mass barrier where $m_2 = 0.368$. The enhancing mass barrier distorts the distribution in exactly the opposite fashion when compared to the reflecting mass barrier. The effect of the enhancing mass barrier is to simultaneously pull the distribution towards higher $k_{x,2}$ and lower $k_{\perp,2}$. Comparison of Fig. 4.19(a) with Fig. 4.17(a) demonstrates that the distribution is certainly being pulled towards lower $k_{\perp,2}$; so much so that the distribution is starting to look Maxwellian. Closer examination of the contour lines in Fig. 4.19(a) shows the distortion that results from the extension in $k_{x,2}$, which is a clear deviation from a Maxwellian form. Further examination of Fig. 4.19(b) in comparison to Fig. 4.17(b) exemplifies the distortion to the ensemble due to the enhancing mass barrier. As similarly occurred with the reflecting case, the volume in Fig. 4.19 appears smaller than the volume in Fig. 4.17. However, there is now a multiplicative constant of 4 (for this enhancing mass barrier) when computing the flux, leading to a net increase in the total flux.

Figs. 4.16 through 4.19 clearly chronicle the effects that tunneling and the mass barrier have upon the electron ensemble distribution entering the neutral base. The one clear conclusion from

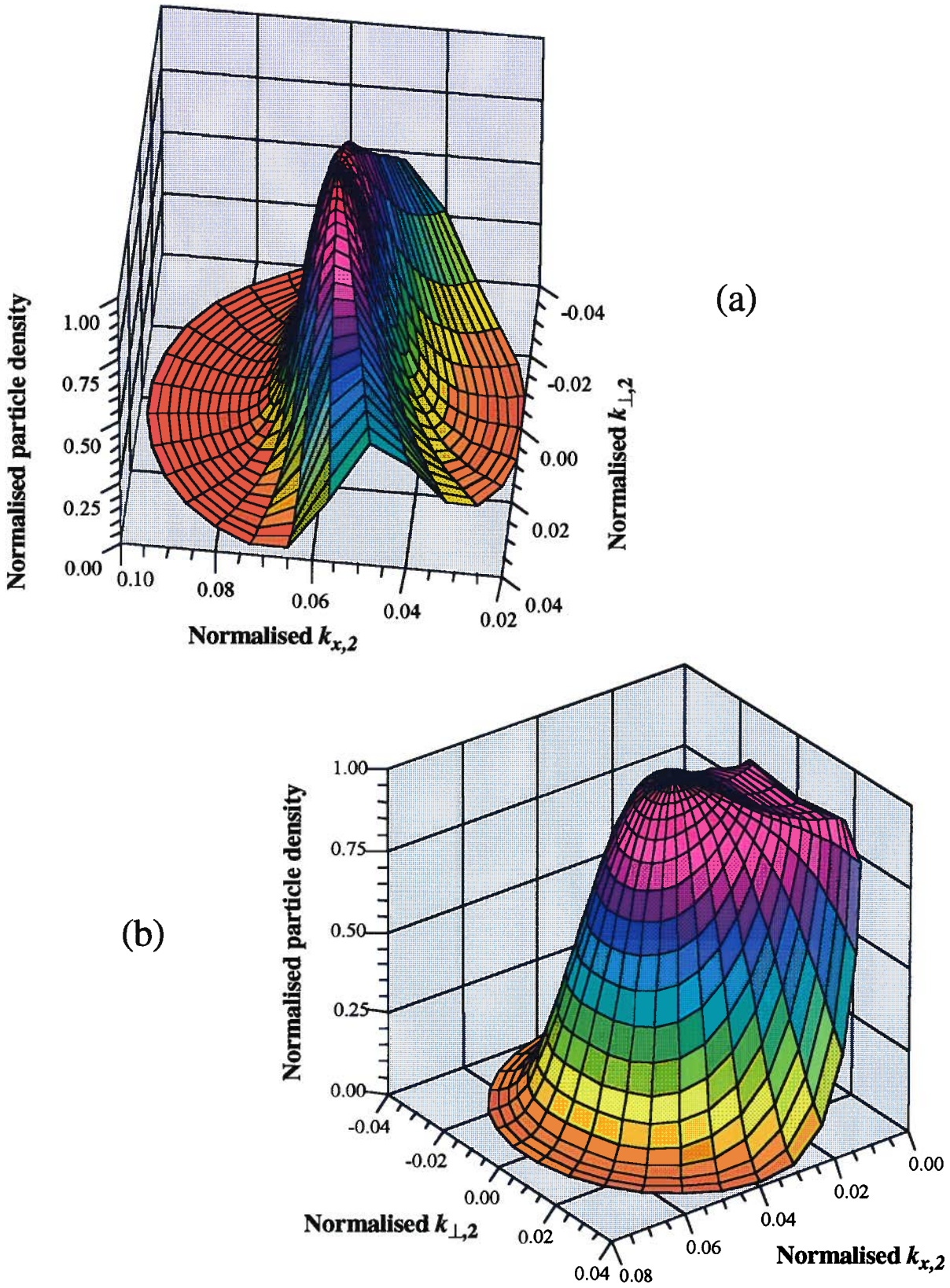


Fig. 4.18. Replot of Fig. 4.17 but this time including a reflecting mass barrier where $m_2 = 0.023$ and $m_1 = 0.092$. (a) $V_{BE} = 1.4$V. The plot has been rotated 45° relative to Fig. 4.17(a) to clearly display the distortion in the $k_{x,2}$ direction. (b) $V_{BE} = 0.9$V. Again notice the extreme distortion compared to Fig. 4.17(b) for $k_{x,2}$ less than the peak.

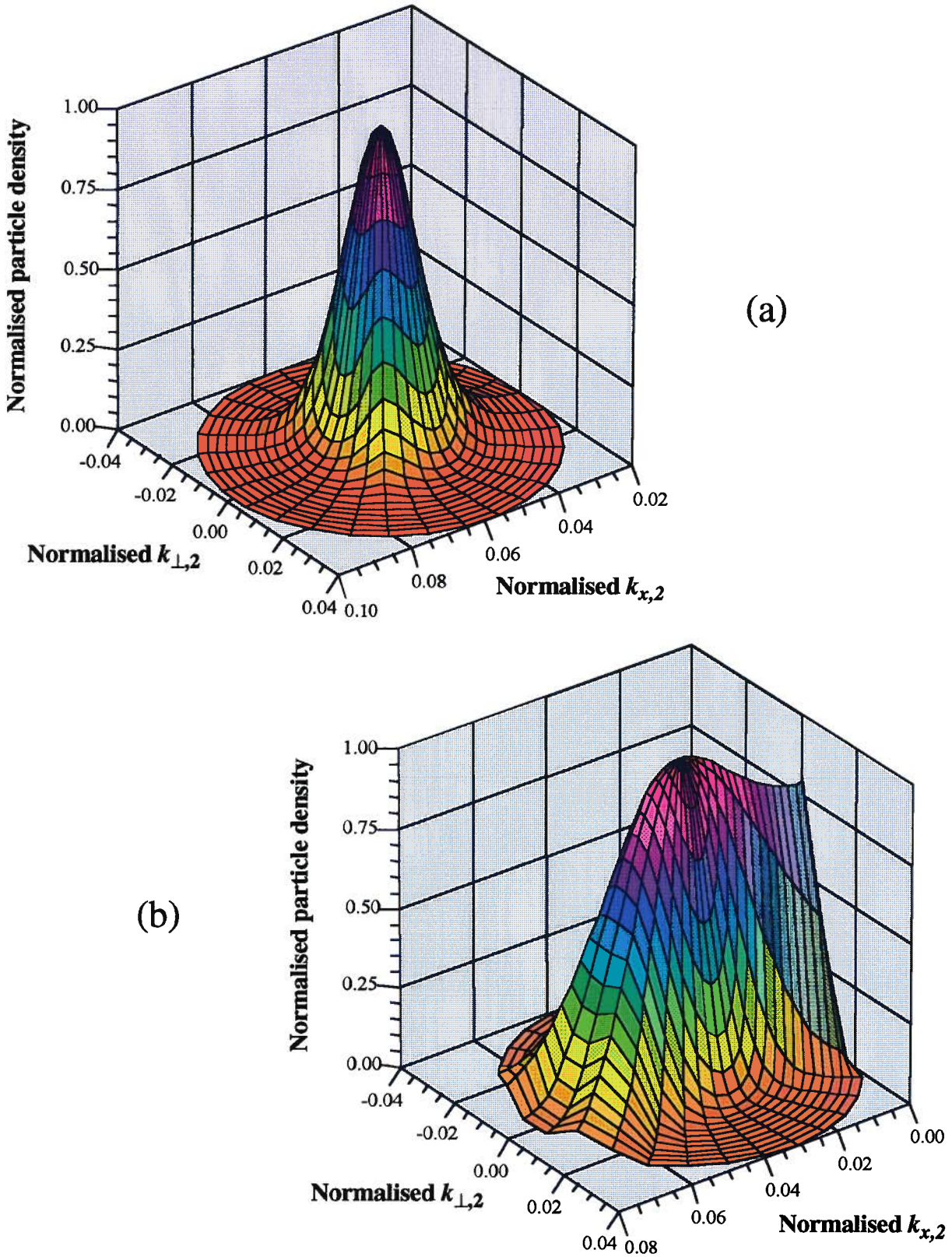


Fig. 4.19. Replot of Fig. 4.17 but this time including an enhancing mass barrier where m_2 = 0.368 and m_1 = 0.092 (the reciprocal to Fig. 4.18). (a) V_{BE} = 1.4V. The distribution looks Maxwellian but comparing to Fig. 4.17(a) shows it to be distorted towards larger $k_{x,2}$. (b) V_{BE} = 0.9V. Clearly $k_{x,2}$ has been extended and $k_{\perp,2}$ has been squashed relative to Fig. 4.17(b).

the analysis of this section is that one cannot assume that the ensemble distribution entering the base has any resemblance to a Maxwellian or hemi-Maxwellian in either a normal or hot condition. Also, the change in the shape of the distribution over bias cannot be accounted for in a simple fashion (such as a constant multiplier). Further, it is the effect of tunneling that contributes most to the distortion of the ensemble distribution, with the mass barrier playing an important but generally subservient role. This fact returns us back to the starting comments of this chapter, *i.e.*, that a failure to account for tunneling through the CBS can lead to considerable error in the analysis of abrupt HBTs. In any event, the analytic models presented in this chapter can be used to construct the correct electron ensemble distribution entering the neutral base. This correct neutral base ensemble distribution can then be used as a boundary condition in a subsequent BTE solution of the transport through the neutral base.

The models presented in this chapter have assumed the condition of ballistic motion throughout the EB SCR. This assumption is relatively solid given that the EB SCR is generally quite narrow and as such is much smaller than the mean free path of the particle. Before going on to talk about the effects of non-ballistic motion throughout the EB SCR, it is important to pause for a moment to discuss the lower boundary of V_b used to calculate the flux through the CBS. Re-examination of Figs. 4.1 and 4.2 show that F_f and F_r are calculated by assuming that a hemi-Maxwellian distribution is launched into the EB SCR from both $x = -x_n$ and x_p respectively. The final flux exiting the EB SCR is then determined by considering how tunneling through the CBS, as well as reflection by V_b and distortion due to the mass barrier, alters the course of the forward and reverse directed hemi-Maxwellians. To assume that a hemi-Maxwellian form exists at both $x = -x_n$ and x_p, the distributions at these two points in space must be fully thermalised and characterised by the lattice temperature T. This is a reasonable assumption given that $x = -x_n$ and x_p are the depletion edges of the EB SCR, and as such are outside of where non-equilibrium effects would begin to occur. It is for this reason that the flux is considered to be injected from $x = -x_n$ and x_p, leading to the potential boundary of V_b (which is equal to E_c at $x = x_p$) to enter the neutral base.

The above argument corrects what Grinberg et al. [51] have suggested. In [51], the injection to the left is from $x = 0$, not from $x = x_p$. The point $x = 0$ is inside of the EB SCR and coincides with the peak electric field. As such, the ensemble distribution at $x = 0$ is expected to be at its largest departure from equilibrium when compared to any other point within the EB SCR. Further-

more, to consider the point $x = 0$ as the boundary condition, one would have to imagine that the electron could ballistically tunnel a few hundred angstroms through the CBS and then suddenly thermalise at $x = 0$, where it could then be carried into the neutral base by diffusion. Clearly, it is not reasonable to assume that $x = 0$ is the source of a thermalised Maxwellian distribution.

By adopting Grinberg's proposals within [51], the lower limit of integration for the calculation of F would be reduced from V_b to V_b - V_p (= $V_{pk} - \Delta E_c$; see Fig. 4.2). The effect of this change would be to increase F as the base potential has been lowered and will thus reflect fewer particles. For HBTs where the base doping is more than 30-fold larger than the emitter doping, then V_p will be very small and the error of adopting the proposals within [51] will be accordingly small. However, as the doping of the EB junction becomes even slightly more symmetric, the error of using [51] will become increasingly large. Furthermore, as the temperature is reduced to the point where U_{max} occurs below V_b, there will be an exponential change to F for a linear change to V_p. Thus, under low temperature conditions the methods contained within [51] for the inclusion of tunneling will be in error even for a highly asymmetric doping junction (see Fig. 4.20).

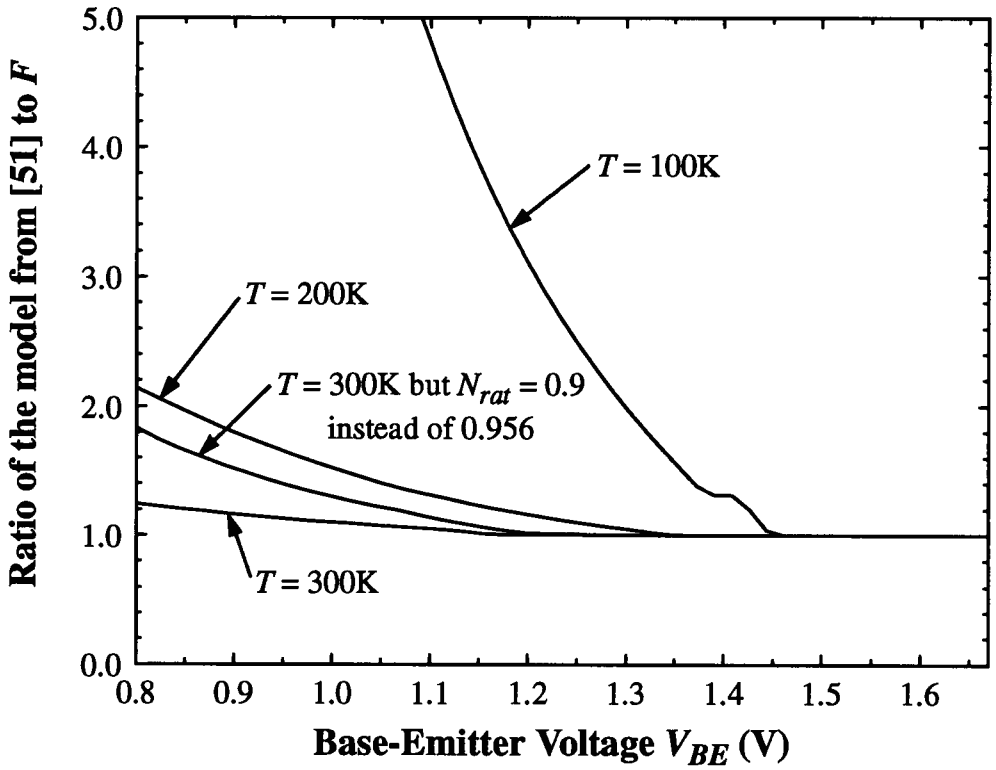


Fig. 4.20. Relative difference between the results obtained from the methods proposed in [51] to the model for F from this chapter. The device is based upon the same $Al_{0.3}Ga_{0.7}As$ HBT used in this section. Note how the reduction to V_b as proposed in [51] leads to an overestimation in the transport through the CBS, and therefore, to an overestimation of I_C.

Finally, is it reasonable to consider ballistic motion throughout the entire EB SCR? Certainly, to consider collisions to the particles while in the process of tunneling would be difficult. However, models that are similar to, but simpler than, the models presented in this chapter are able to explain the terminal characteristics of abrupt HBTs [22,25] because they include the effects of tunneling. Other more complex models, such as Monte Carlo simulation, which do not include the effect of tunneling, grossly underestimate I_C. Since ballistic motion is assumed in eqns (4.50)-(4.53) when accommodating tunneling, and these models explain experimental findings, then experimental evidence tends to corroborate the assumption of ballistic motion throughout the EB SCR. For if there were even a moderate chance of only a single thermalising collision within the EB SCR, then the tunneling current would be drastically altered (any reduction or increase to the energy of the particle will cause a correspondingly rapid reduction or increase in the tunneling probability). Since experimental evidence does not support this, at most, there is a small probability of a thermalising collision within the EB SCR. This justifies the assumption of ballistic motion throughout the EB SCR.

4.7 Conclusion

To conclude this chapter, a summary of the past 40 years' work in this area of electron transport through a SCR is in order. The reason for this summary is to give due credit to all of the individuals who have made contributions, and to demonstrate how a large majority of this past work is disjoint from both the study of HBTs and itself. To begin with, Miller and Good [86] set out the requirements for the WKB approximation to the Schrödinger equation in 1953, which formed the basis for the study by Murphy and Good [68] in 1956 of electron emission from metals into vacuum due to thermionic injection and tunneling (which they term field emission). [68] lead to the formation of the general charge transport model of eqn (4.2). The seminal work of Stratton [69] extends [68] by considering electron emission from semiconductors into vacuum, including the effect of a mass barrier based upon a spherical effective mass. The main concern in [68,69] is the incorporation of image force corrections which alter the tunneling potential and greatly increases the tunneling current. In [69], tunneling is only considered within the vacuum and not within the semiconductor, and does not consider the effect of a base barrier potential V_b (as V_b is far too negative to enter into the problem). Stratton and Padovani [75] apply [69] to Schottky barriers, and include tunneling within the semiconductor but still do not concern themselves with the effect of V_b.

Also, [75] does not include the mass-barrier effect considered in [69]. In parallel to the work of [69,75], Christov independently repeats the work [70,73]. The work done in [69,70,73,75] is meant for the study of Schottky diodes, and is more concerned with surface effects (image force correction) than anything else. Furthermore, the potential profile being considered is linear and not the parabolic one found within the SCR; however, [69] does allude to the solution of an arbitrary potential energy profile through the use of a Taylor expansion. The work up to this point forms the foundation for the study of Schottky diodes and band-offsets between metals and semiconductors.

Crowell [76] derives the Richardson constant for a completely general effective mass tensor, but fails to rigorously derive the result by not presenting the relevant Jacobians. Instead, the work in [76] relies on simple arguments to obtain results that, while applicable to the study of pure thermionic emission, are not clearly applicable when tunneling is considered. Crowell [77] continues the work in [76] in an effort to determine the correct effective mass to apply to a Schottky diode between two materials characterised by different effective masses. The work in [77], much like that done in [76], is not mathematically rigorous, and as a result fails to obtain a vanishing transport current under equilibrium conditions. Grinberg [82] solves this problem but only if thermionic emission is considered and not tunneling. The work of this chapter extends [82] by including tunneling and thermionic injection (eqn (4.60)) through a rigorous mathematical treatment.

Finally, Crowell and Rideout [78] solve for tunneling through the parabolic potential barrier of the SCR, but do not include the effect of a mass barrier. They present the final transform (eqn (4.70)) used to evaluate the tunneling integral of eqn (4.67), but do not present its development (eqns (4.67)-(4.69)), nor do they provide for a spatially varying permittivity ε or the effect of V_b. Eqns (4.50)-(4.53) derive for the first time charge transport through the EB SCR, including thermionic emission and tunneling, between two semiconductors characterised by different effective mass tensors and ε. Furthermore, the effect of V_b is properly included. The most important aspect of the work contained within this chapter is that for the first time all of the essential physical constructs of the EB junction within an abrupt HBT have been considered. The results of these considerations are analytic models, based upon the solution of eqns (4.50)-(4.53), to simulate the transport of flux through the EB SCR. Since there were no special features of a specific material system employed within this chapter, the results of this chapter are applicable to any material system. Finally, the developments presented here have focussed upon electron transport, but apply equally well to the transport of holes with basically little change to the models.

CHAPTER 5

Recombination Currents

As was discussed in Chapter 3, one of the most important parameters of an HBT is the current gain β. Whether one is designing Digital or Analogue circuits within an IC, an accurate understanding of β is essential to the successful operation of the circuit. Chapter 4 dealt with the calculation of transport through the CBS (in an npn device), which is often the determining factor for I_C in abrupt HBTs [18,25]. This chapter will finish off the model for I_C by using the general models of Chapter 2 to include the effect of neutral base transport along with transport through the CBS. More specifically to the calculation of β, this chapter presents the physics underlying the creation of base current. Included in the analysis to follow is the interaction of I_B with I_C that was alluded to in Chapter 2, and which occurs when transport through the CBS is responsible for current-limited-flow (*i.e.*, control of I_C).

This chapter includes the modelling of four different components of the hole current that result in the base terminal current. These components are: 1) Shockley-Read-Hall (SRH) recombination within the EB SCR; 2) Auger recombination within the EB SCR; 3) radiative recombination within the EB SCR; 4) neutral base recombination through all of the processes just detailed. The back injection of carriers (*i.e.*, holes for the npn HBT being considered) from the base into the emitter is not accounted for because this back injection is effectively suppressed by the characteristics of the wide bandgap material that forms the emitter; however, inclusion of back injection is a trivial extension to the results that follow.

Analytic models for the four previously mentioned recombination processes that are responsible for the creation of I_B will be presented. It is shown that these analytic expressions for the four base current components can be reduced to the familiar diode equations with two parameters - namely the saturation current J_S and the injection index n. Even though the physical mechanisms that control the base current in the presence of a heterojunction differ markedly from the homojunction case, one can still recover a simple diode model for the final representation. It is within this analysis that a surprising result regarding the injection index n is made. Standard theoretical calculations give a value of $n = 2$ for the SRH current. However, it was found that $n = 2$ applies only in the limit of a wide, or symmetrically doped, EB SCR. For HBTs of interest, where the base doping is very high compared to the emitter doping (*i.e.*, asymmetrically doped), a value of $n = 1$ is applicable under certain operating conditions.

Most of the work that is to be presented in this chapter has been previously published by this author and Dr. D.L. Pulfrey [24]. Within the context of this published work, HBTs constructed within the $Al_xGa_{1-x}As$ material system were studied. The results of this chapter are general, however, and can be applied to other material systems as well. For the case of indirect material systems such as Si_xGe_{1-x}, the only major change is that the radiative recombination rate is small enough to be ignored in comparison to SRH and Auger recombination.

5.1 Electron Quasi-Fermi Energy Splitting ΔE_{fn}

The presence of an abrupt EB heterojunction in an npn HBT can lead to the splitting of the electron quasi-Fermi energy E_{fn}, as first discussed by Perlman and Feucht [50], and shown in Fig. 5.1. This splitting of E_{fn} (*i.e.*, ΔE_{fn}) has been alluded to in Chapter 2 and was found to be the driving force for the transport current through the CBS (as was proven in Section 4.3, eqn (4.63)). Fig. 5.1 shows ΔE_{fn} and its position within the EB SCR. ΔE_{fn} results due to a departure from quasi-equilibrium, where the transport flux through the CBS is no longer a small perturbation to the forward and reverse equilibrium fluxes that are everywhere present within a semiconductor [50,18].

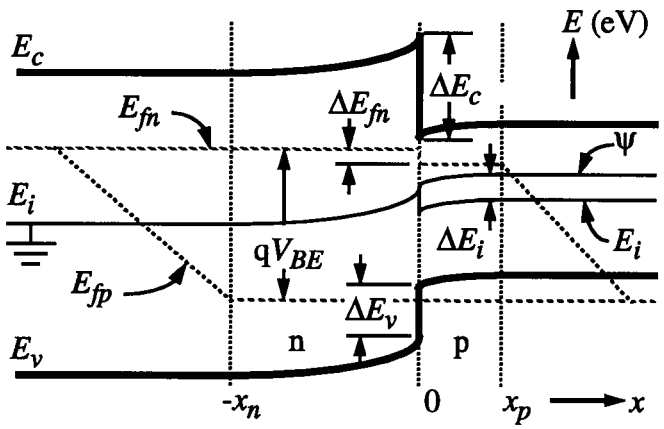


Fig. 5.1. Band diagram of the EB SCR showing the effect of the abrupt heterojunction on E_{fn} under an applied forward bias (reprint of Fig. 2.2). ψ is the solution to the Poisson equation and is therefore continuous. Both the reference energy position and the intrinsic or mid-bandgap energy E_i are also shown.

Section 4.3 and eqn (4.63) clearly bring out ΔE_{fn}, but do not locate the position of ΔE_{fn} if it is indeed abrupt. Perlman and Feucht [50] have addressed the spatial variation of ΔE_{fn} and found that in general ΔE_{fn} occurs abruptly and coincidentally with the position of the EB Heterojunction (as is shown in Fig. 5.1). Finally, the hole quasi-Fermi energy E_{fp} has no discontinuity and is os-

tensibly constant throughout the EB SCR. The reason for the lack of a ΔE_{fp} within the EB SCR is because transport through the neutral emitter and not the EB SCR dictates the back injection current (this is proven at the end of Section 5.3 once the neutral emitter transport current is derived). Essentially, because the EB SCR is not responsible for the current-limited-flow of holes into the emitter, there is no ΔE_{fp} present within this region of the device.

Traditionally, in the modelling of current transport in HBTs, ΔE_{fn} has been implicit in the calculation of the collector current density J_C and the neutral-base recombination current density J_{NB} [20,51,87-89]. The calculation has proceeded via a balancing of J_C and J_{NB} against the combined thermionic/tunnel current J_{ThT} crossing through the CBS at the abrupt junction; *i.e.*,

$$J_{ThT} = J_{NB} + J_C \qquad \rightarrow \Delta E_{fn} \,. \tag{5.1}$$

Further, it has been the usual practice when considering additional base current due to recombination in the EB SCR, to subsequently add this extra current J_{SCR} to the prior-calculated J_{NB}; *i.e.*,

$$J_B = J_{NB}(\Delta E_{fn}) + J_{SCR}. \tag{5.2}$$

Recently, Parikh and Lindholm [90] have emphasized that this calculation of J_B via direct superposition is not strictly correct because the base-side component $J_{SCR,B}$ of J_{SCR} should figure in the original current-balancing equation which is used to compute ΔE_{fn}, and, subsequently, J_C, J_{NB} and $J_{SCR,B}$; *i.e.*, eqns (5.1) and (5.2) should be replaced by

$$J_{ThT} = J_{SCR,B} + J_{NB} + J_C \qquad \rightarrow \Delta E_{fn} \tag{5.3}$$

$$J_{B,B} = J_{SCR,B}(\Delta E_{fn}) + J_{NB}(\Delta E_{fn}) \tag{5.4}$$

where $J_{B,B}$ is that portion of the base current arising from recombination in the metallurgical base (see Fig. 5.2).

It can be appreciated that this more correct, self-consistent computation of $J_{SCR,B}$ will only effect the base current if $J_{SCR,B}$ is comparable to J_{NB} and, furthermore, will only effect the computation of J_C from the balancing equation (*i.e.*, eqn (5.3)) in cases where β is low. To examine these effects is one of the objectives of this chapter and, to ensure that their importance is not underestimated, Auger and radiative recombination in the SCR have been considered, as well as the usual SRH recombination.

The computation of ΔE_{fn} via eqn (5.3) can be done numerically, but an analytical solution would be more insightful, and also very useful in HBT device modelling because ΔE_{fn}, and thus

J_C, J_{NB} and $J_{SCR,B}$ could then all be computed directly from the physical properties of the device and the applied bias. Chapter 2 presented the analytic methods to determine both ΔE_{fn} and the ultimate transport currents that produce J_C and J_B. Therefore, the second objective of this chapter is to develop such an analytical expression for ΔE_{fn}. A final aim is to show that the components of $J_{SCR,B}$, even though they have an extra bias dependence through ΔE_{fn}, can be expressed as diode-like equations. This fact should greatly facilitate the incorporation of these currents into a complete, large-signal representation of the HBT, which may then be implemented in circuit simulators such as SPICE.

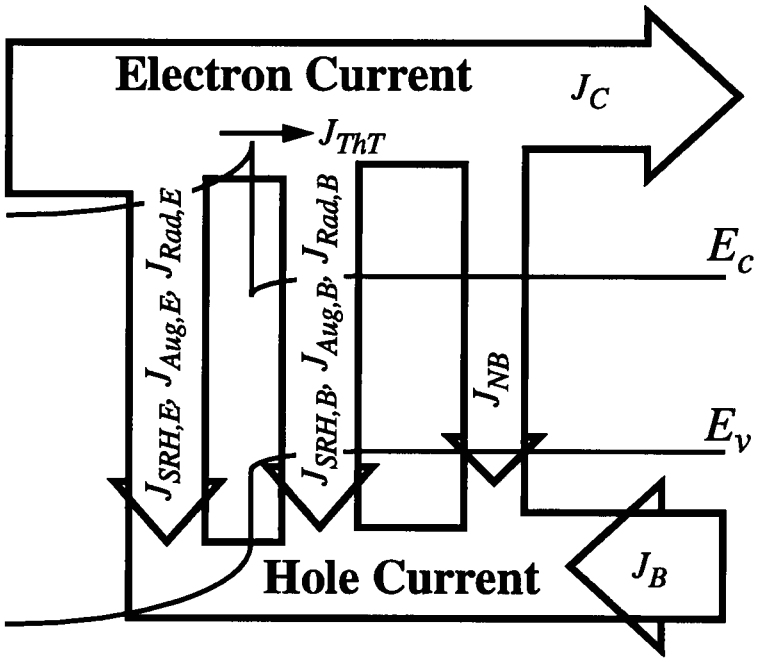


Fig. 5.2. Components of the collector (J_C) and the base (J_B) currents emphasising that J_{ThT} must equal the total of, $J_C + J_{NB} + J_{SRH,B} + J_{Aug,B} + J_{Rad,B}$ when recombination due to Shockley-Read-Hall (SRH), Auger (Aug) and radiative (Rad) processes is considered.

5.2 Modelling the Recombination Processes of HBTs

The "unique relationship" [90] between the collector current, the neutral-base current and the base-side SCR recombination current comes about because all these currents depend upon the electron quasi-Fermi energy splitting at the heterojunction. As this splitting is greatest in the case of an abrupt heterojunction, we consider only this type of junction in this analysis. The junction is taken to be formed by an n-type $Al_{0.3}Ga_{0.7}As$ emitter and a p-type GaAs base (the same as the device in Section 4.5). To reduce the complexity of the algebra, without sacrificing much in the way

of accuracy [90], the permittivities and the effective densities of states have been taken as a constant throughout the entire device.

5.2.1 SRH Recombination

The recombination rate due to SRH recombination can be written as [90,91]

$$R_{SRH} = \frac{n_i}{\tau[\cosh(U_f - E_i/\mathrm{k}T) + b]} \sinh\left(\frac{E_{fn} - E_{fp}}{2\mathrm{k}T}\right) \tag{5.5}$$

where:

$n_i(x)$ is the intrinsic carrier concentration,

$E_{fn}(x)$ is the electron quasi-Fermi energy (see Fig. 5.1),

E_{fp} is the hole quasi-Fermi energy (assumed constant),

$\tau = \sqrt{\tau_{p0}\tau_{n0}}$, where τ_{p0} and τ_{n0} are the hole and electron minority carrier lifetimes, respectively, within the SCR,

$U_f = (E_{fn} + E_{fp})/2\mathrm{k}T + \frac{1}{2}\ln(\tau_{p0}/\tau_{n0})$,

$b = \exp[(E_{fp} - E_{fn})/2\mathrm{k}T]\cdot\cosh[(E_t - E_i)/\mathrm{k}T + \frac{1}{2}\ln(\tau_{p0}/\tau_{n0})]$,

where E_t is the energy level of the single recombination centre assumed in this work, and $E_i(x)$ is the intrinsic Fermi energy. The latter has a discontinuity of ΔE_i at the abrupt heterojunction (see Fig. 5.3), because the bandgap difference between the wide-bandgap emitter and the narrow-bandgap base is generally not distributed evenly between the conduction and valence bands; *i.e.*,

$$\Delta E_i = \frac{\Delta E_{Gap,pn}}{2} + \Delta E_c = \mathrm{k}T\ln\left(\frac{n_{i,n}}{n_{i,p}}\right) + \Delta E_c \tag{5.6}$$

where the subscripts p,n refer to the p-type base and the n-type emitter regions respectively. E_i is related, therefore, to the electrostatic potential energy $\psi(x)$ via

$$E_i(x) = \begin{cases} \psi(x) & x \le 0 \\ \psi(x) - \Delta E_i & x > 0 \end{cases}. \tag{5.7}$$

Here we use the depletion approximation for $\psi(x)$, namely:

$$\psi(x) = \begin{cases} V_{pk}\left(1 + \dfrac{x}{x_n}\right)^2 & x \le 0 \\ q(V_{bi} - V_{BE}) - V_p\left(1 - \dfrac{x}{x_p}\right)^2 & x > 0 \end{cases} \tag{5.8}$$

where, using eqn (4.69)

$$V_{pk} = qN_{rat}(V_{bi} - V_{BE}), \quad V_p = q(1 - N_{rat})(V_{bi} - V_{BE}), \Rightarrow \quad \frac{V_p}{V_{pk}} = \frac{\varepsilon_n N_D}{\varepsilon_p N_A}$$

$$x_n = \sqrt{\frac{2\varepsilon_n V_{pk}}{q^2 N_D}}, \quad x_p = \sqrt{\frac{2\varepsilon_p V_p}{q^2 N_A}} \Rightarrow \quad \frac{x_p}{x_n} = \frac{N_D}{N_A} \tag{5.9}$$

$$\text{where } N_{rat} = \frac{\varepsilon_p N_A}{\varepsilon_p N_A + \varepsilon_n N_D}, \quad V_{bi} = \frac{kT}{q}\ln\left(\frac{N_A N_D}{n_{i,p}^2}\right) + \frac{\Delta E_c}{q} = \frac{kT}{q}\ln\left(\frac{N_A N_D}{n_{i,p} n_{i,n}}\right) + \frac{\Delta E_i}{q}.$$

with V_{BE} being the applied base-emitter voltage, V_{bi} the built-in potential, N_D the emitter doping and N_A the base doping. Eqn (5.9) has included the effects of a non-uniform permittivity for the time being.

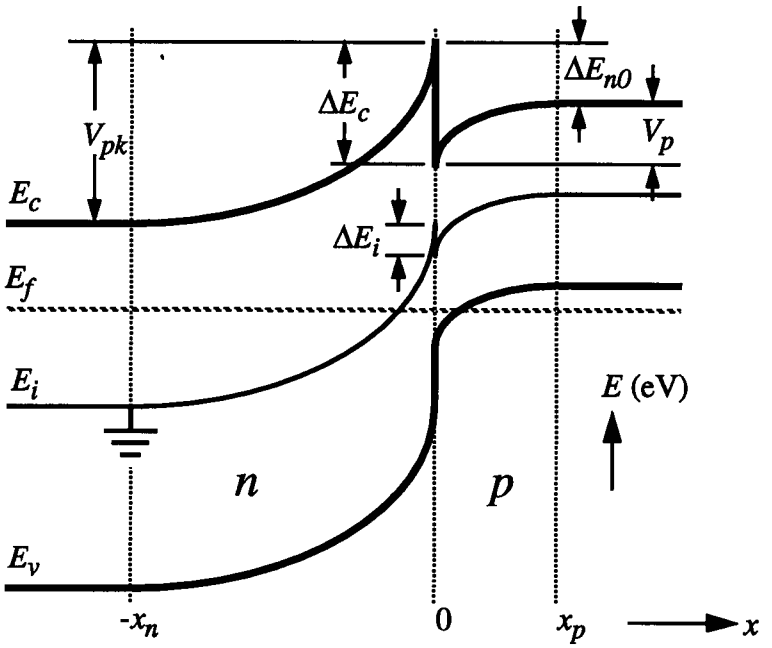


Fig. 5.3. Energy Band diagram for the EB SCR of an HBT under equilibrium conditions. Notice the discontinuity of ΔE_i in the intrinsic energy E_i.

The SCR currents on each side of the heterojunction follow from

$$J_{SRH} = q\int_{0}^{x_p} R_{SRH}dx + q\int_{-x_n}^{0} R_{SRH}dx$$
$$\equiv J_{SRH,B} + J_{SRH,E}. \tag{5.10}$$

This equation can be solved using eqns (5.5)-(5.9), but the solution cannot be made analytic with simple transcendental functions. A closed-form solution demands that some approximation be made for $\psi(x)$. Here we follow the linearisation procedure of Choo [92]; *i.e.*,

$$\psi(x) \approx \psi_{linear}(x) = \frac{q(V_{bi} - V_{BE})}{W_{BE}}(x + x_n), -x_n \le x \le x_p \tag{5.11}$$

where $x_n = W_{BE}N_{rat,d}$, $x_p = W_{BE}(1 - N_{rat,d})$, $W_{BE} = x_n + x_p$, and $N_{rat,d} = N_A/(N_A + N_D)$.

The linearisation of $\psi(x)$ in eqn (5.11) differs from that proposed by Parikh and Lindholm [90]. In [90], the linearisation is based upon a first order expansion of eqn (5.8) about the point where R_{SRH} is a maximum. The problem with this type of expansion is the R_{SRH} maximum must be well localised within the region of integration. If the R_{SRH} maximum is not within the region of integration (as it can be for reasonable operating biases), then the first order expansion proposed in [90] can lead to significant error. Eqn (5.11) alleviates this problem by appealing to the mean-value theorem to define the linearisation. In fact, as V_{BE} approaches V_{bi}, eqn (5.11) becomes exact.

Eqn (5.10) can now be evaluated using eqn (5.11) and

$$E_{fn} - E_{fp} = \begin{cases} qV_{BE} & x \le 0 \\ qV_{BE} - \Delta E_{fn} & x > 0 \end{cases}$$

$$E_{fn} + E_{fp} = \begin{cases} 2kT\ln\left(\frac{N_D}{n_{i,n}}\right) & x \le 0 \\ 2kT\ln\left(\frac{N_D}{n_{i,n}}\right) - qV_{BE} - \Delta E_{fn} & x > 0 \end{cases}$$

to yield

$$J_{SRH,B} = \frac{2qn_{i,p}W_{BE}}{\tau_p\Theta}\sinh\left[q\frac{V_{BE} - \Delta E_{fn}}{2kT}\right]\operatorname{atan}\left(\frac{Z_{0p} - Z_p}{Z_{0p}Z_p + 1}\right)$$
$$J_{SRH,E} = \frac{2qn_{i,n}W_{BE}}{\tau_n\Theta}\sinh\left[\frac{qV_{BE}}{2kT}\right]\operatorname{atan}\left(\frac{Z_n - Z_{0n}}{Z_{0n}Z_n + 1}\right) \tag{5.12}$$

with

$$\Theta = q(V_{bi} - V_{BE})/kT \qquad \tau_x = \sqrt{\tau_{p0,x}\tau_{n0,x}}$$

$$Z_n = \frac{N_D}{n_{i,n}}\sqrt{\frac{\tau_{p0,n}}{\tau_{n0,n}}}\exp\left[-\frac{qV_{BE}}{2kT}\right]$$

$$Z_{0n} = \frac{N_D}{n_{i,n}}\sqrt{\frac{\tau_{p0,n}}{\tau_{n0,n}}}\exp\left[-q\frac{2N_{rat}(V_{bi} - V_{BE}) + V_{BE}}{2kT}\right] \tag{5.13}$$

$$Z_p = \frac{n_{i,p}}{N_A}\sqrt{\frac{\tau_{p0,p}}{\tau_{n0,p}}}\exp\left[\frac{qV_{BE} - \Delta E_{fn}}{2kT}\right]$$

$$Z_{0p} = \frac{N_D}{n_{i,n}}\sqrt{\frac{\tau_{p0,p}}{\tau_{n0,p}}}\exp\left[-\frac{2qN_{rat}(V_{bi} - V_{BE}) + qV_{BE} + \Delta E_{fn} - 2\Delta E_i}{2kT}\right]$$

where it is assumed that E_t and E_i are coincident throughout the device [90], and, therefore, b from eqn (5.5) can be neglected for any reasonable operating conditions. Eqn (5.12) can be obtained from eqn (5.10) by using integral 2.423 #9 in [81]. In all cases, the final subscript of p and n refers to the p-type (base) and n-type (emitter) material regions respectively.

Eqn (5.12) is equivalent to eqns (20) and (21) in Reference [90]. It is, perhaps, in a more appealing form as it can be readily seen to be an extension of the usual equation for SCR recombination in homojunctions. Also, the unique feature to HBTs, quasi-Fermi-energy splitting, is explicitly brought out by the presence of ΔE_{fn} in the expression for $J_{SCR,B}$.

Finally, the linearisation used to obtain eqn (5.11) results in the use of the doping ratio $N_{rat,d}$, and not the voltage ratio N_{rat}, within eqn (5.13). As was stated at the start of Section 5.2, the effect of a non-uniform permittivity is quite small and can be neglected within the larger approximation of a linear $\psi(x)$. For this reason, it is assumed that for all practical devices encountered that $N_{rat} \approx N_{rat,d}$; in fact, for the parameters used in Section 5.4, this is only a 0.4% error.

5.2.2 Auger Recombination

As the doping concentrations increase, Auger recombination becomes an important consideration. There are two Auger processes of interest [93]: 1) a conduction band electron recombines with a heavy-hole, transferring it to the light-hole band; 2) a hole recombines with a conduction band electron, and the energy is transferred to another conduction band electron. In the first case, the recombination rate is proportional to p^2n, while in the second it is proportional to pn^2. When

the equilibrium recombination rates are included, the total Auger recombination rate is:

$$U_{Aug} = (A_n n + A_p p)\ (pn - n_i^2) \tag{5.14}$$

where the constants A_n and A_p are the electron and hole Auger coefficients respectively.

Using the same techniques employed in arriving at eqn (5.5), the above equation can be re-written as:

$$U_{Aug} = n_i^3 \exp\left(\frac{E_{fn} - E_{fp}}{2kT}\right) \sqrt{A_n A_p} \cdot \left[Z_{Aug} + \frac{1}{Z_{Aug}}\right]\left[\exp\left(\frac{E_{fn} - E_{fp}}{kT}\right) - 1\right] \tag{5.15}$$

where

$$Z_{Aug} = \sqrt{\frac{A_n}{A_p}} \exp\left(\frac{E_{fp} + E_{fn} - 2E_i}{2kT}\right).$$

The Auger recombination current is then given by

$$\begin{aligned} J_{Aug} &= q\int_0^{x_p} U_{Aug}\,dx + q\int_{-x_n}^{0} U_{Aug}\,dx \\ &\equiv J_{Aug,B} + J_{Aug,E} \end{aligned} \tag{5.16}$$

which can be solved using eqns (5.15), (5.13), (5.11), (5.9), (5.7) and (5.6) to give:

$$\begin{aligned} J_{Aug,B} &= \frac{2qn_{i,p}^3 W_{BE}}{\Theta Z_p Z_{0p} \tau_p} \exp\left[\frac{qV_{BE} - \Delta E_{fn}}{kT}\right] \sinh\left[\frac{qV_{BE} - \Delta E_{fn}}{2kT}\right] \cdot \\ &\qquad \cdot (Z_{0p} - Z_p)\ (A_{n,p}\tau_{n0,p} Z_p Z_{0p} + A_{p,p}\tau_{p0,p}) \\ J_{Aug,E} &= \frac{2qn_{i,n}^3 W_{BE}}{\Theta Z_n Z_{0n} \tau_n} \exp\left[\frac{qV_{BE}}{kT}\right] \sinh\left[\frac{qV_{BE}}{2kT}\right] (Z_n - Z_{0n})\ (A_{n,n}\tau_{n0,n} Z_n Z_{0n} + A_{p,n}\tau_{p0,n}). \end{aligned} \tag{5.17}$$

Eqn (5.17) gives the Auger recombination currents that are generated from the base and the emitter sides of the SCR.

5.2.3 Radiative Recombination

For materials where there is a direct bandgap, it is important to consider direct band-to-band radiative recombination. The rate at which radiative recombination occurs will be proportional to the *pn* product [94]. When the equilibrium recombination rates are included, the total radiative recombination rate is:

$$U_{Rad} = B\,(pn - n_i^2) \tag{5.18}$$

where the constant B is the radiative recombination coefficient.

The radiative recombination current is then given by

$$\begin{aligned} J_{Rad} &= q\int_0^{x_p} U_{Rad}\,dx + q\int_{-x_n}^{0} U_{Rad}\,dx \\ &\equiv J_{Rad,B} + J_{Rad,E} \end{aligned} \tag{5.19}$$

which can be solved using eqns (5.18), (5.9), (5.7) and (5.6) to give:

$$\begin{aligned} J_{Rad,B} &= q n_{i,p}^2 B_p W_{BE} (1 - N_{rat}) \left[\exp\left(\frac{qV_{BE} - \Delta E_{fn}}{kT}\right) - 1\right] \\ J_{Rad,E} &= q n_{i,n}^2 B_n W_{BE} N_{rat} \left[\exp\left(\frac{qV_{BE}}{kT}\right) - 1\right]. \end{aligned} \tag{5.20}$$

5.3 Current Balancing with the Neutral Region Transport Currents

It is clear from Fig. 5.2 that the electron currents to the right (*i.e.,* the base-side) of the heterojunction must equal the electron current due to the charge transport across the hetero-interface; *i.e.*,

$$J_{ThT} = J_{SCR,B} + J_{NB} + J_C \tag{5.21}$$

where

$$J_{SCR,B} = J_{SRH,B} + J_{Aug,B} + J_{Rad,B}. \tag{5.22}$$

The formulation given in eqns (5.21)-(5.22) was already treated in Section 2.2. Comparison of Fig. 5.2 with Fig. 2.3 shows an exact agreement. Therefore, the current balancing portrayed by eqns (5.21)-(5.22) can be solved using the models given in Section 2.2 if the various transport and recombination currents follow the general functional forms assumed in Chapter 2.

J_{ThT} is the transport current through the CBS that was solved for in Chapter 4. Eqn (4.63) shows that the flux F through the CBS ($\equiv J_{ThT}$) has the functional form assumed in Chapter 2 (see eqn (2.3) for $J_{n,1}$). This immediately allows the models of Chapter 4 to be used in concert with the models of Chapter 2 to solve for the collector and base terminal current densities J_C and J_B respectively. Looking again at eqns (2.3) and (4.63) shows that $J_{n,1}^0 = F_f$, and $\Delta E_{fn,1} = \Delta E_{fn}$. F_f includes both the thermionic emission and tunneling components involved in the transport over and through the CBS. Employing the formalisms of [51], F_f can be written as:

$$F_f = \gamma(V_{BE})\frac{4\pi q\sqrt{m_{y,1}m_{z,1}}\,(kT)^2}{h^3}e^{\frac{\mu_1}{kT}}e^{-\frac{qN_{rat}(V_{bi}-V_{BE})}{kT}} \approx q\gamma\upsilon N_D e^{-\frac{qN_{rat}(V_{bi}-V_{BE})}{kT}} \tag{5.23}$$

where υ is the electron thermal velocity given by

$$\upsilon = \sqrt{\frac{kT}{2\pi m_{x,1}}} \tag{5.24}$$

and $\gamma(V_{BE})$ is the tunneling factor (this is not to be confused with the γ in Chapter 4 used to characterise the mass barrier). With $\gamma = 1$, eqn (5.23) reduces to the thermionic injection current given by the last term in eqn (4.78). Essentially, γ is given by F_f/J_{th} where J_{th} is the thermionic injection current and $F_f = F_{f,CBS}$ given in eqn (4.93). Failure to include γ in eqn (5.23) will result in a severe overestimation of ΔE_{fn} [18] (and an underestimation of the collector current). Finally, μ_1 is the electrochemical potential relative to E_c formed by the doping N_D within the neutral emitter. The approximate solution given in eqn (5.23) is strictly valid only if the emitter is non-degenerately doped.

The neutral-base recombination current J_{NB} and the transport current through the neutral base J_C which must be used in eqn (5.21) follow from the standard, low-level injection solution to the continuity equation. Using the boundary condition that the driving potential at $x = x_p$ (*i.e.*, the start of the neutral base) is $V_{BE} - \Delta E_{fn}$ (see Fig. 5.1), and for the case of a single heterojunction structure operating in the forward active mode, the excess electron concentration near the collector is $\hat{n}(W_{nb}) = 0$, where W_{nb} is the neutral base thickness relative to $x = x_p$, then the expressions for these currents are

$$J_{NB} = \frac{qD_n n_{i,p}^2}{N_A L_{nb}}\frac{\cosh\left(\frac{W_{nb}}{L_{nb}}\right)-1}{\sinh\left(\frac{W_{nb}}{L_{nb}}\right)}\left[e^{\frac{qV_{BE}-\Delta E_{fn}}{kT}}-1\right] \approx \frac{qD_n n_{i,p}^2}{N_A L_{nb}}\frac{\cosh\left(\frac{W_{nb}}{L_{nb}}\right)-1}{\sinh\left(\frac{W_{nb}}{L_{nb}}\right)}\left[e^{\frac{qV_{BE}}{kT}}-1\right]e^{-\frac{\Delta E_{fn}}{kT}} \tag{5.25}$$

and

$$J_C = \frac{qD_n n_{i,p}^2}{N_A L_{nb}}\operatorname{csch}\left(\frac{W_{nb}}{L_{nb}}\right)\left[e^{\frac{qV_{BE}-\Delta E_{fn}}{kT}}-1\right] \approx \frac{qD_n n_{i,p}^2}{N_A L_{nb}}\operatorname{csch}\left(\frac{W_{nb}}{L_{nb}}\right)\left[e^{\frac{qV_{BE}}{kT}}-1\right]e^{-\frac{\Delta E_{fn}}{kT}} \tag{5.26}$$

where D_n is the effective electron diffusivity in the base, and L_{nb} ($=\sqrt{D_n\tau_{nb}}$) is the electron minority carrier diffusion length in the base. Observation of eqns (5.25) and (5.26) show they possess the functional forms of $J^0_{p,3}$ and $J^0_{n,3}$ respectively found in eqn (2.4) (*i.e.*, $J^0_{p,3} = J_{NB}(\Delta E_{fn}=0)$

and $J^0_{n,3} = J_C(\Delta E_{fn}=0)$). The approximate forms of eqns (5.25) and (5.26) introduce a negligible error over almost all bias conditions given the magnitude of $\exp(qV_{BE}/kT)$ compared to unity.

The last remaining task before the models of Section 2.2 can be employed to solve eqns (5.21)-(5.22) is to ensure that $J_{Rad,B}$, $J_{Aug,B}$, and $J_{SRH,B}$ have the same functional form as J_{NB} with respect to ΔE_{fn}. Clearly, $J_{Rad,B}$ in eqn (5.20) can be written in the same approximate form as J_{NB} with respect to ΔE_{fn}. However, it is not clear that the same is true for $J_{Aug,B}$ and $J_{SRH,B}$ in eqns (5.17) and (5.12) respectively. In order to see if $J_{Aug,B}$, $J_{SRH,B}$, and $J_{Rad,B}$ can be rewritten as:

$$\begin{aligned} J_{Aug,B}(V_{BE}, \Delta E_{fn}) &\approx J_{Aug,B}(V_{BE}, 0)\, e^{-\frac{\Delta E_{fn}}{kT}} \\ J_{SRH,B}(V_{BE}, \Delta E_{fn}) &\approx J_{SRH,B}(V_{BE}, 0)\, e^{-\frac{\Delta E_{fn}}{kT}} \\ J_{Rad,B}(V_{BE}, \Delta E_{fn}) &\approx J_{Rad,B}(V_{BE}, 0)\, e^{-\frac{\Delta E_{fn}}{kT}} \end{aligned} \tag{5.27}$$

a plot of the error between the full and the approximate forms in eqn (5.27) is constructed. Fig. 5.4 plots the relative error between the right and left sides of eqn (5.27) for $J_{Aug,B}$, $J_{SRH,B}$, and $J_{Rad,B}$ with V_{BE} fixed at 1.0V. Fig. 5.4 shows that the error in using the approximate relations in eqn (5.27) is less than 10 parts per billion. With such a small error in using eqn (5.27), it is justified to state:

$$\begin{aligned} J^0_{p,3}(V_{BE}) = J_{NB}(V_{BE}, \Delta E_{fn}=0) + J_{SRH,B}(V_{BE}, \Delta E_{fn}=0) + J_{Aug,B}(V_{BE}, \Delta E_{fn}=0) + \\ J_{Rad,B}(V_{BE}, \Delta E_{fn}=0). \end{aligned} \tag{5.28}$$

Eqns (5.21)-(5.22) can now be solved using the models in Section 2.2. The transport current J_T through the device (which is equal to the collector current) is given by eqn (2.7), with $J^0_{n,1} = F_f$ from eqn (5.23), $J^0_{n,3} = J_C(\Delta E_{fn}=0)$ from eqn (5.26), $J^0_{p,3}$ is given by eqn (5.28), $J^0_{p,2} = 0$ (*i.e.*, $\gamma_2 = 1$), and $J^0_{n,2} \gg (J^0_{n,1}, J^0_{n,3})$. Using the above produces:

$$J_T(V_{BE}) = \left[\frac{\gamma_3}{J^0_{n,1}} + \frac{1}{J^0_{n,3}}\right]^{-1} = \begin{cases} F_f(V_{BE})/\gamma_3 & \text{if } J^0_{n,1}/\gamma_3 \ll J^0_{n,3} \\ J_C(V_{BE}, \Delta E_{fn}=0) & \text{if } J^0_{n,1}/\gamma_3 \gg J^0_{n,3} \end{cases} \tag{5.29}$$

where γ_3 is given in eqn (2.7) as

$$\gamma_3 = \frac{J^0_{n,3} + J^0_{p,3}}{J^0_{n,3}} = \left.\frac{J_C + J_{NB} + J_{SRH,B} + J_{Aug,B} + J_{Rad,B}}{J_C}\right|_{\Delta E_{fn}=0}$$

and the V_{BE} dependence has been omitted for clarity. Eqn (5.29) embodies the two different

modes of operation that the HBT can function under; the first condition is where the CBS is responsible for current-limited-flow; while the second condition is the classic BJT regime of operation where the neutral base is responsible for current-limited-flow.

Finally, the base terminal current can be solved directly by using eqn (2.9) to yield:

$$J_B(V_{BE}) = J_T\left(\gamma_3 \frac{F_f + J_{SRH,E} + J_{Aug,E} + J_{Rad,E}}{F_f} - 1\right). \tag{5.30}$$

Or, ΔE_{fn} can be calculated by eqn (2.5) and substituted back into $J_{SCR,B}$ of eqn (5.22) and J_{NB}. J_B is then given by the sum of all the hole currents (*i.e.*, $J_B = J_{SCR,B} + J_{NB} + J_{Aug,E} + J_{SRH,E} + J_{Rad,E}$). The beauty of eqn (5.30) is it solves for the base terminal current without the need to determine the inner driving potential of ΔE_{fn}. However, if a detailed understanding of each component of the base terminal current is desired, then ΔE_{fn} must be solved for explicitly.

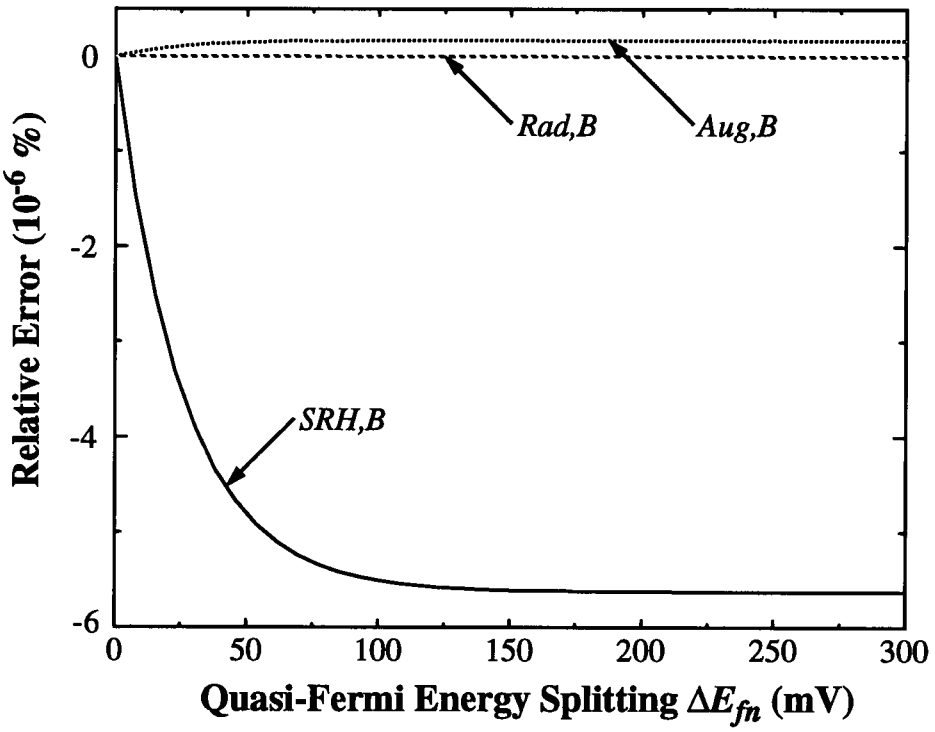


Fig. 5.4. Relative error between the approximate and exact forms given in eqn (5.27). The material parameters are given in Section 5.4, and V_{BE} is fixed at 1.0V.

Before leaving this section, it is important to verify that E_{fp} is indeed constant throughout the EB SCR. If there were a ΔE_{fp} present, it would have to be included in the emitter side hole current $J_{SCR,E}$ just like ΔE_{fn} has been included in $J_{SCR,B}$. Essentially, the same current balancing procedure given by eqns (5.21)-(5.22) needs to be performed regarding the transport of holes from

the neutral base, through the EB SCR, and finally through the neutral emitter. The same models for the electron case can be applied to the hole case, but using the appropriate material parameters for a hole. Using the HBT parameters of Section 5.4, then the hole transport current through the EB SCR is $3.9\times10^{-24}\exp(qV_{BE}/kT)\text{Acm}^{-2}$, and the hole transport current through the neutral emitter assuming a 3000Å emitter cap at a doping of 10^{20}cm^{-3} is $1.2\times10^{-27}\exp(qV_{BE}/kT)\text{Acm}^{-2}$. Clearly, the neutral emitter is the bottleneck to hole transport which validates the claim that ΔE_{fp} is indeed zero through the EB SCR. This does not have to be the case, and a device can be imagined where this is not true, leading to the requirement that hole transport be self-consistently solved with electron transport. It is quite interesting to realise that the valence band discontinuity ΔE_v does not limit the back injection of holes as the literature has lead the device community to believe. The back injection of holes is ostensibly eliminated by the reduced number of minority holes due to a small n_i that is characteristic of a wide bandgap material.

5.4 Full Model Results

The values used for material parameters, unless otherwise stated, are:

N_D: $5\times10^{17}\text{cm}^{-3}$; N_A: $1\times10^{19}\text{cm}^{-3}$; ε_{base}: $12.9\varepsilon_0$; $\varepsilon_{emitter}$: $11.9\varepsilon_0$; $\tau_{n0,n}$: $\tau_{n0,p}$: 5ns; $\tau_{p0,n}$: $\tau_{p0,p}$: 20ns; ΔE_c: 0.24eV; $n_{i,n}$: $4.21\times10^{3}\text{cm}^{-3}$; $n_{i,p}$: $2.25\times10^{6}\text{cm}^{-3}$; $\rightarrow \Delta E_i$: 77.3 meV, V_{bi}: 1.671V, $N_{rat,d}$: 0.952, N_{rat}: 0.956, $x_n(V_{BE}$=1.4 V): 271Å, $x_p(V_{BE}$=1.4 V): 13.6Å; $A_{n,n}$: $7.99\times10^{-32}\text{cm}^6\text{s}^{-1}$; $A_{p,n}$: $5.75\times10^{-31}\text{cm}^6\text{s}^{-1}$; $A_{n,p}$: 1.93×10^{-31} cm^6s^{-1}; $A_{p,p}$: $1.12\times10^{-30}\text{cm}^6\text{s}^{-1}$; B_n: $1.29\times10^{-10}\text{cm}^3\text{s}^{-1}$; B_p: $7.82\times10^{-11}\text{cm}^3\text{s}^{-1}$; D_n: $30\text{cm}^2\text{s}^{-1}$, W_{nb}: 1000Å.

Results for the SCR currents are shown in Fig. 5.5. The slopes of the curves are not constant, owing to the voltage dependence of W_{BE}, but it is clear that all the base-side SCR recombination components have about the same ideality factor (n), and that this is considerably less than that of the emitter-side SCR recombination current (which is dominated by $J_{SRH,E}$). Specifically, at V_{BE} = 1.2 V, $n_{SCR,E}$ = 1.90 and, adding all the base-side currents together, $n_{SCR,B}$ = 1.19. Furthermore, $n_{collector}$ = 1.14 at V_{BE} = 1.2 V due to the effects of ΔE_{fn}. These values are similar to those reported elsewhere [90], and deserve further comment because $J_{SCR,B}$ is so far removed from the "classical" value of n = 2.

With reference to eqn (5.12) for the SRH current, because $n_{i,n}$ is so low and N_{rat} is $\approx$ 1, Z_n and Z_nZ_{0n} are both » 1, leaving the atan term in eqn (5.12) to saturate at $\approx \pi/2$. The voltage depen-

dence of $J_{SCR,E}$ is thus determined by the sinh $qV_{BE}/2kT$ term and n approaches 2. Contrarily, for $J_{SCR,B}$, both Z_p and Z_{0p} are generally « 1, so the atan term modulates the sinh term and reduces the ideality factor from 2 towards 1.

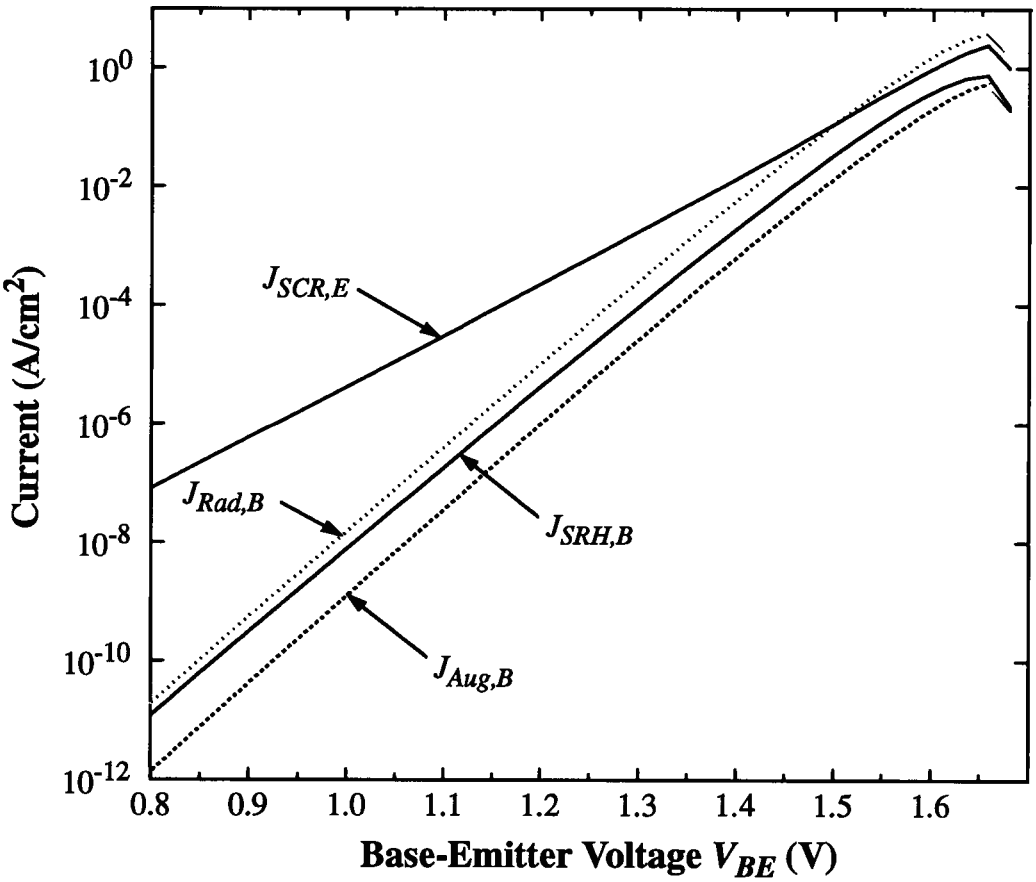


Fig. 5.5. Bias dependence of the SCR current from the emitter side, and the three components of the SCR current from the base side. Material parameters are taken from the start of Section 5.4.

The width of the SCR on the base side of the heterojunction is much less than that on the emitter side, and this fact alone, via N_{rat} in the Z_{0n} and Z_{0p} terms in eqn (5.12), would make $J_{SCR,B} \ll J_{SCR,E}$. However, the much larger n_i on the base side counterbalances this effect and allows the steeper-rising $J_{SCR,B}$ current to exceed $J_{SCR,E}$ beyond some forward bias. In the example shown in Fig. 5.5, this occurs around V_{BE} = 1.45 V. This transfer from an $n \approx 2$ slope to an $n \approx 1$ slope in the SCR current does not occur in a homojunction device as there is no spatial change in n_i to inflate the current in the more highly-doped side of the junction.

In practical HBTs it is possible to imagine that the minority carrier lifetime in the highly-doped base will be less than that in the emitter. Indeed, photoluminescence measurements on ma-

terial doped to $4x10^{19}cm^{-3}$ suggest that $\tau_n \approx 50ps$ [95], and a value of 30ps has been used to model some experimental devices [90]. Fig. 5.6 shows that reducing the base-side τ_n to 50ps causes $J_{SCR,B}$ to exceed $J_{SCR,E}$ at a bias of about 1.15 V. However, lest undue emphasis be placed upon the significance of this change-over, note from Fig. 5.6 that $J_{SCR,B}$ is always less than the quasi-neutral base recombination current J_{NB}. This indicates that, in practical devices, an observed change in base-current ideality factor from $n \approx 2$ to $n \approx 1$, will likely be due to a change from $J_{SCR,E}$-dominated current to a J_{NB}-dominated current. Only in circumstances where it is correct to attribute a much lower minority carrier lifetime to the base-side depletion region only, perhaps due to defects at the interface, can a situation be envisaged where $J_{SCR,B}$ could dominate over J_{NB}, and thus be responsible for the slope change to $n \approx 1$, which is often seen experimentally. The above point about the relative magnitudes of $J_{SCR,B}$ and J_{NB} is an important one as it puts into practical perspective the theoretically-interesting fact that $J_{SCR,B}$ has a different voltage dependence to that of $J_{SCR,E}$.

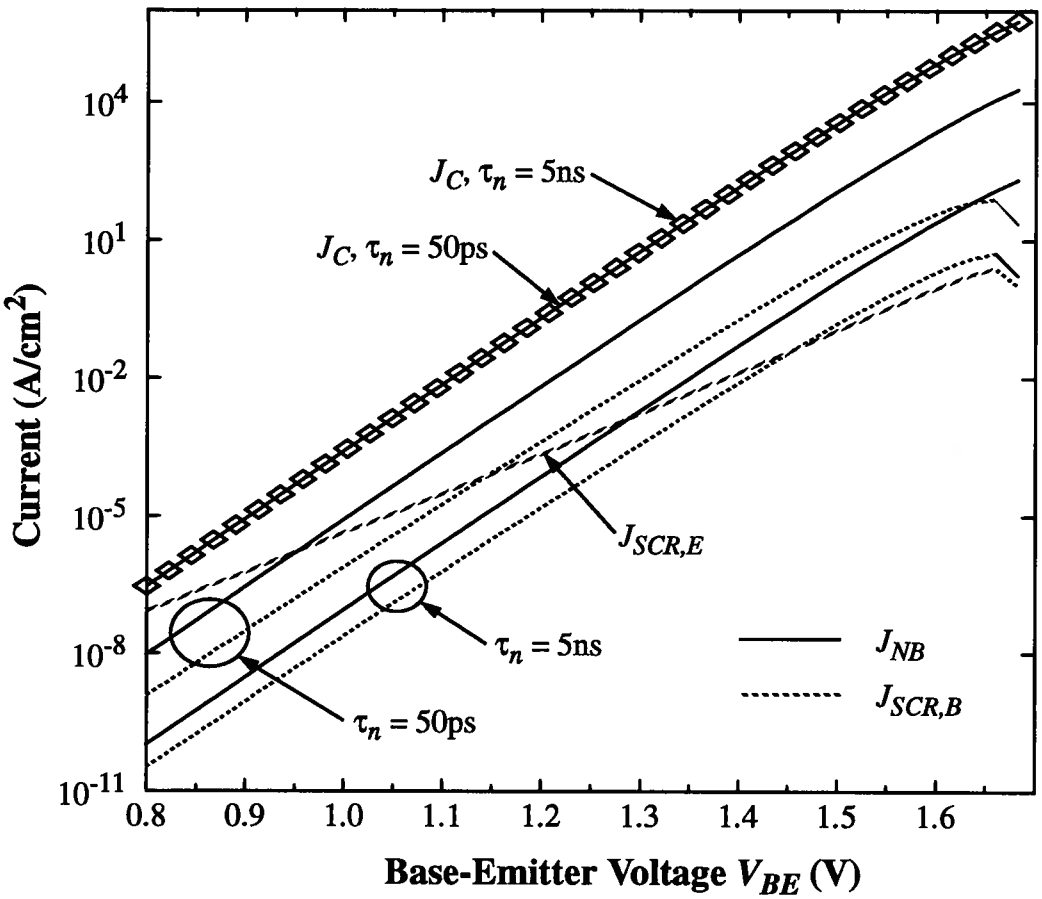


Fig. 5.6. Gummel plot showing the importance of including the emitter- and base-SCR current components in the computation of the total base recombination current. Material parameters are from the start of Section 5.4 for two values of τ_n.

While it is clear from the results of earlier work that ΔE_{fn} must be included in calculating $J_{SCR,B}$ [90], it is, perhaps, not evident how important it is to include $J_{SCR,B}$ in the balancing equation (5.21) to compute ΔE_{fn}. Fig. 5.7 provides an answer for the material properties considered here. By not including $J_{SCR,B}$ in eqn (5.21), yet using the subsequently-calculated ΔE_{fn} to eventually compute $J_{SCR,B}$, leads to a result which is indistinguishable from that of the "full model", where $J_{SCR,B}$ is included in the balancing equation. This is a consequence of $J_{SCR,B}$ being much less than J_{NB} and J_C. However, also from Fig. 5.7, note that it is grossly incorrect to not include ΔE_{fn} in the calculation of $J_{SCR,B}$. Because the electron quasi-Fermi energy splitting is so large for an abrupt junction [18], its omission leads to a large overestimation of $J_{SCR,B}$, and, consequently, to a severe underestimation of the current gain. It is difficult to imagine a practical situation where it might be necessary to include $J_{SCR,B}$ in the actual calculation of ΔE_{fn}. A possible scenario is one in which τ_n in the SCR is less than τ_n in the neutral base, perhaps due to interface defects, and that W_{nb} is much larger than the usual 1000 Å. The latter situation would reduce J_C, and the former would increase $J_{SCR,B}$ with respect to J_{NB}, thus making $J_{SCR,B}$ become more prominent in eqn (5.21). The effect of these changes is shown in Fig. 5.8. Even though the gain has been reduced to a very low value, it appears that there is still no need to include $J_{SCR,B}$ in the balancing equation.

To summarise the results from the analysis of this section: it is necessary to include ΔE_{fn} in the computation of $J_{SCR,B}$; but $J_{SCR,B}$ need not be included in the balancing equation to estimate ΔE_{fn}; and $J_{SCR,B}$ is not very important for devices based upon materials with the properties considered here, because $J_{SCR,B}$ is usually less than either J_{NB} or J_C. Of course, if parameters affecting Auger or radiative recombination in the SCR turn out to be greatly different than the values used here, then $J_{SCR,B}$ could become important.

One instance where $J_{SCR,B}$ will definitely be larger than calculated here is in the case of HBTs which are compositionally graded at the base-emitter junction. The grading gives the junction a more homojunction-like character, so ΔE_{fn} will be reduced, and $J_{SCR,B}$ increased correspondingly. However, because of the lower bandgap of the graded material in the emitter-side of the junction, $n_{i,n}$ is increased and, therefore, $J_{SCR,E}$ also. Thus it is not obvious whether $J_{SCR,B}$ is any more important in graded-junction HBTs than it is in abrupt-junction HBTs. The results of Parikh and Lindholm [90] suggest that $J_{SCR,E}$ remains the dominant current. One situation in

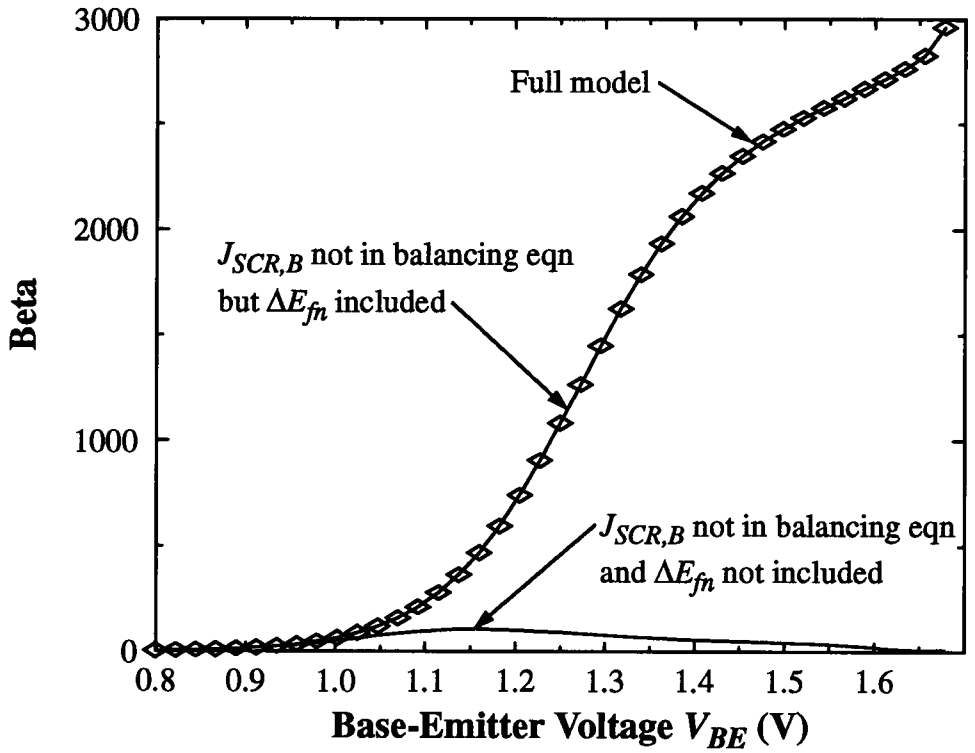


Fig. 5.7. Bias dependence of the current gain β, showing the relative importance of including $J_{SCR,B}$ in the calculation of ΔE_{fn}. Also shown is the dramatic error resulting from not including ΔE_{fn} in the calculation of $J_{SCR,B}$.

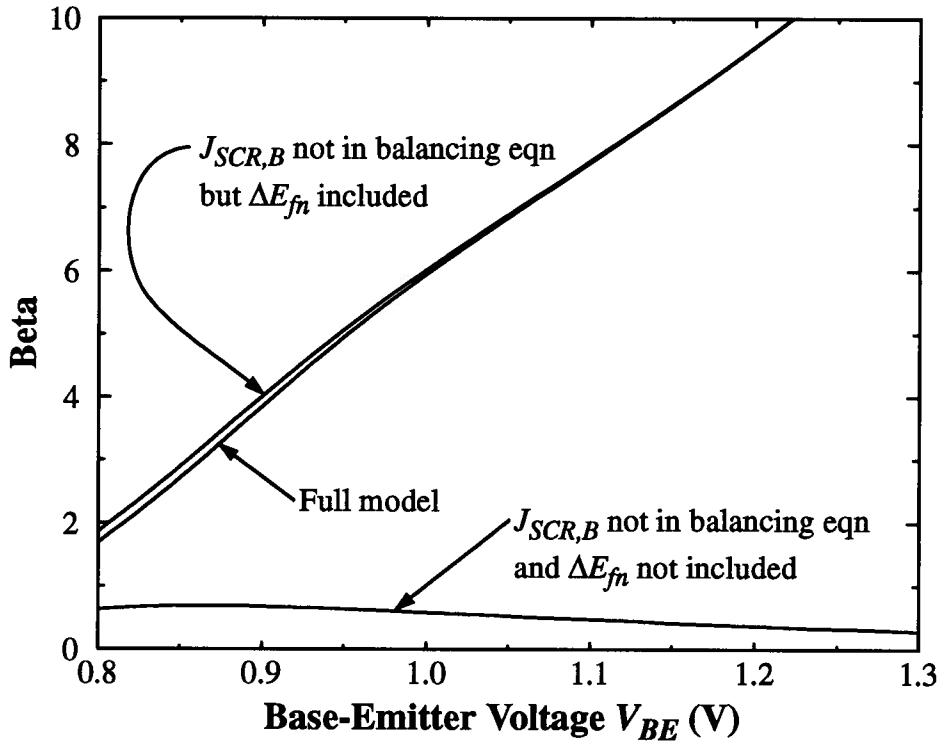


Fig. 5.8. Bias dependence of the current gain β for the case of W_{nb} increased to 5000 Å and τ_n in the SCR reduced to 5 ps. Even in this extreme case there is little error in not including $J_{SCR,B}$ in the balancing equation.

which $J_{SCR,B}$ could be increased without an associated increase in $J_{SCR,E}$ is when recombination at the exposed base surface is important. Providing a reasonable expression for this surface recombination current were available, it could be added to the right-hand side of eqn (5.22) and used in the current balancing to compute ΔE_{fn}. However, as can be deduced from Figs. 5.7 and 5.8, the inclusion of another component of $J_{SCR,B}$ will only effect the estimate of ΔE_{fn} if this new component is comparable in magnitude to J_C.

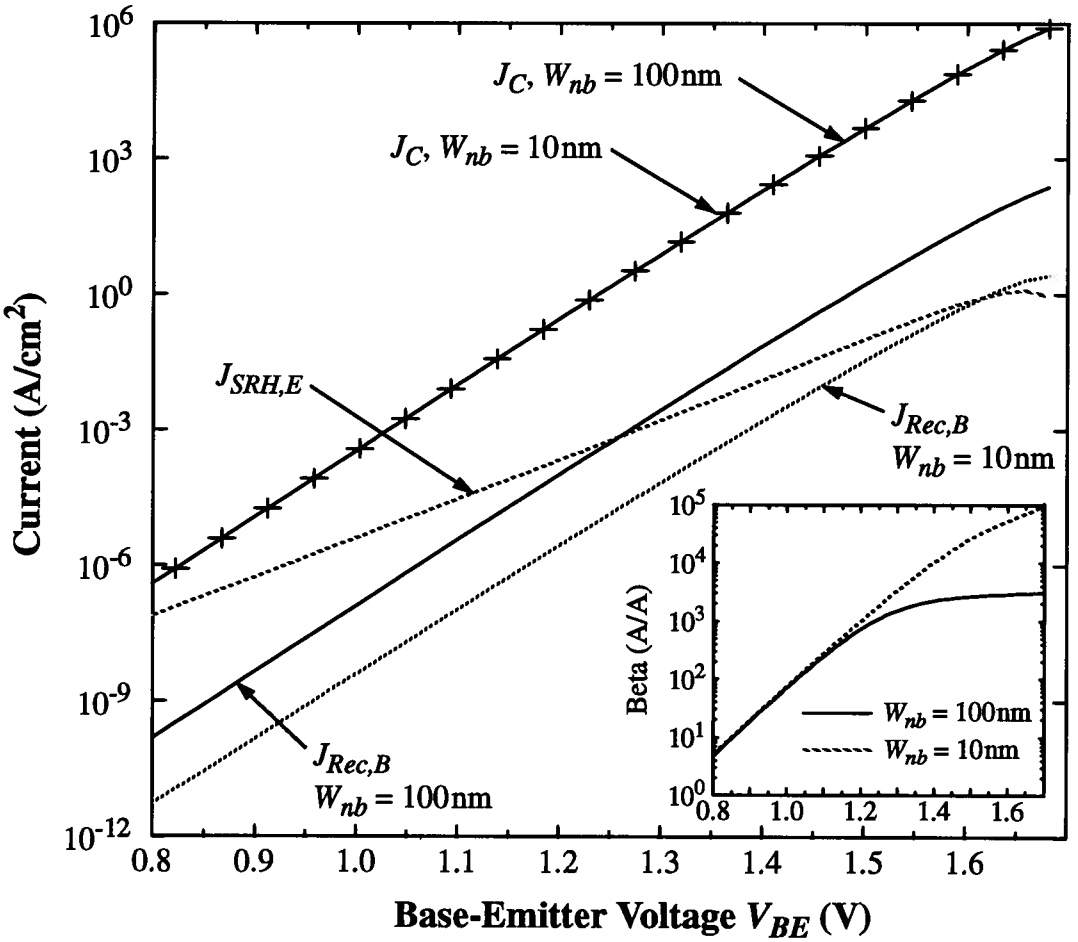


Fig. 5.9. Effect of changing the neutral base thickness W_{nb} when the CBS is responsible for current-limited-flow. Lowering W_{nb} leaves J_C and $J_{SRH,E}$ unchanged, but results in the reduction to the base side recombination current $J_{Rec,B}$ (= $J_{SCR,B}$ + J_{NB}). Under high bias, where $J_{Rec,B}$ dominates, β increases with reductions in W_{nb}. While under low bias, where $J_{SRH,E}$ dominates, β is unaltered by changes in W_{nb}.

Before leaving this section, it is interesting to see how current-limited-flow within the CBS leads to a mixing of the base and collector currents. For the HBT considered, the CBS is indeed responsible for current-limited-flow, so that $J_C \approx F_{f,\mathrm{CBS}}$. Thus, if the neutral base transport cur-

rent were increased by reducing W_{nb}, J_C would remain unchanged because the CBS already represent the bottleneck to charge transport through the device. However, the reduction to W_{nb} does have an effect on the device. Fig. 5.9 shows that the base-side components of the base terminal current are decreased by a reduction to W_{nb}. This decrease occurs due to a reduction of γ_3 in eqn (5.29) because relatively speaking, a shorter neutral base will provide fewer occasions for recombination. Therefore, opposite to what occurs in BJTs, the mixing of the collector and base currents due to current-balancing has coupled W_{nb} to the base instead of the collector current.

Finally, for the sake of completeness, Fig. 5.10 replots the currents displayed thus far using the linearised $\psi(x)$ from eqn (5.11) against the currents obtained with the full potential from the depletion approximation of eqn (5.8). As can be seen in Fig. 5.10, the error is indeed slight, and will be smaller than the uncertainty in the recombination parameters themselves. The linear $\psi(x)$ is not required to solve the radiative recombination current, so there is no approximation used.

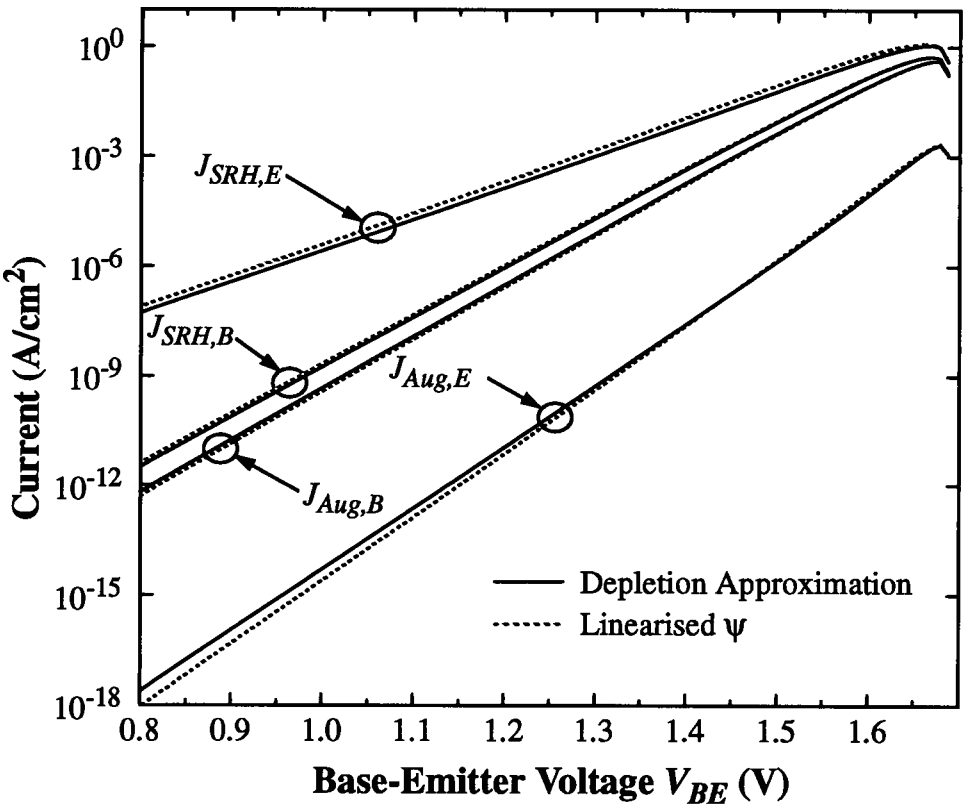


Fig. 5.10. Comparison of the recombination currents when ψ is given by the depletion approximation in eqn (5.8), and when it is given by the linearisation of eqn (5.11) (ΔE_{fn} is included).

5.5 Simple Analytic Diode Equations

For the purpose of including the various SCR recombination current components in HBT device simulators, it would be convenient if a simple closed-form solution for ΔE_{fn} existed. Further, if the various current components could be expressed as diode-like equations, then their representation in circuit simulators such as SPICE would be greatly facilitated. In this section, the approximations that need to be realised to effect these simplifications are discussed.

The starting point for the reduction of eqns (5.12), (5.17) and (5.20) to diode-like expressions is to examine the relative importance of the Z-terms which appear in the expressions for the SRH and Auger recombination currents. Fig. 5.11 shows the results from the full model calculations. From this figure, it appears reasonable to state that $Z_p \ll Z_{0p} \ll 1$, $Z_n \gg 1 \gg Z_{0n}$, and generally $Z_n Z_{0n} \gg 1$. The Z-terms Z_n, Z_{0n}, Z_{0p}, Z_p are representative of the amount of recombination at x_n, 0^-, 0^+ and x_p (see Fig. 5.1), respectively. For the condition $Z_p \ll Z_{0p}$ to remain valid, the depletion region on the base-side must not be vanishingly small. This can be ensured by having the doping density ratio $N_A/N_D \leq 30$. Contrarily, there is a lower limit to the allowable value of N_A/N_D, below which the recombination on the base-side of the depletion region becomes large and the inequality $Z_{0p} \ll 1$ is violated. This limit is $N_A/N_D \approx 3$. Therefore, keeping within the range $3 \leq N_A/N_D \leq 30$, and following the usual practice of expressing W_{BE} and Θ by their equilibrium forms, eqns (5.12) and (5.17) reduce to

$$
\begin{aligned}
J_{SRH,B} &\approx C_S \frac{N_D n_{i,p}}{\tau_{n0,p} n_{i,n}} \exp\left[\frac{\Delta E_i - qN_{rat}V_{bi}}{kT}\right] \exp\left[\frac{qN_{rat}V_{BE} - \Delta E_{fn}}{kT}\right] \\
J_{SRH,E} &\approx C_S \frac{\pi n_{i,n}}{2\tau_n} exp\left[\frac{qV_{BE}}{2kT}\right] \\
J_{Aug,B} &\approx C_S n_{i,p}^2 A_{p,p} N_A \exp\left[\frac{qV_{BE} - \Delta E_{fn}}{kT}\right] \\
J_{Aug,E} &\approx C_S n_{i,n}^2 A_{n,n} N_D exp\left[\frac{qV_{BE}}{kT}\right]
\end{aligned}
\tag{5.31}
$$

where

$$
C_S = kT \sqrt{\frac{2\varepsilon}{qN_A(1 - N_{rat})V_{bi}}}.
$$

Writing the radiative recombination currents in eqn (5.20) in similar form, gives

$$J_{Rad,B} \approx \frac{qC_S V_{bi}}{kT} n_{i,p}^2 B_p (1 - N_{rat}) \exp\left[\frac{qV_{BE} - \Delta E_{fn}}{kT}\right]$$
$$J_{Rad,E} \approx \frac{qC_S V_{bi}}{kT} n_{i,n}^2 B_n N_{rat} \exp\left[\frac{qV_{BE}}{kT}\right]. \tag{5.32}$$

Using these diode-like equations, along with the expressions for J_{NB} in eqn (5.25), J_C in eqn (5.26) and $J_{ThT} = F_f$ in eqn (5.23) in the balancing equation of eqn (5.21), yields a convenient expression for ΔE_{fn}; *i.e.*,

$$e^{\frac{q\Delta E_{fn}}{kT}} = \frac{J_{Recom}(V_{BE}) + q\gamma\upsilon N_D e^{-\frac{qN_{rat}(V_{bi} - V_{BE})}{kT}} + qD_n n_{B0} e^{\frac{qV_{BE}}{kT}} / W_{nb,e}}{J_{S,Recom} + q\gamma\upsilon N_D e^{-\frac{qN_{rat}(V_{bi} - V_{BE})}{kT}} + qD_n n_{B0} / W_{nb,e}} \tag{5.33}$$

where

$$J_{Recom}(V_{BE}) = J_{S,SRH,B} e^{\frac{qV_{BE}}{n_{SRH,B}kT}} + J_{S,Aug,B} e^{\frac{qV_{BE}}{n_{Aug,B}kT}} + J_{S,Rad,B} e^{\frac{qV_{BE}}{n_{Rad,B}kT}}$$

$$J_{S,Recom} = J_{S,SRH,B} + J_{S,Aug,B} + J_{S,Rad,B}$$

$$= C_S \frac{N_D n_{i,p}}{\tau_{n0,p} n_{i,n}} e^{\frac{\Delta E_i - qN_{rat}V_{bi}}{kT}} + C_S n_{i,p}^2 A_{p,p} N_A + \frac{qC_S V_{bi}}{kT} n_{i,p}^2 B_p (1 - N_{rat})$$

$$W_{nb,e} = L_{nb} \tanh\left(\frac{W_{nb}}{L_{nb}}\right)$$

$$n_{B0} = n_{i,p}^2 / N_A.$$

The values for the saturation currents and ideality factors in eqn (5.33) can be found either through a statistical fitting method, or from the analytic diode equations in eqns (5.31)-(5.32). Note that the n factors appearing in eqn (5.33) are independent of ΔE_{fn}. Their values, based upon the diode forms in eqns (5.31)-(5.32) are: $n_{SRH,B} = 1/N_{rat}$; $n_{Aug,B} = 1$; $n_{Rad,B} = 1$.

A comparison of the predictions of the diode forms in eqns (5.31)-(5.32) with results from the full expressions in eqns (5.12),(5.17), and (5.20) is shown in Fig. 5.12. The agreement is very good, with the only discrepancies occurring at very high forward bias. As V_{BE} approaches V_{bi}, the diminishing depletion-region thickness becomes a factor in that the depletion approximation no longer holds. Thus, for values of V_{BE} near V_{bi}, the voltage dependence of W_{BE} needs to be included, and the assumptions regarding the relative magnitudes of the Z functions re-addressed.

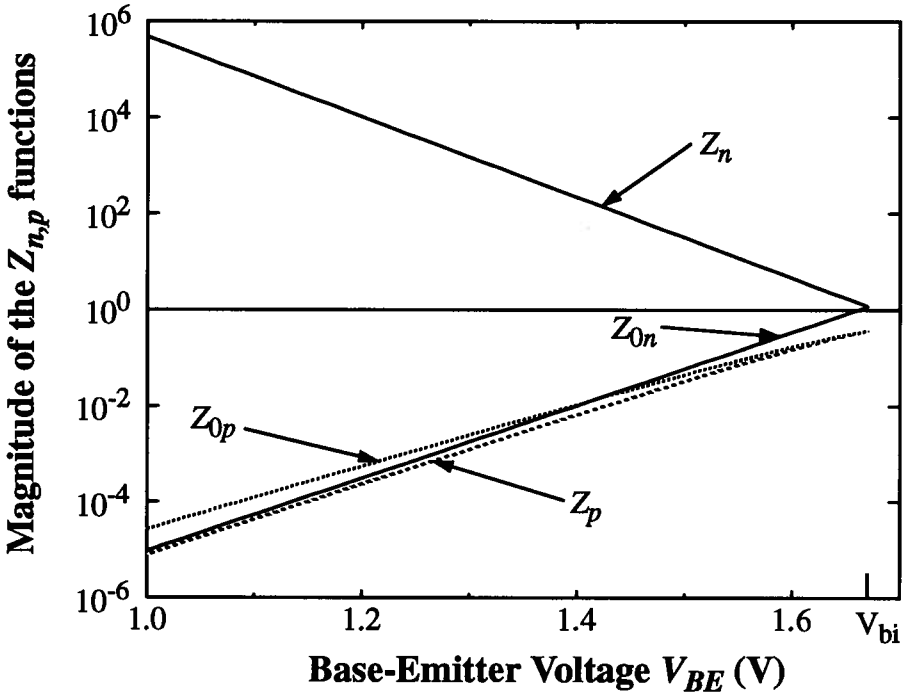


Fig. 5.11. *Z*-functions as computed from eqn (5.13) when using the material parameters from Section 5.4.

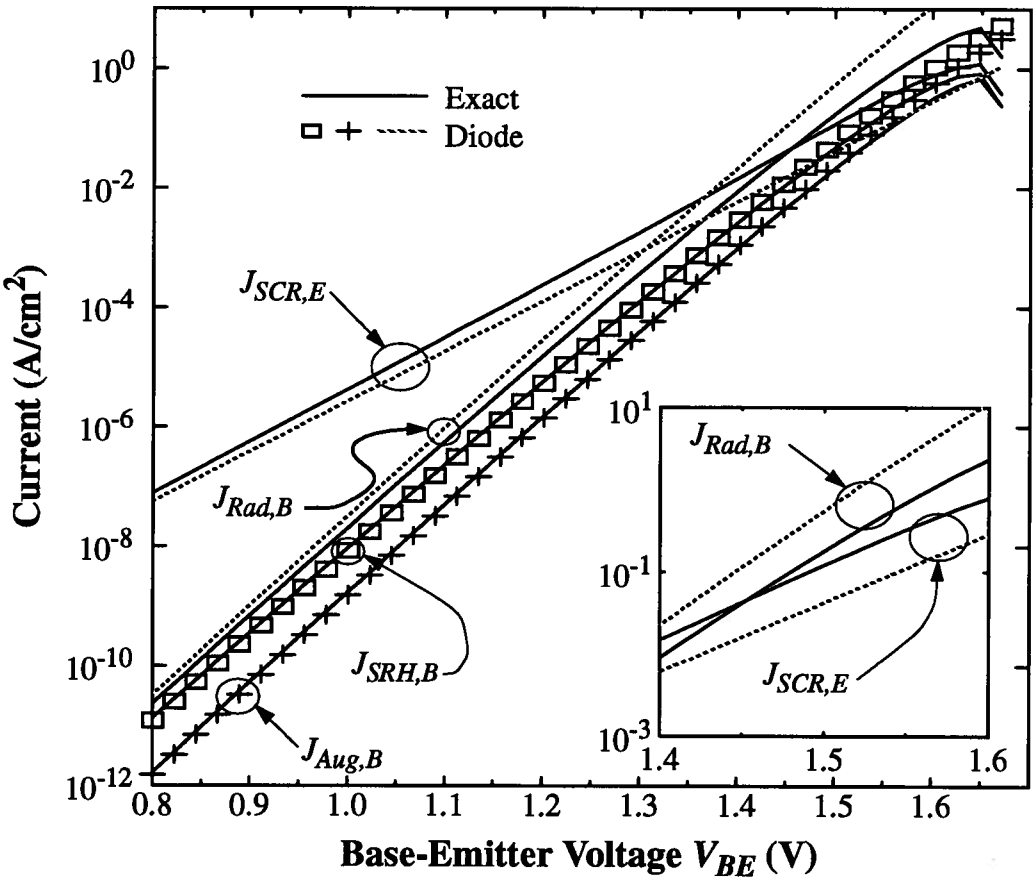


Fig. 5.12. Comparison of the full model and "diode-like" expressions for the SCR currents. The high-bias region of the figure is enlarged in the inset with $J_{SRH,B}$ and $J_{Aug,B}$ omitted for clarity.

If, as found to be the case for the material parameters used here, it is not necessary to include $J_{SCR,B}$ in the balancing equation, then the J_{Recom} and $J_{S,Recom}$ terms can be omitted from eqn (5.33). Finally, for most abrupt HBTs, the CBS represents the bottleneck to charge transport. In addition, for cases where $\beta \gg 1$, then $J_T = J_{ThT}$, and eqns (5.29) and (5.33) can be further simplified to give

$$J_T = J_{ThT} = q\gamma\upsilon N_D e^{-\frac{qN_{rat}(V_{bi}-V_{BE})}{kT}}$$

$$e^{\frac{q\Delta E_{fn}}{kT}} = \frac{D_n n_{B0}}{\gamma\upsilon W_{nb,e} N_D} e^{q\frac{N_{rat}V_{bi}+(1-N_{rat})V_{BE}}{kT}} \tag{5.34}$$

Substituting this expression for ΔE_{fn} into the diode forms in eqns (5.31)-(5.32), gives overall ideality factors for the base-side SRH, Auger and radiative currents of: $1/(2N_{rat} - 1)$, $1/N_{rat}$ and $1/N_{rat}$ respectively (where the bias dependence of the tunneling factor γ is not included).

From this study of space-charge region recombination currents in a typical AlGaAs/GaAs HBT, it can be concluded that:

1. recombination currents in the base-side SCR are generally less than the neutral-base current and, therefore, need not be included in the current-balancing equation used to compute the quasi-Fermi energy splitting ΔE_{fn} at the base-emitter junction;
2. however, when subsequently computing the base-side SCR currents, ΔE_{fn} must be taken into account if the gain is not to be grossly underestimated;
3. the ideality factor for the base-side SCR currents is closer to 1, than to the normally-used value of 2;
4. a simple, yet acceptably-accurate analytical expression for ΔE_{fn} can be derived;
5. the base-side SCR currents can be accurately represented by diode-like expressions, so facilitating their implementation in SPICE-style circuit simulators.

CHAPTER 6

The $Si_{1-x}Ge_x$ HBT

The previous chapters have presented a collection of models for the calculation of the transport and recombination currents in HBTs. Chapter 2 presented the generic models for current transport in an arbitrarily shaped device where there can be any number of sub regions within the defined regions of the emitter, base and collector. Chapter 4 presented the transport models for the movement of carriers through a forward biased *pn*-junction under the influence of a heterojunction. Chapter 5 presented the models for the recombination currents that occur both in the neutral regions of the device (specifically the base and the emitter), and the forward-biased EB SCR. Also included in Chapter 5 were models for the transport of charge through the neutral regions of the device. Finally, Chapter 3 presented the models for the calculation of the base transit time based upon an optimisation of either the base doping, or the base bandgap, or both. In all of the work presented thus far, no assumptions have been made that depended upon a specific attribute of a given material system. Thus, the models contained within this thesis are general, and may be applied to the study of an arbitrary HBT created within an arbitrary material system.

Even though the models presented within this thesis are indeed applicable to any material system, whenever an analysis of a specific model was performed, the material system of $Al_xGa_{1-x}As$ was invariably chosen for the study. The reason for choosing the $Al_xGa_{1-x}As$ material system is that current-day technologies for HBTs prefer this material system. The dominance of the $Al_xGa_{1-x}As$ material system stems mainly from the fact that the lattice mismatch over the usable range of Al content (*i.e.*, $0 \le x \le 0.45$) is under 0.07% [61]. This nearly ideal lattice match allows for an arbitrary film thickness because there will be virtually no strain placed upon the lattice at the heterojunction. Coupled with the lattice-matched characteristic, the $Al_xGa_{1-x}As$ material system can also provide for large changes to the bandgap (ΔE_g) [61]. However, compound semiconductors like GaAs and AlAs have numerous undesirable features when it comes to manufacturing. The $Al_xGa_{1-x}As$ material system lacks a usable native oxide, is a poor thermal conductor, cannot be pulled into wide ingots which results in small wafer diameters, is brittle, suffers from a high defect density, cannot employ ion implantation for bipolar devices, exposed surface layers have high recombination velocities, cannot be used in low-power applications because of the large V_{bi} inherent with large bandgaps, does not etch easily and generally lacks an abrupt end-point detection for etching, and finally is expensive to manufacture. Given all of these manufacturing and electrical drawbacks, however, the lattice-matched attribute is important enough to make $Al_xGa_{1-x}As$ the preferred material system for the construction of HBTs.

Essentially all of the manufacturing issues with regard to the $Al_xGa_{1-x}As$ material system are solved by using the $Si_{1-x}Ge_x$ material system: save one issue. At issue with the $Si_{1-x}Ge_x$ material system is its large lattice mismatch. The Ge lattice is 4.2% larger than the Si lattice [96]. Even if the Ge content is constrained to be under 20% (*i.e.*, $0 \le x \le 0.20$) there would still be a 0.84% lattice mismatch between a $Si_{0.8}Ge_{0.2}$ film and a Si substrate. The issue with a lattice mismatch of around 1% is that to commensurately place an epitaxial film upon a given substrate would result in a strain within the film that would be large enough to tear the film apart [97-99]. If strain were allowed to tear the film and form dislocations, then deep states would form along the heterojunction interface which would greatly enhance recombination. Since the heterojunction will be formed in the middle of the EB SCR of the HBT, a plane of recombination centres at the heterojunction would result in an intolerably high base current; large enough to reduce β below 1.

There is no physical way to alter the bulk lattice constant of a material or alloy. However, if the epitaxial film is grown thin enough and at a low enough temperature, it will conform to the substrate [99]. Under such conditions, the epitaxial layer is said to be commensurately strained to fit the substrate, and the layer itself will be pseudomorphic [99]. Pseudomorphic films are thus strained in order to maintain the in-the-growth-plane crystalline structure of the substrate. The key to obtaining a pseudomorphic film is to ensure that the layer thickness is below the critical thickness h_c [99]. However, in order to maintain a pseudomorphic film, and ensure that it does not relax back to its bulk lattice constant, subsequent exposure of the layer to high temperature environments must be severely limited. In the past 5 years, great progress has been made at IBM in the quality of $Si_{1-x}Ge_x$ pseudomorphic films [31]. These developments have shown great potential regarding operating speeds [100-102], so much so that many other companies including the Japanese at NEC [103] are developing SiGe IC processes. Through the recent successes regarding the high quality growth of pseudomorphic $Si_{1-x}Ge_x$ films, the $Si_{1-x}Ge_x$ material system is fast becoming a practical alternative for the manufacture of HBT-based ICs. In fact, with the massive installed base of Si-based IC manufacturing, coupled with the ability to integrate $Si_{1-x}Ge_x$ films into the process, it is expected that $Si_{1-x}Ge_x$ will rapidly displace $Al_xGa_{1-x}As$ as the preferred material system for the manufacture of HBT-based ICs.

This chapter will apply the general models obtained from the previous chapters to the study of HBTs based within the $Si_{1-x}Ge_x$ material system. Due to the complex nature of $Si_{1-x}Ge_x$ under

the influence of strain, a number of extensions to the work of previous chapters is necessary. Most importantly, due to the indirect nature of the $Si_{1-x}Ge_x$ energy bands, there are six separate conduction band valleys [104] (compared to only one valley in a direct semiconductor such as GaAs). Each of these conduction band valleys will transport electrons. Since strain breaks the degeneracy of the six conduction band valleys, it will become important to consider electron transport within each valley separately. Once the needed extensions to the models of the previous chapters have been determined, a study of current-day SiGe HBTs can be performed. Furthermore, it will be shown that the use of strain can be turned into a tool for the HBT developer, instead of being seen only as a liability in terms of critical layer thickness.

6.1 The Effect of Strain on $Si_{1-x}Ge_x$

The use of pure unstrained crystals of Ge in the formation of SiGe HBTs is possible, but due to the large lattice mismatch (~4%), would result in a high defect density at the heterointerface, severely degrading device performance. Furthermore, if only pure Si or Ge crystals were used in the formation of HBTs, there would be a considerable limitation imposed upon the ability to engineer the bandgap within the HBT. Instead, pseudomorphic $Si_{1-x}Ge_x$ films, that are commensurately strained to become lattice matched to the substrate (which is pure Si in present day devices), are used. These pseudomorphic $Si_{1-x}Ge_x$ layers will remain strained without relaxing as long as the layer thickness remains below the critical thickness h_c [97-99,105]. (For $Si_{0.70}Ge_{0.30}$ grown on {100} Si substrates the critical layer thickness is 600Å, while for $Si_{0.45}Ge_{0.55}$ it is only 100Å). Thus, unlike $Al_xGa_{1-x}As$, which is essentially lattice matched to GaAs and thus has no critical layer thickness, SiGe HBTs can have considerably less freedom in the choices for layer thicknesses.

The key to manufacturing SiGe HBTs is the commensurate growth of strained $Si_{1-x}Ge_x$ layers to the underlying substrate. However, the strain in the plane of growth results in a distortion of the crystal structure that breaks the cubic symmetry and causes the crystal unit cell to become tetragonal. With the breaking of the cubic symmetry comes a change to the dispersion relations for the energy of the Bloch electrons versus wave vector $\boldsymbol{k}$. The most important effect of this symmetry breaking is the relative change to the energy of the conduction band minima and the valence band maxima in $\boldsymbol{k}$-space.

Constant energy surfaces near to the conduction band minima for pure unstrained Si and Ge are shown in Fig. 6.1. Looking at the case of Si, there are six separate but degenerate conduction band minima located along the ⟨100⟩ directions at the Δ point (which is 80% from the zone centre at Γ to the Brillouin zone edge at X). For alloys of $Si_{1-x}Ge_x$, these six minima are dependent both on the alloy content x and on the state of strain. Take, for an example, $Si_{1-x}Ge_x$ grown on a Si substrate with the direction of growth parallel to [001]. As x moves from 0 to 1, the $Si_{1-x}Ge_x$ layer moves from an unstrained cubic structure to a compressively strained tetragonal structure [99,105]. As the strain decreases from zero (compression being negative strain), we find that the degeneracy of the six minima is lifted [105-108]. The two minima aligned to the normal of the interface plane (*i.e.*, parallel to the direction of growth) remain degenerate and are raised in energy, while the other four minima parallel to the interface plane also remain degenerate but are lowered in energy. For the case of $Si_{1-x}Ge_x$ grown on a Ge substrate with the direction of growth still parallel to [001], the situation is reversed. As x moves from 1 to 0, the $Si_{1-x}Ge_x$ layer moves from an unstrained cubic structure to an expanded, tensile-strained tetragonal structure. For this case of tensile strain, as the strain increases from zero, the two minima normal to the interface plane are lowered in energy, while the other four minima parallel to the interface plane are raised in energy. Thus, we find that there are now two types of Δ conduction band minima in a strained $Si_{1-x}Ge_x$ film; those parallel (which will be termed E_c^4) and those perpendicular (which will be termed E_c^2) to the interface plane. Therefore, depending on the sign of the strain tensor (*i.e.*, either compressive or tensile), either the E_c^4 or the E_c^2 bands will form the ultimate conduction band.

The valence band also suffers considerable change due to the symmetry breaking caused by strain. The valence band of pure, unstrained Si and Ge (or for that matter, all semiconductors), is composed of what should be three degenerate bands. These three bands are the light-hole (lh), heavy-hole (hh) and split-off bands (so). When the interaction of the electron's internal angular momentum (spin), is coupled with its orbital angular momentum (termed spin-orbit coupling), the degeneracy of the so band is lifted [109,110]. The resultant interaction leaves the lh and hh bands degenerate with the so band maxima moved to a lower energy (see Fig. 6.2). The symmetry breaking caused by strain goes on to lift the degeneracy of the lh and hh bands. As in the conduction band, the valence band maxima is dependent both on the alloy content x and on the state of strain [105-108]. Returning to the case of $Si_{1-x}Ge_x$ grown on a Si substrate, with the direction of growth parallel to [001], as x moves from 0 to 1 the $Si_{1-x}Ge_x$ layer experiences an increasing compressive

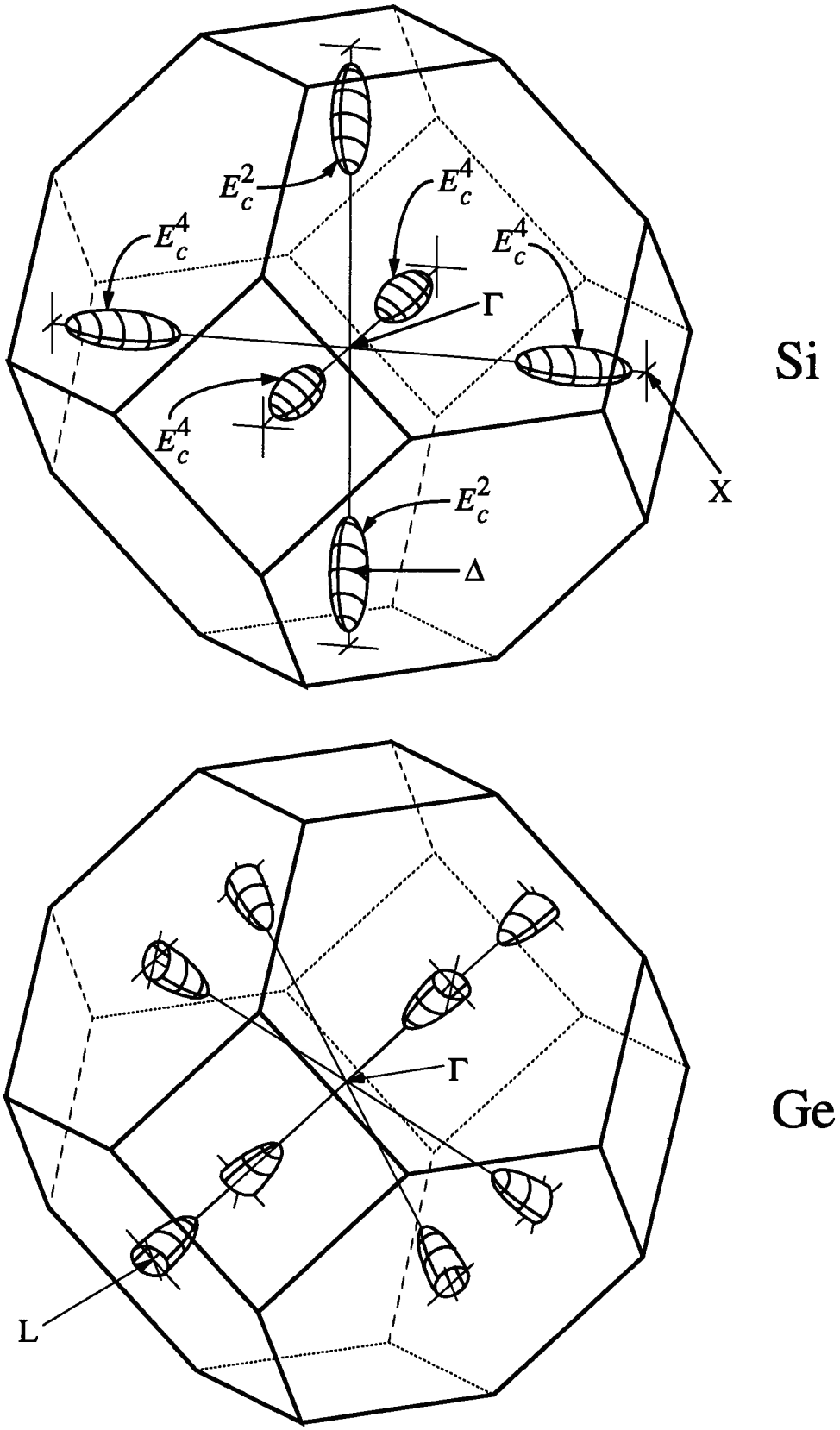


Fig. 6.1. First Brillouin zone showing (in k-space) the constant energy surfaces near the bottom of the conduction band for Si and Ge. Also shown are the designations for the symmetry points and the degenerate bands E_c^4 and E_c^2 in strained $Si_{1-x}Ge_x$ with the growth direction along [001].

strain. The result of compressive strain is an increase in the maxima of the hh band relative to the lh band, accompanied by a decrease in the maxima of the so band relative to the lh band. Under tensile strain, however, the effect is reversed for the lh and hh bands (but not the so band). Returning to the case of $Si_{1-x}Ge_x$ grown on a Ge substrate, with the direction of growth still parallel to [001], as x moves from 1 to 0 the $Si_{1-x}Ge_x$ layer experiences an increasing tensile strain. The result of tensile strain is an increase in the maxima of the lh band relative to the hh band. However, there is still a decrease in the maxima of the so band relative to the hh band. Therefore, strain eliminates the degeneracies of all the valence bands, with the so band always moved to lower energies. However, depending on the sign of the strain tensor (*i.e.*, either compressive or tensile), either the lh or the hh band will form the ultimate valence band.

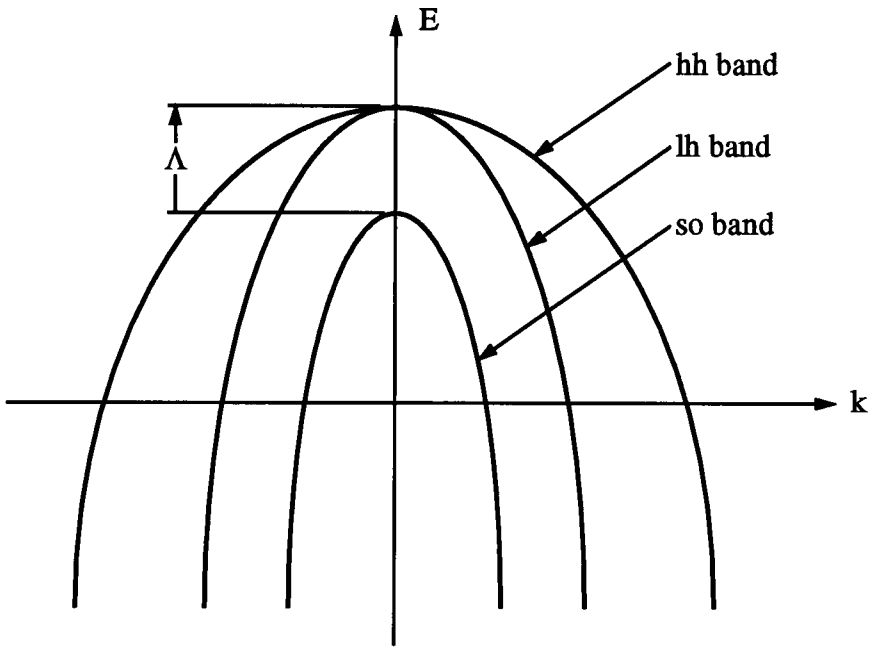


Fig. 6.2. Valence bands in unstrained $Si_{1-x}Ge_x$. The light hole (lh) and heavy hole (hh) bands remain degenerate for all values of Ge alloy composition x (only in the bulk state where there is no strain present). However, the split off (so) band maxima changes in energy with alloy content, where $\Delta(x) = 0.044 + 0.246x$eV.

The previous paragraphs have outlined that the energy of the conduction band minima and the valence band maxima change under the effect of strain, while their position in $\boldsymbol{k}$-space remains unaltered. However, it is also important to ascertain the effect of strain on the shape of the band in $\boldsymbol{k}$-space, as this will set the effective mass which determines the velocity of the carrier and its probability for tunneling. Considering the valence band first, the effective mass for the lh and hh

bands in pure unstrained Si and Ge are quite different. Therefore, as the Ge alloy content in the $Si_{1-x}Ge_x$ layer changes, there must be a change to the shape of the band in $\boldsymbol{k}$-space regardless of the strain state. To account for this varying shape of the lh and hh bands, a linear interpolation between the experimental values for the lh and hh masses in Si and Ge is used to arrive at the appropriate masses for the $Si_{1-x}Ge_x$ layer [111]. It is further assumed that the effect of strain is negligible with regard to the shape of the band in $\boldsymbol{k}$-space. This leads to:

$$\begin{aligned} m_{hh} &= 0.49 - 0.21x \\ m_{lh} &= 0.16 - 0.116x \end{aligned} \tag{6.1}$$

where x is the Ge alloy content, and the masses are a fraction of the electron rest mass m_e (the Si and Ge hole masses are based upon [96]). The lh and hh effective masses are maintained separately instead of combining them into an effective density of states mass because under the influence of strain, the degeneracy breaking will result in a change to the effective density of states mass (see Section 6.3).

For the conduction band, it is assumed that the conduction sub-bands E_c^4 and E_c^2 do not change shape with either a change in the Ge alloy content or the state of strain [107,112]. To first order in the strain tensor there must be a change to the effective mass for the electrons because the reciprocal lattice vector is being changed. However, this change will be relatively small as the maximum change to the reciprocal lattice vector is 4.2% over the entire range of Ge alloy content. As for the effect of the Ge alloy content, it is important to realise that Si and Ge (and therefore $Si_{1-x}Ge_x$) have conduction band minima at Δ and at L. The difference between Si and Ge is that the Δ minima form the ultimate conduction band in Si while the L minima form the ultimate conduction band in Ge. For $Si_{1-x}Ge_x$ the Δ minima typically form the ultimate conduction band. However, if the Ge alloy content is high enough, then the ever-present L minima within the $Si_{1-x}Ge_x$ alloy will form the ultimate conduction band [113]. It is therefore postulated [107,112] that the electron effective mass for the E_c^4 and E_c^2 bands are the same as that for the Δ minima in Si, while the effective masses for the L minima are the same as the Ge effective mass; *i.e.*, [96];

$$\begin{aligned} m_t(\Delta) &= 0.19 \\ m_l(\Delta) &= 0.98 \\ m_t(\mathrm{L}) &= 0.082 \\ m_l(\mathrm{L}) &= 1.64 \end{aligned} \tag{6.2}$$

Therefore, there is no change to the effective electron mass, for a given band, with either a change

to the Ge alloy content or the state of strain. However, in similar fashion to the valence band, the effective density of states mass will change with strain depending on which band forms the ultimate conduction band.

The qualitative features that strain and Ge alloy content impart to the $Si_{1-x}Ge_x$ layer have been presented. Using empirical deformation-potential theory [114-116], the quantitative features are now presented. The reason for using empirical deformation-potential theory, where the deformation potentials are measured and not derived from first principles, is that current-day solid-state quantum mechanics is not sophisticated enough to predict the desired results with any reasonable tolerance (errors on the order of 1eV are standard). To this end, the problem of including the strain state and the Ge alloy content is broken down into two independent problems. First of all, experimental measurements of the $Si_{1-x}Ge_x$ bulk bandgap (*i.e.*, unstrained) are performed over the entire range of $0 \le x \le 1$ to produce the function $E_g(x)$. Thus, $E_g(x)$ contains all of the Ge alloy effects. Then, empirical deformation-potential theory is used to determine the amount of degeneracy splitting that occurs within the sub-bands of the conduction and valence bands due to the addition of strain. Adding together $E_g(x)$ with the results from deformation-potential theory produces the total change to the various bands within the $Si_{1-x}Ge_x$ layer.

Beginning with the calculation of $E_g(x)$, in the seminal works of [113,117] the necessary experimental measurements on the bandgap of bulk and strained $Si_{1-x}Ge_x$ have been performed. It has been found that for $x < 0.85$, the Δ minima form the ultimate conduction band minima in $Si_{1-x}Ge_x$. However, in the range $0.85 \le x \le 1$, the L minima form the ultimate conduction band minima in bulk $Si_{1-x}Ge_x$. Concentrating on the Δ minima alone, then using a quadratic fit to the data in [113] produces:

$$E_g(x) = \begin{cases} E_{g,\,\mathrm{Si}} - 0.51446x + 0.31164x^2 & x \le 0.732 \\ E_{g,\,\mathrm{Si}} - 0.1501 - 0.0813x & x > 0.732 \end{cases} \tag{6.3}$$

where x is the Ge alloy content, $E_{g,\mathrm{Si}}$ is the bulk Si bandgap, and all values are in eV.

Eqn (6.3) gives the $Si_{1-x}Ge_x$ bulk bandgap from the top of the valence band to the bottom of the Δ minima in the conduction band. Caution must be exercised when using eqn (6.3) for $x > 0.85$ as the L minima will form the ultimate conduction band in bulk $Si_{1-x}Ge_x$ material. However, the strain imparted to the $Si_{1-x}Ge_x$ layers used in HBTs is generally sufficient to reduce some of the Δ minima below the L minima even as x approaches 1 (*i.e.*, pure Ge) [106]. For this reason it will be assumed that the L minima can be ignored. However, the energy of the L minima change much

more rapidly for a given change to x than the Δ minima do. Thus, it would be possible to achieve larger band offsets using the L minima versus the Δ minima, or to achieve the same band offsets but with a smaller change in x (which would help address the critical layer thickness problem). The drawback to using the L minima is the substrate would have to be essentially Ge, and not Si, grown along $\langle 111 \rangle$. But, given the much higher mobilities in Ge versus Si, then SiGe HBTs based upon the L minima should outperform current SiGe HBTs based upon the Δ minima.

$E_g(x)$ in eqn (6.3) solves the first problem of including the alloy effects into the conduction and valence bands of $Si_{1-x}Ge_x$. The second problem of including the effect of strain is now addressed. Fig. 6.3 shows the effect of in-plane biaxial tension and compression. Fig. 6.3(a) shows the case where the substrate lattice constant a_s is larger than the alloy lattice constant a_a. The commensurate growth of the alloy layer to the substrate forces the in-plane alloy lattice constant to match a_s. In so doing, a biaxial in-plane tension results in the pseudomorphic alloy film. In an attempt to lower the energy contained within the film, the out-of-plane alloy lattice constant compresses below a_a. The pseudomorphic alloy layer will then have a larger in-plane lattice constant when compared to the out-of-plane alloy lattice constant, leading to a tetragonal crystal instead of a cubic one. Contrarily, Fig. 6.3(b) shows the case where the substrate lattice constant a_s is smaller than the alloy lattice constant a_a. The commensurate growth of the alloy layer to the substrate forces the in-plane alloy lattice constant to match a_s. In so doing, a biaxial in-plane compression results in the pseudomorphic alloy film. In an attempt to lower the energy contained within the film, the out-of-plane alloy lattice constant expands past a_a. The pseudomorphic alloy layer will now have a smaller in-plane lattice constant when compared to the out-of-plane alloy lattice constant, which again leads to a tetragonal crystal instead of a cubic one. It is the fact that the pseudomorphic alloy layer has broken the cubic symmetry of the original lattice that leads to the changes in the conduction and the valence bands.

The initial applied stress tensor to the alloy layer can be viewed as a uniaxial stress accompanied by a uniform hydrostatic pressure applied over the entire cell. If the in-plane interface is parallel to the x-y plane, with the direction of growth parallel to the z-direction, then the initial applied stress is [108]:

$$\text{Applied stress} = \begin{bmatrix} \tau & & \\ & \tau & \\ & & 0 \end{bmatrix} = \tau\bar{1} + \begin{bmatrix} 0 & & \\ & 0 & \\ & & -\tau \end{bmatrix} \tag{6.4}$$

where the growth is in the [001] direction, and a blank location in the tensor is zero. The first term on the right of eqn (6.4) is the hydrostatic pressure applied to the overall cell, while the second term is the uniaxial stress, of opposite direction to the biaxial stress, applied to the out-of-plane lattice constant. Therefore, the symmetry breaking of the alloy's unit cell occurs along the direction of growth (*i.e.*, the z-direction). Thus, any changes to the energies of the Δ conduction band minima will leave the Δ minima along [001] and the [00$\bar{1}$] directions degenerate (*i.e.*, E_c^2), as well as the Δ minima along [010], [0$\bar{1}$0], [100] and the [$\bar{1}$00] directions degenerate (*i.e.*, E_c^4).

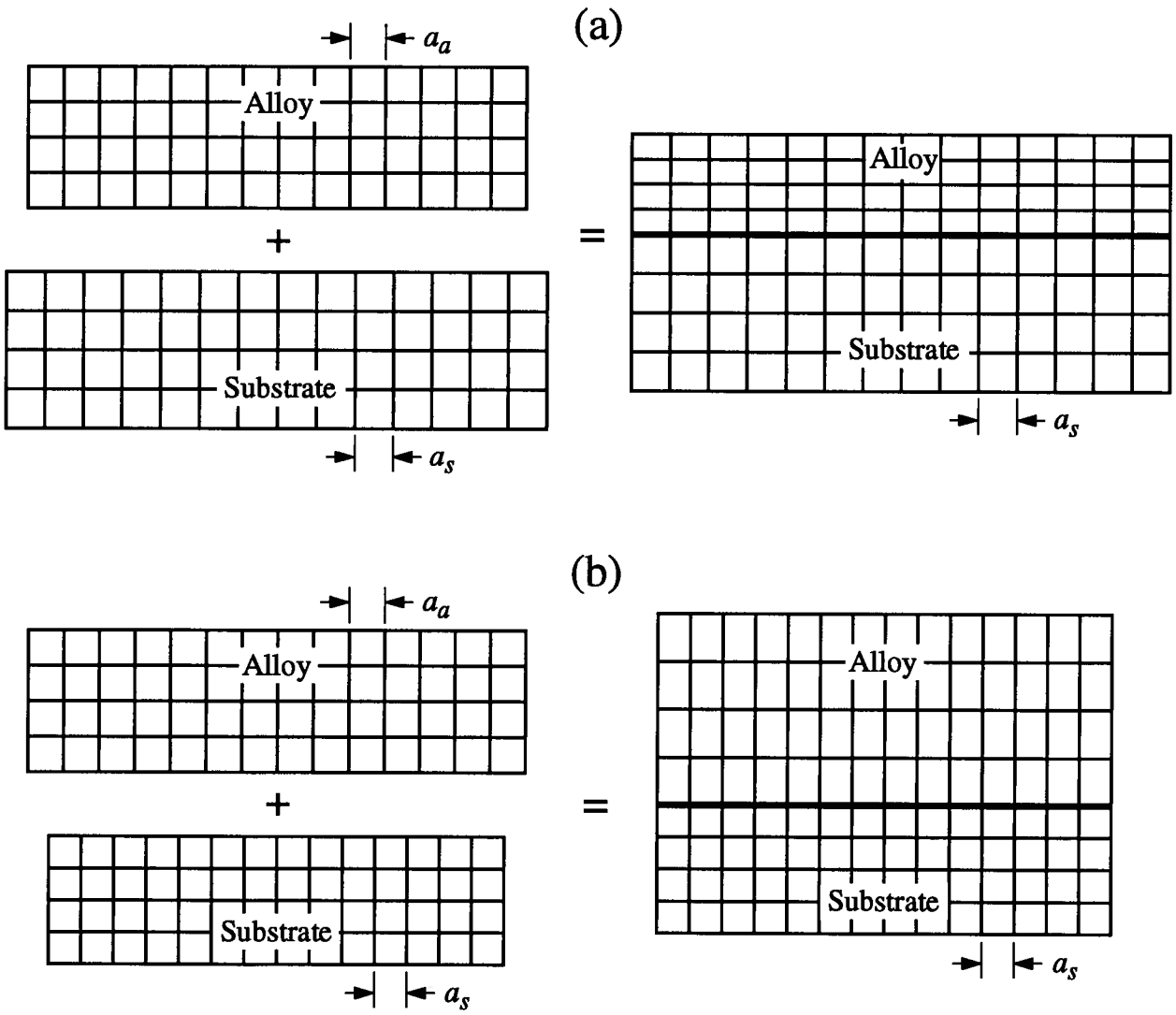


Fig. 6.3. Commensurate growth of the $Si_{1-x_a}Ge_{x_a}$ alloy layer to the $Si_{1-x_s}Ge_{x_s}$ substrate, leading to a pseudomorphic alloy film. (a) the substrate lattice constant a_s is larger than the alloy lattice constant a_a. The resultant biaxial tension, which results from a_a expanding to fit a_s, distorts the out of plane alloy lattice constant by compressing it. (b) the substrate lattice constant a_s is smaller than the alloy lattice constant a_a. The resultant biaxial compression, which results from a_a compressing to fit a_s, distorts the out of plane alloy lattice constant by expanding it.

The final diagonal components e_{xx}, e_{yy}, and e_{zz} of the strain tensor, after the layer becomes pseudomorphic, are given by the relative difference between the final pseudomorphic lattice constants and the initial bulk values [106,107,112,114]. Given that we are dealing with systems that are lattice matched to {100} substrates, and that the direction of growth is [001], then the strain tensor is:

$$\bar{\mathbf{e}} = \begin{bmatrix} e_{xx} & & \\ & e_{yy} & \\ & & e_{zz} \end{bmatrix} = \begin{bmatrix} \frac{a_s - a_a}{a_a} & & \\ & \frac{a_s - a_a}{a_a} & \\ & & e_{xx} + \left(\frac{1+\nu}{1-\nu}\right)\frac{a_a - a_s}{a_s} \end{bmatrix} \tag{6.5}$$

where ν is the Poisson ratio (which is equal to 0.273 for Ge and 0.280 for Si [107], so on average is 0.277 for $Si_{1-x}Ge_x$). The lattice constants a_a and a_s are obtained by a linear interpolation between the bulk lattice constants for Si and Ge giving:

$$\begin{aligned} a_a &= 5.43 + 0.23x_a \text{Å} \\ a_s &= 5.43 + 0.23x_s \text{Å} \end{aligned} \tag{6.6}$$

where x_a is the Ge content in the alloy layer, and x_s is the Ge content in the substrate layer.

In order to determine how E_c^4 and E_c^2 respond to strain it is instructive to define an average conduction band energy $\overline{E}_c$. The reason for defining $\overline{E}_c$ is that depending on the direction of strain, either E_c^4 or E_c^2 will form the ultimate conduction band E_c; so using E_c as the reference would become mathematically cumbersome. $\overline{E}_c$ is given by the weighted average of E_c^4 and E_c^2; *i.e.*,

$$\overline{E}_c = \frac{4E_c^4 + 2E_c^2}{6} = \frac{2E_c^4 + E_c^2}{3}. \tag{6.7}$$

Using deformation-potential theory, the change to $\overline{E}_c$ (*i.e.*, $\Delta\overline{E}_c$) due to strain is [106,107,112,114]:

$$\Delta\overline{E}_c = (\Xi_d + \frac{1}{3}\Xi_u)\,\bar{\mathbf{1}}:\bar{\mathbf{e}} = (\Xi_d + \frac{1}{3}\Xi_u)\,(e_{xx} + e_{yy} + e_{zz}), \tag{6.8}$$

where Ξ_d and Ξ_u are the dilation and uniaxial deformation potentials respectively. Further, the change in energy for a specific Δ conduction band minima is given by:

$$\Delta E_c^{[i]} = [\Xi_d \bar{\mathbf{1}} + \Xi_u \{\hat{a}_i \hat{a}_i\}] : \bar{\mathbf{e}} \tag{6.9}$$

where $\hat{a}_i$ is the unit vector parallel to the i'th Δ conduction band minima, and $\{\}$ denotes dyadic product. For example, the change in the energy of the Δ conduction band minima along [100] is given by:

$$\Delta E_c^{[100]} = \left(\begin{bmatrix} \Xi_d & & \\ & \Xi_d & \\ & & \Xi_d \end{bmatrix} + \begin{bmatrix} \Xi_u & & \\ & 0 & \\ & & 0 \end{bmatrix} \right) : \bar{\mathbf{e}} = \Xi_d (e_{yy} + e_{zz}) + (\Xi_d + \Xi_u) e_{xx}.$$

Finally, using eqns (6.9), (6.8) and (6.5) where $e_{xx} = e_{yy}$, then the energy difference between E_c^4 and $\overline{E}_c$, as well as E_c^2 and $\overline{E}_c$ are given by:

$$\begin{aligned} E_c^4 = E_c^{[\pm 100]} = E_c^{[0 \pm 10]} = \Delta E_c^{[100]} - \Delta \overline{E}_c &= \Xi_u \left(\{ \hat{a}_i \hat{a}_i \} - \frac{1}{3} \bar{1} \right) : \bar{\mathbf{e}} \\ &= \Xi_u \left(\begin{bmatrix} 1 & & \\ & 0 & \\ & & 0 \end{bmatrix} - \begin{bmatrix} 1/3 & & \\ & 1/3 & \\ & & 1/3 \end{bmatrix} \right) : \bar{\mathbf{e}} = \Xi_u \left(\frac{2}{3} e_{xx} - \frac{1}{3} (e_{yy} + e_{zz}) \right) \\ &= -\frac{1}{3} \Xi_u e_\perp \end{aligned} \tag{6.10}$$

and

$$\begin{aligned} E_c^2 = E_c^{[00 \pm 1]} = \Delta E_c^{[001]} - \Delta \overline{E}_c &= \Xi_u \left(\{ \hat{a}_i \hat{a}_i \} - \frac{1}{3} \bar{1} \right) : \bar{\mathbf{e}} \\ &= \Xi_u \left(\begin{bmatrix} 0 & & \\ & 0 & \\ & & 1 \end{bmatrix} - \begin{bmatrix} 1/3 & & \\ & 1/3 & \\ & & 1/3 \end{bmatrix} \right) : \bar{\mathbf{e}} = \Xi_u \left(-\frac{1}{3} (e_{xx} + e_{yy}) + \frac{2}{3} e_{zz} \right) \\ &= \frac{2}{3} \Xi_u e_\perp, \end{aligned} \tag{6.11}$$

where $e_\perp \equiv e_{zz} - e_{xx}$, and $\Xi_u(x_a) = 9.16 + 0.26 x_a$eV [106]. Eqns (6.10) and (6.11) give the change due to strain in the energy of the band minima for E_c^4 and E_c^2 respectively, relative to $\overline{E}_c$. Thus, E_c^4 and E_c^2 are used both as a label and as a material parameter.

Observation of eqns (6.10) and (6.11) confirm the general statements given earlier in the section regarding the changes to the conduction band due to strain. For compressive strain in the alloy layer, $x_a > x_s$ so that $a_a > a_s$ and eqn (6.5) has it that $e_\perp \equiv e_{zz} - e_{xx} > 0$. Since $\Xi_u > 0$, then eqns (6.10) and (6.11) have it that $E_c^4 < 0$ and $E_c^2 > 0$, which confirms that under compression E_c^4 forms the ultimate conduction band. Contrarily, for tensile strain in the alloy layer, $x_a < x_s$ so that $a_a < a_s$ and eqn (6.5) has it that $e_\perp < 0$. Then eqns (6.10) and (6.11) have it that $E_c^4 > 0$ and $E_c^2 < 0$, which confirms that under tension E_c^2 forms the ultimate conduction band.

With the changes to the conduction band due to strain determined, the valence band is now solved for. The designations for the hh, lh and so valence bands are based upon the valence band

strain Hamiltonian [118]. To this end, it has been determined that the quantum numbers for total angular momentum J as well as magnetic moment (spin) m_j remain unchanged with the application of strain. This leads to the hh band designation of $|J=\frac{3}{2};\ m_j=\pm\frac{3}{2}\rangle$ or $|\frac{3}{2};\ \pm\frac{3}{2}\rangle$ for short; the lh band designation of $|\frac{3}{2};\ \pm\frac{1}{2}\rangle$; and the so band designation of $|\frac{1}{2};\ \pm\frac{1}{2}\rangle$. The solution of the valence band strain Hamiltonian [118,107] produces:

$$
\begin{aligned}
E_v^{|\frac{3}{2};\pm\frac{3}{2}\rangle} &= E_v^{\mathrm{hh}} = \frac{2}{3}D_u(x_a)e_\perp = \frac{2}{3}D_u(x_a)\,(e_{zz}-e_{xx}) \\
E_v^{|\frac{3}{2};\pm\frac{1}{2}\rangle} &= E_v^{\mathrm{lh}} = -\frac{1}{2}\,(E_v^{\mathrm{hh}}+\Lambda(x_a)) + \frac{1}{2}\sqrt{9\,(E_v^{\mathrm{hh}})^2+\Lambda^2(x_a)-2E_v^{\mathrm{hh}}\Lambda(x_a)} \\
E_v^{|\frac{1}{2};\pm\frac{1}{2}\rangle} &= E_v^{\mathrm{so}} = -\frac{1}{2}\,(E_v^{\mathrm{hh}}+\Lambda(x_a)) - \frac{1}{2}\sqrt{9\,(E_v^{\mathrm{hh}})^2+\Lambda^2(x_a)-2E_v^{\mathrm{hh}}\Lambda(x_a)}
\end{aligned}
\tag{6.12}
$$

where $D_u(x_a)$ is the valence band deformation potential equal to 3.15 + 1.14x_aeV [106], and $\Lambda(x_a)$ is the split off energy, defined in Fig. 6.2, which is equal to 0.044 + 0.246x_aeV [107].

Using a similar technique to the one used for the solution of the conduction band, an average valence band energy $\overline{E}_v$ is defined and subsequently used as the reference point for all valence band energies; *i.e.*,

$$
\overline{E}_v = \frac{E_v^{\mathrm{hh}}+E_v^{\mathrm{lh}}+E_v^{\mathrm{so}}}{3} = -\frac{1}{3}\Lambda(x_a). \tag{6.13}
$$

It is interesting to note that $\overline{E}_v$ defined in eqn (6.13) is independent of the applied strain. Also, because $\overline{E}_v$ is not zero, the valence band energies in eqn (6.12) are not using $\overline{E}_v$ as their energy reference (substituting eqns (6.11) and (6.10) into (6.7) gives $\overline{E}_c = 0$, which shows that $\overline{E}_c$ is indeed the energy reference for the conduction band). Observation of eqn (6.12) shows that under the condition of zero strain (*i.e.*, $e_\perp = 0$), then $E_v^{\mathrm{hh}} = E_v^{\mathrm{lh}} = 0$, and $E_v^{\mathrm{so}} = -\Lambda(x_a)$. Therefore, the energy reference for eqn (6.12) is not $\overline{E}_v$ but the valence band edge of bulk $Si_{1-x}Ge_x$. The reason for using $\overline{E}_v$ will become obvious when the band offsets at a heterojunction are determined in Section 6.2.

Eqns (6.10)-(6.12) determine the effect of uniaxial strain, due to the second term on the right-hand-side of eqn (6.4), on the conduction and valence bands of $Si_{1-x}Ge_x$. Eqn (6.3) determines the effect of the Ge alloy content. Finally, the effect of the hydrostatic force, due to the first term on the right-hand-side of eqn (6.4), is determined. The hydrostatic force results in either a net decrease or increase in the total volume of the crystal's unit cell. A volume change in the unit cell will be accompanied by a change in the absolute energy of the conduction and valence bands.

This net change in absolute energy is best determined by calculating the differential change to $\overline{E}_c$ and $\overline{E}_v$ (*i.e.*, $\Delta\overline{E}_c$ and $\Delta\overline{E}_v$). Eqn (6.8) solves for $\Delta\overline{E}_c$, and in a similar fashion $\Delta\overline{E}_v = a\bar{1}:\bar{\mathbf{e}}$, where a is another deformation potential that is characteristic of the material [105,107]. Put together, the hydrostatic change to the bulk $Si_{1-x}Ge_x$ bandgap is:

$$\Delta\overline{E}_g = \Delta\overline{E}_c - \Delta\overline{E}_v = (\Xi_d + \frac{1}{3}\Xi_u - a)\,\bar{1}:\bar{\mathbf{e}} = (\Xi_d + \frac{1}{3}\Xi_u - a)\,(e_{xx} + e_{yy} + e_{zz}) \qquad (6.14)$$

where $\Xi_d + \frac{1}{3}\Xi_u - a = 1.5 - 0.19x_a$eV [106].

Eqns (6.1)-(6.3), and (6.10)-(6.14) together determine the effect of Ge alloy content and strain on the conduction and valence bands. Specifically, the bandgap of a $Si_{1-x_a}Ge_{x_a}$ alloy layer commensurately strained to a $Si_{1-x_s}Ge_{x_s}$ substrate is:

$$E_g(x_a, x_s) = E_g(x_a) + \Delta\overline{E}_g + \min(E_c^4, E_c^2) - \max(E_v^{hh}, E_v^{lh}). \qquad (6.15)$$

It must be remembered that for $x_a > 0.85$ it is possible for the L conduction band minima to become the ultimate conduction band. Therefore, the use of eqn (6.15) is valid for $x_a > 0.85$ only if there is sufficient strain to ensure that the Δ minima, and not the L minima, still form the ultimate conduction band.

Fig. 6.4 plots the $Si_{1-x_a}Ge_{x_a}$ bandgap for a variety of substrate cases. The most striking feature of Fig. 6.4 is the effect of strain on the bandgap. Comparing the bulk material bandgap to any of the other strained cases shows that the Ge alloy content of the $Si_{1-x_a}Ge_{x_a}$ layer plays a far smaller role than strain does in determining the bandgap. In fact, observation of the line for a pure Si substrate shows that $Si_{0.45}Ge_{0.55}$ lattice matched to {100} Si has a bandgap of 0.66eV, which is that of bulk Ge. The strange shape concerning the lines for material strained to substrates of $Si_{0.75}Ge_{0.25}$, $Si_{0.50}Ge_{0.50}$, and $Si_{0.25}Ge_{0.75}$ is due to the fact the material is shifting from a case of in-plane tension to compression. Take the example of a $Si_{0.50}Ge_{0.50}$ substrate. When the pseudomorphic layer has a Ge mole fraction in the range of $0 \le x_a \le 0.50$, the layer is under in-plane tension as the substrate has a larger lattice constant. Thus, as x_a increases towards 0.50 the tension is decreasing and the bandgap will increase, with E_c^2 forming the ultimate conduction band. When $x_a = 0.50$ there is no strain and the bandgap will be given by the bulk value. Finally, as x_a increases past 0.50, the strain switches from in-plane tension to compression. When this change in the direction of the strain occurs, E_c^4 forms the ultimate conduction band (this is why there is a corner in the plot, however, the E_c^2 and E_c^4 bandgaps continue on in a smooth fashion but do not form the ultimate bandgap). As x_a increases past 0.50 the amount of in-plane compression

continues to increase which reduces that bandgap once again. The essential feature of strain is that it always reduces the bandgap from the bulk unstrained value.

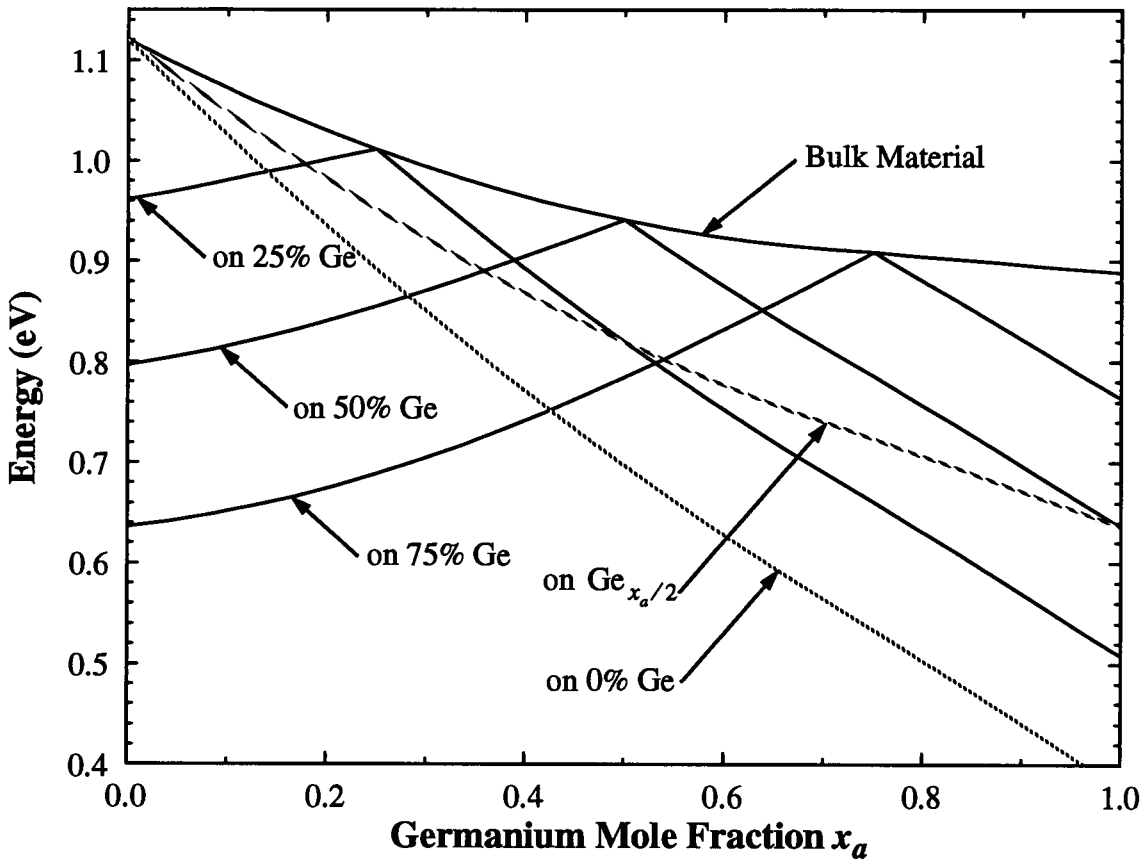


Fig. 6.4. $Si_{1-x_a}Ge_{x_a}$ bandgap when grown commensurately to a variety of substrates oriented along ⟨100⟩. All values reflect the energy from the top of the valence band to the lowest Δ minima in the conduction band. The bulk material bandgap is for reference and is valid only for $x_a < 0.85$; for $x_a > 0.85$ the bulk material line is not the ultimate bandgap but the bandgap to the Δ minima.

Eqns (6.10)-(6.11) give the conduction band energies of E_c^4 and E_c^2 relative to $\overline{E}_c$. Examination of eqns (6.10)-(6.11) shows that under zero strain, when $e_\perp = 0$, $E_c^4 = E_c^2 = 0$. Thus, $\overline{E}_c$ is the position of the ultimate conduction band in the absence of strain. If the position of the unstrained conduction band is known, eqns (6.10)-(6.11) will yield the offset to the conduction band due to any strain in the layer. Fig. 6.5 plots E_c^4 and E_c^2 relative to $\overline{E}_c$ using similar substrates as found in Fig. 6.4. Observation of Fig. 6.5 shows the changes in E_c^4 and E_c^2 to be quite linear in terms of strain. Furthermore, whenever the pseudomorphic layer is under compression then E_c^4 forms the conduction band, but when the layer is in tension then E_c^2 forms the conduction band. Finally, E_c^2 changes more rapidly than does E_c^4 for a given increase in the amount of strain.

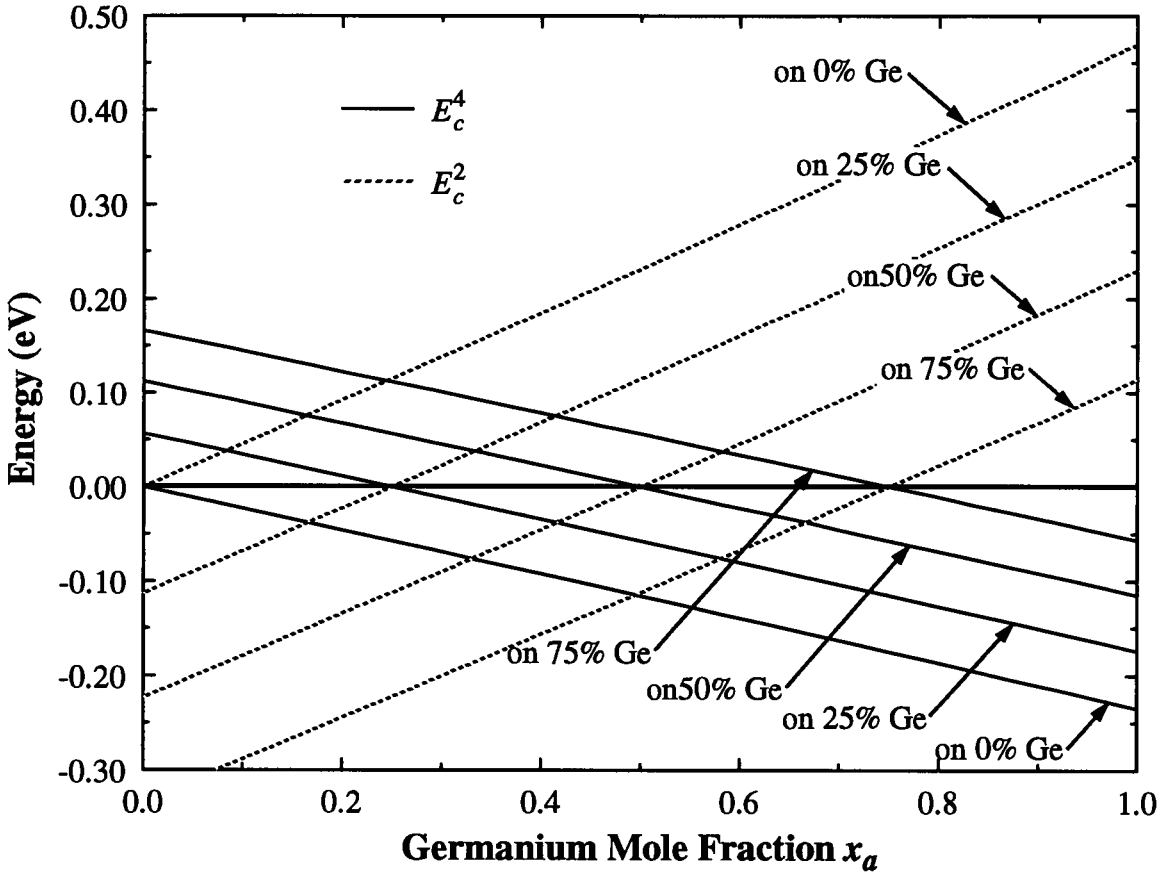


Fig. 6.5. E_c^4 and E_c^2 conduction band energies relative to the unstrained conduction band edge $\overline{E}_c$ for $Si_{1-x_a}Ge_{x_a}$ commensurately grown to a variety of substrates oriented along ⟨100⟩. The ultimate conduction band edge will be formed by the band with the lowest energy.

As was stated earlier, eqn (6.12) gives the energy offset of the hh, lh, and so bands relative to the unstrained valence band edge. Fig. 6.6 plots the hh and lh bands relative to the unstrained valence band using similar substrates as found in Figs. 6.4 and 6.5. The so band is not plotted because strain simply continues to lower the band peak even further, meaning that the so band will not be of any consequence regarding the transport of holes. Comparison of Fig. 6.6 with Fig. 6.5 shows that unlike the conduction bands, the valence bands respond in a non-linear fashion with respect to an applied strain. Furthermore, there is not as large a change in the energy of the valence bands due to strain as there is in the conduction bands. Finally, whenever the $Si_{1-x_a}Ge_{x_a}$ layer is under compression, then the hh band will form the ultimate conduction band; while under tension, the lh band will form the ultimate conduction band.

Finally, it is instructive to present a surface plot of constant energy in $\boldsymbol{k}$-space, depicting the conduction bands in $Si_{1-x_a}Ge_{x_a}$ under the influence of strain. Fig. 6.7 plots the surface of constant

energy that envelopes the six Δ minima in $Si_{0.83}Ge_{0.17}$ commensurately strained to (001) Si. Since the pseudomorphic $Si_{0.83}Ge_{0.17}$ layer is under an in-plane compressive strain, then Fig. 6.5 shows that E_c^4 will form the ultimate conduction band. The constant energy surface used in Fig. 6.7 is set at 209meV above the minimum in the E_c^4 band. The energy separation between E_c^4 and E_c^2 for the case considered is 116meV. As a result of the choice for the energy surface, the ellipses that represent E_c^2 are reduced by 33% compared to the ellipses that represent E_c^4. If a more realistic surface energy of $2kT$ (= 52meV at room temperature) were used instead of 200meV, then the E_c^2 band would not be seen at all. This demonstrates the profound effect that strain imparts to the $Si_{1-x}Ge_x$ layer.

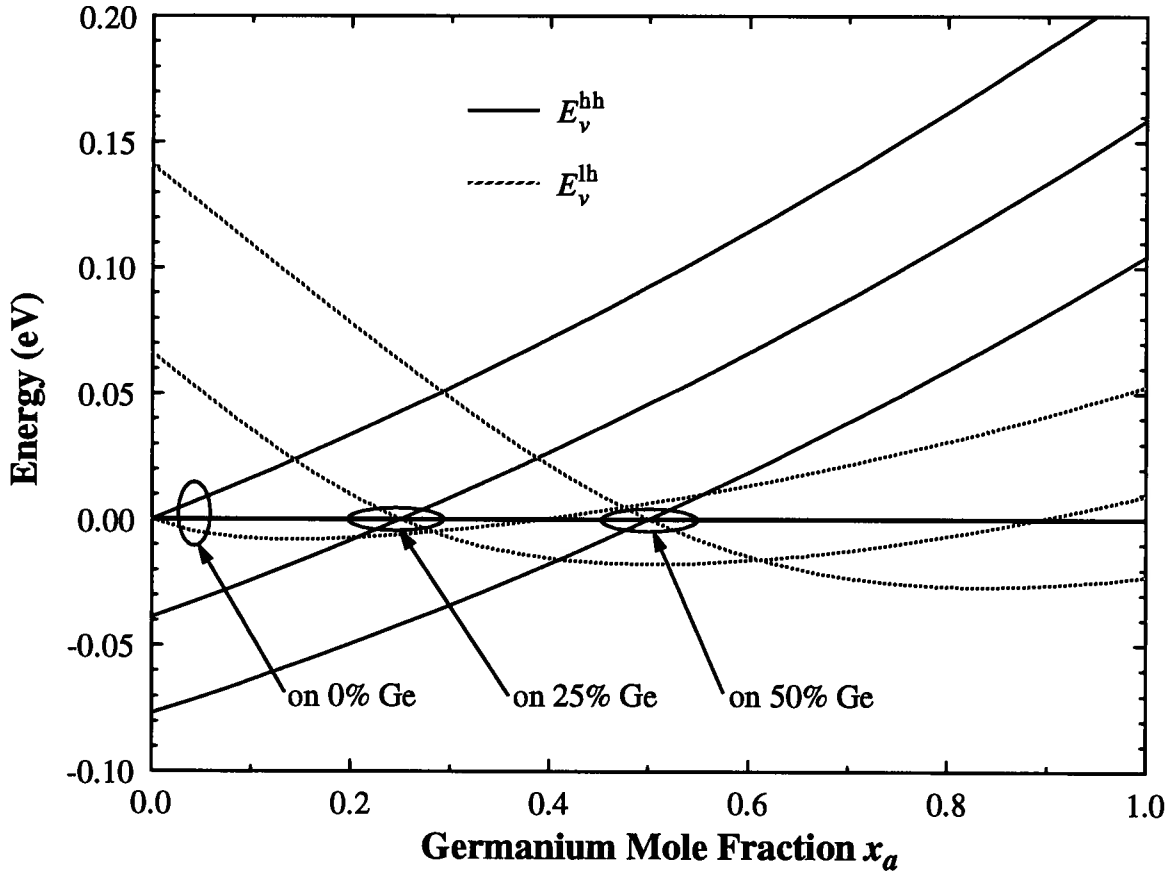


Fig. 6.6. E_v^{hh} and E_v^{lh} valence band energies relative to the unstrained valence band edge for $Si_{1-x_a}Ge_{x_a}$ commensurately grown to a variety of substrates oriented along ⟨100⟩. The ultimate valence band edge will be formed by the band with the highest energy.

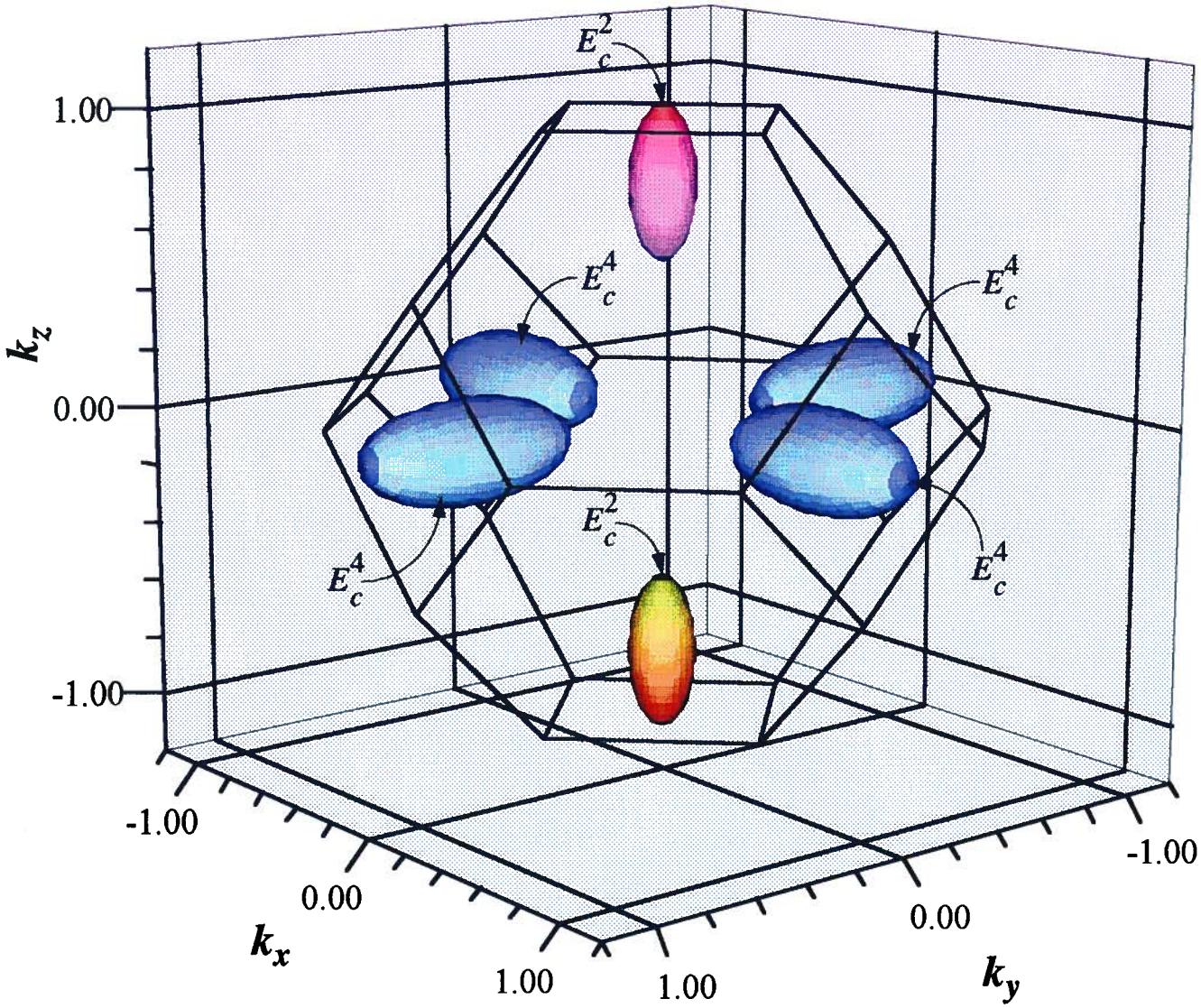


Fig. 6.7. Constant energy surface plot depicting the E_c^4 and E_c^2 bands in $Si_{0.83}Ge_{0.17}$ commensurately strained to (001) Si. The $\boldsymbol{k}$-wave vectors are normalised to one-half the length of the reciprocal lattice vector. The constant energy surface is set at 209meV above the minimum in the E_c^4 band. The E_c^2 band lies 116meV above the E_c^4 band. The E_c^4 ellipses have a longitudinal extent of 0.8, while the E_c^2 ellipses have a longitudinal extent that is 33% less than the E_c^4 band, or 0.53.

This section essentially presents a concise review of the relevant theories regarding the movement of the conduction and valence bands in $Si_{1-x}Ge_x$ under the influence of strain. Furthermore, the most recent material parameters regarding deformation-potentials have been included. However, there is still considerable change occurring to the relevant material parameters of $Si_{1-x}Ge_x$ at this time. As $Si_{1-x}Ge_x$ becomes a more important material in mainstream commercial ICs, the need to ultimately obtain the relevant material parameters will force the solid-state community to finalise on the parameters. This process will most likely follow the course that

$Al_xGa_{1-x}As$ took, in which a decade passed before the solid-state community settled on a firm set of material parameters. In any event, this section has clearly shown the profound effect that strain has on $Si_{1-x}Ge_x$; so much so that strain produces more of an effect on the bandgap than does the Ge alloy content.

6.2 Band Offsets in $Si_{1-x}Ge_x$

Section 6.1 presented all of the relevant material parameters to describe the conduction and valence bands of a $Si_{1-x_a}Ge_{x_a}$ alloy layer commensurately strained on top of a {100} $Si_{1-x_s}Ge_{x_s}$ substrate. This section will present the band offset models that predict the valence band and conduction band discontinuity at an abrupt heterojunction. Therefore, when the results of this section are combined with the results of Section 6.1, all of the relevant models for $Si_{1-x}Ge_x$ regarding the position of the conduction and valence bands within a device can be determined.

The seminal theoretical work on the band alignments between $Si_{1-x_{al}}Ge_{x_{al}}$ and $Si_{1-x_{ar}}Ge_{x_{ar}}$ (where the *l,r* subscripts refer to the left and right films respectively), when commensurately strained on top of a {100} $Si_{1-x_s}Ge_{x_s}$ substrate, was done by Van de Walle and Martin [106,119,120]. They analysed a SiGe system in one dimension using a quantum mechanical model. To remove the issue of boundary conditions that would destroy the crystalline periodicity required to establish Bloch functions, they developed a supercell structure. The supercell structure had a unit lattice cell that was constructed of n Si atoms followed by n Ge atoms. By extending this unit supercell to infinity, though the Born-von Karman boundary conditions, Van de Walle and Martin were able to obtain the band offsets. In order to establish that the size of the supercell was large enough to ensure bulk material properties away from the heterojunction, the band offsets were determined for a variety of n. Van de Walle and Martin established that for $n > 5$ the material was bulk-like away from the heterojunction. In fact, the shape of the Bloch electron's wave function became bulk-like after moving only one lattice constant away from the heterojunction. Therefore, Van de Walle and Martin concluded that the perturbing effect of the abrupt heterojunction was indeed localised to the space immediately surrounding the interface.

The main conclusion from the work of Van de Walle and Martin is that the average valence band offset $\Delta\overline{E}_v$ between a pseudomorphic Si to Ge heterojunction, whether commensurately strained to either a {100} Si or Ge substrate, is a constant of 0.54±0.04eV (where the Si $\overline{E}_v$ is lower in energy that the Ge $\overline{E}_v$). Numerous other individuals [121-124] have gone on to perform ex-

perimental measurements of $\Delta\overline{E}_v$ with variations that are always lower than 0.54eV, and which are as low as 0.2eV. Recently, experimental measurements by Yu [125] have given $\Delta\overline{E}_v$ = 0.49±0.13eV. However, after performing an array of measurements on a variety of substrates (thereby changing the strain), it was found that $\Delta\overline{E}_v$ varied slightly with strain. The final results from Yu [125] were:

$$\Delta\overline{E}_v(x_{al}|\,x_{ar}) = (0.55 - 0.12x_s)\,(x_{al} - x_{ar})\,, \tag{6.16}$$

where x_{al}, x_{ar} and x_s refer to the Ge mole fraction in the left, right and substrate crystals respectively. Finally, Fig. 6.8 defines all of the energies and the offsets.

At issue with eqn (6.16) is the considerable appeal to linear interpolation between material parameters for bulk Si and bulk Ge. To complicate things further, the material parameters that govern the conduction band and valence band movements due to strain have considerable variability depending on which experimental method is used to obtain the results. At the moment there is no clear set of material parameters to use in order to determine the band offsets and movements within SiGe. The complexity of the SiGe system is quite high, however, it is essential that the material science community finalise on a set of material parameters and models so that SiGe HBTs may be accurately simulated.

Use of eqn (6.16) produces conduction band offsets ΔE_c that are far too large. Experimental measurements of ΔE_c [105,111,126,127] show that there should be no more than ≈±30meV of offset between $Si_{1-x_{al}}Ge_{x_{al}}$ and $Si_{1-x_{ar}}Ge_{x_{ar}}$ grown on a pure Si {100} substrate, where x_{al} and x_{ar} can take on any value in the range of 0 to 1. Furthermore, recent measurements by Gan et. al. [128] have shown that ΔE_v should equal $0.64x_{al}$eV when $x_{ar} = x_s = 0$. Use of eqn (6.16) produces $\Delta E_v = 0.80x_{al}$eV. By reducing $\Delta\overline{E}_v$ from 0.55eV back down to 0.49eV in eqn (6.16) produces:

$$\Delta\overline{E}_v(x_{al}|\,x_{ar}) = (0.49 - 0.12x_s)\,(x_{al} - x_{ar})\,. \tag{6.17}$$

Use of eqn (6.17) instead of eqn (6.16) reduces ΔE_c to be no more than +48meV and -42meV (as compared to +30meV and -100meV), while also giving $\Delta E_v = 0.74x_{al}$eV. Finally, if $D_u(x_a)$ in eqn (6.12) is changed to $2.04 + 1.77x_a$eV [107] then ΔE_c remains unchanged and $\Delta E_v = 0.68x_{al}$eV. The use of eqn (6.17) instead of eqn (6.16) is within the experimental error of the measurements in [125]. Further, eqn (6.17) when combined with $D_u(x_a) = 2.04 + 1.77x_a$eV produces conduction and valence band offsets that match experimental observations closer than when the material values proposed within [125] are employed. Thus, there is no clear set of parameters as of yet for the

modelling of the SiGe material system. However, the differences between the various models presented here is within 50meV. Therefore, in terms of the studies to be presented later on in this chapter, a small discrepancy of 50meV will simply cause a slight variation in the Ge alloy content of the various layers, but will not effect the ultimate function of the HBT.

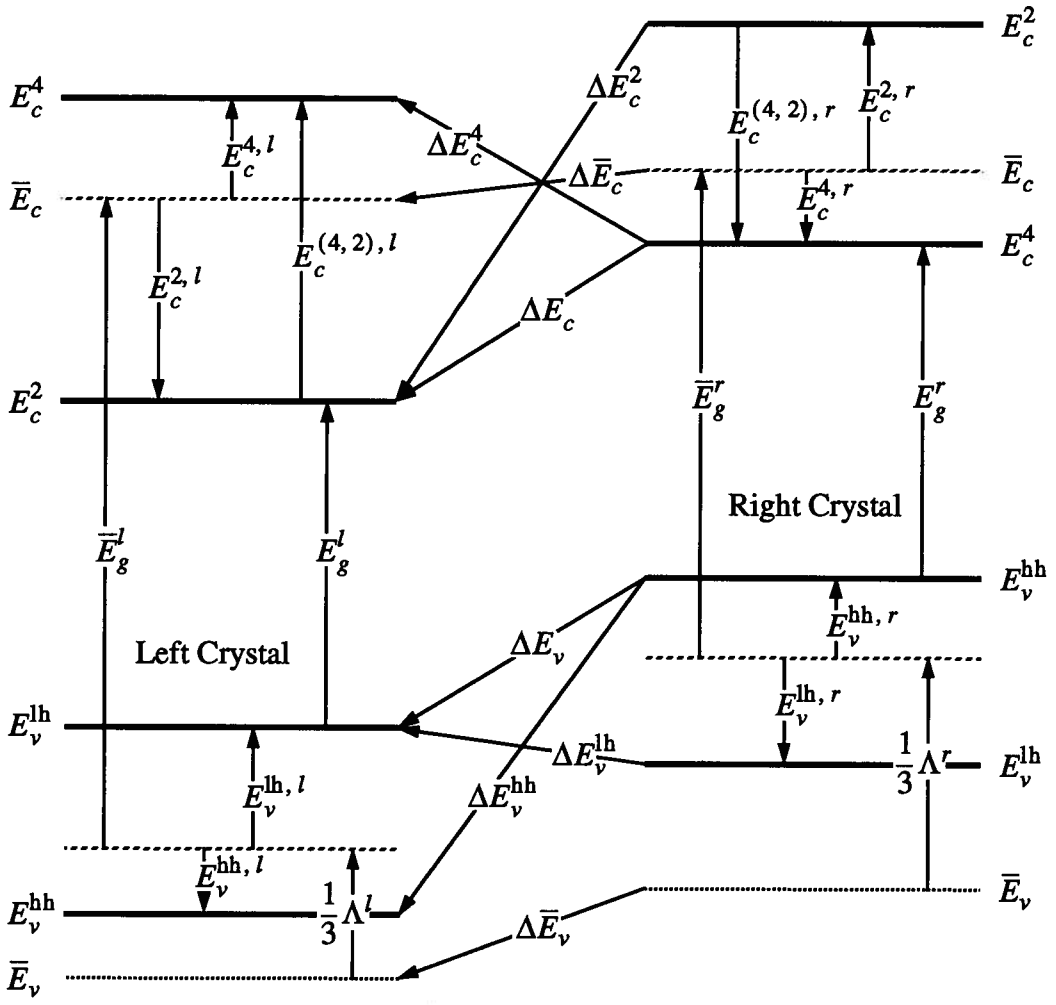


Fig. 6.8. Conduction and valence band energies including all of the band offsets for a $Si_{1-x_{al}}Ge_{x_{al}}$ to a $Si_{1-x_{ar}}Ge_{x_{ar}}$ heterojunction commensurately strained to a {100} $Si_{1-x_s}Ge_{x_s}$ substrate. The designation of *l* and *r* refer to the left and right crystal respectively, where all Δ energies are referred to the crystal on the right.

Eqn (6.17) provides the critical model that relates the band energies of two different SiGe crystals across an abrupt heterojunction. Once $\Delta\overline{E}_v$ is known, then by using Fig. 6.8, all of the other relevant offsets can be determined by appealing to the models of Section 6.1. Using eqns (6.3), (6.10)-(6.15), and (6.17) along with the aid of Fig. 6.8 yields:

$$
\begin{aligned}
\Delta E_v &= \Delta \bar{E}_v + \frac{1}{3}(\Lambda^l - \Lambda^r) + [\max(E_v^{\mathrm{lh},\,l}, E_v^{\mathrm{hh},\,l}) - \max(E_v^{\mathrm{lh},\,r}, E_v^{\mathrm{hh},\,r})] \\
\Delta E_v^{\mathrm{lh}} &= \Delta \bar{E}_v + \frac{1}{3}(\Lambda^l - \Lambda^r) + (E_v^{\mathrm{lh},\,l} - E_v^{\mathrm{lh},\,r}) \\
\Delta E_v^{\mathrm{hh}} &= \Delta \bar{E}_v + \frac{1}{3}(\Lambda^l - \Lambda^r) + (E_v^{\mathrm{hh},\,l} - E_v^{\mathrm{hh},\,r}) \\
\Delta \bar{E}_c &= \Delta \bar{E}_v + \frac{1}{3}(\Lambda^l - \Lambda^r) + (\bar{E}_g^l - \bar{E}_g^r) \\
\Delta E_c &= \Delta \bar{E}_c + [\min(E_c^{2,\,l}, E_c^{4,\,l}) - \min(E_c^{2,\,r}, E_c^{4,\,r})] \\
\Delta E_c^2 &= \Delta \bar{E}_c + (E_c^{2,\,l} - E_c^{2,\,r}) \\
\Delta E_c^4 &= \Delta \bar{E}_c + (E_c^{4,\,l} - E_c^{4,\,r})
\end{aligned}
\tag{6.18}
$$

where the average bandgap $\bar{E}_g$ is equal to the bulk alloy bandgap given in eqn (6.3), plus the hydrostatic change to the bandgap $\Delta \bar{E}_g$ given in eqn (6.14). Therefore, eqn (6.18) provides all of the necessary information to calculate any of the band offsets within the SiGe material system.

Although no one equation that forms the model of the SiGe material system is of a complex nature, the cumulative effect of each sub-model leads to a complex system as is evident from eqn (6.18). However, it is possible to arrive at a set of Taylor expansions for the models that govern the band movements within the conduction band. Unfortunately, the valence band models (*i.e.*, eqn (6.12)) contain a square root dependence that proves impossible to approximate. Given the non-linear nature of the strain tensor, is it is not possible to achieve a simple linear approximation for the conduction bands. By performing a multivariate Taylor expansion of the conduction band models in eqn (6.18), up to and including second order terms, yields:

$$
\begin{aligned}
\Delta \bar{E}_c &\approx 0.1429\,(x_{ar} - x_{al})\,x_s - 0.32789x_{ar}^2 + 0.02155x_{ar} + 0.32985x_{al}^2 - 0.02252x_{al} \\
\Delta E_c^2 &\approx 0.1751\,(x_{ar} - x_{al})\,x_s - 0.34084x_{ar}^2 - 0.43481x_{ar} + 0.34281x_{al}^2 + 0.43384x_{al} \\
\Delta E_c^4 &\approx 0.1268\,(x_{ar} - x_{al})\,x_s - 0.32141x_{ar}^2 + 0.24973x_{ar} + 0.32338x_{al}^2 - 0.25070x_{al} \\
E_c^{4,\,2} &\approx -0.02723x_s^2 + 0.04836x_a x_s - 0.01943x_a^2 + 0.68368x_s - 0.68454x_a
\end{aligned}
\tag{6.19}
$$

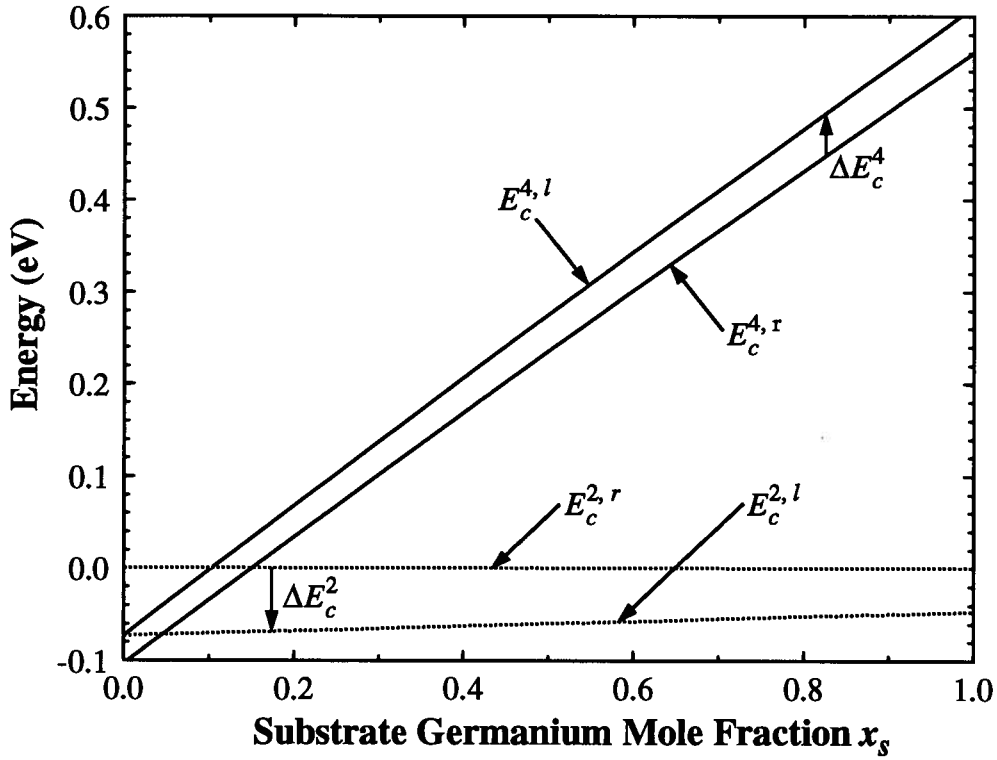
where all results are in eV, and $E_c^{4,\,2} = E_c^4 - E_c^2$. Eqn (6.19) is accurate to within 1% of the full model given in eqn (6.18) over the entire allowed range for x_{al}, x_{ar}, and x_s. The multivariate Taylor expansions were centered around $x_{al} = 0$, $x_{ar} = 0.5$, and $x_s = 0.5$. Thus, eqn (6.19) should strictly be used with $x_{al} < x_{ar}$; however, if this is not true, then simply interchange x_{al} and x_{ar} and

multiply the result by -1. If the interchange of variables is not performed for $x_{al} > x_{ar}$, then the error in eqn (6.19) will rise to 1.5%.

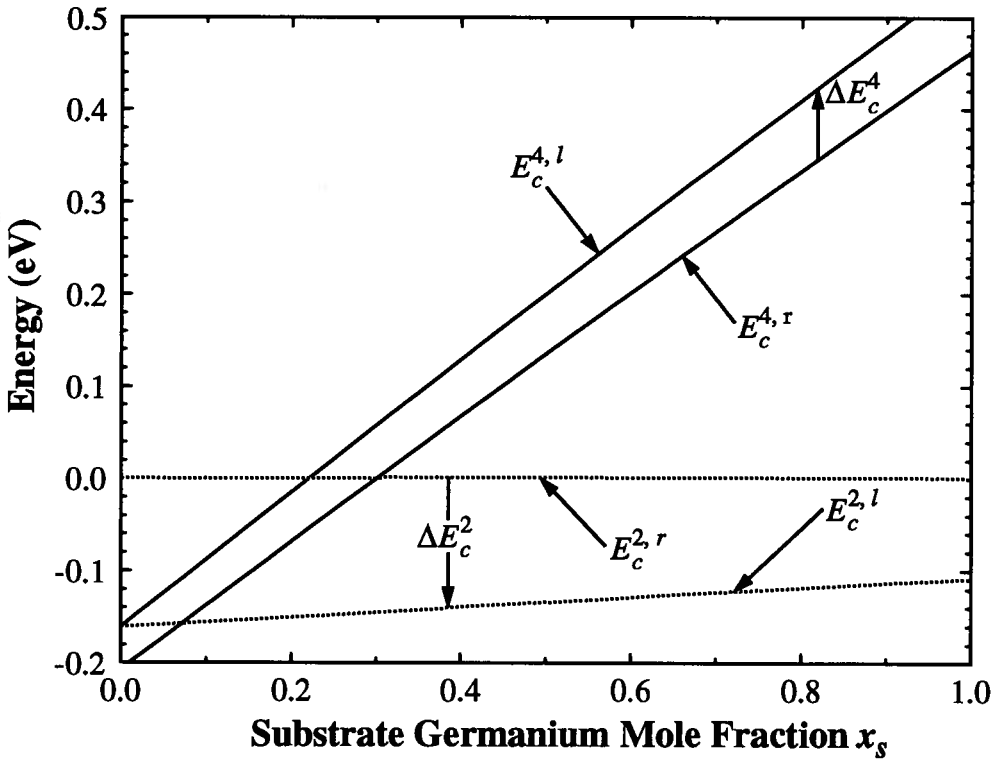
Examination of eqn (6.19) provides insight into the conduction bands of SiGe. Considering $E_c^{4,2}$ first, the two last linear terms in x_a and x_s are the dominant terms. Therefore, to a crude approximation, $E_c^{4,2} \approx 0.684(x_s - x_a)$; which corrects the proposal of $E_c^{4,2} \approx -0.6x_a$ by People [105] and Pejcinovic [28] who considers only a Si substrate. Examination of the other models in eqn (6.19) shows a linear dependence upon the substrate Ge alloy content x_s. It is by no coincidence that the coefficient that governs the x_s dependence in ΔE_c^4 is 0.1268, as compared to the coefficient of 0.12 in eqn (6.17). The largest portion of the substrate dependence in eqn (6.19) is due to the model for $\Delta \overline{E}_v$. Therefore, the material science community must determine for certain the effects of substrate strain, in order than SiGe devices can be developed where substrate strain is utilised. Finally, the non-linear terms in eqn (6.19) stem mainly from the non-linear dependence that the bulk bandgap has on the Ge alloy content.

In terms of the conduction band, Fig. 6.9 plots E_c^4 and E_c^2 to the left and right of a heterojunction under the proviso that the entire system is commensurately strained to a {100} $Si_{1-x_s}Ge_{x_s}$ substrate. The first thing to note is that ΔE_c^4 is generally smaller than ΔE_c^2, and is of such a nature that in going from the left to the right there is a downwards step. The reason for not classifying this as either a type I or II heterojunction is that the bandgap is not a monotonic function of strain, as is evidenced in Fig. 6.4. Thus, classification in terms of type I or II would require detailed knowledge of the strain state, which would destroy the simplicity of the type I or II designation. However, when going from a pure Si crystal to a $Si_{1-x}Ge_x$ crystal there is always a small downwards ΔE_c^4. Contrarily, ΔE_c^2 is in general quite large, much larger than ΔE_c^4, and is of an upwards nature in going from a pure Si crystal to a $Si_{1-x}Ge_x$ crystal. Most importantly, Fig. 6.9 clearly demonstrates that the character of the conduction band can change between E_c^4 and E_c^2 when crossing a heterojunction. Fig. 6.10 goes on to show that ΔE_c indeed has a complex nature when strain is brought into the picture. There are three distinct regions in Fig. 6.10: 1) when $x_s < (x_{al}, x_{ar})$ then ΔE_c is governed by $E_c^{4,l}$ to $E_c^{4,r}$; 2) when $x_{al} < x_s < x_{ar}$ then ΔE_c is governed by $E_c^{2,l}$ to $E_c^{4,r}$; 3) when $x_s > (x_{al}, x_{ar})$ then ΔE_c is governed by $E_c^{2,l}$ to $E_c^{2,r}$.

To conclude this section ΔE_v is plotted in Fig. 6.11. The various parameters are identical to the ones in Fig. 6.10. As with ΔE_c, ΔE_v also displays the same type of complex features which are

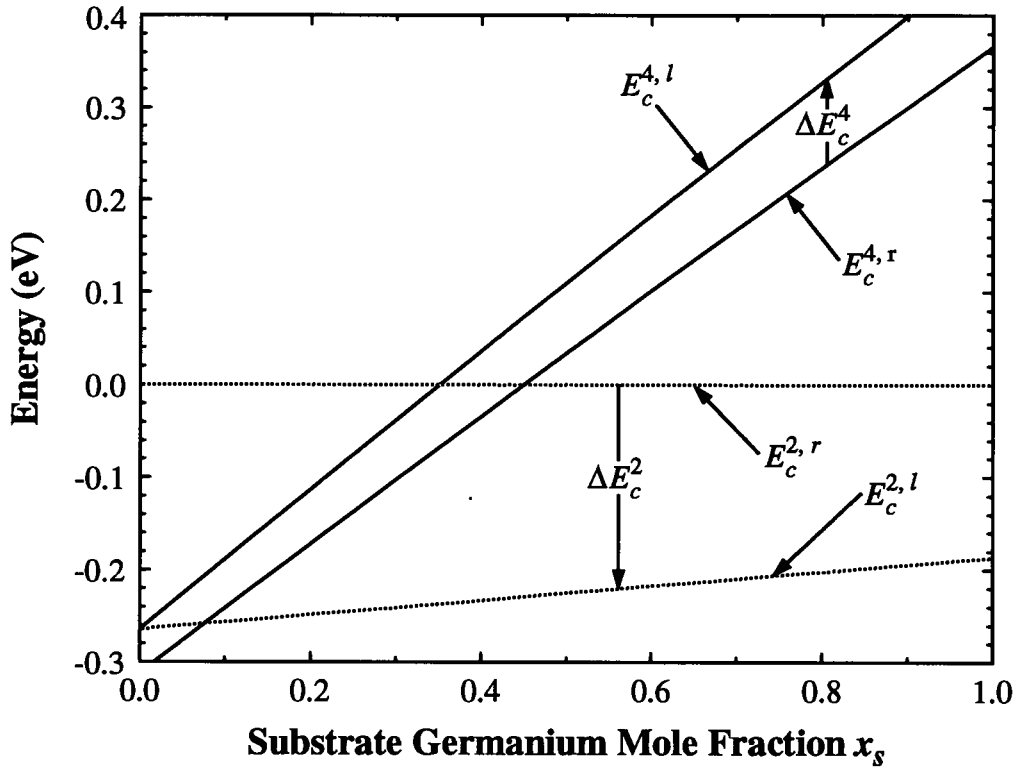


(a)

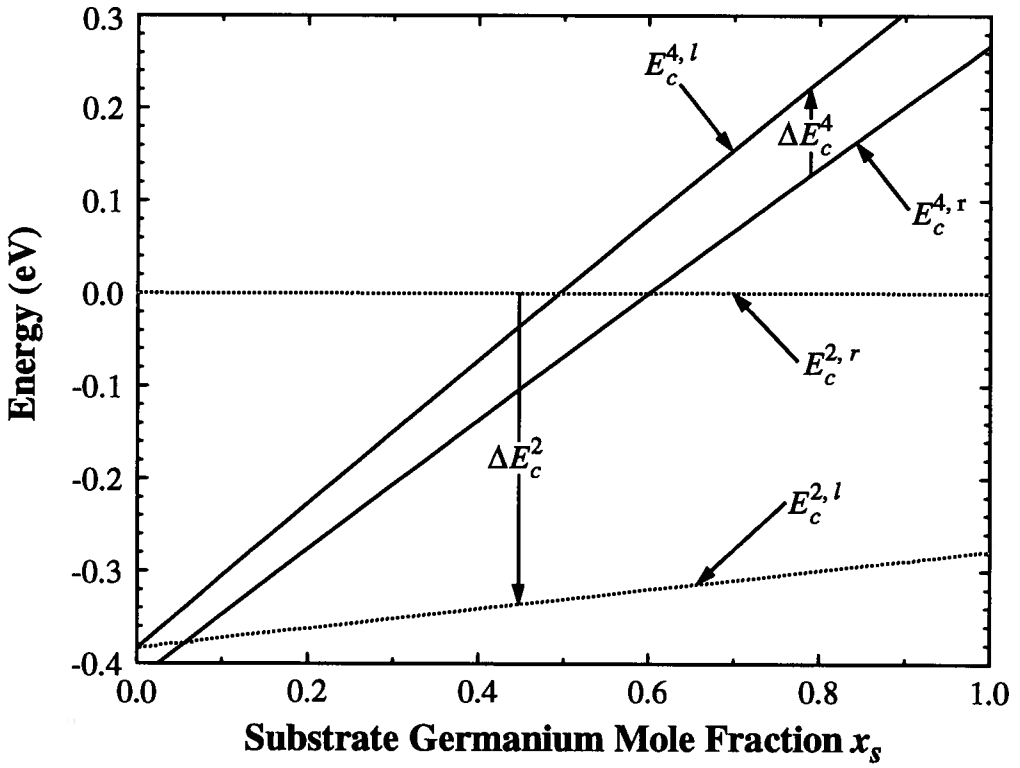


(b)

Fig. 6.9. E_c^4 and E_c^2 conduction band minima to the left and right of an abrupt heterojunction when commensurately grown atop a {100} $Si_{1-x_s}Ge_{x_s}$ substrate. All energies are relative to E_c^2 on the right hand side of the heterojunction. (a) $x_{al} = 0$, $x_{ar} = 0.15$; (b) $x_{al} = 0$, $x_{ar} = 0.30$.



(c)



(d)

Fig. 6.9. Continuation of Fig. 6.9 from the previous page. (c) $x_{al} = 0$, $x_{ar} = 0.45$; (d) $x_{al} = 0$, $x_{ar} = 0.60$.

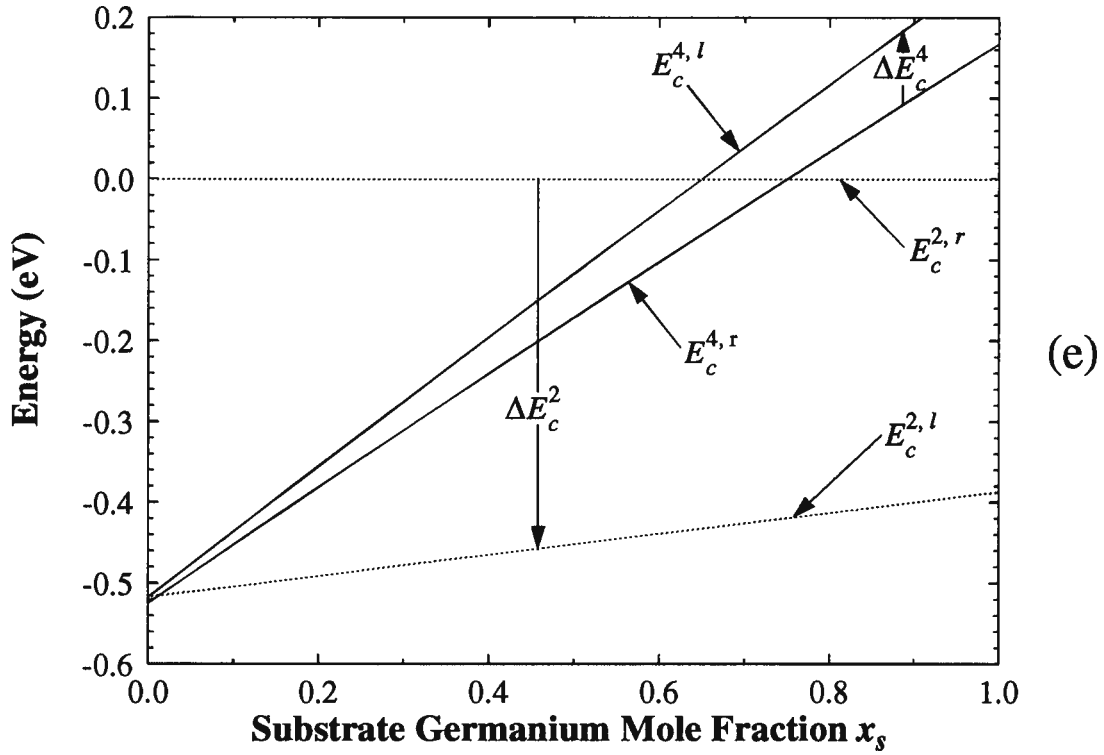


Fig. 6.9. Continuation of Fig. 6.9 from the previous page. (e) $x_{al} = 0$, $x_{ar} = 0.75$.

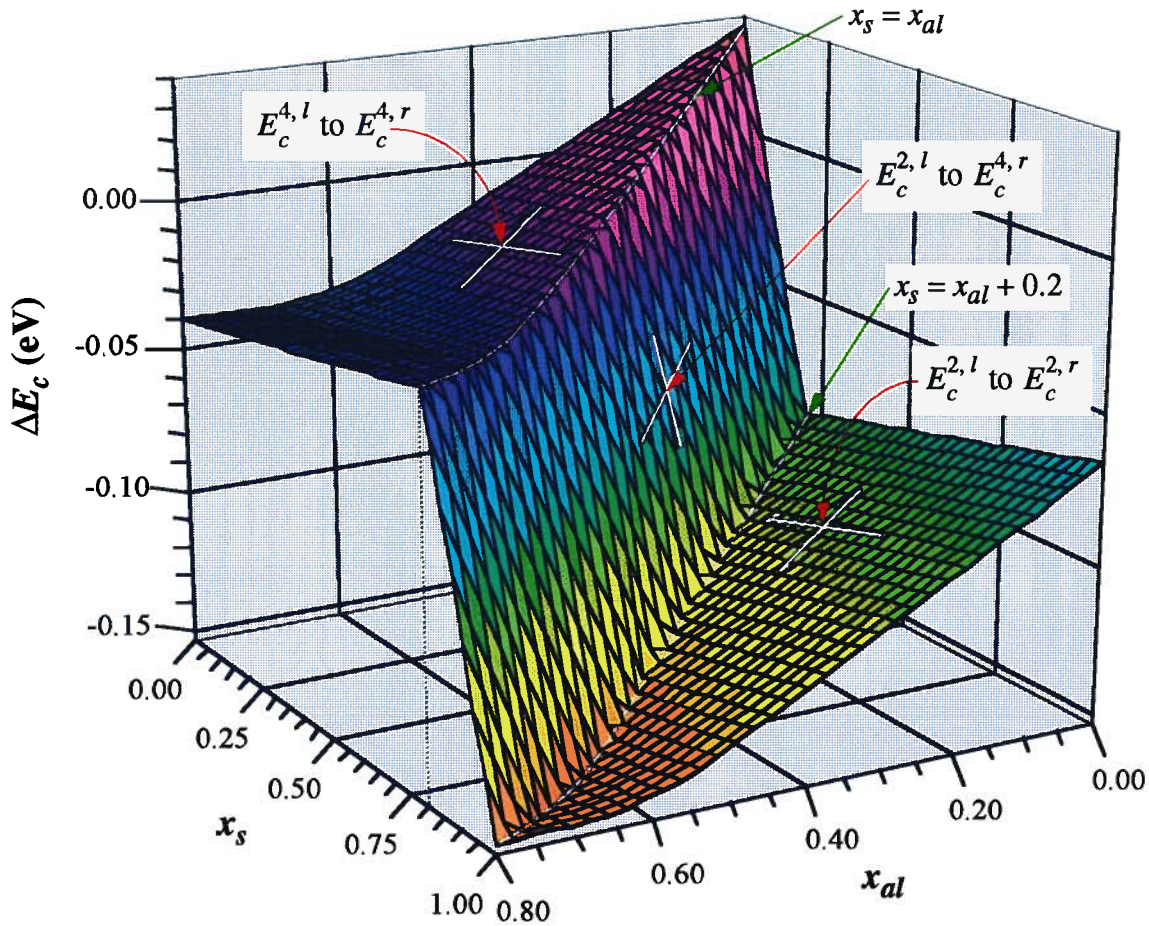


Fig. 6.10. ΔE_c when $x_{ar} = x_{al} + 0.20$, and x_{al} and x_s are varied. The right side is the reference.

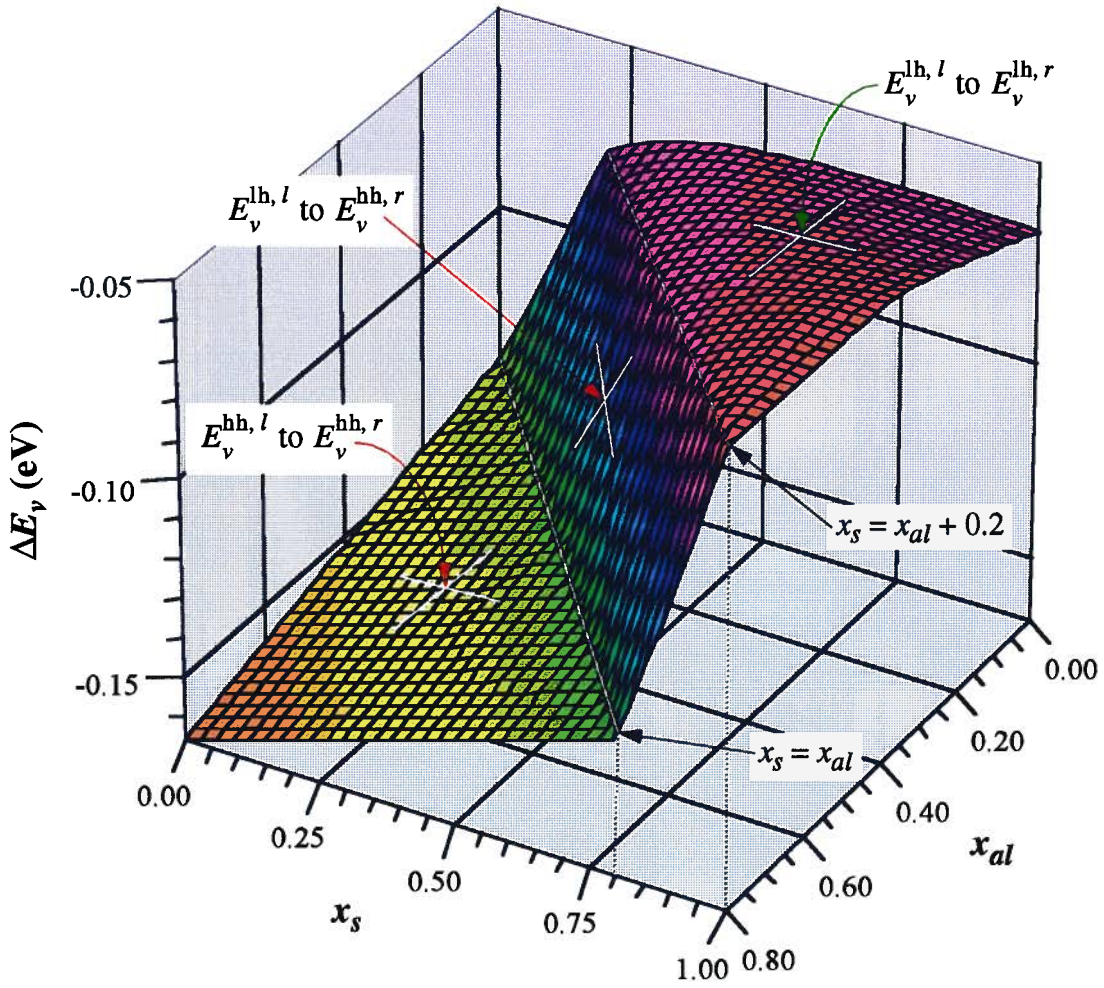


Fig. 6.11. ΔE_v when $x_{ar} = x_{al} + 0.20$, and x_{al} and x_s are varied. The right side is the reference

due to the ultimate valence band changing from E_v^{hh} to E_v^{lh}. Just like Fig. 6.10, there are three distinct regions in Fig. 6.11: 1) when $x_s < (x_{al}, x_{ar})$ then ΔE_v is governed by $E_v^{hh,\,l}$ to $E_v^{hh,\,r}$; 2) when $x_{al} < x_s < x_{ar}$ then ΔE_v is governed by $E_v^{lh,\,l}$ to $E_v^{hh,\,r}$; 3) when $x_s > (x_{al}, x_{ar})$ then ΔE_v is governed by $E_v^{lh,\,l}$ to $E_v^{lh,\,r}$.

6.3 Electron Transport in Strained $Si_{1-x}Ge_x$

Sections 6.1 and 6.2 present the necessary $Si_{1-x}Ge_x$ material models to determine the overall band diagram, including offsets at abrupt heterojunctions, within any SiGe solid-state device. This section will focus on determining the transport models for electrons and holes within the $Si_{1-x}Ge_x$ material system. Essentially, the models presented in all of the previous chapters are applicable to the study of SiGe-based devices. For example, Chapter 4 presented the EB SCR transport models

which included the effects of tunneling and the mass barrier. Therefore, Chapter 4 can be applied to a SiGe device to determine if the EB SCR will generate current-limited-flow. However, care must be exercised in the application of Chapter 4, and indeed all of the other chapters, as there is a multi-band model for the $Si_{1-x}Ge_x$ material system. This section will discuss and present the transport models for the multiband $Si_{1-x}Ge_x$ material system.

From the work in the previous two sections, it is clear that the conduction and valence bands are both broken down into two distinct sub-bands (the so valence band is ignored as it is always lower in energy than the lh and hh bands, especially under strain, and is of such a low carrier mass [96] that hole transport can be ignored). Unlike the $Al_xGa_{1-x}As$ material system, where the higher energy satellite band never forms the ultimate conduction band, E_c^4 and E_c^2 in the $Si_{1-x}Ge_x$ material system can both form the ultimate conduction band. Thus, it is possible to have near equilibrium transport occur within both E_c^4 and E_c^2 at spatially separate points with the device; this is in contrast to the $Al_xGa_{1-x}As$ material system where transport in the satellite band need only be considered under extreme non-equilibrium injection conditions. Further, this multiband transport can also occur in the valence band of the $Si_{1-x}Ge_x$ material system. Given the strange band offsets depicted in Figs. 6.9 to 6.11, it will be shown that transport within the $Si_{1-x}Ge_x$ material system can offer a rich set of possibilities, both in terms of commercial HBT optimisation, and as a tool for research into the mechanics of transport within solids.

Considering the valence band first of all, Fig. 6.2 shows that E_v^{hh} and E_v^{lh} are degenerate under the condition of no strain. More importantly, the maxima in both E_v^{hh} and E_v^{lh} occur at the same point in $\boldsymbol{k}$-space. Even under strain, the maxima in E_v^{hh} and E_v^{lh} remain coincident in their $\boldsymbol{k}$-space location. Therefore, there is very little issue regarding the conservation of crystal momentum in moving between the lh and hh bands. If the mass barrier that would occur at the heterojunction for holes is neglected, then to a good approximation one need only consider the ultimate valence band in terms of hole transport. However, if the strain is small, so that the energy separation between E_v^{hh} and E_v^{lh} is less than ~2kT, then transport within both bands needs to be considered. As is attested by eqn (6.1), the mass barrier for holes cannot be neglected as γ from eqn (4.80) is typically -2 but can be as small as -10. With $\gamma = -2$, fully two-thirds of the current crossing the mass barrier could be reflected, leading to a 3-fold reduction in the transport current. A 3-fold reduction in the transport current would be equivalent to having an upwards step in energy of

28.5meV at room temperature. Therefore, when ΔE_v is less than ~2kT one must consider parallel transport within E_v^{hh} and E_v^{lh}. But, no matter how large or small ΔE_v is, the calculation of the valence band effective density of states N_v must include both E_v^{hh} and E_v^{lh} due to the large difference in the lh and hh effective mass.

The complexity of the valence band stems from the coincident ***k***-space location of the band maxima for E_v^{hh} and E_v^{lh}. Examination of Fig. 6.7 shows that the $Si_{1-x}Ge_x$ conduction band minima are not coincident in ***k***-space. Thus, in order to move between any of the six Δ minima in $Si_{1-x}Ge_x$, crystal momentum must be conserved. There are two scattering processes that are responsible for intervalley scattering between the six conduction band Δ minima in $Si_{1-x}Ge_x$ [129] (see Fig. 6.12); g scattering moves electrons between two bands that are along a common major k-axis, such as the [001] and [00$\bar{1}$] bands that form E_c^2; while f scattering moves electrons between two bands that are not along a common major k-axis, such as the [100] and [010] bands within E_c^4. Given the proximity of the Δ minima to the Brillouin zone edge, an Umklapp process can easily take place, leading to g scattering, because of the relatively small ***k***-space separation that must be conserved. On the other hand, f scattering involves a ***k***-space conservation that is over one-half of the reciprocal lattice length. Therefore, it is found that f scattering rates are almost 10-fold lower than g scattering rates [129]. In terms of the E_c^4 and E_c^2 band groupings, g scattering will not result in movement between the E_c^4 and E_c^2 bands. Finally, for small distances, such as those that are typical of the EB SCR and neutral base width, f scattering is small enough to be ignored [108,130]. These two results regarding intervalley scattering allow the E_c^4 and E_c^2 bands to be treated independently, allowing for a large simplification as compared to the valence sub-bands.

The arguments of the previous paragraph, justifying the independence of the E_c^4 and E_c^2 bands, must be considered in the light of an abrupt heterojunction. At an abrupt heterojunction, one would expect that a powerful Bragg plane could exist that would be perpendicular to the direction of charge transport across the heterojunction. Such a powerful Bragg plane could enhance f scattering, leading to a coupling between the E_c^4 and E_c^2 bands. Consideration of the k-vector involved in f scattering relative to the Bragg plane, shows the two are separated by 45°. With a 45° degree separation, it would not be expected that Bragg plane scattering at an abrupt heterojunction would lead to a significant increase in the f scattering rate [108]. Therefore, the independence of the E_c^4 and E_c^2 bands should be maintained even at an abrupt heterojunction. This leads to the for-

mation of a selection rule regarding transport in $Si_{1-x}Ge_x$ that prohibits a mixing between the electrons in E_c^4 and E_c^2.

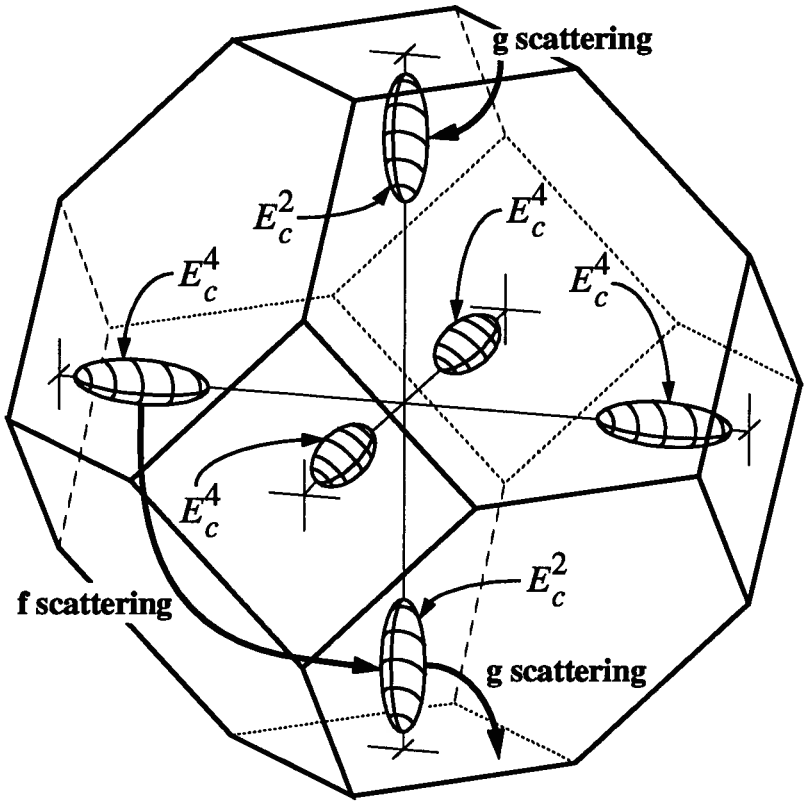


Fig. 6.12. Diagram of the Δ conduction band minima involved in f and g intervalley scattering. For clarity, only 1 of the 3 g scattering processes, and 1 of the 12 f scattering processes is shown.

With the E_c^4 and E_c^2 bands treated independently of each other, the task of modelling electron transport within the $Si_{1-x}Ge_x$ material system begins with calculation of the electron effective densities of states, N_c^4 and N_c^2 respectively. The density of states for band n is given by [131]:

$$g_c^n(E) = n \oint_{\text{B.Z.}} \frac{\mathrm{d}^3k}{4\pi^3} \delta(E - E_c^n(\boldsymbol{k})) \tag{6.20}$$

where n = 2 or 4 in the case of $Si_{1-x}Ge_x$ strained to a {100} substrate, and B.Z. means Brillouin zone. The pre-multiplying factor of n in eqn (6.20) results from the degeneracy of the E_c^4 and E_c^2 bands and the fortuitous designation where n is equal to 2 or 4. Then the effective density of states, assuming that the band-width is E_b and that Boltzmann statistics can be used, is equal to:

$$N_c^n = \int_0^{E_b} \mathrm{d}E\, g_c^n(E) \mathrm{e}^{-\frac{E}{kT}} = n \oint_{\text{B.Z.}} \frac{\mathrm{d}^3k}{4\pi^3} \int_0^{E_b} \mathrm{d}E\, \delta(E - E_c^n(\boldsymbol{k})) \mathrm{e}^{-\frac{E}{kT}} = n \oint_{\text{B.Z.}} \frac{\mathrm{d}^3k}{4\pi^3} \mathrm{e}^{-\frac{E_c^n(\boldsymbol{k})}{kT}}. \tag{6.21}$$

Eqn (6.21) can be easily integrated with little error by assuming that the limit of integration can be extended past the Brillouin zone to infinity; *i.e.*,

$$N_c^n = \frac{n}{4\pi^3} e^{-\frac{E_c^n}{kT}} \int_{-\infty}^{\infty} dk_x e^{-\frac{\hbar^2 k_x^2}{2m_l kT}} \int_{-\infty}^{\infty} dk_y e^{-\frac{\hbar^2 k_y^2}{2m_l kT}} \int_{-\infty}^{\infty} dk_z e^{-\frac{\hbar^2 k_z^2}{2m_t kT}} = 2ne^{-\frac{E_c^n}{kT}} \left(\frac{m^* kT}{2\pi\hbar^2}\right)^{3/2} \tag{6.22}$$

where

$$m^* = (m_l m_t m_t)^{1/3}.$$

The appearance of the term $\exp(-E_c^n/kT)$ in eqn (6.22) is due to the fact that the reference energy for the conduction sub bands is not located at the band minima, but at $\overline{E}_c$. One could have maintained the reference energy at the band minima, but then N_c^4 and N_c^2 would have different energy references and eqns (6.10)-(6.11) could not be used directly within eqn (6.22). Furthermore, by employing a common energy reference of $\overline{E}_c$, the total conduction band effective density of states is:

$$N_c = N_c^4 + N_c^2 = 4\left(\frac{m^* kT}{2\pi\hbar^2}\right)^{3/2} \left(e^{-\frac{E_c^2}{kT}} + 2e^{-\frac{E_c^4}{kT}}\right). \tag{6.23}$$

Finally, it is possible to reflect N_c from the energy reference of $\overline{E}_c$ back to the ultimate conduction band minima by multiplying eqn (6.23) with $\exp(\min(E_c^4, E_c^2)/kT)$.

The exact same methods used to determine N_c^4 and N_c^2 can be applied to the calculation of the hole effective density of states within the valence sub-bands, leading to:

$$N_v^{hh} = 2e^{\frac{E_v^{hh}}{kT}} \left(\frac{m_{hh} kT}{2\pi\hbar^2}\right)^{3/2} \text{ and } N_v^{lh} = 2e^{\frac{E_v^{lh}}{kT}} \left(\frac{m_{lh} kT}{2\pi\hbar^2}\right)^{3/2}. \tag{6.24}$$

Then, owing to the different effective masses for the lh and hh, the total valence band effective density of states is given by:

$$N_v = N_v^{hh} + N_v^{lh} = 2\left(\frac{kT}{2\pi\hbar^2}\right)^{3/2} \left((m_{hh})^{3/2} e^{\frac{E_v^{hh}}{kT}} + (m_{lh})^{3/2} e^{\frac{E_v^{lh}}{kT}}\right). \tag{6.25}$$

In a similar fashion to the conduction band, the reference energy for the valence band is not located at the ultimate valence band maxima, but at the location of the valence band maxima under the condition of no strain. To reflect N_v back to the energy of the ultimate valence band maxima, multiply eqn (6.25) by $\exp(-\max(E_v^{hh}, E_v^{lh})/kT)$.

Eqns (6.22)-(6.25) present the conduction and valence band effective density of states for the $Si_{1-x}Ge_x$ material system. These equations represent an extension to the traditional definitions for effective density of states, necessitated by the complex band structure of $Si_{1-x}Ge_x$ under the influence of symmetry-breaking strain. Finally, the electron and hole concentrations n and p respectively are defined using eqns (6.25) and (6.23) in the usual non-degenerate manner, to yield:

$$n = N_c e^{\frac{E_{f,n}}{kT}} \text{ and } p = N_v e^{-\frac{E_{f,p}}{kT}}, \tag{6.26}$$

where $E_{f,n}$ is the electron quasi-Fermi energy relative to $\overline{E}_c$, and $E_{f,p}$ is the hole quasi-Fermi energy relative to the unstrained valence band maxima. After allowing for the fact that the conduction and valence band energy references are separated by $\overline{E}_g$, as is shown in Fig. 6.8, then:

$$\begin{aligned} n_i^2 = pn &= N_c N_v e^{-\frac{\overline{E}_g}{kT}} \\ &= \left(\frac{kT}{\pi\hbar^2}\right)^3 (m^*)^{3/2} \left(e^{-\frac{E_c^2}{kT}} + 2e^{-\frac{E_c^4}{kT}}\right)\left((m_{hh})^{3/2} e^{\frac{E_v^{hh}}{kT}} + (m_{lh})^{3/2} e^{\frac{E_v^{lh}}{kT}}\right) e^{-\frac{\overline{E}_g}{kT}}, \end{aligned} \tag{6.27}$$

where the average bandgap $\overline{E}_g$ is equal to the bulk alloy bandgap given in eqn (6.3), plus the hydrostatic change to the bandgap $\Delta\overline{E}_g$ given in eqn (6.14). Unlike eqns (6.22)-(6.26), n_i^2 given in eqn (6.27) does not reference itself to an abstract energy reference, but is the standard definition for the intrinsic carrier concentration.

With the effective density of states defined for the conduction and valence sub-bands in eqns (6.22)-(6.25), along with the carrier concentrations and n_i^2 given in eqns (6.26)-(6.27), it is possible to define the built-in potential V_{bi} of a *pn*-junction. Looking at Fig. 6.13, then clearly:

$$V_{bi} = \frac{1}{q}\left(E_{f,n} - \left(E_{f,p} - \overline{E}_{g,p}\right)\right) + \frac{1}{q}(\chi_p - \chi_n) = \frac{kT}{q}\ln\left(\frac{N_D N_A}{n_{i,p}^2}\frac{N_{c,p}}{N_{c,n}}\right) + \frac{1}{q}(\chi_p - \chi_n). \tag{6.28}$$

Comparison of eqn (6.28) with eqn (4.69) shows, apart from the effect of a spatially varying effective density of states (which is neglected in eqn (4.69)), exact agreement if $\chi_p - \chi_n = \Delta E_c$. V_{bi} is the variation in the vacuum potential across the SCR extrapolated back to equilibrium conditions. Thus, the electron affinities χ_p and χ_n on the p- and n-sides of the junction are evaluated at $x = x_p$ and $x = -x_n$ respectively. If χ_p and χ_n are spatially varying, then as a changing applied bias moves x_p and x_n, V_{bi} will also vary with applied bias. It is well known that Anderson's electron affinity rule for the calculation of ΔE_c is not correct. However, at some distance far from the hetero-

junction, χ_p and χ_n must become bulk-like. The question becomes how rapidly do χ_p and χ_n return to their bulk values? The deviation of ΔE_c from $\chi_p - \chi_n$ has been attributed to such things as a complex rearrangement of charge surrounding the heterojunction. Thus, if this rearrangement of charge is abrupt, as is potentially suggested by Van de Walle and Martin [106, 119,120], then χ_p and χ_n would definitely change over the width of the SCR; leading to an extra driving force for the transport of charge than is not taken into account by any known theories. If this rapid variation in χ_p and χ_n turns out to be true, then V_{bi} will not be a constant as is given in eqn (4.69), but is instead given by eqn (6.28) with χ_p and χ_n being a function of position. Finally, V_{bi} contains all of the desired information regarding χ_p, χ_n, and thus ΔE_c. Therefore, if the *pn*-junction could be driven up to and past V_{bi}, without resistive effects dominating the transport current, then information regarding χ_p, χ_n, and thus ΔE_c could be extracted. This possibility of operation near and past V_{bi} will be considered in Section 6.4.

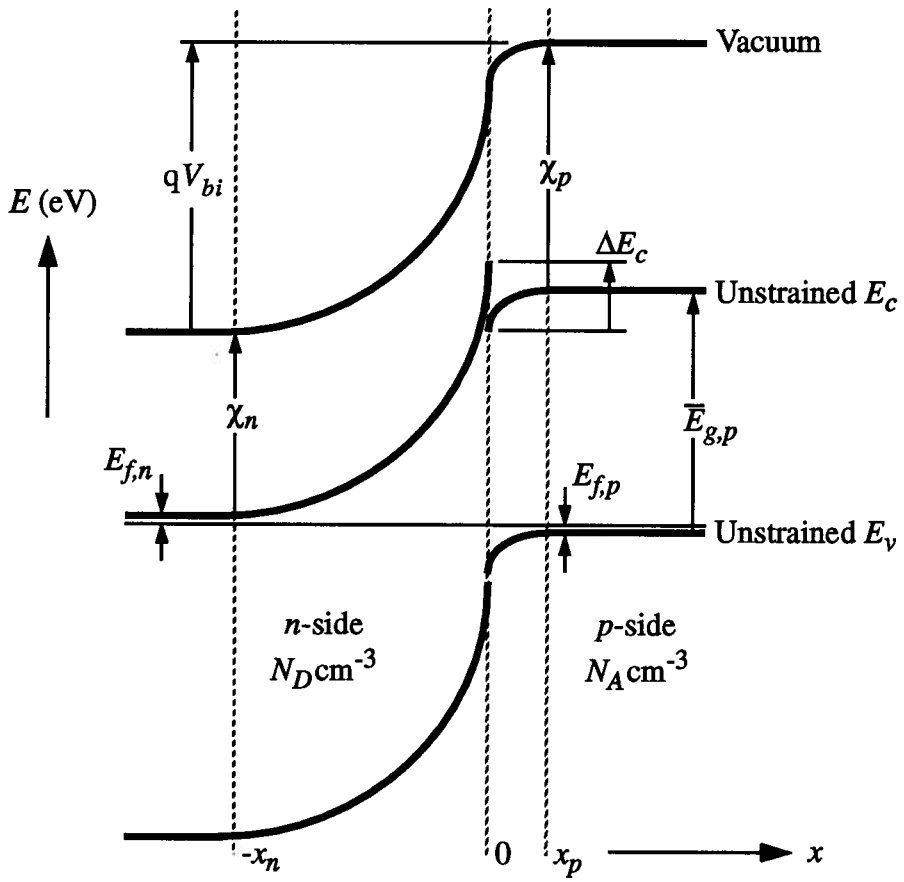


Fig. 6.13. Equilibrium band diagram of a *np*-junction, showing the relevant energies and potentials. $E_{f,n}$ is referenced to E_c while $E_{f,p}$ is referenced to E_v. Note that the Vacuum potential is continuous while E_c and E_v are not.

Concentrating once again on the conduction band, the final models for electron transport can be determined. By the previous arguments, electron transport in the EB SCR and the neutral base can be modelled as two parallel conduction paths via E_c^4 and E_c^2. It is further assumed, at least with the current-day knowledge of the $Si_{1-x}Ge_x$ material system, that eqn (6.2) is correct, which precludes the formation of a mass barrier. Therefore, transport through the EB SCR would be given by the sum of E_c^4 and E_c^2 conduction solved by the standard transport model given by eqns (4.78)-(4.79) and (4.92).

To this end, the correct parameters to use in the standard EB SCR transport model regarding E_c^4 conduction are:

$$m_{x,1} = m_t(\Delta) \qquad m_{y,1} = m_l(\Delta) \qquad m_{z,1} = m_t(\Delta) \qquad \mu_1 = \text{relative to } E_c^4$$

$$F_{f,s0}^4 = \frac{4\pi q^2\sqrt{m_{y,1}m_{z,1}}\,kT}{h^3}e^{\frac{\mu_1}{kT}} = \frac{4\pi q^2\sqrt{m_{y,1}m_{z,1}}\,kT}{h^3}\left(4\frac{N_D}{N_c}e^{-\frac{E_c^4}{kT}}\right) = \frac{q^2}{\sqrt{2\pi m_t(\Delta)kT}}\left(N_D\frac{N_c^4}{N_c}\right)$$

band degeneracy

$$= \frac{q^2N_D}{\sqrt{2\pi m_t(\Delta)kT}}\left(\frac{2e^{-\frac{E_c^4}{kT}}}{e^{-\frac{E_c^2}{kT}}+2e^{-\frac{E_c^4}{kT}}}\right) \tag{6.29}$$

While the correct parameters to use in the standard EB SCR transport model regarding E_c^2 conduction are:

$$m_{x,1} = m_l(\Delta) \qquad m_{y,1} = m_t(\Delta) \qquad m_{z,1} = m_t(\Delta) \qquad \mu_1 = \text{relative to } E_c^2$$

$$F_{f,s0}^2 = \frac{4\pi q^2\sqrt{m_{y,1}m_{z,1}}\,kT}{h^3}e^{\frac{\mu_1}{kT}} = \frac{4\pi q^2\sqrt{m_{y,1}m_{z,1}}\,kT}{h^3}\left(2\frac{N_D}{N_c}e^{-\frac{E_c^4}{kT}}\right) = \frac{q^2}{\sqrt{2\pi m_l(\Delta)kT}}\left(N_D\frac{N_c^4}{N_c}\right)$$

band degeneracy

$$= \frac{q^2N_D}{\sqrt{2\pi m_l(\Delta)kT}}\left(\frac{e^{-\frac{E_c^2}{kT}}}{e^{-\frac{E_c^2}{kT}}+2e^{-\frac{E_c^4}{kT}}}\right) \tag{6.30}$$

Examination of $F_{f,s0}$ in eqns (6.29) and (6.30) reveals the exact context of parallel transport within E_c^4 and E_c^2. There are N_D majority electrons that are distributed between the E_c^4 and E_c^2 bands, depending upon the energy separation between the two. Since there are twice as many Δ minima in E_c^4 as compared to E_c^2, there will be preferential transport within E_c^4, all other things being equal.

Finally, the electrons within E_c^4 and E_c^2 move with a velocity that is proportional to the square-root of $1/m_t$ and $1/m_l$ respectively. Therefore, neglecting the energy separation between E_c^4 and E_c^2, the E_c^4 band will carry $2\sqrt{m_l/m_t} = 4.54$ times the current compared to E_c^2. Furthermore, because E_c^4 has the light transverse mass parallel to the direction of transport, as compared to the heavy longitudinal mass for E_c^2, not only is the mobility higher [132] but the probability of tunneling through a given barrier will be much higher for E_c^4 compared to E_c^2.

Regarding transport within the neutral base, the independence between E_c^4 and E_c^2 can only be maintained if the neutral base width is small enough to preclude coupling via the f scattering process. For current-day SiGe HBTs the neutral base width W_{nb} is under 1000Å and is rapidly approaching 300Å [10,26,28,100-103,133]. With such a small neutral basewidth it is reasonable to maintain the separation between E_c^4 and E_c^2 used for the modelling of EB SCR transport. With E_c^4 and E_c^2 treated independently, the neutral base transport current within either one of the E_c^j sub-bands is given by Kroemer [38] as:

$$J_{T,NB}^{j} = \frac{q\overline{D}_n^j}{\int_{W_{nb}} \frac{N_A(x)}{n_{i,j}^2(x)}dx}\left(e^{\frac{qV_{BE}-\Delta E_{fn}^j}{kT}} - 1\right) \tag{6.31}$$

where $j = 2$ or 4. It should be noted that eqn (6.31) is an extension of Kroemer's work which was based upon Shockley boundary conditions. The reason for generalising the diffusion coefficient $\overline{D}_n^j$, as was discussed in the previous paragraph, stems from the fact that the mobilities within the E_c^4 and E_c^2 bands will be different due to their highly anisotropic nature [132]. This leads to the conclusion that $\overline{D}_n^4 > \overline{D}_n^2$ because $m_t < m_l$. Further, each sub-band will have its own intrinsic carrier concentration $n_{i,j}^2$, which is determined in the same way as N_c^4, N_c^2 and the total n_i^2 to yield:

$$n_{i,j}^2 = N_c^j N_v e^{-\frac{\overline{E}_g}{kT}} = \frac{N_c^j}{N_c} n_i^2 \qquad \Rightarrow n_i^2 = n_{i,4}^2 + n_{i,2}^2. \tag{6.32}$$

Finally, due to the independence of the E_c^4 and E_c^2 bands, a separate quasi-fermi energy must be present in order to account for the driving force within each sub-band. For this reason, there is ΔE_{fn}^4 to characterise transport within the E_c^4 band, and ΔE_{fn}^2 to characterise transport within the E_c^2 band.

The final model for electron transport within the SiGe HBT is achieved using exactly the same methods employed in Section 5.3 for the derivation of eqn (5.29). Eqn (5.29) is based upon

the general models of Section 2.2. Applying the models of Section 2.2 to the solution of transport within the conduction sub-bands yields for the E_c^4 band:

$$J_T^4 = \left[\frac{1}{F_{f,s}^4} + \frac{1}{J_{T,NB}^4}\bigg|_{\Delta E_{fn}^4 = 0} \right]^{-1}, \tag{6.33}$$

where, based upon the findings of Chapter 5, the failure to include recombination effects specifically in the calculation of the total transport current J_T^4 will produce an error that is of order $1/\beta$. To reiterate, $F_{f,s}^4$ is the EB SCR transport current solved by the standard transport model given by eqns (4.78)-(4.79) and (4.92) with the pertinent parameters obtained from eqn (6.29). In a similar fashion, transport within the E_c^2 band is:

$$J_T^2 = \left[\frac{1}{F_{f,s}^2} + \frac{1}{J_{T,NB}^2}\bigg|_{\Delta E_{fn}^2 = 0} \right]^{-1}, \tag{6.34}$$

where $F_{f,s}^2$ is once again the EB SCR transport current solved by the standard transport model given by eqns (4.78)-(4.79) and (4.92), but with the pertinent parameters obtained from eqn (6.30). Then, the total electron transport J_T through the HBT is given by the sum of J_T^4 and J_T^2.

To conclude this section, transport within the valence sub-bands is addressed. As was discussed earlier in this section, the coincident nature of E_v^{hh} and E_v^{lh} in terms of $\boldsymbol{k}$-space location prohibits an independent treatment, such as was done for the conduction sub-bands, of the two valence sub-bands. Fig. 6.2 clearly shows that the valence band of unstrained SiGe, and for that matter all semiconductors, is a multi-band system. To this end, transport within the unstrained valence band is determined by appealing to a single total effective mass that correctly produces the total valence band effective density of states. Then, by way of experimental measurement, a single mobility is extracted to characterise the valence band as a whole. This method breaks down for the case of strained SiGe, as the degeneracy of E_v^{hh} and E_v^{lh} is lifted and the energy separation is dependent upon the amount of strain present. This prohibits the use of a single effective mass and fixed mobility to characterise the valence band of strained SiGe. Yet, the valence sub-bands cannot be treated independently for the purpose of determining charge transport, as was done for the conduction sub-bands.

Essentially, the only way to solve transport within the strained SiGe valence band is to resort to Monte Carlo simulation. However, as was pointed out at the start of Chapter 4, Monte Car-

lo simulators cannot presently model the non-local effects of tunneling. To this end, the following two assumptions are made: 1) hole transport within the EB SCR is considered ballistically due to the small width of the SCR, but the holes will always attempt to minimise their energy by moving to the highest sub-band; 2) due to the strong intervalley scattering that occurs between E_v^{hh} and E_v^{lh}, because of their coincidence in $\boldsymbol{k}$-space, transport within the wider neutral regions of the HBT is treated using a single equivalent valence band.

The implication of the second assumption is straightforward; transport is treated in the standard single equivalent valence band approach. The only consideration that must be made in treating the valence band as a single valley is the mobility will change with strain. In a region where the Ge alloy content is not uniform the strain will change with position, which will move E_v^{hh} and E_v^{lh} either closer or further apart in terms of energy. Since the lh mass is much smaller than the hh mass, considerable change to the mobility of the material will occur as E_v^{hh} and E_v^{lh} move closer and further apart. This leads to a complex and spatially non-uniform mobility that is only due to the energy separation of the valence sub-bands. Other effects such as impurity and alloy scattering would also have to be considered.

The implications of the first assumption are even more interesting than those of the second. For the purpose of tunneling, the lightest mass will produce the largest tunneling flux. But, conservation of transverse momentum must be ensured for a hole to change bands, which leads to the mass barrier results of Chapter 4. However, the hole will attempt to take the path of least resistance by minimising its energy; it may either continue on in the sub-band it currently occupies, or change bands in an attempt to minimise its energy while taking into account the possible loss or gain due to the mass boundary effect. The complexity of transport within the SiGe valence band stems purely from the large difference in the lh and hh masses. If the lh and hh masses were the same, then transport would occur along the highest energy sub-band (in terms of electron energies), with a spatially varying N_v to consider.

The model for the EB SCR in Chapter 4 is simple in that the heterojunction is abrupt; thereby producing two regions, separated by a single mass barrier, where the material parameters within each region are a constant (see Fig. 4.2). As a result of this, the relative separation between E_v^{hh} and E_v^{lh} will not change, except at the mass boundary. Therefore, for the calculation of the EB SCR transport current for holes in the $Si_{1-x}Ge_x$ material system:

- initially consider the E_v^{hh} and E_v^{lh} bands independently, injecting a hemi-Maxwellian of holes into the EB SCR, characterised by the individual mass of the band.
- Using the standard flux model, given by eqns (4.78), (4.79) and (4.92), calculate the standard flux $F_{f,s}$ ($F_{f,s}$ does not include the mass barrier) using the appropriate mass from eqn (6.1), and

$$F_{f,s0}^{j} = \frac{q^2}{\sqrt{2\pi m_j kT}}\left(N_A \frac{N_v^j}{N_v}\right),$$

where j is either hh or lh. Within the standard flux model, the base barrier potential V_b is no longer the one from the originating band, but is given by the maximum of E_v^{hh} and E_v^{lh} in the neutral region (this is where the minimisation of hole energy enters the calculation).

- If the mass barrier effects are not considered, then the problem ends here. But, the mass barrier can be quite large in the valence band, producing a potentially non-negligible effect. However, the mass barrier effects are only important if the aforementioned calculation of the standard flux has the holes changing between E_v^{hh} and E_v^{lh}. If the holes do change bands then eqns (4.85)-(4.86) are used in the case of an enhancing mass barrier; where as eqn (4.87) is used for the reflecting mass barrier, but with the infinite upper limit of integration replaced with the V_b that is appropriate to the sub-band that injected the holes.

The physical explanation of the valence band transport model is: holes ballistically travel through the EB SCR, perhaps tunneling through a Valence Band Spike (VBS), by way of independent E_v^{hh} and E_v^{lh} bands. Upon reaching the mass barrier the holes attempt to occupy the lowest energy band, and do so by exchanging sub-bands, if necessary, while taking into account any losses or gains due to the mass barrier. Depending upon the construction of the HBT, the emerging fluxes from the EB SCR, contained within E_v^{hh} and E_v^{lh}, will generally be characterised by different driving forces of ΔE_{fp}^{hh} and ΔE_{fp}^{lh} respectively. However, due to the strong intervalley scattering that occurs between the valence sub-bands upon reaching the neutral region, a common quasi-equilibrium condition of ΔE_{fp} will result for both E_v^{hh} and E_v^{lh}. Therefore, the final transport model for holes is:

$$J_{T,holes} = \left[\frac{1}{F_f^{hh} + F_f^{lh}} + \frac{1}{J_{T,neutral}\Big|_{\Delta E_{fp}=0}}\right]^{-1}, \tag{6.35}$$

where F_f^{hh} and F_f^{lh} are the full EB SCR transport models, and $J_{T,neutral}$ is the neutral region transport current calculated by eqn (6.31) using n_i^2 from eqn (6.27).

6.4 The Accumulation Regime Beyond the Built-In Potential

Chapter 4, and therefore Section 6.3, have both dealt with transport for an applied bias V_{BE} that is less than the built-in potential V_{bi}. For the case of a band diagram where there is a negative step, as shown in Figs. 6.13 and 4.2, as V_{BE} approaches V_{bi}, a current density of ~10^6A/cm^2 will flow (this is based upon an emitter doping that is ~10^{18}cm^{-3}). At a current density of 10^6A/cm^2, resistive effects will dominate the device and limit the internal forward bias to be much less than the external applied bias. For example, with an emitter area of 1μm^2, there would be a current of 10mA at a current density of 10^6A/cm^2. Even with an unrealistically low emitter contact resistance of 50Ωμm^2, there would be a 500mV drop to the external applied bias before it even reached the junction. It is for this simple reason that observation of the device with a forward bias near, and certainly beyond, V_{bi} is not really experimentally possible.

As is evidenced by the plot of ΔE_c in Fig. 6.10, along with ΔE_c^4 and ΔE_c^2 shown in Fig. 6.9, there exists the possibility of constructing a positive-going potential step (see Fig. 6.14a) in the path of the electrons trying to surmount the potential barrier of the EB SCR. A positive potential step would force the electrons to surmount the entire barrier, because unlike the CBS there is no way to tunnel though the step. Therefore, if the step potential were as large as 240meV (*i.e.*, ΔE_c = -240meV), then by eqn (4.79) the charge flowing through the EB SCR at room temperature would be reduced by a factor of exp(-240/25.9) ≈ 10^{-4}. Therefore, when V_{BE} approaches V_{bi}, the current density will have dropped to only 10^2A/cm^2. A current density of 10^2A/cm^2 will certainly be observable, and would even allow for V_{BE} to exceed V_{bi}.

Before going on to present a physical demonstration of operation beyond V_{bi}, the transport theory for this domain of operation is first developed. When V_{BE} is exactly equal to V_{bi}, and if the resistive effects are negligible, then the band diagram will be flat except at abrupt heterojunctions or regions of spatially non-uniform Ge alloy content (see Fig. 6.14b). For this reason, the point at which V_{BE} is exactly equal to V_{bi} is termed flat-band (in much the same manner as the flat-band condition in MOSFETs). At flat-band there will be no space charge present. As V_{BE} is increased past V_{bi} (see Fig. 6.14c), an accumulation region of mobile electrons on the n-side, as well as mobile holes on the p-side, of the heterojunction will begin to form (as has been the case throughout, a coincident hetero- and metallurgical-junction is assumed). This is contrast to the standard case where $V_{BE} < V_{bi}$, and a depletion region forms where the space charge is composed of immobile

ion cores from the dopant atoms. For this reason, operation past V_{bi} is termed the accumulation regime. Finally, as V_{BE} is increased further, the accumulation of charge will proceed exponentially, with a net reduction to the potential step, and therefore, a continued exponential increase in the EB SCR transport current.

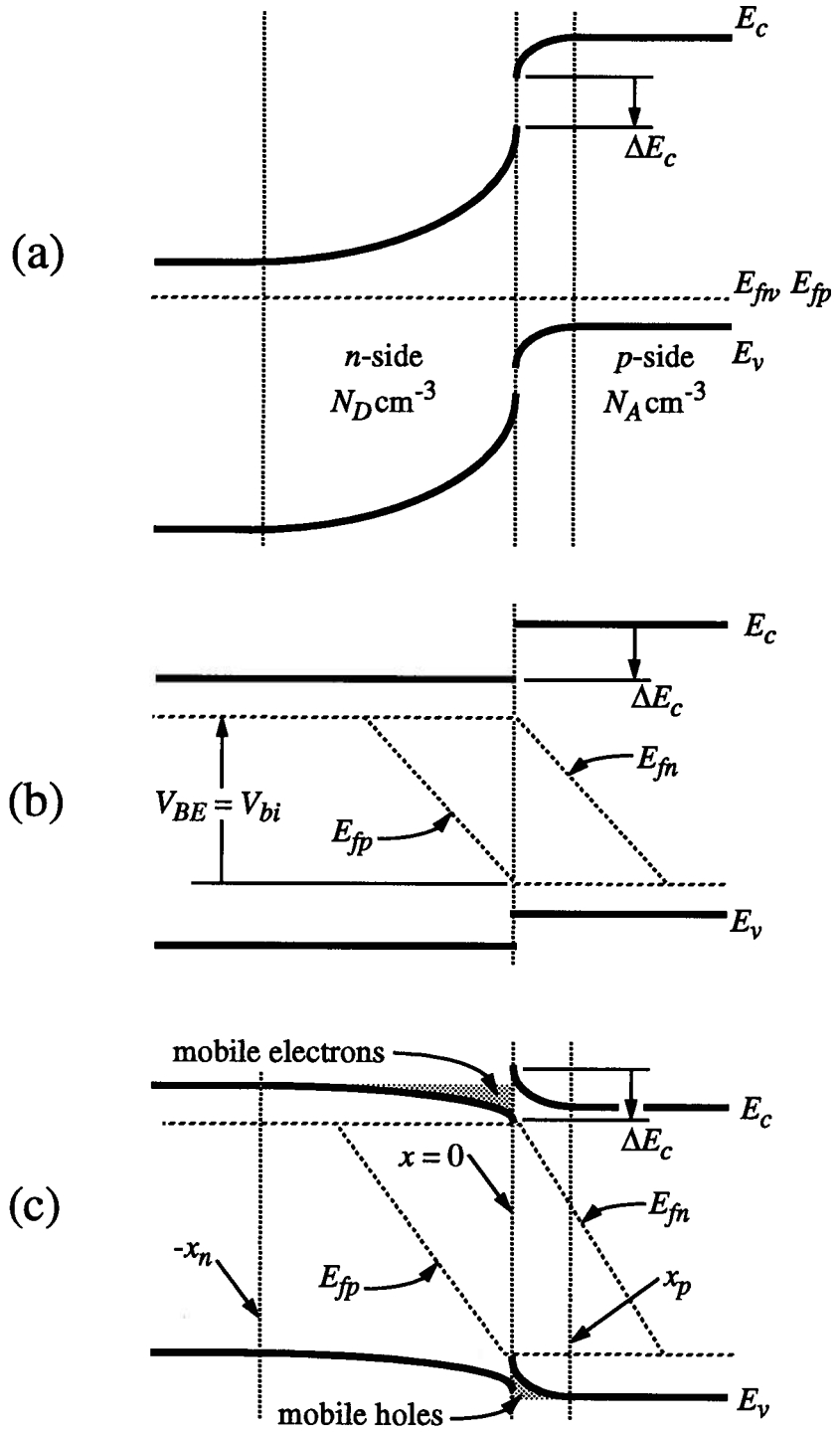


Fig. 6.14. Band diagram for a *np*-junction with a positive step potential (*i.e.*, $\Delta E_c < 0$). (a) equilibrium; (b) flat-band where $V_{BE} = V_{bi}$; (c) accumulation region where $V_{BE} > V_{bi}$.

A reasonable first approximation to the complex accumulation regime begins by assuming that for operation just beyond V_{bi} the accumulation layer is non-degenerate. Based upon this assumption, and neglecting the effects of a non-uniform ε, the Poisson equation in one dimension becomes

$$\frac{d^2\psi}{dx^2} = \begin{cases} \frac{q^2}{\varepsilon}\left(N_D - N_D e^{-\frac{\psi_n}{kT}}\right) & \text{on the n-side} \\ -\frac{q^2}{\varepsilon}\left(N_A - N_A e^{\frac{\psi_p}{kT}}\right) & \text{on the p-side} \end{cases} \tag{6.36}$$

where ψ_n and ψ_p are in terms of electron energy (*i.e.*, the negative of potential energy). Eqn (6.36) is solved on the *n*-side (the *p*-side solution can be obtained directly from the *n*-side solution by symmetry arguments) to yield the following implicit transcendental function:

$$x = \pm\sqrt{\frac{\varepsilon}{2}}\frac{1}{q}\int \sqrt{\frac{e^{\frac{\psi_n}{kT}}}{N_D\left[(\psi_n + A_1)e^{\frac{\psi_n}{kT}} + kT\right]}}\, d\psi_n + A_2, \tag{6.37}$$

where A_1 and A_2 are arbitrary constants. There is no way to reduce eqn (6.37) down to a function of simple transcendental functions, nor will it be possible to invert the result. However, it is reasonable to assume that the charge in the accumulation layer will overwhelm the background dopant ion potential. With this assumption eqn (6.36) is recast to:

$$\frac{d^2\psi}{dx^2} \approx \begin{cases} -\frac{q^2}{\varepsilon}N_D e^{-\frac{\psi_n}{kT}} & \text{on the n-side} \\ \frac{q^2}{\varepsilon}N_A e^{-\frac{\psi_p}{kT}} & \text{on the p-side} \end{cases} \tag{6.38}$$

whose *n*-side solution is

$$x = \pm\sqrt{\frac{\varepsilon}{2}}\frac{1}{q}\int \sqrt{\frac{e^{\frac{\psi_n}{kT}}}{N_D\left[A_1 e^{\frac{\psi_n}{kT}} + kT\right]}}\, d\psi_n + A_2. \tag{6.39}$$

It is interesting to note that the only difference between the approximate solution of eqn (6.39) and the full solution of eqn (6.37) is the extra term containing ψ_n in the denominator of the radi-

cal. This linear term in ψ_n produces the asymptotic solution to the underlying depletion space charge. Since eqn (6.39) assumes that the depletion space charge is negligible in comparison to the accumulation charge, the linear term in ψ_n is lost.

The solution of eqn (6.39) begins with the determination of A_1. If the neutral assumption is employed at $-x_n$ (the boundary to the accumulation region), then because the doping N_D is a constant the electric field will vanish. Since the electric field is given by $(1/q)d\psi_n/dx$, then taking the derivative of eqn (6.39) with respect to ψ_n, inverting it, and setting it equal to zero with $\psi_n = 0$ at $x = -x_n$ yields $A_1 = -kT$. With $A_1 = -kT$, eqn (6.39) is solved using the change of variables

$$y = 2e^{\frac{\psi_n}{kT}} - 1$$

to produce:

$$x = \pm\frac{1}{q}\sqrt{\frac{\varepsilon kT}{2N_D}}\,\mathrm{asin}\left(2e^{\frac{\psi_n}{kT}} - 1\right) + A_2 \;\Rightarrow e^{\frac{\psi_n}{kT}} = \frac{1}{2}\sin\left(\mp\frac{x}{\frac{1}{q}\sqrt{\frac{\varepsilon kT}{2N_D}}} - A_2\right) + \frac{1}{2}\,. \tag{6.40}$$

Finally, applying the energy reference of $\psi_n = 0$ at $x = -x_n$ to eqn (6.40), and choosing the positive x-direction produces:

$$e^{\frac{\psi_n}{kT}} = \cos^2\left(\frac{x + x_n}{2a_{1,n}}\right) \tag{6.41}$$

where

$$a_{1,n} = \frac{1}{q}\sqrt{\frac{\varepsilon kT}{2N_D}}.$$

By appealing to the symmetry of the problem, the p-side solution of eqn (6.38) is:

$$e^{-\frac{\psi_p}{kT}} = \cos^2\left(\frac{x - x_p}{2a_{1,p}}\right) \tag{6.42}$$

where

$$a_{1,p} = \frac{1}{q}\sqrt{\frac{\varepsilon kT}{2N_A}}.$$

It is important to realise that ψ_n is set equal to 0 at $x = -x_n$, and ψ_p is set equal to 0 at $x = x_p$. However, the form of the Poisson equation requires that when ψ_n joins up with ψ_p at the heterojunction (*i.e.*, at $x = 0$), the joint be analytic up to first derivatives. Given that we are solving for accumulation and not depletion, then continuity of ψ_n and ψ_p requires that:

$$\psi_p(0) - \psi_n(0) = q\,(V_{BE} - V_{bi}) \quad \Rightarrow e^{-q\frac{V_{BE}-V_{bi}}{2kT}} = \cos\left(\frac{x_n}{2a_{1,n}}\right)\cos\left(\frac{x_p}{2a_{1,p}}\right). \tag{6.43}$$

Further, continuity of the electric field requires that:

$$\left.\frac{d\psi_n}{dx}\right|_{0^-} = \left.\frac{d\psi_p}{dx}\right|_{0^+} \quad \Rightarrow -\frac{kT}{a_{1,n}}\tan\left(\frac{x_n}{2a_{1,n}}\right) = -\frac{kT}{a_{1,p}}\tan\left(\frac{x_p}{2a_{1,p}}\right). \tag{6.44}$$

It is a straightforward task involving considerable bookkeeping to solve simultaneously eqns (6.43) and (6.44) for x_n and x_p. Eqns (6.43) and (6.44) form a quadratic equation involving the squared cosines of $x_n/2a_{1,n}$ and $x_p/2a_{1,p}$. Choosing the positive roots of the solution for eqns (6.43) and (6.44) yields:

$$\begin{aligned} x_n &= 2a_{1,n}\,\text{acos}\left(\sqrt{\frac{\sqrt{4N_AN_De^{\Delta V_{BE}/V_T} + (N_A - N_D)^2} + N_A - N_D}{2N_Ae^{\Delta V_{BE}/V_T}}}\right) \\ x_p &= 2a_{1,p}\,\text{acos}\left(\sqrt{\frac{\sqrt{4N_AN_De^{\Delta V_{BE}/V_T} + (N_D - N_A)^2} + N_D - N_A}{2N_De^{\Delta V_{BE}/V_T}}}\right) \end{aligned} \tag{6.45}$$

where $\Delta V_{BE} = V_{BE} - V_{bi}$, and $V_T = kT/q$.

The accumulation regime solution of eqn (6.45) is certainly much more complex than eqn (5.9) for the depletion regime. However, the accumulation regime shares many similarities with the depletion regime. In fact, when V_{BE} is within the immediate neighbourhood of V_{bi} (*i.e.*, small ΔV_{BE}), then a Taylor expansion of eqn (6.45) about the point $\Delta V_{BE} = 0$ yields exactly the same equations for x_n and x_p that is obtained from the depletion regime. Further examination of eqn (6.45), however, shows that as ΔV_{BE} increases, x_n and x_p quickly saturate at a constant value of $\pi a_{1,n}$ and $\pi a_{1,p}$ respectively. This saturation of the SCR width is a feature of the rapid accumulation of mobile charge that screens out the applied bias with essentially no further increase to the extent of the SCR. This result is also the point at which the assumption of a non-degenerate accumulation layer will fail; so care must be exercised in the absolute application of eqn (6.45) for large ΔV_{BE}.

A useful metric from the depletion regime was the ratio of x_n to the total SCR width $x_n + x_p$. Due to the complex nature of x_n and x_p in the accumulation regime, this same metric will not be a simple constant. However, by appealing to a Taylor expansion about $\Delta V_{BE} = 0$, and the asymptotic limit for large ΔV_{BE}, it is found that:

$$N_{rat} = \frac{x_n}{x_n + x_p} \approx \begin{cases} \dfrac{N_A}{N_A + N_D} + \dfrac{N_A N_D (N_D - N_A)}{3V_T (N_D + N_A)^3} \Delta V_{BE} & 0 \le \Delta V_{BE} \le V_{knee,D} \\ \dfrac{\sqrt{N_A}}{\sqrt{N_A} + \sqrt{N_D}} & \Delta V_{BE} > V_{knee,D} \end{cases} \tag{6.46}$$

where

$$V_{knee,D} = \frac{3V_T (N_D + N_A)^2}{(\sqrt{N_A N_D} + N_A)\,(\sqrt{N_A N_D} + N_D)}.$$

In a similar fashion, the metric for the splitting of ΔV_{BE} between the n- and p-sides of the junction yields:

$$V_{rat} = \frac{|\psi_n(0)|}{q\Delta V_{BE}} \approx \begin{cases} \dfrac{N_A}{N_A + N_D} + \dfrac{N_A N_D (N_D - N_A)}{2V_T (N_D + N_A)^3} \Delta V_{BE} & 0 \le \Delta V_{BE} \le V_{knee,V} \\ \dfrac{1}{2} & \Delta V_{BE} > V_{knee,V} \end{cases} \tag{6.47}$$

where

$$V_{knee,V} = \frac{V_T (N_D + N_A)^2}{N_A N_D}.$$

N_{rat} in eqn (6.46), as was stated a few paragraphs earlier, shares many of the same features as N_{rat} in eqn (5.9) under the depletion regime. Now, V_{rat} in the depletion regime is exactly the same as N_{rat}, owing to the spatial uniformity of the space charge due to the immobile dopant ions. However, under the accumulation regime, V_{rat} in eqn (6.47) starts out the same as N_{rat}, but due to the mobile nature of the accumulation space charge, quickly results in an equal portioning of the excess applied potential ΔV_{BE} between the n- and p-sides of the junction. Therefore, the potential distribution in the accumulation regime differs markedly from what is found in the depletion regime. Finally, Fig. 6.15 plots N_{rat} and V_{rat} in both exact and approximate form, as well as x_n and x_p, in order to gain a familiarity with the accumulation regime.

Eqns (6.46) and (6.47) provide very useful tools for the solution of charge transport within the accumulation regime. Fig. 6.14c shows that within the accumulation regime, the positive step potential has produced a CBS; but unlike the negative step potential within the depletion regime (see Fig. 4.2), the CBS now appears on the other side of the heterojunction. Taking the standard HBT case where $N_A \gg N_D$, then $x_p \ll x_n$ and for small ΔV_{BE} one also finds $\psi_p(0) \ll \psi_n(0)$. These

two findings mean that the CBS within the accumulation regime will be very narrow, and very weak in terms of a potential to be tunneled through. Strictly speaking, the transport current through the CBS in the accumulation regime requires that the general transport model of eqns (4.51) and (4.53) be solved using $W_{CBS} = 1$, and W_N obtained from eqn (4.6) with the accumulation potential of eqn (6.42). However, with the parameters used in Fig. 6.15, when ΔV_{BE} = 120mV, then the CBS stands only 28meV tall, and 17Å wide at the base. Clearly, this small CBS will allow a significant current to pass though it. In any event, the largest that the CBS barrier could be, by assuming $W_N = 0$, would be an energy of $|\Delta E_c| - q\Delta V_{BE} V_{rat}$; and the smallest that the CBS barrier could be, assuming that $W_N = 1$, would be an energy of $|\Delta E_c| - q\Delta V_{BE}$ (see Fig. 6.16). Therefore, with $V_{rat} \approx 1$ for small ΔV_{BE} (given the typical HBT doping), then the upper and lower bounds for the effect of the CBS will be fairly close together.

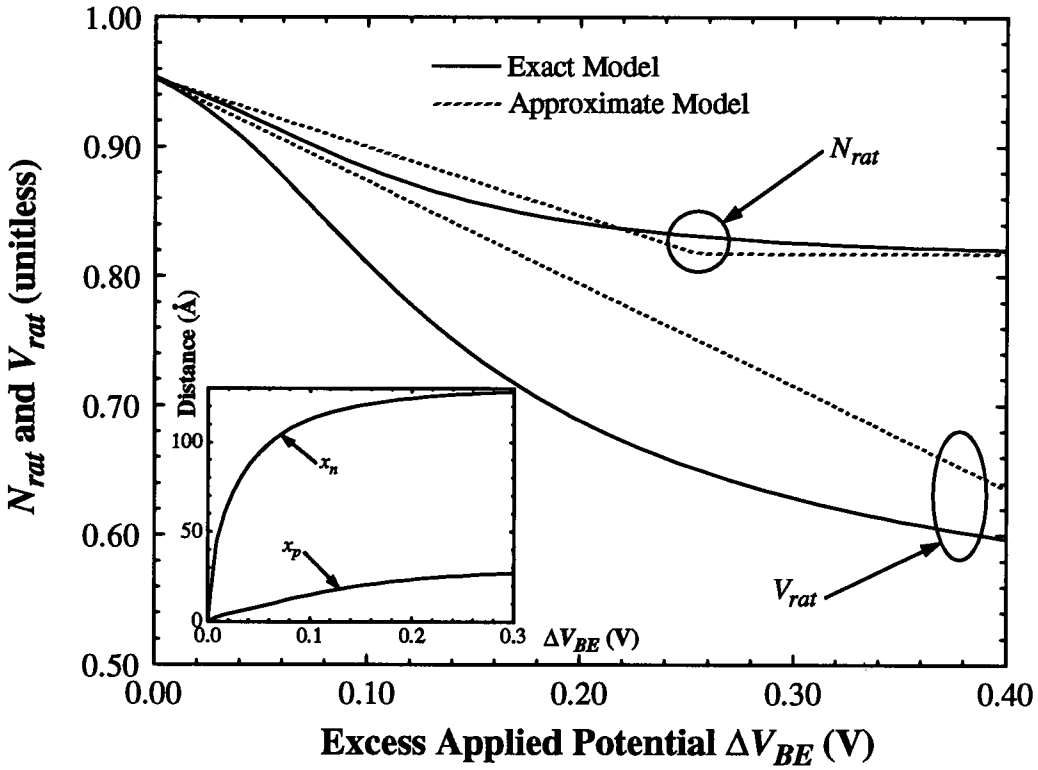


Fig. 6.15. The exact and approximate forms for N_{rat} and V_{rat} from eqns (6.46)-(6.47). The material parameters are: N_D: 5×10^{17}cm^{-3}; N_A: 1×10^{19}cm^{-3}; ε: 12.0.

One of the essential results of Chapter 4 was that the peak emission flux density occurred at a fixed energy relative to the height of the CBS. This result occurred only because of the parabolic nature of the potential profile within the depletion regime. Given the fairly simple model present-

ed for the accumulation regime, where degenerate effects have not been accounted for, there is little point in solving the general transport models of Chapter 4. Instead, based on the arguments of the previous paragraph, it seems reasonable to characterise the accumulation CBS by an effective energy height. Finally, only thermionic emission over this effective CBS will be considered. Given the result from Chapter 4 that was mentioned at the start of this paragraph, the effective height of the CBS is given by:

$$\begin{aligned} E_{\mathrm{CBS}} &= |\Delta E_c| - \mathrm{q}\Delta V_{BE} + \mathrm{q}\,(1 - V_{rat})\,U_{\max}\Delta V_{BE} \\ &= |\Delta E_c| - \mathrm{q}\Delta V_{BE}\,(1 - U_{\max} + U_{\max}V_{rat}) \end{aligned} \tag{6.48}$$

where $0 \leq U_{\max} \leq 1$, and $U_{\max}$ will be taken as a phenomenological constant. Strictly, based upon the analysis of Chapter 4, $U_{\max}$ will have a temperature dependence. However, as a first approximation, $U_{\max}$ can be taken as a constant independent of temperature. Then, the transport current under the accumulation regime is simply given by the thermionic term from eqn (4.79) as:

$$F_{f,s} = F_{f,s0}V_t \mathrm{e}^{-\frac{E_{\mathrm{CBS}}}{\mathrm{k}T}}, \tag{6.49}$$

where both sub-bands within the valence and conduction bands need to be considered in the case of the $Si_{1-x}Ge_x$ material system.

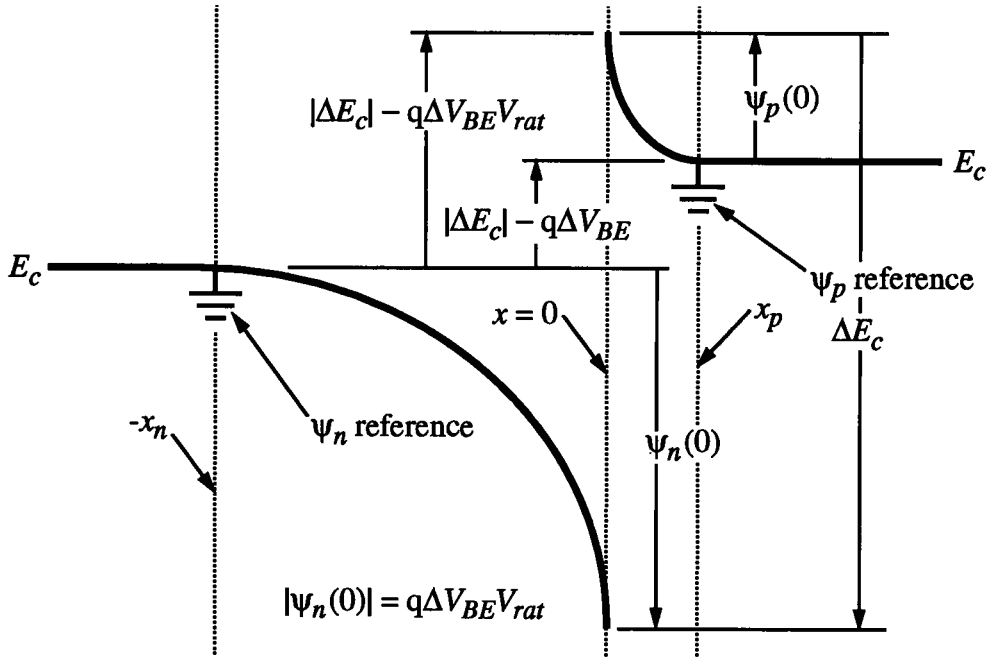


Fig. 6.16. Diagram of the CBS that forms under the accumulation regime. Only the conduction band is shown, but a similar structure can occur in the valence band. Note: this is for one sub-band.

6.5 Conventional and Novel $Si_{1-x}Ge_x$ HBT Structures

The $Si_{1-x}Ge_x$ material system represents a further step on the road to bandgap engineering. Unlike the $Al_xGa_{1-x}As$ material system, the $Si_{1-x}Ge_x$ material system allows one to essentially manipulate ΔE_g and ΔE_c (and thereby ΔE_v) independently. This independence between ΔE_g and ΔE_c is achieved through two independent parameters: 1) the Ge mole fraction x_a in the pseudomorphic strained alloy layer; 2) the amount of compressive or tensile strain applied to the pseudomorphic alloy layer by the substrate (*i.e.*, the substrate Ge mole fraction x_s). The addition of strain is the key to the rich possibilities regarding bandgap Engineering offered by the $Si_{1-x}Ge_x$ material system. Sections 6.1 through 6.4 have set out the various material models and transport models to study the flow of charge within a SiGe HBT. This section will apply the results of these previous sections to the study of current-day SiGe HBTs structures, as well as some other novel structures.

The study of highly strained pseudomorphic layers cannot be properly performed without consideration of the critical layer thickness h_c. As was stated early on in this chapter, the potential strain in the $Si_{1-x}Ge_x$ material system can be quite large, owing to the 4.2% lattice mismatch between Si and Ge. As the in-plane strain is increased (see Fig. 6.3), the maximum thickness of the alloy layer decreases in an essentially exponential fashion. The determination of h_c has been the focus of numerous studies and controversies [97,99,105]. At present, there is still debate as to the exact model for h_c versus in-plane alloy strain, but the work of People [105] is at least a reasonable reference point. In [105], the critical layer thickness is given as:

$$h_c = \left(\frac{1-\nu}{1+\nu}\right)\left(\frac{1}{20\pi\sqrt{2}}\right)\left(\frac{b^2}{a_a}\right)\left(\frac{1}{f^2}\ln\left(\frac{h_c}{b}\right)\right) \tag{6.50}$$

where h_c is in Å, $b = 4$Å (the magnitude of the Burger's vector), ν is the Poisson ratio from eqn (6.5), a_a is the unstrained (bulk) alloy lattice constant from eqn (6.6), and f is the alloy strain given by $(a_a - a_s)/a_s$ (where a_s is the substrate lattice constant). Substituting all of these parameters into eqn (6.50) gives:

$$h_c = \frac{1.928}{(5.43+0.23a_a)}\left(\frac{5.43+0.23a_s}{a_a-a_s}\right)^2 \ln\left(\frac{h_c}{4}\right). \tag{6.51}$$

Eqn (6.51) is an implicit phenomenological equation that People has fit to the best available data for h_c (see Fig. 6.17). Detailed information, such as what temperature and duration can a pseudomorphic layer tolerate before relaxing is still not conclusively known.

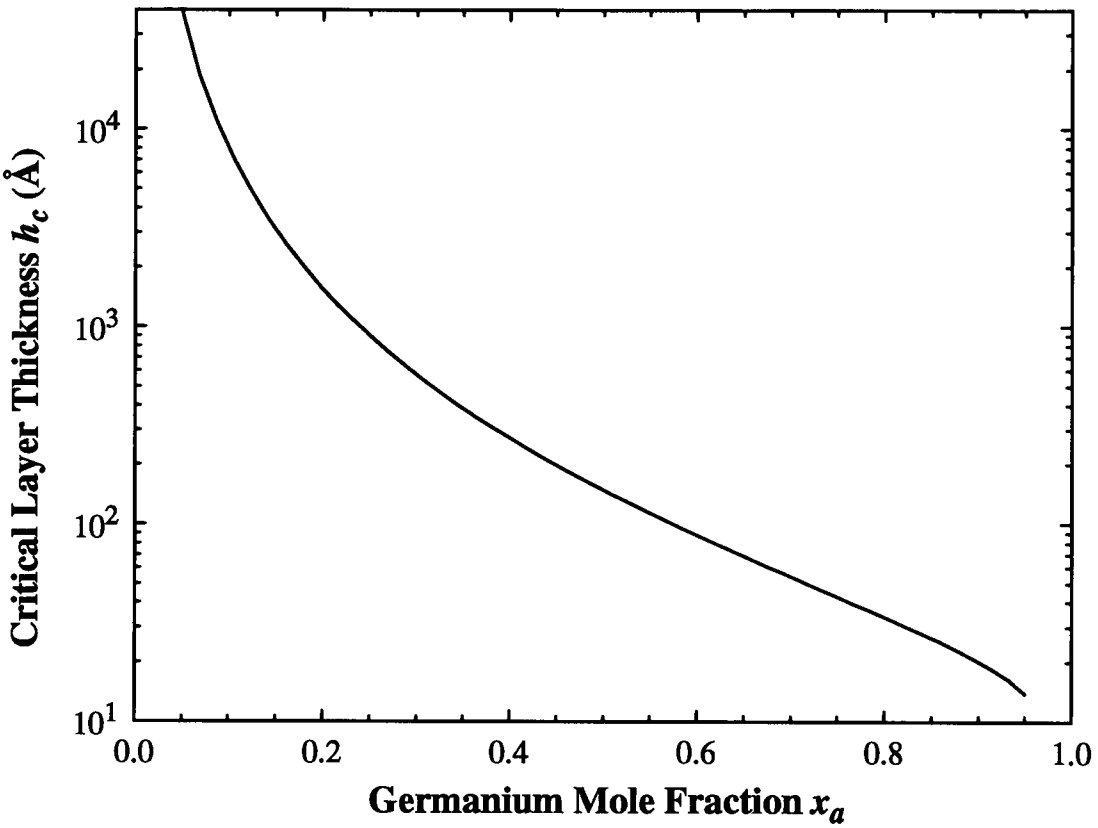
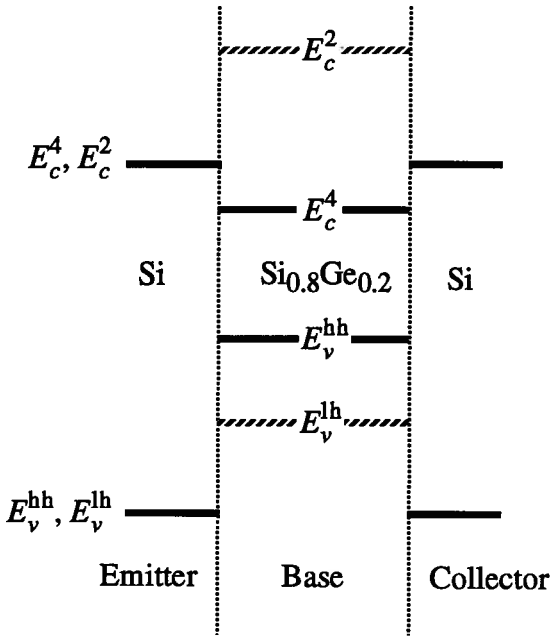


Fig. 6.17. Critical layer thickness for a $Si_{1-x_a}Ge_{x_a}$ layer on a {100} Si substrate. If the substrate is $Si_{1-x_s}Ge_{x_s}$ instead, then a good approximation is to find $|x_a - x_s|$ and use this on the above plot.

Current-day SiGe HBTs, of which [100-103] are examples, have all been based on a substrate that is {100} Si. The emitter and collector regions are pure Si, and the base is the only region made up of $Si_{1-x_a}Ge_{x_a}$. The essential premise for this type of SiGe HBT stems directly from the early work of Kroemer [2,46,47] and Shockley [1] who called for a wide-bandgap emitter injecting into a narrow-bandgap base. Within this physical construct, the Ge alloy content x_a in the base is either fixed at some constant value, or a drift field is created in the base by increasing x_a as one proceeds from the emitter towards the base.

Starting with a constant x_a in the base of 0.2, then eqn (6.51) gives h_c = 1550Å. Because the HBT is lattice matched to a pure Si substrate, all regions of the device except the base have E_c^4 and E_c^2 degenerate, as well as E_v^{hh} and E_v^{lh} degenerate. However, compressive strain in the base produces $E_c^{4,2}$ = -138meV, meaning that the ultimate conduction band in the base is E_c^4-like. Further, compressive strain in the base makes the ultimate valence band E_v^{hh}-like, with $E_v^{hh,lh}$ = 34meV. Fig. 6.18 presents the band diagram for the above device, with the relevant material parameters noted. Observation of Fig. 6.18 clearly shows that electron transport will occur via E_c^4. Since ΔE_c^2

= -100meV, while ΔE_c^4 is 37meV, ostensibly all of the electrons contained by the E_c^2 band in the emitter (which is 33% of the total number of majority electrons) will be reflected by ΔE_c^2 and not contribute to electron transport. Thus, if the EB SCR determines the transport current, then after including the different effective masses, I_C would be 18% less than expected from a simple examination of the device that does not account for the independence of E_c^4 and E_c^2. However, if the neutral base region determines the transport current, then I_C would be larger than expected given that $\overline{D}_n^4$ is higher than the bulk value. In order to determine if it is the EB SCR or the neutral base that is responsible for current-limited-flow, the detailed construction of the device must be considered. For the devices in [100-102], where $N_D \gg N_A$, then the neutral base is narrowly responsible for current-limited-flow; although, inclusion of bandgap narrowing effects could lead to the EB SCR being responsible for current-limited-flow. However, for the devices in [10,134], where $N_D \ll N_A$, then depending on how bandgap narrowing in the base splits between E_c and E_v, the EB SCR will be responsible for current-limited-flow; resulting in a much smaller increase to I_C than would be expected from neutral base transport considerations alone. This analysis of current-day SiGe HBTs shows that a failure to correctly model both E_c^4 and E_c^2, including EB SCR limitations, could lead to an incorrect understanding of transport within the device.



$\Delta E_c^4 = 37\text{meV}$ $\quad \Delta E_v^{hh} = -138\text{meV}$

$\Delta E_c^2 = -100\text{meV}$ $\quad \Delta E_v^{lh} = -104\text{meV}$

$\Delta E_c = 37\text{meV}$ $\quad \Delta E_v = -138\text{meV}$

$E_{g,e} = 1120\text{meV}$ $\quad n_{i,e}^2 = 6.94\times10^9\text{cm}^{-3}$

$E_{g,b} = 945\text{meV}$ $\quad n_{i,b}^2 = 1.47\times10^{11}\text{cm}^{-3}$

Without EB SCR limitations, E_c^2 will transport 0.25% of the current in the neutral base, leaving E_c^4 to transport the remaining 99.75% of the current.

Fig. 6.18. Band diagram for an HBT with 20% Ge in the base, lattice matched to Si. The base is the reference. The effect of the EB and CB SCR potential is not shown for clarity.

For SiGe HBTs, where the emitter and base are E_c^4-like, ΔE_c is too small to produce a CBS (see Fig. 6.10). Therefore, unlike AlGaAs HBTs, when the EB SCR limits the transport current in SiGe HBTs, then log I_C versus V_{BE} will look identical to the case where the neutral base limits the transport current (*i.e.*, the injection index will be unity). Thus, there will be no overt tell-tale sign in SiGe HBTs that the transport current is not being controlled by the neutral base. However, I_C will indeed be smaller than expected due to the EB SCR limitation, plus, the Early voltage should become theoretically infinite as basewidth modulation should no longer effect I_C [135].

The SiGe HBT where x_a is varied across the base represents the device that has piqued the interests of the semiconductor community. By generating an aiding field in the base through a monotonically increasing x_a from the emitter to the base (and hence a decreasing E_g), an f_T as high as 113GHz has been obtained [102]. In order to achieve this remarkable metric the device was fabricated with as large a Δx_a in the base as possible; minimising the base transit time. To this end, x_a was 0 at the emitter and was linearly ramped up to 0.25 at the collector. The result is a band diagram as depicted in Fig. 6.19a. Since the neutral base closest to the emitter is pure Si, then one has essentially a homojunction for the EB SCR, and it is expected that the neutral base will limit the transport current (see Fig. 6.19b). The base region, given the shape of the E_c^4 and E_c^2 bands, produces a demanded current that differs between the sub-bands by a factor of 8.3; *i.e.*, the current in E_c^4 will be 8.3-fold larger than E_c^2. This is not an overwhelming amount, which shows that 11% of the collector current is carried by the slower E_c^2 band. In fact, using eqn (3.8) shows that τ_B for E_c^4 is reduced 4.6-fold compared to τ_{B0}, while τ_B for E_c^2 is reduced only 1.5-fold compared to τ_{B0} (where τ_{B0} is the τ_B given in eqn (3.6)). Assuming that the final base transit time is given by the average of the results from each band weighted with the relative current carried by the band, then the effective reduction to τ_B compared to τ_{B0} is $(0.89/4.6 + 0.11/1.5)^{-1} = 3.8$-fold. If the two sub-bands were considered as one single band then τ_B would have been wrongly reduced 4.4-fold relative to τ_{B0}, and I_C overestimated by 13%. In the above calculations the effect of bandgap narrowing has not been accounted for. Inclusion of base bandgap narrowing could cause the EB SCR to limit the transport current (again, depending on how the bandgap narrowing splits between E_c and E_v), which would greatly effect the current partitioning between the conduction sub-bands. Furthermore, the anisotropic nature of E_c^4 and E_c^2 has also not been accounted for, which would increase τ_B even further given that $\overline{D}_n^2$ would be greatly reduced.

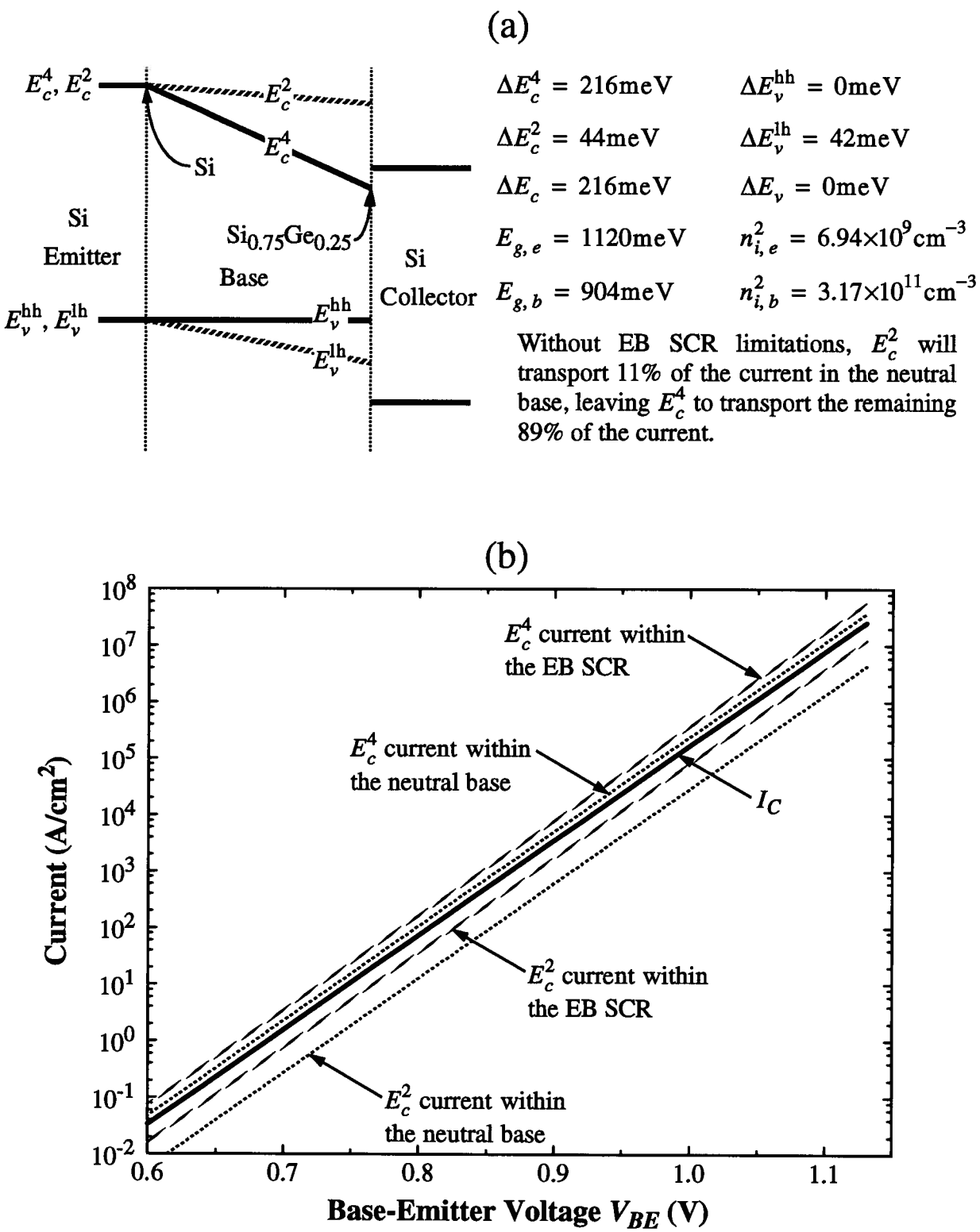


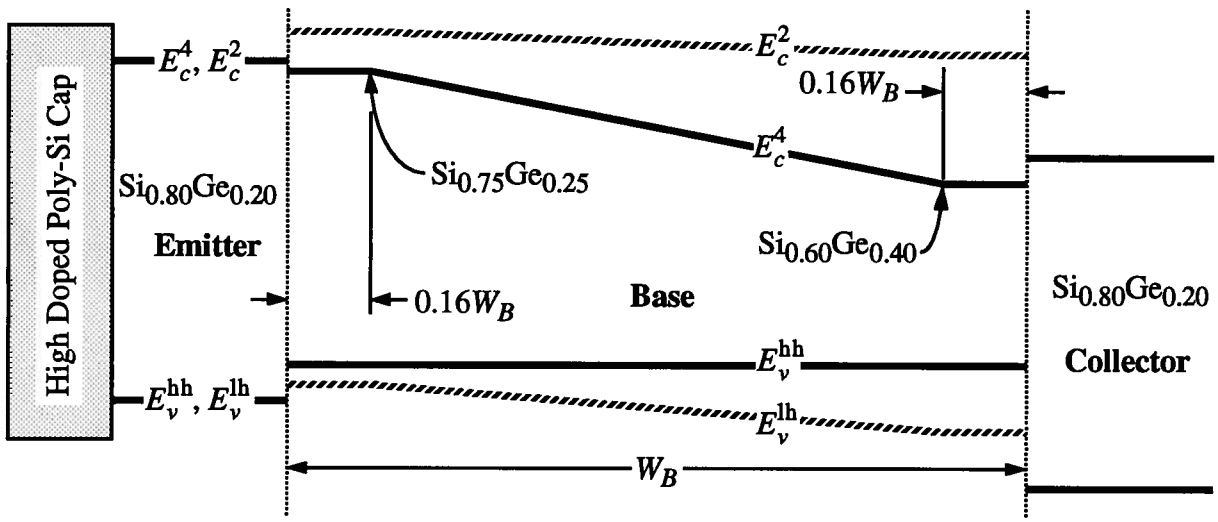
Fig. 6.19. (a) Band diagram for an HBT with 25% linear grading of Ge in the base, lattice matched to Si. ΔE is from the 25% Ge point in the base to the emitter. Note: the base bandgap has a slightly parabolic nature due to the Ge alloy effects. (b) Transport currents through the various regions of the HBT, including the collector current. N_A=5×10^{18}cm^{-3}, N_D=1×10^{20}cm^{-3}, and W_B=700Å. Given that E_c^4 transport within the EB SCR is not substantially larger than transport through the neutral base, I_C is subsequently 31% lower than expected from neutral base transport considerations alone. Thus, the neutral base is controlling I_C but the EB SCR does have an effect.

The previous analysis of conventional SiGe HBT structures is not intended to be exhaustive, but it clearly demonstrates that the $Si_{1-x}Ge_x$ material system cannot be characterised by an effective conduction band. In order to properly model a SiGe HBT, the rich nature of the E_c^4 and E_c^2 bands must be included via the models developed in Sections 6.1 to 6.4. Further, the assumption of Shockley boundary conditions (*i.e.*, that the EB SCR is not responsible for current-limited-flow) can come under question in the design of SiGe HBTs. Finally, the importance of considering transport through the entire device becomes even more important when optimisation, or the extraction of material parameters, is sought after: for if the transport is being dictated by a region other than the one being considered, the result will be a an erroneous conclusion regarding either the correct path for optimisation or the material parameter being extracted.

The main problem with the $Si_{1-x}Ge_x$ material system is that the band offsets tend to be quite small because of the limits imposed on the Ge content by the critical layer thickness. For this reason, it is still common to see $N_D \gg N_A$ in order to maintain a usable β. As the neutral base width is reduced, then N_A must increase in order to offset a rapid decrease in f_{max}. However, increasing N_A must be accompanied by an increase in N_D or the gain will drop. With N_D near the solid-solubility limit this is not really possible. Further, with N_A and N_D increasing, the EB capacitance will increase, and a tunnel diode could form. The device in [134] attempted to solve this with a constant 22% Ge base content. By having a narrow bandgap in the base, the subsequent increase to $n_{i,b}$ can be traded off for a higher base Gummel number. However, this precludes a graded base, as the EB heterojunction is required to maintain the gain, and the critical layer thickness will not allow for a higher Ge content (this is the alloy budget of Section 3.2). Therefore, in order to continue decreasing the neutral basewidth without compromising f_{max} or f_T, a way must be found to include higher Ge contents in the base.

The answer to the problem of the previous paragraph is to lattice match the HBT to a $Si_{1-x_s}Ge_{x_s}$ substrate, where $x_s > 0$. Consider a 500Å $Si_{0.8}Ge_{0.2}$ emitter with a poly-Si cap, a base graded from $Si_{0.75}Ge_{0.25}$ at the emitter to $Si_{0.6}Ge_{0.4}$ at the collector, all lattice matched to a $Si_{0.8}Ge_{0.2}$ collector and substrate (see Fig. 6.20). The base grading is started at 25% Ge instead of 20% in order to increase the transport current in E_c^4 relative to E_c^2, thereby reducing the parasitic effect on τ_B found from the HBT in Fig. 6.19. Then, the 15% Ge base grading provides the aiding field to keep the base transit time small. However, unlike the HBT in Fig. 6.19, the optimum aug-

mented-linear doping of Fig. 3.8 is used instead of the sub-optimum linear grading. The optimum base profile, due to the constant Ge regions near the emitter and the base, also increases the Early voltage and decreases the anomalous change to I_C due to the reverse Early voltage effect [11]. The 500Å $Si_{0.8}Ge_{0.2}$ emitter next to the base ensures that the EB SCR will be free of dislocations that will occur at the boundary to the poly-Si cap; plus it serves as an efficient source of E_c^4 electrons. Finally, the poly-Si emitter cap provides stress relief to the system and a wide bandgap to kill the back injection of holes. With the wide bandgap of the poly-Si cap controlling the gain, N_A can be significantly increased in order to increase f_{max}, while N_D can be decreased in order to decrease the EB SCR capacitance. The result is a 264-fold increase in I_C compared to a similar bulk Si device, with τ_B reduced 2.9-fold compared to τ_{B0}. These results are based upon the neutral base controlling I_C. As N_A is increased to the point where bandgap narrowing becomes quite large, it is expected that the EB SCR will dictate I_C and limit the expected increase to β.



$\Delta E_c^4 = 120\text{meV}$ $\Delta E_v^{hh} = -34\text{meV}$ $n_{i,e}^2 = 3.68\times10^{10}\text{cm}^{-3}$ $n_{i,b}^2 = 5.05\times10^{11}\text{cm}^{-3}$

$\Delta E_c^2 = -18\text{meV}$ $\Delta E_v^{lh} = 9\text{meV}$ $E_{g,e} = 990\text{meV}$

$\Delta E_c = 120\text{meV}$ $\Delta E_v = -34\text{meV}$ $E_{g,b} = 876\text{meV}$

Without EB SCR limitations, E_c^2 will transport 4.5% of the current in the neutral base, leaving E_c^4 to transport the remaining 95.5% of the current.

Fig. 6.20. Novel SiGe HBT based on a 20% Ge substrate. The incorporation of the optimum base grading provides the maximum reduction to τ_B possible. The poly-Si emitter cap provides the wide bandgap necessary to control hole back injection, while lattice matching to a 20% Ge substrate allows a 40% Ge content in the base without being restricted by h_c. ΔE is from the 40% Ge point in the base to the emitter.

The operation of the novel transistor being proposed rests on two requirements: 1) that high quality $Si_{1-x_s}Ge_{x_s}$ substrates can be formed; 2) that the poly-Si cap will indeed control the back injection of holes. The ability to grow high quality $Si_{1-x_s}Ge_{x_s}$ substrates is currently an issue. At present, bulk epitaxial $Si_{1-x_s}Ge_{x_s}$ layers on top of Si substrates have defect densities ranging from $10^4 cm^{-2}$ to $10^6 cm^{-2}$ [31]. This is too high to produce commercially yielding LSI ICs. However, given the infancy of epitaxially growing bulk $Si_{1-x_s}Ge_{x_s}$ layers on Si, in time it is expected that the process will mature and the defect density will fall. The other option is to pull raw $Si_{1-x_s}Ge_{x_s}$ ingots so that the starting wafer contains the desired substrate. In either case, for the study being presented here, it is sufficient to demonstrate the usefulness of using non-Si substrates in order to provide the impetus to grow low defect bulk $Si_{1-x_s}Ge_{x_s}$ substrates on Si. The second question, regarding the efficacy of the poly-Si cap to control hole back injection, can only be answered by experimentation. However, recent work by Kondo et. al. [136,137] for poly-Si to Si shows that the interface is not characterised by a high recombination velocity, and that the bandgap is, if anything, larger than in bulk Si. Thus, n_i in the poly-Si layer will be small compared to the n_i in the base, controlling the back injection of holes and β. Finally, the band alignment of the poly-Si layer to the $Si_{0.8}Ge_{0.2}$ emitter will only be an issue if the resulting ΔE_c is large enough to limit the electron transport current through the entire device. Based upon Si lattice matched to $Si_{0.8}Ge_{0.2}$, ΔE_c should not exceed -90meV, which would not reduce the transport current given the high doping that would exist in the poly-Si layer. Therefore, it is expected that the poly-Si cap will control the hole back injection of the proposed SiGe HBT.

This section concludes by examining an intriguing HBT structure that invokes all of the models of this chapter. Beginning with Fig. 6.9c for $x_{al} = 0$ and $x_{ar} = 0.45$, examination of substrates where $0 \le x_s \le 0.35$ is very interesting. Let the left side be the emitter and the right side the base. The emitter is under tensile strain so that the ultimate conduction band is E_c^2-like. Contrarily, with the substrate range being considered, the base is under compressive strain and the ultimate conduction band is E_c^4-like. Just because the emitter conduction band is E_c^2 does not preclude electrons from existing in E_c^4. In fact, given the band alignments for $0 \le x_s \le 0.35$, more electrons from E_c^4, rather than E_c^2, will be able to go from the emitter into the base. Essentially, the band with the lowest energy in both the emitter and the base will be the one that transports the current. With $x_s \le 0.35$, E_c^4 will be responsible for current transport as E_c^2 in the base is larger than E_c^4 in the emitter.

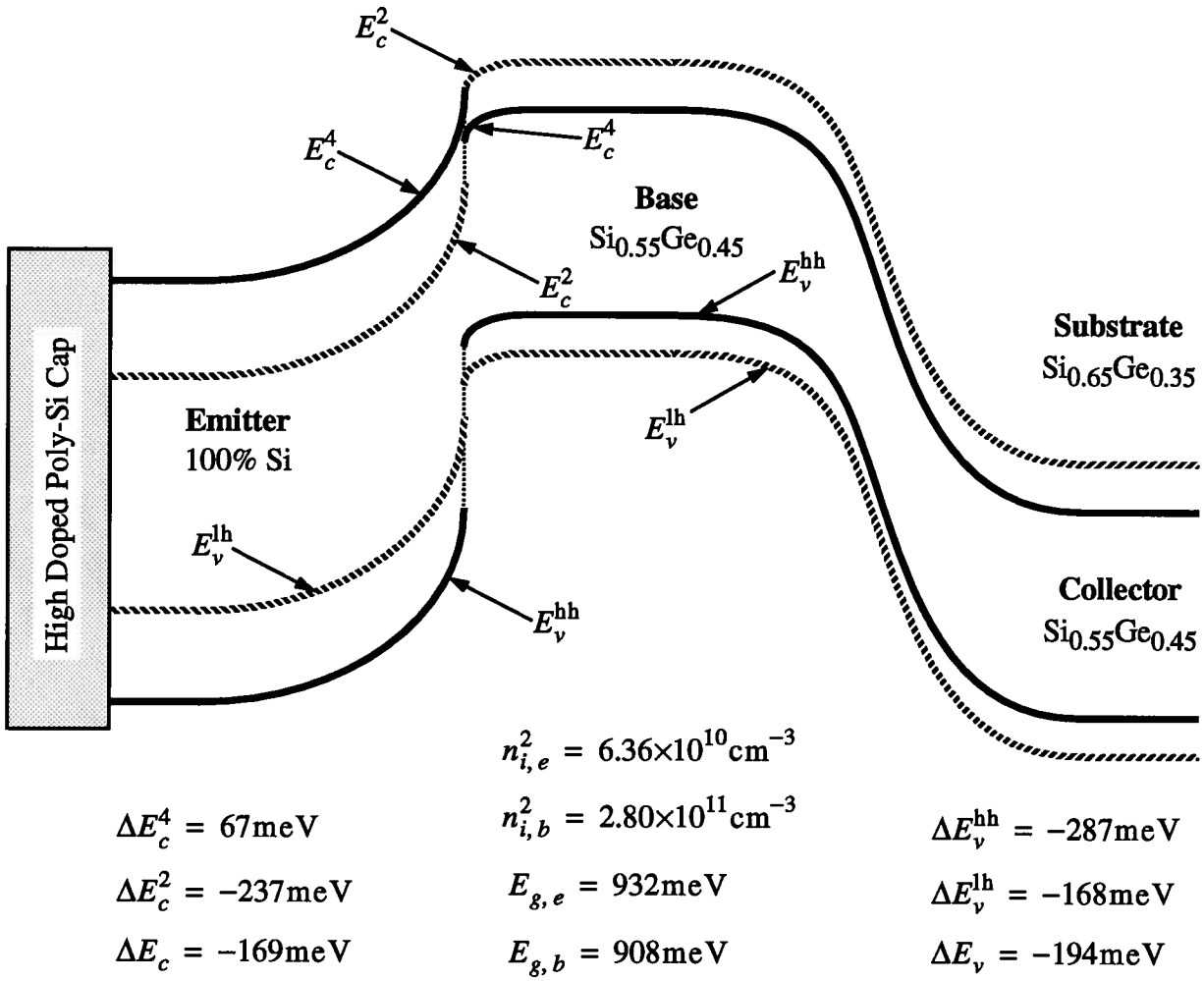


Fig. 6.21. Band diagram showing the conduction and valence sub-bands for an HBT where x_{al} = 0, x_{ar} = 0.45, x_s = 0.35, N_A=1×10^{19}cm^{-3}, N_D=5×10^{17}cm^{-3}, and W_b=700Å.

Fig. 6.21 plots the band diagram, including SCR effects, for an HBT where x_{al} = 0, x_{ar} = 0.45, x_s = 0.35, N_A=1×10^{19}cm^{-3}, N_D=5×10^{17}cm^{-3}, and W_b=700Å. As is the case for the HBT in Fig. 6.20, there is a high doped poly-Si cap on top of the emitter to provide stress relief and control the back injection of holes. What is interesting to note for the device in Fig. 6.21 is the emitter and base have essentially the same bandgap. Thus, there is no wide-gap emitter injecting into a narrow-gap base that is common to traditional HBT designs. Instead, the HBT is controlled by the band offsets and n_i for the given sub-band within the neutral regions. Fig. 6.22 plots the EB SCR currents, the neutral base transport currents, and the final collector current that will occur within the device of Fig. 6.21. It is important to realise that V_{bi} = 0.673V due to the positive ΔE_c of this device. For $V_{BE} < V_{bi}$ transport occurs via E_c^4 through a small CBS, but with neutral base transport essentially controlling I_C. Thus, electron transport within the emitter is occurring in a band

that does not form the ultimate conduction band. Now, when $V_{BE} > V_{bi}$, the HBT is operating within the accumulation regime. Due to ΔE_c = -169meV, EB SCR transport within E_c^4 is reduced to only 10^2A/cm^2 when $V_{BE} = V_{bi}$. Furthermore, because ΔE_c^4 = 67meV, any increase in V_{BE} past V_{bi} will do nothing to increase the EB SCR current as there is no barrier to surmount, leaving only the thermal movement of majority carriers to dictate the current. Thus, E_c^4 transport is now controlled by the EB SCR and not the neutral base. However, with the accumulation model of Section 6.4, E_c^2 transport becomes the dominant path that controls I_C when V_{BE} increases past V_{bi}; leading to transport in the base that occurs within a band that does not form the ultimate conduction band. The final result is a very interesting log I_C versus V_{BE} relationship that is due to the interaction between the two conduction sub-bands.

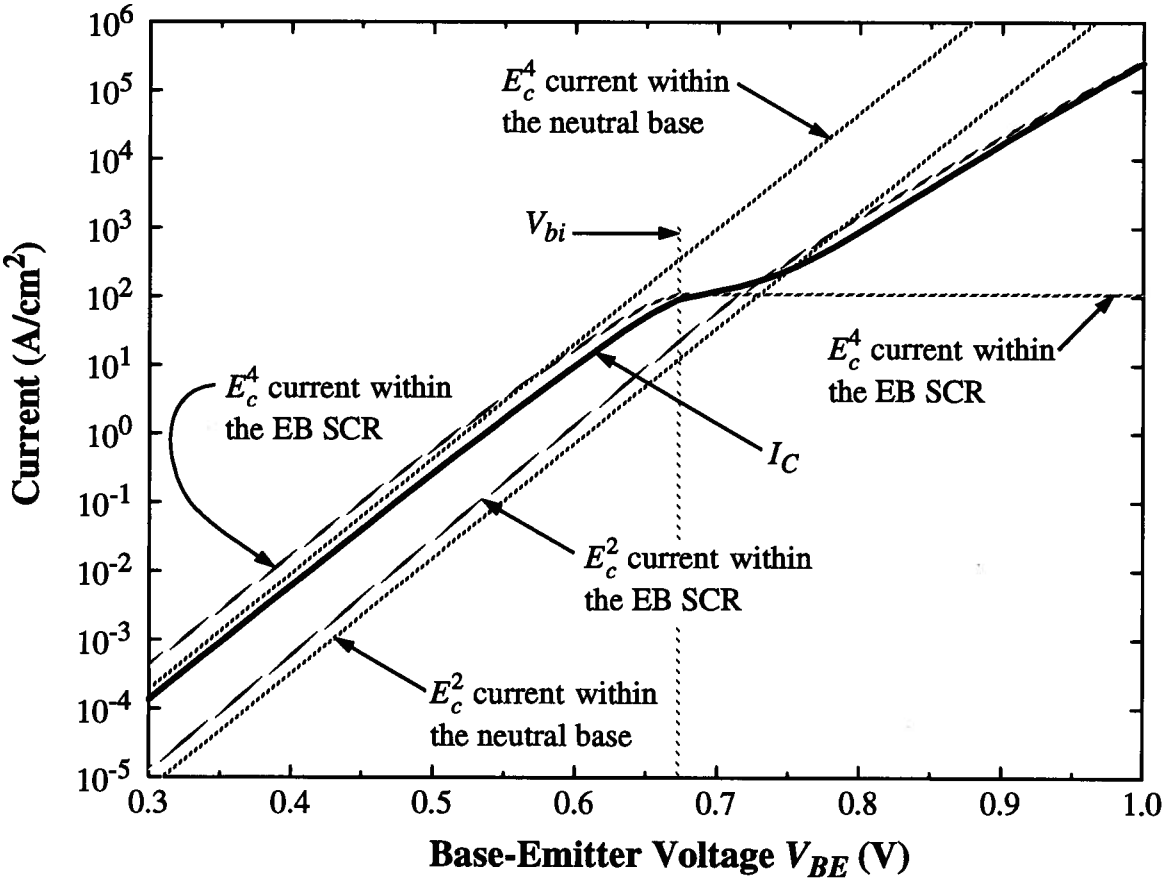


Fig. 6.22. Transport currents within the various regions of the HBT given in Fig. 6.21.

The HBT of Fig. 6.21 may have some practical uses as a current source due to its flat I_C versus V_{BE} relationship near V_{bi}; however, it is probably more useful as a tool to investigate transport

properties and band offsets within the $Si_{1-x}Ge_x$ material system. Careful consideration of h_c for this HBT reveals some interesting results. Because the Ge content in the base is only 10% higher than in the substrate, then h_c = 8054 Å. With such a large h_c it is conceivable that the base and intrinsic collector regions could be formed without a heterojunction, thereby achieving an SHBT instead of the DHBT common to SiGe devices. Furthermore, it is not unreasonable to imagine that the base and intrinsic collector could be formed in only 3000 to 5000 Å, leaving considerable room to the maximum h_c, which should help to increase the thermal budget for the base layer. The issue of DHBT devices is not a real concern in npn HBTs, due to the small ΔE_c, but would be of considerable appeal in making a pnp device. Finally, the result of a large h_c for the base and collector regions is a significant lowering of the emitter h_c to 407 Å. However, an h_c = 407 Å would be wide enough to contain the emitter extent of the EB SCR. Therefore, the critical layer thickness has been moved from out of the EB-SCR and into the neutral emitter, which will have less of an effect on device performance if dislocations due to strain relaxation occur.

In conclusion to this chapter the following results regarding the $Si_{1-x}Ge_x$ material system have been achieved:

- A review of the literature, including the best material models, for the effect of strain on the conduction and valence sub-bands has been performed.
- The band offset theory of Van de Walle and Martin, including the material models of Yu and Gan et. al., have been reviewed with the most consistent set of material parameters chosen to fit the experimental data available to date. To this end, a simple set of equations has been found to accurately describe the conduction band.
- A theory regarding transport within the conduction E_c^4 and E_c^2 bands, and the valence E_v^{hh} and E_v^{lh} bands has been developed. The theory presented does not resort to an effective conduction and valence band, but considers carrier transport within both sub-bands. Included in this development is the full effective density of states and the intrinsic carrier concentration for all of the sub-bands.
- A theory for the operation of an HBT past the built-in potential has been developed.
- Finally, the models of this chapter, which are based upon the models of all the previous chapters, have been used to study current-day SiGe HBTs and a few other novel structures. The

most important result of this study is that the neutral base will no longer be the sole region controlling I_C as the neutral base width continues to shrink and the Ge grading in the base increases: the limitations of EB SCR transport must be considered. Furthermore, there is a significant error in both the calculation of charge flow and transit time by considering the sub-bands as a single effective conduction or valence band.

CHAPTER 7

Summary and Future Work

To begin with, Chapter 2 has presented a unique and general model (eqns (2.7) and (2.9)) for the simulation of HBTs. This model forms the framework for simulating charge transport within the entire HBT by providing a means to break the modelling effort into separate physical regions; each region characterised by its own unique physical transport process. Furthermore, the model presented in Chapter 2 allows for the existence of recombination sinks within each region; furthering the general nature of the model. Due to the abstract nature of eqns (2.7) and (2.9), it is possible to apply the model of Chapter 2 both to the microscopic transport of charge (*i.e.*, to transport over atomic distances), and to the macroscopic transport of charge (*i.e.*, to transport over distances large enough to treat the electrons as a continuous flux, such as is done in drift-diffusion analysis). In so doing it may be possible to determine the point at which rapid spatial changes in the conduction or valence bands produce transport conditions that deviate from the models of drift and diffusion (such as can occur within an SCR, and certainly at the heterojunction where ΔE_c forms). This may allow for a solution to a question posed by Dr. Mike Jackson of UBC as to the condition for which thermionic injection begins and drift-diffusion ends. However, the most logical extension to the work of Chapter 2 is to remove the restriction that E_{fp} (for an npn HBT) be a constant throughout the EB SCR.

Chapter 3 presents some interesting ideas for optimising the metrics of an HBT by exploiting the concept of current-limited flow outside of the neutral base. It would be a reasonable extension to the ideas of Chapter 3 to simulate and measure a number of HBT designs that exploit the optimisations that have been alluded to. Chapter 3 has also gone on to determine the simultaneous optimisation of the base bandgap and the base doping profiles for the minimisation of τ_B. This work has, however, neglected the effect of a non-constant mobility with respect to doping variations. Numerical work [63] has shown that the optimum profiles which include the full $\mu_n(N_A)$ do not appear to be too complex, and certainly have a shape that is expected from consideration of the functional form of $\mu_n(N_A)$ itself. Therefore, it is expected that the analytic optimum profile, shown in Fig. 3.9, for the minimisation of τ_B could be extended to include either the full $\mu_n(N_A)$ or a judicious approximation to it.

Chapter 4 derives the model of charge transport with the EB SCR, including the effects of tunneling and momentum conservation across a mass boundary. To this end, the general models of eqns (4.50)-(4.53) were presented. Chapter 4 goes on to derive analytic approximation to eqns

(4.50)-(4.53). However, for the purpose of deriving analytic results, the mass boundary is considered in an isotropic fashion, but with the effective mass maintained as a diagonal tensor and not a simple scalar. Thus, a logical extension to the analytic work of Chapter 4 is to remove the assumption of an isotropic mass boundary and resolve eqns (4.50)-(4.53) in an analytic form.

Other extensions to the work of Chapter 4 are certainly alluded to in Section 4.6. By plotting the ensemble electron density entering the neutral base of the HBT, it was clear that the distribution could not be considered as a Maxwellian or even a hemi-Maxwellian. These distortions from a hemi-Maxwellian form are due to the effect of tunneling though the CBS. Since accurate simulation of transport through a narrow base (in terms of mean free path [43]) demands a full solution to the BTE, then a way must be found to incorporate the non-local effect of tunneling into the BTE. A possible extension to the work of Chapter 4 is to connect the EB SCR transport models of the chapter to a BTE solver for the neutral base; thereby allowing for the inclusion of tunneling within the BTE via a hybrid model.

The modelling of charge transport in Chapter 4, due to tunneling through the CBS contained within the EB SCR, is formulated upon ballistic considerations. It is common to consider tunneling electrons in a ballistic fashion, if for no other reason than to simplify the calculation of the tunneling probabilities. This position of neglecting thermalising collisions of the electron while undergoing tunneling is often substantiated on the grounds that tunneling distances are generally less than 100 or 200Å, and are therefore significantly less than the mean free path. However, if any collisions did occur while the electron is in the midst of tunneling, then the tunneling probability would be essentially reduced to zero. Thus, a potential extension to the work of Chapter 4 is to consider non-ballistic tunneling. The ultimate outcome of such non-ballistic tunneling considerations would be the development of a Monte Carlo simulator that can incorporate non-local effects (*i.e.*, tunneling).

A final extension to the work of Chapter 4 can be found by careful observation of Fig. 4.9 and eqn (4.74). Φ_{max} occurs at U'_{max}, which for a fixed temperature is a constant. Furthermore, the flux density $\Phi_{f,s}$ is fairly well centred about U'_{max}, and will become even more localised as the temperature is reduced. Therefore, the tunneling current through the CBS can be thought of as occurring at an energy of $qU'_{max}(V_{bi} - V_{BE})N_{rat}$ relative to the conduction band minimum in the emitter. Now, the tunneling current is very sensitive to the forward-directed effective mass, which

is dependent upon the full nature of the dispersion relation $E_c(k)$. Then, with the CBS responsible for controlling I_C, by measuring I_C the tunneling current through the CBS can be determined. Finally, by extracting the effective mass through a matching of the measured I_C to the tunneling models of Chapter 4, it should be possible to infer $E_c(k)$. Therefore, it should be possible to extend the work of Chapter 4 by developing an electrical spectroscopy method for the determination of $E_c(k)$.

Chapter 5 presents the models for the recombination currents that occur within both the EB SCR and the neutral base. Specifically, the need to balance the total current entering a region with the net current leaving plus any charge that has recombined within the region, is considered. This leads to a mixing of the base and collector currents of an HBT. The result of this mixing is a new connection between the physical construction of the HBT and it's terminal characteristics. Regarding future work, the basis for all of the recombination models (SRH, Auger, and radiative) used within Chapter 5 is essentially drift-diffusion. By the arguments of Chapter 4, drift-diffusion analysis is not applicable within the EB SCR. Therefore, combined with the extension being proposed for Chapter 4 (regarding integration with the BTE), the recombination currents should be recomputed from a particle scattering cross-section point of view. This would place the calculation of the recombination currents on par with the quantum mechanical view of a tunneling electron.

Chapter 6 reviews the various material models that are required to understand the composition of the conduction and valence bands within pseudomorphically strained $Si_{1-x}Ge_x$. Further, the band offset models for the determination of ΔE_c and ΔE_v at an abrupt heterojunction are also presented. Using these material models, transport models which include the two conduction sub-bands E_c^4, E_c^2 and the two valence sub-bands E_v^{hh}, E_v^{lh}, are developed. It is shown that the multi-band nature of strained $Si_{1-x}Ge_x$ must be considered, even in present-day HBTs, lest considerable error regarding both the quantitative and qualitative aspects of charge transport be made. Regarding future work, it is imperative that a final and consistent set of material parameters for $Si_{1-x}Ge_x$ be obtained. Without a firm understanding of the material parameters, it is impossible to accurately determine the transport current. With this in mind, Chapter 6 presents a number of novel HBT structures, including a study of some present-day HBTs. In order to ascertain the validity of the models developed within Chapter 6, these SiGe HBTs should be fabricated and tested against these theories

Finally, Chapter 6 only considers substrates aligned to ⟨100⟩. However, there could be considerable performance gains for growth along ⟨111⟩. Traditionally, BJTs have used ⟨111⟩ aligned substrates because epitaxial growth is the fastest for this orientation. ⟨100⟩ aligned substrates have come about because of the need to minimise surface states at the Si/SiO_2 interface in MOSFETs. One of the most interesting features of strained $Si_{1-x}Ge_x$ is the possibility of only having charge transport occur parallel to the small transverse mass for electrons. The anisotropic nature of Si produces a 5-fold difference between the transverse and longitudinal mass for electrons. Thus, a significant improvement to tunneling and mobility can be had if the electrons predominantly move with the transverse mass. This would be further increased by using the ⟨111⟩ conduction bands instead of the ⟨100⟩ bands. In fact, the ⟨111⟩ bands have a 20-fold difference between the transverse and longitudinal mass for electrons, with the transverse mass near that of GaAs. Therefore, a logical extension to the work of Chapter 6 would be the development of ⟨111⟩ aligned transport models. Finally, with the ability to set a large effective mass band at an arbitrary energy above a light effective mass band, it should in theory be possible to produce negative differential mobility, in terms of μ versus electric field, within strained $Si_{1-x}Ge_x$; leading to the possibility of devices, such as Gunn diodes, which can only be presently made in materials such as GaAs. Therefore, a further extension to the work of Chapter 6 is to investigate the feasibility of generating and utilising strained $Si_{1-x}Ge_x$ films that produce negative differential mobility versus electric field.

As a final parting comment regarding future work, it is clear that with the rapid progress continuing in the development of ICs, device dimensions will continue to shrink at an exponential rate. Obviously, this will take devices down into the atomic realm where distances cover only 10 Angstroms and not a thousand. Even with present-day devices, where relevant dimensions are 500 to 1000Å, quantum mechanical effects are important (as can be seen from the consideration of tunneling in Chapter 4). As dimensions reduce to 10Å, clearly, classical mechanics will have no part. For this reason, work on hydrodynamic models, which are really only a second order perturbative solution of the BTE (drift-diffusion being the zero-th and first), will have very limited usefulness. Instead, a "full" quantum mechanical model will be required. But then what is meant by a "full" model? With relevant dimensions of 10Å, it will not even be possible to utilise Bloch's theorem because there will truly be no dimension over which the crystal can be considered as bulk. Furthermore, considering only the conduction electrons in a quantum mechanical fashion, and not

the core electrons, will not be acceptable at 10Å. Thus, by "full" model, it is meant that all electron, protons, and neutrons be considered in a quantum mechanical fashion, without even the simple luxury of assuming Bloch solutions. Obviously, such a "full" model is not even remotely possible today. However, with computing power increasing exponentially, and the number of atoms in the device decreasing exponentially, it will be interesting to see how long it will be before such "full" models come into existence.

In any event, the pursuit of better models which incorporate evermore quantum mechanics must continue in lock-step with the advancement in processing technology. This will enable the high technology sector to understand current day devices and visualise future ones.

References

[1] W. Shockley, *U.S. Patent 2 569 347*, filed June 26, 1948 and issued September 25, 1951.

[2] H. Kroemer, "Theory of a Wide-Gap Emitter for Transistors", *Proc. IRE*, 1535-1538, 1957.

[3] T. Nittono, et al., "A New Self-Aligned AlGaAs/GaAs HBT Based on Refractory Emitter and Base Electrodes", *IEEE Electron Dev. Lett.*, vol. 10, 506-507, November 1989.

[4] H. Ichino, et al., "A 10Gb/s Decision Circuit Using AlGaAs/GaAs HBT Technology", *1990 ISSCC*, 188-189.

[5] K. Wang, et al., "A 15GHz Gate Array Implementation with AlGaAs/GaAs Heterojunction Bipolar Transistors", *1991 ISSCC*, 154-155.

[6] P. Enquist, L. Ramberg, L. Eastman, "Comparison of Compositionally Graded Abrupt Emitter-Base Junctions Used in the Heterojunction Bipolar Transistor", *J. Appl. Phys*, 2663-2669, April 1987.

[7] A. Grinberg, S. Luryi, "On the Thermionic-Diffusion Theory of Minority Transport In Heterostructure Bipolar Transistors", *IEEE Trans. Electron Dev.*, vol. 40, 859-866, May 1993.

[8] A. Grinberg, S. Luryi, "Coherent Transistors", *IEEE Trans. Electron Dev.*, vol. 40, 1512-1522, August 1993.

[9] D. Herbert, "Quasi-Ballistic Corrections to Base Transit Time in Bipolar Transistors", *Semicon. Science & Tech.*, vol. 6, n 5, 405-407, May 1991.

[10] T. Kamins, et al., "Small-Geometry, High-Performance, Si-$Si_{1-x}Ge_x$ Heterojunction Bipolar Transistors", *IEEE Electron Dev. Lett.*, vol. 10, 503-505, November 1989.

[11] E. Crabbé, et al., "Current Gain Rolloff in Graded-Base SiGe Heterojunction Bipolar Transistors", *IEEE Electron Dev. Lett.*, vol. 14, 193-195, April 1993.

[12] A. Levi, T. Chiu, "Room-Temperature Operation of Hot-Electron Transistors", *Appl. Phys. Lett.*, 984-986, September 1987.

[13] M. Hafizi, et al., "39.5-GHz Static Frequency Divider Implemented in AlInAs/GaInAs HBT Technology", *IEEE Electron Dev. Lett.*, vol. 13, 612-614, December 1992.

[14] J. Hayes, et al., "Base Transport Dynamics in a Heterojunction Bipolar Transistor", *Appl. Phys. Lett.*, 1481-1483, November 1986.

[15] N. Ashcroft, D. Mermin, Solid State Physics, Philadelphia: *Saunders College*, 1976, Chapters 12, 13, 16.

[16] T. Tang, S. Ramaswamy, J. Nam, "An Improved Hydrodynamic Transport Model for Silicon", *IEEE Trans. Electron Dev.*, vol. 40, 1469-1477, August 1993.

[17] A. Benvenuti, "Hierarchical PDE Simulation of Nonequilibrium Transport Effects in Semiconductor Devices", *NUPAD IV 1992*, 155-160.

[18] D. Pulfrey, S. Searles, "Electron Quasi-Fermi Level Splitting at the Base-Emitter Junction of AlGaAs/GaAs HBTs", *IEEE Trans. Electron Dev.*, vol. 40, 1183-1185, June 1993.

[19] A. Das, M. Lundstrom, "Numerical Study of Emitter-Base Junction Designs for AlGaAs/GaAs Heterojunction Bipolar Transistors", *IEEE Trans. Electron Dev.*, vol. 35, 863-870, July 1988.

[20] M. Lundstrom, "An Ebers-Moll Model for the Heterostructure Bipolar Transistor", *Solid-State Electronics*, vol. 29, 1173-1179, 1986.

[21] M. Lundstrom, "Boundary Conditions for pn Heterojunctions", *Solid-State Electronics*, vol. 27, 491-496, 1984.

[22] D. Pulfrey, "Development of Models for Heterojunction Bipolar Transistors", BNR Contract: Final Report, July 27, 1993.

[23] W. Liu, et al., "Current Transport Mechanism in GaInP/GaAs Heterojunction Bipolar Transistors", *IEEE Trans. Electron Dev.*, vol. 40, 1378-1383, August 1993.

[24] S. Searles, D. Pulfrey, "An Analysis of Space-Charge-Region Recombination in HBTs", *IEEE Trans. Electron Dev.*, vol. 41, 476-483, April 1994.

[25] S. Searles, D. Pulfrey, "Tunneling and its Inclusion in Analytical Models for Abrupt HBTs", *1994 International Workshop on Computational Electronics*, WeP8.

[26] C. King, J. Hoyt, J. Gibbons, "Bandgap and Transport Properties of $Si_{1-x}Ge_x$ by Analysis of Nearly Ideal $Si/Si_{1-x}Ge_x/Si$ Heterojunction Bipolar Transistors", *IEEE Trans. Electron Dev.*, vol. 36, 2093-2104, October 1989.

[27] R. People, J. Bean, "Band Alignments of Coherently Strained Ge_xSi_{1-x}/Si Heterostructures on $\langle 001 \rangle$ Ge_ySi_{1-y}", *Appl. Phys. Let.*, vol. 48, 538-540, February 1986.

[28] B. Pejcinovic, et al., "Numerical Simulation and Comparison of Si BJT's and $Si_{1-x}Ge_x$ HBT's", *IEEE Trans. Electron Dev.*, vol. 36, 2129-2136, October 1989.

[29] RF Micro Devices Inc, TRW Space and Electronics Group, "HBT Amplifiers Break $1 Price Barrier", *Microwave Journal*, 120-123, February 1995.

[30] Edited by S. Sze, High-Speed Semiconductor Devices, Toronto: *John Wiley & Sons*, 1990, Chapter 1.

[31] B. Meyerson, "UHV/CVD Growth of Si and Si:Ge Alloys: Chemistry, Physics, and Device Applications", *Proc. of the IEEE*, vol. 80, no. 10, 1592-1608, October 1992.

[32] J. Cressler, "The SiGe bipolar transistor", *IEEE Spectrum*, vol. 32, 49-55, March 1995.

[33] J. Warnock, "Silicon Bipolar Device Structures for Digital Applications: Technology Trends and Future Directions", *IEEE Trans. Electron Dev.*, vol. 42, 377-389, March 1995.

[34] T. Nakamura, H. Nishizawa, "Recent Progress in Bipolar Transistor Technology", *IEEE Trans. Electron Dev.,* vol. 42, 390-398, March 1995.

[35] D. Harame et al., "Si/SiGe Epitaxial-Base Transistors- Part I: Materials, Physics and Circuits", *IEEE Trans. Electron Dev.,* vol. 42, 455-468, March 1995.

[36] D. Harame et al., "Si/SiGe Epitaxial-Base Transistors- Part II: Process Integration and Analog Applications", *IEEE Trans. Electron Dev.,* vol. 42, 469-482, March 1995.

[37] F. Sato, et al., "Sub-20ps ECL Circuits with High-Performance Super Self-Aligned Selectively Grown SiGe Base (SSSB) Bipolar Transistors", *IEEE Trans. Electron Dev.,* vol. 42, 483-488, March 1995.

[38] H. Kroemer, "Two Integral Relations Pertaining to the Electron Transport through a Bipolar Transistor with a Nonuniform Energy Gap in the Base Region", *Solid-State Electronics*, vol. 28, 1101-1103, 1985.

[39] C. Maziar, et al., "Monte Carlo Evaluation of Electron Transport in Heterojunction Bipolar Transistor Base Structures", *IEEE Trans. Electron Dev.,* vol. 33, 881-888, July 1986.

[40] M. Heiblum, et al., "Direct Observation of Ballistic Transport in GaAs", *Physical Review Lett.*, vol. 55, n 20, 2200-2203, November 1985.

[41] M. Stettler, M. Alam, M. Lundstrom, "A Critical Examination of the Assumptions Underlying Macroscopic Transport Equations for Silicon Devices", *IEEE Trans. Electron Dev.,* vol. 40, 733-740, April 1993.

[42] A. Grinberg, S. Luryi, "Diffusion in a Short Base", *Solid-State Electronics*, vol. 35, no. 9, 1299-1309, 1992.

[43] A. St. Denis, D. Pulfrey, "An Analytical Expression for the Current in Short-Base Transistors", *Solid-State Electronics*, accepted for publication in December, 1994.

[44] W. Shockley, "The Path to the Conception of the Junction Transistor, "*IEEE Trans. Electron Dev.,* vol. 23, 597-620, 1976.

[45] F. Capasso, "Band-Gap Engineering: From Physics and Materials to New Semiconductor Devices", *Science*, vol. 235, 172-176, January 1987.

[46] H. Kroemer, "Heterostructure Bipolar Transistors and Integrated Circuits", *Proc. of the IEEE*, vol. 70, no. 1, 13-25, January 1982.

[47] H. Kroemer, "Heterostructure bipolar transistors: What should we build", *J. of Vac. Sci. Tech. B*, vol. 1, no. 2, 126-130, April-June 1983.

[48] M. Heiblum, et al., "Direct Observation of Ballistic Transport in GaAs", *Phys. Rev. Lett.*, vol. 55, 2200-2203, 1985.

[49] F. Berz, "The Bethe Condition for Thermionic Emission Near an Absorbing Boundary", *Solid-State Electronics*, vol. 28, no. 10, 1007-1013, 1985.

[50] S. Perlman, D. Feucht, "p-n Heterojunctions", *Solid-State Electronics*, vol. 7, 911-923, 1964.

[51] A. Grinberg, et al., "An investigation of the effect of graded-layers and tunneling on the performance of AlGaAs/GaAs heterojunction bipolar transistors", *IEEE Trans. Electron Dev.*, vol. 31, 1758-1765, 1984.

[52] A. Marty, G. Rey, J. Bailbe, "Electrical Behavior of an NPN GaAlAs/GaAs Heterojunction Transistor", *Solid-State Electronics*, vol. 22, 549-557, 1979.

[53] R, Warner, B. Grung, Transistors: Fundamentals for the Integrated-Circuit Engineer, Toronto: *John Wiley & Sons*, 1983, p. 491.

[54] J. Moll, I. Ross, "The Dependence of Transistor Parameters on the Distribution of Base Layer Resistivity", *Proc. IRE*, 72-78, January 1956.

[55] H. Gummel, "A Charge Control Relation for Bipolar Transistors", *Bell Sys. Tech. J.*, 115-120, January 1970.

[56] H. Gummel, H. Poon, "An Integral Charge Control Model of Bipolar Transistors", *Bell Sys. Tech. J.*, 827-852, May 1970.

[57] J. Early, "Effects of Space-Charge Layer Widening In Junction Transistors", *Proc. IRE*, 1401-1406, 1952.

[58] C. Kirk Jr., "A Theory of Transistor Cutoff Frequency (f_T) Falloff at High Current Densities", *IRE Trans. Elec. Dev.*, vol. 9, 164-173, 1962.

[59] S. Searles, A Study and Modeling of Two Dimensional Effects in Bipolar Transistors, *Carleton University*, 1989, M.Eng. Thesis, Chapter 3.

[60] A. Marshack, "Optimum Doping Distribution for Minimum Base Transit Time", *IEEE Trans. Electron Dev.*, vol. 14, 190-194, 1967.

[61] M. Shur, Physics of Semiconductor Devices, New Jersey: *Prentice-Hall*, 1990, p. 626.

[62] C. Fox, An introduction to the Calculus of Variations, New York: *Dover Publications*, 1987, 31-33.

[63] S. Winterton, S. Searles, C. Peters, N. Tarr, D. Pulfrey, "Distribution of Base Dopant for Transit Time Minimization in a Bipolar Transistor", *IEEE Trans. Electron Dev.*, submitted October 1994.

[64] J. McGregor, T. Manku, D. Roulston, "Bipolar Transistor Base Bandgap Grading for Minimum Delay", *Solid-State Electronics*, vol. 34, 421-422, 1991.

[65] J. McGregor, T. Manku, D. Roulston, "Retraction: Bipolar Transistor Base Bandgap Grading for Minimum Delay", *Solid-State Electronics*, vol. 35, p. 1383, 1992.

[66] E. Prinz, J. Sturm, "Analytical Modeling of Current Gain - Early Voltage Products in Si/$Si_{1-x}Ge_x$/Si Heterojunction Bipolar Transistors", *IEDM 1991*, 33.2.1-33.2.4.

[67] W. Shockley, "Theory of *p-n* Junctions in Semiconductors and *p-n* Junction Transistors", *Bell Sys. Tech. J.*, vol. 28, 435-489, 1949.

[68] E. Murphy, R. Good Jr., "Thermionic Emission, Field Emission, and the Transition Region", *Phys. Rev.*, vol. 102, no. 6, 1464-1473, 1956.

[69] R. Stratton, "Theory of Field Emission from Semiconductors", *Phys. Rev.*, vol. 125, no. 1, 67-82, January 1962.

[70] S. Christov, "Unified Theory of Thermionic and Field Emission from Semiconductors", *Phys. Stat. Sol.*, vol. 21, 159-173, 1967.

[71] L. Landau, E. Lifshitz, Quantum Mechanics (Non-Relativistic Theory), Toronto: *Pergamon Press*, 1991, Chapter 7.

[72] P. Wallace, Mathematical Analysis of Physical Problems, New York: *Dover Publications*, 1984, 55-58.

[73] S. Christov, "General Theory of Electron Emission from Metals", *Phys. Stat. Sol.*, vol. 17, 11-27, 1966.

[74] N. Ashcroft, D. Mermin, Solid State Physics, Philadelphia: *Saunders College*, 1976, 221-223, and also K. Symon, Mechanics, Reading Massachusetts: *Addison-Wesley*, 1971, p. 395.

[75] F. Padovani, R. Stratton, "Field and Thermionic-Field Emission in Schottky Barriers", *Solid-State Electronics*, vol. 9, 695-707, 1966.

[76] C. Crowell, "The Richardson Constant for Thermionic Emission in Schottky Barrier Diodes", *Solid-State Electronics*, vol. 8, 395-399, 1965.

[77] C. Crowell, "Richardson Constant and Tunneling Effective Mass for Thermionic and Thermionic-Field Emission in Schottky Barrier Diodes", *Solid-State Electronics*, vol. 12, 55-59, 1969.

[78] C. Crowell, V. Rideout, "Normalized Thermionic-Field (T-F) Emission in Metal-Semiconductor (Schottky) Barriers", *Solid-State Electronics*, vol. 12, 89-105, 1969.

[79] S. Perlman, D. Feucht, "p-n Heterojunctions", *Solid-State Electronics*, vol. 7, 911-923, 1964.

[80] M. Shur, Physics of Semiconductor Devices, New Jersey: *Prentice-Hall*, 1990, p. 225.

[81] I. Gradshteyn, I. Ryzhik, Table of Integrals, Series, and Products, Toronto: *Academic Press*, 1980.

[82] A. Grinberg, "Thermionic Emission in Heterosystems With Different Effective Electronic Masses", *Phys. Rev. B*, vol. 33, no. 10, 7256-7258, 1986.

[83] J. Gunn, "Transport of Electrons in a Strong Built-In Field", *J. Appl. Phys.*, vol. 39, 4602-4604, 1968.

[84] B. Gokhale, "Numerical Solutions for a One-Dimensional Silicon n-p-n Transistor", *IEEE Trans. Electron Dev.*, vol. 17, 594-602, 1970.

[85] J. Higman, K. Hess, "Comment of the Use of the Temperature Concept for Non-Linear Transport Problems in Semiconductor p-n Junctions", *Solid-State Electronics*, vol. 29, 915-918, 1986.

[86] S. Miller, R. Good Jr., "A WKB-Type Approximation to the Schrödinger Equation", *Phys. Rev.*, vol. 91, no. 1, 174-179, 1953.

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

[88] S. Ho, D. Pulfrey, "The effect of base grading on the gain and high-frequency performance of AlGaAs/GaAs heterojunction bipolar transistors", *IEEE Trans. Electron Dev.*, vol. 36, 2173-2182, 1989.

[89] K. Horio, H. Yanai, "Numerical modeling of heterojunctions including the thermionic emission mechanism at the heterojunction interface", *IEEE Trans. Electron Dev.*, vol. 37, 1093-1098, 1990.

[90] C. Parikh, F. Lindholm, "Space-Charge Region Recombination in Heterojunction Bipolar Transistors", *IEEE Trans. Electron Dev.*, vol. 39, 2197-2205, 1992.

[91] C. Sah, R. Noyce, W. Shockley, "Carrier Generation and Recombination in p-n Junctions and p-n Junction Characteristics", *Proc. IRE*, vol. 45, 1228-1243, 1957.

[92] S. Choo, "Carrier Generation-Recombination in the Space Charge Region of an Asymmetrical p-n Junction", *Solid-State Electronics*, vol. 11, 1069-1077, 1968.

[93] M. Takeshima, "Effect of Auger Recombination on Laser Operation in $Ga_{1-x}Al_xAs$", *J. Appl. Physics,* vol. 58, 3846-3850, 1985.

[94] R. Hall, "Recombination Processes in Semiconductors", *Proc. IEEE*, vol. B-106, Suppl. 15-18, 923-931, 1959.

[95] A. Heberle, et al., "Minority-carrier lifetime in heavily doped GaAs:C", *Extended Abstracts, Int. Conf. on Solid-State Devices and Materials*, Tsukuba, 290-292, 1992.

[96] M. Shur, Physics of Semiconductor Devices, New Jersey: *Prentice-Hall*, 1990, 623-624.

[97] J. Matthews, A. Blakeslee, "Defects in Epitaxial Multilayers-I: Misfit Dislocations", *J. Cryst. Growth*, vol. 27, 118-125, 1974.

[98] J. Matthews, "Defects Associated with the Accommodation of Misfit Between Crystals", *J. Vac. Sci. Tech.*, vol. 12, 126-133, 1975.

[99] J. Bean, et al., "Ge_xSi_{1-x}/Si Strained-Layer Superlattice Grown By Molecular Beam Epitaxy", *J. Vac. Sci. Tech. A*, vol. 2, no. 2, 437-440, 1984.

[100] G. Patton, et al., "75-GHz f_T SiGe-Base Heterojunction Bipolar Transistors", *IEEE Electron Dev. Lett.*, vol. 11, 171-173, April 1990.

[101] E. Crabbé, et al., "73-GHz Self-Aligned SiGe-Base Bipolar Transistors With Phosphorus-Doped Polysilicon Emitters", *IEEE Electron Dev. Lett.*, vol. 13, 259-261, May 1992.

[102] E. Crabbé, et al., "113-GHz f_T Graded-Base SiGe HBTs", *Proc. 51st Dev. Res. Conf.*, IIA-3, 1993.

[103] F. Sato, et al., "Sub-20ps ECL Circuits with High-Performance Super Self-Aligned Selectively Grown SiGe Base (SSSB) Bipolar Transistors", *IEEE Trans. Electron Dev.*, vol. 42, no. 3, 483-488, 1995.

[104] N. Ashcroft, D. Mermin, Solid State Physics, Philadelphia: *Saunders College*, 1976, 568-570.

[105] R. People, "Physics and Applications of Ge_xSi_{1-x}/Si Strained-Layer Heterostructures", *IEEE J. Quantum Elec.*, vol. QE-22, 1696-1710, September 1986.

[106] C. Van de Walle, R. Martin, "Theoretical Calculations of Heterojunction Discontinuities in the Si/Ge System", *Phys. Rev. B*, vol. 34, 5621-5634, October 1986.

[107] R. People, "Indirect Band Gap of Coherently Strained Ge_xSi_{1-x} Bulk Alloys on ⟨001⟩ Silicon Substrates", *Phys. Rev. B*, vol. 32, 1405-1408, July 1985.

[108] C. Zeller, G. Abstreiter, "Electric Subbands in Si/SiGe Strained Layer Superlattices", *Zeitschrift Phys. B*, vol. 64, 137-143, 1986.

[109] N. Ashcroft, D. Mermin, Solid State Physics, Philadelphia: *Saunders College*, 1976, p. 169.

[110] W. Harrison, Electronic Structure and the Properties of Solids, New York: *Dover Publications*, 1989, 160-161.

[111] R. People, et al., "Modulation Doping in Ge_xSi_{1-x}/Si Strained Layer Heterostructures", *Appl. Phys. Lett.*, vol. 45, 1231-1233, December 1984.

[112] J. Singh, Physics of Semiconductors and Their Heterostructures, New York: *McGraw-Hill Inc.*, 1993, p. 238.

[113] R. Braunstein, A. Moore, F. Herman, "Intrinsic Optical Absorption in Germanium-Silicon Alloys", *Physical Review*, vol. 109, no.3, 695-710, 1958.

[114] C. Herring, E. Vogt, "Transport and Deformation-Potential Theory for Many-Valley Semiconductors with Anisotropic Scattering", *Physical Review*, vol. 101, 944-961, 1956.

[115] I. Balslev, "Influence of Uniaxial Stress on the Indirect Absorption Edge in Silicon and Germanium", *Physical Review*, vol. 143, 636-647, 1966.

[116] J. Singh, Physics of Semiconductors and Their Heterostructures, New York: *McGraw-Hill Inc.*, 1993, Chapter 7.

[117] D. Lang, et al., "Measurement of the Band Gap of Ge_xSi_{1-x}/Si Strained-Layer Heterostructures", *Appl. Phys. Lett.*, vol. 47, 1333-1335, December 1985.

[118] H. Hasegawa, "Theory of Cyclotron Resonance in Strained Silicon Crystals", *Physical Review*, vol. 129, 1029-1040, 1963.

[119] C. Van de Walle, R. Martin, "Theoretical Calculations of Semiconductor Heterojunction Discontinuities", *J. Vac. Sci. Tech. B*, vol. 4, no. 4, 436-440, 1984.

[120] C. Van de Walle, R. Martin, "Theoretical Study of Si/Ge Interfaces", *J. Vac. Sci. Tech. B*, vol. 3, n 4, 1256-1259, 1985.

[121] T. Keuch. M. Mäenpää, S. Lau, "Epitaxial Growth of Ge on ⟨100⟩ Si by a Simple Chemical Vapor Deposition Technique", *Appl. Phys. Lett.*, vol. 39, 245-247, 1981.

[122] G. Margaritondo, et al., "Nature of Band Discontinuities at Semiconductor Heterojunction Interfaces", *Solid State Comm.*, vol. 43, 163-166, 1982.

[123] P. Mahowald, et al., "Heterojunction Band Discontinuity at the Si-Ge (111) Interface", *J. Vac. Sci. Tech. B*, vol. 3, 1252-1255, 1985.

[124] G. Schwartz, "Core-Level Photoemission Measurements of Valence-Band Offsets in Highly Strained Heterojunctions: Si-Ge System", *Physical Review B*, vol. 39, 1235-1241, 1989.

[125] E. Yu, Physics and Applications of Semiconductor Heterostructures: I Measurement of Band Offsets in Semiconductor Heterojunctions; II Theoretical and Experimental Studies of Tunneling in Semiconductor Heterostructure Devices, *California Institute of Technology*, 1991, Ph.D. Thesis, Chapter 4.

[126] G. Abstreiter, H. Brugger, T. Wolf, "Strain-Induced Two-Dimensional Electron Gas in Selectively Doped Si/Si_xGe_{1-x} Superlattices", *Phys. Rev. Lett.*, vol. 54, 2441-2444, June 1985.

[127] W. Ni, J. Knall, G. Hansson, "Strain-Affected Band Offsets at Si/Si_xGe_{1-x} (100) Heterojunction Interfaces Studied With X-Ray Photoemission", Surface Sci., vol. 189/190, 379-384, 1987.

[128] C. Gan, et al., "$Si_{1-x}Ge_x$/Si Valence Band Discontinuity Measurements Using a Semiconductor-Insulator-Semiconductor (SIS) Heterostructure", *IEEE Trans. Electron Dev.*, vol. 41, no. 12, 2430-2439, 1994.

[129] J. Singh, Physics of Semiconductors and Their Heterostructures, New York: *McGraw-Hill Inc.*, 1993, Chapter 12.

[130] T. Yamada, et al., "In-Plane Transport Properties of Si/Si_xGe_{1-x} Structure and its FET Performance by Computer Simulation", *IEEE Trans. Electron Dev.*, vol. 41, no. 9, 1513-1522, 1994.

[131] N. Ashcroft, D. Mermin, Solid State Physics, Philadelphia: *Saunders College*, 1976, Chapter 8.

[132] L.Kay, T. Tang, "Monte Carlo Calculation of Strained and Unstrained Electron Mobilities in Si_xGe_{1-x} Using an Improved Ionized-Impurity Model", *J. Appl. Physics*, vol. 70, 1483-1488, 1991.

[133] S. Iyer, et al., "Heterojunction Bipolar Transistors Using Si-Ge Alloys", *IEEE Trans. Electron Dev.*, vol. 36, 2043-2063, October 1989.

[134] J. Burghatz, et al., "APCVD-Grown Self-Aligned SiGe-Base HBT's, *1993 IEEE BCTM*, 55-62, 1993.

[135] A. Grinberg, S. Luryi, "Dynamic Early Effect in Heterojunction Bipolar Transistors", *IEEE Electron Dev. Lett.*, vol. 14, no. 6, 292-294, May 1993.

[136] M. Kondo, T. Kobayashi, Y. Tamaki, "Hetro-Emitter-Like Characteristics of Phosphorous Doped Polysilicon Emitter Transistors - Part I: Band Structure in the Polysilicon Emitter Obtained from Electrical Measurements", *IEEE Trans. Electron Dev.*, vol. 42, no. 3, 419-426, 1995.

[137] M. Kondo, T. Kobayashi, Y. Tamaki, "Hetro-Emitter-Like Characteristics of Phosphorous Doped Polysilicon Emitter Transistors - Part II: Band Deformation Due to Residual Stress in the Polysilicon Emitter", *IEEE Trans. Electron Dev.*, vol. 42, no. 3, 427-435, 1995.

Appendix A

Ramped $N_{AB}(x)$ to Minimise τ_B

The proof of eqn (3.17) begins by solving eqn (3.10) for τ_B using the doping profile depicted in Fig. 3.4. To this end, it is seen that the doping profile of Fig. 3.4 is actually a subset of the profile depicted in Fig. 3.3 with $h_1 = 0$ and $h_2 = h$. Using the symbolic math tool MACSYMA©, eqn (3.10) yields the following result for τ_B based upon the distribution presented in Fig. 3.3:

(d5) $$Ne\, e^{\frac{\log(U)\, x - h1 \log(U)}{h1 - h2}}$$

(c6) integrate(\ne/\u,x,h2,1);

(d6) $$\frac{(1 - h2)\, Ne}{U}$$

(c7) integrate(d5,x,h1,h2);

Is U - 1 zero or nonzero?
nonzero;

(d7) $$Ne \left(\frac{h2 - h1}{\log(U)} - \frac{(h2 - h1)\, e^{\frac{h1 \log(U)}{h2 - h1} - \frac{h2 \log(U)}{h2 - h1}}}{\log(U)} \right)$$

(c8) integrate(\ne/\u,xx,1);

(d8) $$\frac{Ne\ (1 - x)}{U}$$

(c9) integrate(d5,xx,h2);

Is U - 1 zero or nonzero?
nonzero;

(d9) $$Ne \left(\frac{(h2 - h1)\, e^{\frac{h1 \log(U)}{h2 - h1} - \frac{\log(U)\, x}{h2 - h1}}}{\log(U)} - \frac{(h2 - h1)\, e^{\frac{h1 \log(U)}{h2 - h1} - \frac{h2 \log(U)}{h2 - h1}}}{\log(U)} \right)$$

(c10) integrate(\ne ,x,x,h1);

(d10) $$Ne\ (h1 - x)$$

Eqn (d5) is the exponential doping profile for $h_1 \le x \le h_2$, and it ensures that there are no jump discontinuities at the break points h_1 and h_2 between the exponential doping profile and the regions of constant doping. Then, eqns (d6)-(d10) collect together the various sub-integrals required to solve eqn (3.10). It should be noted that the doping at $x = 0$ is N_e, at $x = 1$ is N_c, and that $U = N_e/N_c$. Using eqns (d5)-(d10), eqn (3.10) produces:

(c11) tau = ratsimp(radcan(integrate(1/\ne*(d10+d7+d6),x,0,h1)+integrate(radcan(1/d5*(d9+d6)),x,h1,h2)+ integrate(\u/\ne*(d8),x,h2,1));

Is U - 1 zero or nonzero?

nonzero;

(d11)
$$\tau=\frac{\left(\left(h2^2-2\,h2+1\right)U^{\frac{h2}{h2-h1}}+U^{\frac{h1}{h2-h1}}\left(h1^2\,U-2\,h1\,h2+2\,h1\right)\right)\log^2(U)+\left(\left(2\,h2-2\,h1\right)U^{\frac{h2}{h2-h1}}+\left(2\,h2^2+\left(-4\,h1-2\right)h2+2\,h1^2+2\,h1\right)U^{\frac{h1}{h2-h1}}\right)*\log(U)+\left(-2\,h2^2+4\,h1\,h2-2\,h1^2\right)U^{\frac{h2}{h2-h1}}+\left(2\,h2^2-4\,h1\,h2+2\,h1^2\right)U^{\frac{h1}{h2-h1}}}{2\,U^{\frac{h2}{h2-h1}}\log^2(U)}$$

Eqn (d11) is the general model for τ_B from the optimum doping profile of Fig. 3.3.

Using the optimum equation for τ_B given in eqn (d11), then the τ_B needed for the proof of eqn (3.17) is obtained by setting $h_1 = 0$ and $h_2 = h$; *i.e.*,

(c12) ev(d11,h1=0,h2=h);

(d12)
$$\tau=\frac{\left(h^2-2\,h+1\right)U\log^2(U)+\left(2\,h\,U+2\,h^2-2\,h\right)\log(U)-2\,h^2\,U+2\,h^2}{2\,U\log^2(U)}$$

Eqn (d12) can then be solved for the h that minimises τ_B. Differentiating eqn (d12) with respect to h, setting equal to zero, and solving for h produces:

(c13) ratsimp(diff(rhs(d12),h));

(d13)
$$\frac{\left(h-1\right)U\log^2(U)+\left(U+2\,h-1\right)\log(U)-2\,h\,U+2\,h}{U\log^2(U)}$$

(c14) solve(d13=0,h);

(d14)
$$\left[h=\frac{U\log^2(U)+\left(1-U\right)\log(U)}{U\log^2(U)+2\log(U)-2\,U+2}\right]$$

(c16) ratsimp(radcan(ev(d12,d14)));

(d16)
$$\tau=\frac{2\,U^2\log(U)-3\,U^2+4\,U-1}{2\,U^2\log^2(U)+4\,U\log(U)-4\,U^2+4\,U}$$

where (d14) is the same as h in eqn (3.17), and eqn (d16) is the same as τ_B in eqn (3.17) once the

factor of 1/2 is included within τ_{B0}. This completes the proof of eqn (3.17) for the ramped $N_{AB}(x)$ to minimise τ_B. It should be noted that the output displayed within this Appendix comes directly from MACSYMA©. As such, there is occasion to perform some intermediates steps that are not instructive to the proof but are more of a bookkeeping function for MACSYMA© itself. This is why some of the d-equations are missing.

Finally, it can be shown that an intriguing symmetry exists in the ramped doping profile. If the profile is changed from that shown in Fig. 3.4 so that the exponential region follows the constant doping region, then it is found that τ_B remains unchanged from what is given in eqn (3.17), and $h \rightarrow 1-h$. Returning back to eqn (d11), the necessary change to the doping profile is accomplished by setting $h_1 = h$ and $h_2 = 1$ in the optimum equation for τ_B given in eqn (d11); *i.e.*,

(c17) expand(ev(d11,h1=h,h2=1));

(d17)
$$\tau = \frac{h^2 U^{\frac{h}{1-h}-\frac{1}{1-h}}}{\log(U)} - \frac{h U^{\frac{h}{1-h}-\frac{1}{1-h}}}{\log(U)} - \frac{h}{\log(U)} + \frac{1}{\log(U)} + \frac{h^2 U^{\frac{h}{1-h}-\frac{1}{1-h}}}{\log^2(U)} - \frac{2h U^{\frac{h}{1-h}-\frac{1}{1-h}}}{\log^2(U)} + \frac{U^{\frac{h}{1-h}-\frac{1}{1-h}}}{\log^2(U)} - \frac{h^2}{\log^2(U)} + \frac{2h}{\log^2(U)} - \frac{1}{\log^2(U)} + \frac{h^2 U^{\frac{h}{1-h}-\frac{1}{1-h}+1}}{2}$$

(c18) substpart(xthru(map(radcan,piece)),d17,2);

(d18)
$$\tau = \frac{U\left(h^2\log^2(U) + 2\left(-h^2+2h-1\right)\right) + 2(1-h)U\log(U) + 2\left(h^2-h\right)\log(U) + 2\left(h^2-2h+1\right)}{2U\log^2(U)}$$

Eqn (d18) is the τ_B for the symmetric doping profile used to develop eqn (d12). As was done with eqn (d12), the optimum value for h is found by differentiation eqn (d18) with respect to h, setting equal to zero and solving; *i.e.*,

(c19) diff(rhs(d18),h)=0;

(d19)
$$\frac{U\left(2h\log^2(U) + 2(2-2h)\right) - 2U\log(U) + 2(2h-1)\log(U) + 2(2h-2)}{2U\log^2(U)} = 0$$

(c20) solve(d19,h);

(d20)
$$\left[h = \frac{(U+1)\log(U) - 2U + 2}{U\log^2(U) + 2\log(U) - 2U + 2}\right]$$

Eqn (d20) is that h which renders eqn (d18) a minimum. Substituting (d20) back into (d18) yields the minimum τ_B; *i.e.*,

(c21) ratsimp(radcan(ev(d18,d20)));

(d21)
$$\tau = \frac{2\,U^2\log(U) - 3\,U^2 + 4\,U - 1}{2\,U^2\log^2(U) + 4\,U\log(U) - 4\,U^2 + 4\,U}$$

Eqn (d21) is exactly the same as eqn (d16), showing that a symmetric change to the doping profile produces no change to the transit time. It can finally be shown that the symmetric change to the doping profile results in $h \rightarrow 1 - h$ by adding together the h from eqn (d14) and (d20); *i.e.*,

(c26) rhs(first(d20))+d14;

(d26)
$$\frac{U\log^2(U) + (1-U)\log(U)}{U\log^2(U) + 2\log(U) - 2\,U + 2} + \frac{(U+1)\log(U) - 2\,U + 2}{U\log^2(U) + 2\log(U) - 2\,U + 2}$$

(c27) ratsimp(combine(d26));

(d27) 1

Eqn (d27) proves that the symmetric change to the doping profile of Fig. 3.4 does indeed result in $h \rightarrow 1 - h$.

Appendix B

Optimum $N_{AB}(x)$ to Minimise τ_B

The proof of eqn (3.18) begins by solving eqn (3.10) for τ_B using the doping profile depicted in Fig. 3.3. However, this task has already been accomplished in Appendix A as eqn (d11). Using eqn (d11) for the optimum τ_B, the pair h_1 and h_2 which minimise eqn (d11) is found. Using the symbolic math tool MACSYMA©, the partial derivatives of eqn (d11) with respect to h_1 and h_2 are taken; *i.e.*,

(c29) ratsimp(diff(rhs(d11),h1));

(d29)
$$\frac{\left(U^{\frac{h1}{h2-h1}}(h1\,U-h2+1)\log^2(U)+\left((-2\,h2+2\,h1+1)\,U^{\frac{h1}{h2-h1}}-U^{\frac{h2}{h2-h1}}\right)\log(U)+(2\,h2-2\,h1)\,U^{\frac{h2}{h2-h1}}+(2\,h1-2\,h2)\,U^{\frac{h1}{h2-h1}}\right)}{U^{\frac{h2}{h2-h1}}\log^2(U)}$$

(c30) ratsimp(diff(rhs(d11),h2));

(d30)
$$\frac{\left(\left((h2-1)\,U^{\frac{h2}{h2-h1}}-h1\,U^{\frac{h1}{h2-h1}}\right)\log^2(U)+\left(U^{\frac{h2}{h2-h1}}+(2\,h2-2\,h1-1)\,U^{\frac{h1}{h2-h1}}\right)\log(U)+(2\,h1-2\,h2)*U^{\frac{h2}{h2-h1}}+(2\,h2-2\,h1)\,U^{\frac{h1}{h2-h1}}\right)}{U^{\frac{h2}{h2-h1}}\log^2(U)}$$

Eqns (d29) and (d30) present the simultaneous set of equations, once both are set equal to zero, that must be solved to determine the pair h_1 and h_2 which minimise eqn (d11). Given the highly non-linear form of these two equations it is not clear that an analytic solution is possible. Therefore, before attempting to solve eqns (d29) and (d30), a numerical solution will be found so that a "feel" may be developed that will hopefully guide the steps to follow.

Using MACSYMA©, a numerical Newton-Raphson solution to eqns (d29) and (d30) is found for three different cases of U; *i.e.*,

(c40) newton(ev([d29,d30],\u=3.9d0),[h1,h2],[0.25d0,0.75d0]);

C:\MACSYMA2\share\newton.fas being loaded.
C:\MACSYMA2\matrix\bla_lu.fas being loaded.
C:\MACSYMA2\matrix\blinalgl.fas being loaded.

(d40) [h1 = 0.29753257250992d0, h2 = 0.70246742749006d0]

(c41) d40[1]+d40[2];

(d41) h2 + h1 = 1.0d0

(c42) newton(ev([d29,d30],\u=50.4d0),[h1,h2],[0.25d0,0.75d0]);

(d42) [h1 = 0.16891917072612d0, h2 = 0.83108082927388d0]

(c43) d42[1]+d42[2];

(d43) h2 + h1 = 1.0d0

(c44) newton(ev([d29,d30],\u=2000.4d0),[h1,h2],[0.25d0,0.75d0]);

(d44) [h1 = 0.10415470580558d0, h2 = 0.89584529419442d0]

(c45) d44[1]+d44[2];

(d45) h2 + h1 = 1.0d0

The numerical results of eqns (d40)-(d45) indicate that $h_1 + h_2 = 1$. In order to prove that $h_1 + h_2 = 1$ is indeed a solution of eqns (d29) and (d30), the following is performed: substitute $h_2 = 1 - h_1$ into both eqns (d29) and (d30); then, if the resulting eqns differ at most by a multiplicative constant, then it is proven that $h_1 + h_2 = 1$ is indeed a solution of eqns (d29) and (d30).

Using MACSYMA© to perform the above test yields:

(c31) ev(d29,h2=1-h1);

(d31)
$$\frac{\left(\begin{array}{c} U^{\frac{h1}{1-2h1}}\,(h1\,U+h1)\,\log^2(U) \\ +\left(\left(2\,h1-2\,(1-h1)+1\right)U^{\frac{h1}{1-2h1}}-U^{\frac{1-h1}{1-2h1}}\right)\log(U) \\ +\left(2\,h1-2\,(1-h1)\right)U^{\frac{h1}{1-2h1}}+\left(2\,(1-h1)-2\,h1\right)U^{\frac{1-h1}{1-2h1}} \end{array}\right)}{U^{\frac{1-h1}{1-2h1}}\log^2(U)}$$

(c32) ev(d30,h2=1-h1);

(d32)
$$\frac{\left(\left(-h1\,U^{\frac{h1}{1-2h1}}-h1\,U^{\frac{1-h1}{1-2h1}}\right)\log^2(U)+\left(\left(-2h1+2(1-h1)-1\right)U^{\frac{h1}{1-2h1}}+U^{\frac{1-h1}{1-2h1}}\right)\log(U)+\left(2(1-h1)-2h1\right)U^{\frac{h1}{1-2h1}}+\left(2h1-2(1-h1)\right)U^{\frac{1-h1}{1-2h1}}\right)}{U^{\frac{1-h1}{1-2h1}}\log^2(U)}$$

(c33) ratsimp(combine(d31+d32));

(d33)
$$-\frac{h1\,U^{\frac{2h1}{2h1-1}}-h1\,U^{\frac{1}{2h1-1}+1}}{U^{\frac{2h1}{2h1-1}}}$$

(c34) radcan(expand(d33));

(d34) 0

Eqns (d31) and (d32), after substituting $h_2 = 1 - h_1$, are equal and opposite. Thus, these two equations would differ by a multiplicative constant of "-1". Eqns (d33) and (d34) prove that $h_2 = 1 - h_1$ by showing the sum of eqns (d31) and (d32) vanishes. This result immediately asserts that there is only one independent equation to solve for. The solution for h_1 being:

(c35) distrib(expand(d31));

(d35)
$$\frac{4\,h1\,U^{\frac{2h1}{1-2h1}-\frac{1}{1-2h1}}}{\log(U)}-\frac{U^{\frac{2h1}{1-2h1}-\frac{1}{1-2h1}}}{\log(U)}-\frac{1}{\log(U)}+\frac{4\,h1\,U^{\frac{2h1}{1-2h1}-\frac{1}{1-2h1}}}{\log^2(U)}-\frac{2\,U^{\frac{2h1}{1-2h1}-\frac{1}{1-2h1}}}{\log^2(U)}-\frac{4\,h1}{\log^2(U)}+\frac{2}{\log^2(U)}+h1\,U^{\frac{2h1}{1-2h1}-\frac{1}{1-2h1}+1}+h1\,U^{\frac{2h1}{1-2h1}-\frac{1}{1-2h1}}$$

(c36) map(radcan,d35);

(d36)
$$\frac{4\,h1}{U\log(U)}-\frac{1}{U\log(U)}-\frac{1}{\log(U)}+\frac{4\,h1}{U\log^2(U)}-\frac{2}{U\log^2(U)}-\frac{4\,h1}{\log^2(U)}+\frac{2}{\log^2(U)}+\frac{h1}{U}+h1$$

(c37) solve(d36=0,h1);

(d37)
$$\left[h1=\frac{1}{\log(U)+2}\right]$$

Eqn (d37) proves eqn (3.18) for the optimum h_1, along with the result from eqn (d34) which proves eqn (3.18) for the optimum h_2. Finally, using the optimum h_1 and h_2, the optimum τ_B is found by substituting back into eqn (d11) found in Appendix A; *i.e.*,

(c38) radcan(ev(d11,h2=1-h1));

(d38)
$$\frac{\left(\begin{array}{c} U^{\frac{2\,h1+1}{2\,h1-1}}\left(h1^2\,U+h1^2\right)\log^2(U) \\ +\left(\left(h1-2\,h1^2\right)U^{\frac{4\,h1}{2\,h1-1}}+U^{\frac{2\,h1+1}{2\,h1-1}}\left(\left(2\,h1^2-3\,h1+1\right)U+4\,h1^2-2\,h1\right)\right)\log(U) \\ +U^{\frac{2\,h1+1}{2\,h1-1}}\left(\left(-4\,h1^2+4\,h1-1\right)U+4\,h1^2-4\,h1+1\right)\end{array}\right)}{U^{\frac{4\,h1}{2\,h1-1}}\log^2(U)}$$

(c39) radcan(ev(d38,d37));

(d39)
$$\tau=\frac{1}{\log(U)+2}$$

Eqn (d39) is the same as τ_B in eqn (3.18) once the factor of 1/2 is included within τ_{B0}. This completes the proof of eqn (3.18) for the optimum $N_{AB}(x)$ to minimise τ_B.