12-1-1989

Simulation of Heterojunction Bipolar Transistors in Two Dimensions

Paul Emerson Dodd
Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/ecetr

https://docs.lib.purdue.edu/ecetr/689

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
Simulation of Heterojunction Bipolar Transistors in Two Dimensions

P. E. Dodd

TR-EE 89-68
December, 1989

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907
SIMULATION OF
HETEROJUNCTION BIPOLAR TRANSISTORS
IN TWO DIMENSIONS

Final Report: December, 1989
Supported by the Research Triangle Institute
Subcontract Number 1-415T-3329

Paul E. Dodd
Mark S. Lundstrom

Purdue University
School of Electrical Engineering
Technical Report: TR-EE 89-68
West Lafayette, Indiana 47907
SIMULATION OF
HETEROJUNCTION BIPOLAR TRANSISTORS
IN TWO DIMENSIONS

A Thesis
Submitted to the Faculty
of
Purdue University

by
Paul Emerson Dodd

In Partial Fulfillment of the
Requirements for the Degree
of
Master of Science in Electrical Engineering

December 1989
To Grandma with love and memories of summer conversations
ACKNOWLEDGEMENTS

The production of this thesis would not have been possible without the guidance and assistance of many people. Foremost among these is my major professor, Mark S. Lundstrom, whose patience and vast store of useful knowledge have been greatly appreciated and frequently called upon. I wish also to thank Professor Jeff Gray for his willingness to help and ability to debug (in ten minutes) code that has had me mired for days. Thanks to Professor Elliott for taking the time to be on my advisory committee.

I would like also to acknowledge the not inconsiderable contributions of my colleagues. In addition to being highly valued friends, Mark Stettler and Amitava Das have been continual sources of thought stimulation. Michael McLennan and Skip Egley deserve a special nod of recognition for their time and effort freely given whenever consulted.

I would be inexcusably remiss were I not to acknowledge the constant faith and love of my family. My parents' support has been wonderful. I feel especially lucky also to have a brother who I consider to be a good friend.
# TABLE OF CONTENTS

**LIST OF FIGURES** ................................................................. vii

**ABSTRACT** ........................................................................................................ x

**CHAPTER 1: INTRODUCTION** ................................................................. 1

1.1. Introduction and Rudiments of HBT Operation ........................................ 1
1.2. History of the HBT ............................................................................. 6
1.3. History of HBT Simulation .................................................................. 9
1.4. Research Objectives and Thesis Organization ....................................... 12

**CHAPTER 2: MODEL FORMULATION** ..................................................... 13

2.1. Introduction .................................................................................... 13
2.2. The Steady-State Semiconductor Equations ......................................... 13
2.3. Numerical Analysis Techniques .......................................................... 17
   2.3.1. Discretisation .......................................................................... 17
   2.3.2. Solution .................................................................................. 24
2.4. Physical Models .................................................................................. 25
   2.4.1. Bulk Recombination ............................................................... 25
   2.4.2. Surface Recombination ......................................................... 26
   2.4.3. Carrier Mobilities .................................................................. 28
   2.4.4. Bandgap and Bandgap Narrowing ....................................... 31
   2.4.5. Sundries ................................................................................ 33
2.5. Summary ............................................................................................ 34

**CHAPTER 3: MODEL DEVELOPMENT** .................................................... 35
3.1. Introduction.............................................................................................................35
3.2. Extensions for Transistor Analysis..........................................................................35
3.3. Treatment of Nonrectangular Geometries..............................................................36
3.4. Quasi-Static Frequency Evaluation........................................................................43
3.5. The Transient Problem............................................................................................46
   3.5.1. The Time-Dependent Semiconductor Equations...............................................46
   3.5.2. Discretisation and Solution..................................................................................47
3.6. Modeling Hot Electron Transport Phenomena.........................................................48
   3.6.1. Generalized Drift-Diffusion Equations...............................................................49
   3.6.2. Localized Energy Balance Equations.................................................................50
   3.6.3. Balance Equation Solution Method.....................................................................52
3.7. Summary..................................................................................................................52

CHAPTER 4: MODEL VERIFICATION.....................................................................................56

4.1. Introduction..............................................................................................................56
4.2. Treatment of Non-Planar Devices............................................................................56
4.3. Transient Analyses..................................................................................................58
   4.3.1. Short-Circuit Current Decay..............................................................................58
   4.3.2. Transistor Switching..........................................................................................61
4.4. Hot Electron Diode..................................................................................................61
4.5. Summary..................................................................................................................64

CHAPTER 5: MODEL APPLICATIONS.....................................................................................65

5.1. Introduction..............................................................................................................65
5.2. HBT Characteristics...............................................................................................65
5.3. Placement of Base Contacts in HBT's......................................................................68
5.4. Base Currents in Abrupt and Graded HBT's..........................................................74
5.5. Summary..................................................................................................................78

CHAPTER 6: CONCLUSION....................................................................................................79

6.1. Summary and Conclusions......................................................................................79
6.2. Recommendations....................................................................................................81
LIST OF REFERENCES .......................................................................................... 83

APPENDIX: ANALYTIC SOLUTION TO THE ONE-DIMENSIONAL P⁺–N DIODE UNDER FORWARD BIAS ................................................................................. 93
LIST OF FIGURES

Figure | Page
--- | ---
1.1 Minority carrier injection across a) np+ homojunction, and b) Np+ heterojunction | 2
1.2 Current components in an HBT | 3
1.3 Unity current gain cutoff frequency as a function of year reported in representative HBT's | 8
1.4 Gate propagation delay as a function of year reported in ring oscillators fabricated using HBT's | 8
2.1 Boundary surface types for a typical transistor | 16
2.2 Two–dimensional finite difference grid for a transistor | 18
2.3 Graphic illustration of the center difference approximation to a first order derivative | 20
2.4 Node spacing, subscript definitions, and evaluation points for interior of a two–dimensional finite (center) difference mesh | 23
2.5 Node spacing and evaluation points for top boundary of a two-dimensional finite (center) difference mesh | 23
2.6 Majority carrier mobilities in Al_xGa_{1-x}As | 32
2.7 Bandgap narrowing in p–GaAs | 33
3.1 Bipolar transistor nonequilibrium solution flowchart for PUPHS2D | 37
3.2 Cross–sections of a) planar bipolar transistor, and b) mesa–etched bipolar transistor | 39
3.3 Placement of insulating material for prevention of lateral carrier injection across base–emitter sidewall | 41
<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4 p-I-N energy band diagrams along section A-A' of Figure 3.3 under a) equilibrium, and b) 1.2 volts of forward bias</td>
<td>42</td>
</tr>
<tr>
<td>3.5 Small-signal hybrid-π equivalent circuit for bipolar transistors</td>
<td>45</td>
</tr>
<tr>
<td>3.6 Quasi-static frequency analysis flowchart for bipolar transistors</td>
<td>45</td>
</tr>
<tr>
<td>3.7 Rigorous and Einstein relation diffusion coefficients versus electric field in an homogeneous GaAs slab</td>
<td>51</td>
</tr>
<tr>
<td>3.8 Local energy balance equation solution flowchart</td>
<td>53</td>
</tr>
<tr>
<td>3.9 Energy relaxation time in GaAs as a function of electron energy [90]</td>
<td>54</td>
</tr>
<tr>
<td>3.10 Mobility shape function as a function of electron energy [90]</td>
<td>54</td>
</tr>
<tr>
<td>4.1 Analytic and numeric I-V characteristics of a 1-D homojunction diode</td>
<td>57</td>
</tr>
<tr>
<td>4.2 I-V characteristics of 1-D and 2-D homojunction diodes</td>
<td>57</td>
</tr>
<tr>
<td>4.3 Electron concentration in 2-D GaAs homojunction diode</td>
<td>59</td>
</tr>
<tr>
<td>4.4 Short-circuit current decay in a GaAs solar cell</td>
<td>60</td>
</tr>
<tr>
<td>4.5 Transient collector current in an Al_xGa_{1-x}As HBT</td>
<td>62</td>
</tr>
<tr>
<td>4.6 Electron energy and effective velocity in a GaAs n⁺-i-n⁺ diode</td>
<td>63</td>
</tr>
<tr>
<td>4.7 Electron concentration and mobility in the GaAs n⁺-i-n⁺ diode</td>
<td>63</td>
</tr>
<tr>
<td>5.1 Typical heterojunction bipolar transistor</td>
<td>66</td>
</tr>
<tr>
<td>5.2 Electron concentration in an HBT at V_{BE} = 1.5 V, V_{CE} = 2.0 V</td>
<td>67</td>
</tr>
<tr>
<td>5.3 Recombination rate in an HBT at V_{BE} = 1.5 V, V_{CE} = 2.0 V</td>
<td>69</td>
</tr>
<tr>
<td>5.4 Base and collector current densities in an HBT</td>
<td>70</td>
</tr>
<tr>
<td>5.5 Cutoff frequency behavior of an HBT</td>
<td>70</td>
</tr>
<tr>
<td>5.6 Current gain dependence on emitter mesa–base contact spacing. Points shown are for J_C = 10^4 amps/cm²</td>
<td>71</td>
</tr>
<tr>
<td>5.7 Recombination rate in passivated HBT with L = 1.0 μm, V_{BE} = 1.35 V, V_{CE} = 2.0 V</td>
<td>72</td>
</tr>
</tbody>
</table>
5.8 Recombination rate in passivated HBT with $L = 0.2 \, \mu m, \; V_{BE} = 1.35 \, V, \; V_{CE} = 2.0 \, V$.................................................................73

5.9 Base currents in HBT's with abrupt and graded base–emitter heterojunctions. The experimental data are from [95].................................76
ABSTRACT

Dodd, Paul Emerson. M.S.E.E., Purdue University, December 1989. Simulation of Heterojunction Bipolar Transistors in Two Dimensions. Major Professor: Dr. Mark S. Lundstrom.

This work describes the formulation and development of a two-dimensional drift-diffusion simulation program for accurate modeling of heterojunction bipolar transistors (HBT's). The model described is a versatile tool for studying HBT's, allowing the user to determine the terminal characteristics and physical operation of devices. Nonplanar structures can be treated, response to transient conditions can be computed, and the high-frequency characteristics of transistors may be projected. The formulation of an electron energy balance equation is presented and included in the model in an attempt to more accurately compute high-field transport characteristics. The model is applied to some common design questions and experimental results are reproduced.
CHAPTER I: INTRODUCTION

1.1. Introduction and Rudiments of HBT Operation

Ever since the debut of the bipolar junction transistor (BJT) more than forty years ago, engineers have sought methods of raising the ultimate limits of device capability. At first, solid state engineers were content to improve performance levels by improving material quality and fabrication precision, and by scaling toward smaller devices. The use of self-aligned processing techniques and polysilicon emitters has increased silicon BJT performance considerably [1,2]. Cutting-edge silicon BJT's are becoming less and less governed by external parasitics such as high contact resistances and are approaching fundamental limits determined by physical properties of the semiconductor. These constraints are imposed by such factors as transit times and intrinsic junction capacitances. Thus, today's device designer appears to have two paths for improved performance: 1) design a new device with a higher performance potential, or 2) use a material with better physical properties.

The heterojunction bipolar transistor (HBT), while hardly a new device, does indeed have a higher ultimate performance potential than the BJT [3]. In addition, the most popular material for HBT work is currently the Al$_x$Ga$_{1-x}$As–GaAs system, which has physical properties that may be exploited to give higher limits than silicon.

The basic design principle of the HBT is to use an emitter with a wider bandgap than the base and collector. This has the effect of suppressing minority carrier injection from the base into the emitter of the transistor, while minority carriers are readily injected from the emitter into the base. Figure 1.1 illustrates carrier injection across an np$^+$ homojunction and an Np$^+$ heterojunction (the capital letter is used to indicate the wider bandgap...
Figure 1.1 Minority carrier injection across a) np+ homojunction, and b) Np+ heterojunction.
Figure 1.2 shows the current components in an Npn HBT operating in the forward active bias regime. For simplicity, the device is assumed to have a graded emitter-base metallurgical junction. There are four components in the base-emitter region: $I_n$ is the electron current injected into the base, $I_p$ is the hole current injected into the emitter, $I_r$ is the current due to recombination in the quasi-neutral base, and $I_s$ is the current due to recombination in the base-emitter depletion region. The terminal currents can be written in terms of these current components:

$$I_e = I_n + I_p + I_s, \quad (1.1)$$
$$I_b = I_p + I_r + I_s, \quad (1.2)$$
$$I_c = I_n - I_r. \quad (1.3)$$

Current gain is then given by:

$$I_c = \frac{I_n - I_r}{I_b}. \quad (1.4)$$
A statistic of interest for bipolar transistors is the common emitter current gain:

\[
\beta = \frac{I_c}{I_b} = \frac{I_n - I_t}{I_p + I_r + I_s}.
\]  

(1.4)

Neglecting recombination currents, the current gain is simply the ratio of the injected electron current to the injected hole current. A modified pn homojunction theory may be used to express the injected currents at a base-emitter bias of \(V_{BE}\) [4]:

\[
I_n = \frac{q A D_{nb}}{W_B} \frac{n_{ib}^2}{N_{AB}} \exp \left( \frac{qV_{BE}}{kT} \right)
\]

(1.5)

and

\[
I_p = \frac{q A D_{pe}}{W_E} \frac{n_{ie}^2}{N_{DE}} \exp \left( \frac{qV_{BE}}{kT} \right)
\]

(1.6)

where \(N_{DE}\) and \(N_{AB}\) are dopant densities, \(D_{pe}\) and \(D_{nb}\) are minority carrier diffusion lengths, and \(n_{ie}^2\) and \(n_{ib}^2\) are intrinsic carrier concentrations in the emitter and base, respectively. \(W_E\) and \(W_B\) are the emitter and base widths, and \(A\) is the junction area. In writing (1.6) it has been assumed that the emitter width is much less than the hole diffusion length in the emitter, since HBT's almost invariably have shallow emitters. Taking the ratio of the injected currents gives the current gain:

\[
\beta = \frac{I_n}{I_p} = \frac{N_{DE} W_E D_{nb} n_{ib}^2}{N_{AB} W_B D_{pe} n_{ie}^2}.
\]

(1.7)

Using a bandgap narrowing formulation, the intrinsic carrier concentrations may be related [5]:

\[
\frac{n_{ib}^2}{n_{ie}^2} = \exp \left( \frac{\Delta E_G}{kT} \right).
\]

(1.8)

Here \(\Delta E_G\) is the bandgap difference between the base and the emitter. Thus the current gain is simply

\[
\beta = \frac{N_{DE} W_E D_{nb}}{N_{AB} W_B D_{pe}} \exp \left( \frac{\Delta E_G}{kT} \right).
\]

(1.9)
For the BJT, the exponential term reduces to 1 and the familiar equation for the current gain in an ideal transistor results. The "WD products" are likely to be of approximately the same order of magnitude, so that the emitter–base doping ratio governs the current gain of the BJT. It is for this reason that BJTs have heavily doped emitters and lightly doped bases. Unfortunately, the light doping of the base leads to high base resistance, which couples with the large capacitance of the asymmetrically-doped base–emitter junction to produce a large RC time constant that seriously degrades the high frequency performance of the BJT. In an HBT, even a small bandgap difference of a few kT is sufficient to produce high current gains due to the exponential dependence of $\beta$ on $\Delta E_C$. The exponential factor allows the HBT to maintain acceptable $\beta$ values when using an inverted doping profile– the base can actually be doped more heavily than the emitter. In an Al$_x$Ga$_{1-x}$As–GaAs HBT, the electron mobility is much higher than the hole mobility, so that the diffusion constant ratio in (1.9) will also enhance the gain. The BJT relies on the heavily doped emitter to maintain injection efficiency, whereas good injection efficiency is assured in the HBT by the heterojunction. The performance advantage of the HBT lies primarily in the ability to use an inverted doping profile, which lowers the base resistance and thereby improves the high speed capabilities of the transistor.

In practice, high current gains are harder to achieve than is indicated by this simple development. The recombination currents play a very important role in determining the device characteristics. The presence of spikes in the conduction band profile further reduces current gain. A conduction band discontinuity, $\Delta E_C$ as shown in Figure 1.1, suppresses the injection of electrons into the base [6], and the injected electron current may be written as:

$$I_n = \frac{q A D_{nb}}{W_B} \frac{n_{ib}^2}{N_{AB}} \exp\left(\frac{qV_{BE} - \Delta E_C}{kT}\right).$$

(1.10)

Current gain is then given by

$$\beta = \frac{N_{DE} W_E D_{nb}}{N_{AB} W_B D_{pe}} \exp\left(\frac{\Delta E_C - \Delta E_V}{kT}\right),$$

(1.11)

or, since $\Delta E_C - \Delta E_V = \Delta E_V$:
$$\beta = \frac{N_{DE}}{N_{AB}} \frac{W_E}{W_B} \frac{D_{nb}}{D_{pe}} \exp \left( \frac{\Delta E_V}{kT} \right).$$

(1.12)

Most contemporary HBT's incorporate some type of grading scheme to reduce or eliminate conduction band discontinuities for maximum emitter injection efficiency.

1.2. History of the HBT

The inherent performance advantages of heterojunction bipolar transistors over conventional bipolar junction transistors have been recognized from the outset of transistor design [7,8]. Until recently, however, the precise control of dopant and material junctions necessary for fabrication of HBT's was not available. The advent of epitaxial growth techniques brought about the eventual realization of the HBT.

The first HBT's to be built successfully were reported in 1969 by Jadus and Feucht [9]. These transistors utilized a GaAs emitter grown epitaxially by vapor deposition on Ge. Although the base–emitter junctions exhibited properties which were far from ideal, the transistors did display a current gain of up to 15.

Liquid phase epitaxy (LPE) as a fabrication technique was demonstrated by Nelson in 1963 [10], and was extended to Al$_x$Ga$_{1-x}$As heterojunction growth in 1967 [11]. By 1972, some of the first LPE–grown Al$_x$Ga$_{1-x}$As–GaAs HBT's were described [12]. Current gains of 25 were achieved at low current densities in these large devices. More recently, Bailbe et al. [13] reported LPE–grown Al$_x$Ga$_{1-x}$As–GaAs HBT's with current gains as high as 850 and cutoff frequencies of around 1 GHz as determined by s–parameter measurements.

The current growth techniques of choice are molecular beam epitaxy (MBE) and metalorganic chemical vapor deposition (MOCVD). Both of these techniques offer extremely precise control (on the order of monolayers) of the doping and material profiles [14–16]. MBE–grown GaAs–Ge HBT's with current gains of 45 have just been reported [17]. The base–emitter ideality factor of these transistors was 1.04 at room temperature, which can be contrasted with the ideality factor of 2.1 reported by Jadus in the first HBT's. This is an indication of the excellent material interface quality which can be achieved with MBE.
The majority of recent HBT work has concentrated on the Al\textsubscript{x}Ga\textsubscript{1-x}As system. As the fabrication processes have become more refined, these HBT's have shown continuously improved high speed performance [18-26] (Figs. 1.3 and 1.4). Cutoff frequencies as high as 105 Ghz have been achieved in transistors with novel collector designs [27]. In digital circuit applications, propagation delays of less than 6 ps/gate have been observed in emitter-coupled logic ring oscillators [28]. While these results are a couple of orders of magnitude better than silicon BJT's, fundamental physical limits such as the base transit time are already becoming important in determining device characteristics.

At the present time, one of the hottest areas of HBT research is the use of alternate material systems. The most important of these is probably the growth of strained Ge\textsubscript{x}Si\textsubscript{1-x} double-heterojunction bipolar transistors (DHBT's). In these devices, a very thin base layer of Ge\textsubscript{x}Si\textsubscript{1-x} (or a superlattice base with many thin Ge\textsubscript{x}Si\textsubscript{1-x} layers [29]) is sandwiched between a Si collector and emitter. The Ge\textsubscript{x}Si\textsubscript{1-x} layer must be thin to prevent the formation of dislocations at the interfaces [30], due to the high degree of lattice mismatch. Several groups have reported successful fabrication of these HBT's, with current gains as high as 400 in devices with inverted base-emitter doping profiles [31,32]. Si–Ge\textsubscript{x}Si\textsubscript{1-x} HBT's are particularly exciting because of their ability to take immediate advantage of highly developed silicon processing techniques. Projections of the speed potential of Si–Ge\textsubscript{x}Si\textsubscript{1-x} HBT's have suggested theoretical cutoff frequencies in excess of 75 GHz [33].

Several other material systems have also been receiving much attention of late. Among these are the In\textsubscript{0.52}Al\textsubscript{0.48}As–In\textsubscript{0.53}Ga\textsubscript{0.47}As and InP–In\textsubscript{1-x}Ga\textsubscript{x}As systems [34–36]. These systems are especially promising for several reasons. The electron mobilities in these materials are high, and the narrower bandgaps of these materials (with respect to GaAs) indicate that in digital circuits, a smaller voltage swing is required to switch between states. This should lead to both faster switching times and reduced power consumption. In addition, large intervalley separations (Γ→L and Γ→X) in the conduction band of In\textsubscript{1-x}Ga\textsubscript{x}As allow ballistic injection of electrons into the base of these transistors without significant intervalley transfer. Cutoff frequencies of 165 Ghz and propagation delays of less than 15 pS have already
Figure 1.3  Unity current gain cutoff frequency as a function of year reported in representative HBT's.

Figure 1.4  Gate propagation delay as a function of year reported in ring oscillators fabricated using HBT's.
been observed in InP-In$_{1-x}$Ga$_x$As HBT's, and cutoff frequencies higher than 380 GHz have been predicted [37].

1.3. History of HBT Simulation

While progress in design and fabrication of HBT's has shot forward at a somewhat alarming rate, the difficult task of modeling these devices has generally lagged behind. One-dimensional computer-aided DC solutions to homojunction transistors were demonstrated by Gummel in 1964 [38]. By the time that MBE and MOCVD made routine fabrication of HBT's feasible, silicon device simulation was widely used and understood (see [39,40] and references therein). In 1977 Sutherland and Hauser were the first to use numerical techniques to analyze heterojunction devices, in this case solar cells [41]. It was shown that the basic formulation for homojunction devices was easily generalized to include the effects of a position-dependent band structure. This was extremely important, for it allowed existing drift-diffusion homojunction codes to be easily modified for the simulation of heterostructures. The formulation was later expanded to treat degenerate semiconductors via Fermi-Dirac statistics for heterostructures in equilibrium and nonequilibrium [42-44]. The effects of dopant deionization, nonparabolicity of the Γ valley, and multiple conduction band minima have also been addressed [45].

Application of computer simulation to the HBT was reported in 1982 by Asbeck et al., who converted an existing one-dimensional steady state npn BJT model to treat HBT's [46]. The model included field-dependent mobilities to reproduce the proper steady state velocity-field characteristics, and there was a primitive method for including velocity overshoot effects. The program did not, however, handle degenerate carrier statistics or partial deionization of dopants.

In 1984, Toshiba reported the use of a one-dimensional transient drift-diffusion simulator for evaluation of switching performance in HBT's [47]. This program allowed the inclusion of external base and collector resistors in order to give realistic predictions of switching times. Velocity overshoot was neglected, under the assumption that overshoot occurs only in a localized area and does not adversely affect the total device characteristics. 1985 saw the
Toshiba group incorporating their device simulator directly into their circuit simulator, allowing simulation of complete circuits without the errors normally introduced by modeling devices as lumped circuit elements [48].

Researchers at NTT described a two-dimensional DC drift-diffusion modeling program in 1984 and used it to confirm the high frequency potential of Al\textsubscript{x}Ga\textsubscript{1-x}As-GaAs HBT's [49]. While velocity overshoot was again neglected, some simulations were performed assuming ballistic transport across the base of the transistor.

Although the drift-diffusion model is the most widely used and understood tool for semiconductor device simulation, it unfortunately fails to predict nonstationary transport effects. As a derivative of the Boltzmann Transport Equation (BTE), it also fails to reflect the quantum-mechanical nature of carrier transport. The continuous push toward smaller devices has led to a need to address these shortcomings, and to the development of alternative modeling techniques.

Monte Carlo techniques have evolved as a way to model nonstationary transport phenomena [50]. Several groups have used Monte Carlo techniques to model these effects and determine how best to take advantage of them [51–53]. Unfortunately, since the method involves keeping statistics on a large number of carriers undergoing random collisions, the Monte Carlo method is very expensive in terms of computer time. The simulation of a complete transistor requires tracking a prohibitive number of carriers in order to attain statistical significance. This typically limits the Monte Carlo technique to use as an aid in studying only part of the transistor, for instance the emitter-base junction.

Another method of modeling nonstationary transport is the use of balance equations (also known as the hydrodynamic or energy transport model). In this method, the first three moments of the BTE are taken, yielding the particle, momentum, and energy conservation equations. To solve these equations it is generally necessary to make many assumptions (for instance invocation of the relaxation time approximation). An interesting comparison of results making use of differing degrees of assumptions has recently been published [54]. As the drift-diffusion model is pushed to its limits, more people are trying the hydrodynamic method of solution [55–57].
Given that several alternative methods and levels of sophistication for device simulation exist, one is forced to ask the inevitable question— which is the best method and how much do I need? The craft of device simulation requires the modeler to have an intimate knowledge of the limits of the available tools and to temper this with an understanding of the cost effectiveness of using each tool. The trick is to use the simplest model which will correctly predict device behavior. The question to be addressed here is what is the minimum model necessary for today's HBT designer. It is evident that a two-dimensional model is essential, as perimeter effects and two-dimensional carrier flow are important aspects of HBT's. A good method for treating surface recombination is also of great importance, as the surface properties often play a large role in determining HBT performance. It is clear that predicting high-field effects is also a critical part of HBT simulation. In particular, velocity saturation must be handled, while velocity overshoot is probably somewhat less influential in determining characteristics of the typical HBT. As bandgap narrowing effects in highly-doped materials become understood and quantified they must also be included in useful models, as they can drastically affect device parameters such as current gain. High speed performance of HBT's is of primary importance to the designer and a transistor simulator must possess the capability to predict either cutoff frequencies or switching times. The venerable drift-diffusion model can satisfy all of these needs if suitably enhanced, and is easily the most cost-effective method in terms of development and computer time. A well equipped simulation toolbox should therefore include such a drift-diffusion model as the workhorse. For studying the effect of non-stationary transport in localized areas of the device, a Monte Carlo simulator beautifully complements the drift-diffusion model. To closely examine phenomena such as tunneling and reflection of carriers that may occur in the presence of conduction band spikes, a quantum mechanical simulation tool is helpful.

Regardless of the modeling methodology used, the ultimate responsibility will always rest on the user of the simulator to intelligently interpret the results and know when the assumptions inherent to the method are being violated. Otherwise, as was pointed out by Tang and Laux [58], "... computationally sophisticated 2-D or even 3-D device simulations
are rendered merely expensive, and perhaps misleading, curve-fitting programs."

1.4. Research Objectives and Thesis Organization

The objective of this research is to develop an existing two-dimensional heterostructure solar cell analysis program into an effective tool for study and design of heterojunction bipolar transistors. The code will then be exercised to show its suitability for HBT analysis and design. Control structure modifications will be necessary for investigation of transistors. Facilities for high frequency and switching evaluation of transistors will also be added. Modeling of high-field phenomena will also be addressed.

Chapter 1 presented a brief introduction to the operation of the heterojunction bipolar transistor, and included a synopsis of the evolution of the HBT. A brief history and overview of different modeling techniques was also given.

Chapter 2 will discuss the mathematical formulation of the simulation tool used throughout this work, as well as the techniques used to arrive at a solution. The physical parameters used in the model will be surveyed.

The extensions of this work to the model will be described in Chapter 3. This includes modifications to facilitate bipolar transistor analysis, a method for treating nonrectangular geometries, quasi-static frequency evaluation, time-dependent solutions, and a high-field model.

Verification of the model is the topic of Chapter 4, and several simulation results will be presented and compared with theoretical expectations. Transient solutions and a high-field transport example will be studied.

Chapter 5 will consist of applications of the model to HBT design and the results will be compared to recent experimental data. A summary and recommendations will comprise Chapter 6.
CHAPTER 2: MODEL FORMULATION

2.1. Introduction

As has already been discussed, after twenty-five years of use the drift-diffusion model is still an effective method for analyzing semiconductor devices. A two-dimensional drift-diffusion silicon solar cell analysis program (SCAP2D) was developed at Purdue by Lundstrom and Gray several years ago [59,60]. A couple of years later, Schuelke wrote a one-dimensional drift-diffusion program (PUPHS1D) for evaluation of compositionally nonuniform structures [61]. The heterostructure formulation and material models of Schuelke were combined with the solution techniques of the 2-D silicon code by DeMoulin to form the Purdue University Program for Heterostructure Simulation in 2 Dimensions (PUPHS2D) [62].

This chapter covers the mathematical formulation of PUPHS2D: the equations to be solved, the numerical technique applied for solution, and a brief survey of the material models used in the code.

2.2. The Steady-State Semiconductor Equations

The behavior of carriers in a semiconductor device is described by the so-called basic semiconductor device equations: the Poisson equation and the hole and electron current continuity equations. In steady-state conditions, these equations take the form:

\[ \nabla \cdot \vec{D} = \rho , \]  
\[ \nabla \cdot \vec{J}_n = q ( R - G ) , \]  
\[ \nabla \cdot \vec{J}_p = q ( G - R ) . \]

In these equations, \( \vec{D} \) is the displacement field, \( \rho \) is the volume charge density, \( J_n \) and \( J_p \) are the hole and electron current densities, and \( q \) is the...
magnitude of charge of an electron, and $R$ and $G$ are the recombination and generation rates per unit volume.

To solve these equations simultaneously, it is necessary to include the constitutive relationship for each. For the Poisson equation, this relates the charge density to the doping and carrier concentrations:

$$\rho = q (N_D - N_A + p - n) ,$$

where $N_A$ and $N_D$ are the ionized acceptor and donor densities, and $n$ and $p$ are the electron and hole densities. Substitution of this result into the Poisson equation gives

$$\nabla \cdot \vec{D} = \nabla \cdot \varepsilon \vec{E} = q (N_D - N_A + p - n) .$$

For compositionally uniform semiconductors this reduces to the familiar result (using $\vec{E} = -\nabla V$):

$$\nabla^2 V = \frac{q}{\varepsilon} (N_A - N_D + n - p) .$$

The constitutive relationships required for the current continuity equations are the aptly-named drift-diffusion equations, which describe the electron and hole currents in terms of a drift and a diffusion component. For homostructures the drift-diffusion current equations are:

$$\vec{J}_n = -qn\mu_n\nabla V + qD_n \nabla n ,$$

$$\vec{J}_p = -qp\mu_p\nabla V - qD_p \nabla p .$$

Compositional nonuniformity affects the drift component of the currents by introducing a quasi-field due to band-edge variations. The diffusion component of the currents is also modified because of the nonuniform densities-of-states. Fortunately both effects can be modeled by appropriate definitions of hole and electron "band parameters" as quasi-potentials [44]. The current equations are then written as:

$$\vec{J}_n = -qn\mu_n V(V + V_n) + qD_n \nabla n ,$$

$$\vec{J}_p = -qp\mu_p V(V - V_p) - qD_p \nabla p .$$

$V_n$ and $V_p$ are the band parameters and are defined by:
\[ qV_n = \chi - \chi_{\text{ref}} + kT \ln \left( \frac{N_C}{N_{C\text{ref}}} \right), \tag{2.11} \]

\[ qV_p = -(\chi - \chi_{\text{ref}}) - (E_G - E_{G\text{ref}}) + kT \ln \left( \frac{N_V}{N_{V\text{ref}}} \right). \tag{2.12} \]

In these equations \( \chi \) is the electron affinity, \( N_C \) is the conduction band density of states, \( N_V \) is the valence band density of states, and \( E_G \) is the energy gap of the semiconductor. The subscript "ref" indicates that the parameter is evaluated in some reference material. For degenerate semiconductors, the inclusion of Fermi-Dirac statistics in (2.11) and (2.12) is accomplished through the use of the generalized Einstein relation, which for the electron diffusion coefficient produces:

\[ D_n = \frac{kT}{q} \frac{\frac{n}{dn}/d\eta_c}{\mu_n}, \tag{2.13} \]

where \( \eta_c = (E_F - E_C)/kT \) and \( n = N_C \mathcal{F}_{1/2}(\eta_c) \). \( \mathcal{F}_{1/2}(\eta_c) \) is the Fermi-Dirac integral of order 1/2. PUPHS2D presently assumes Boltzmann statistics.

We now have three equations in terms of three unknowns (\( p, n, \) and \( V \)) and a large number of constants which are determined by the actual device being modeled. The problem is still not fully specified; boundary conditions are needed to arrive at a unique solution. There are three distinct types of surfaces at which we may need boundary conditions: ohmic contacts, free surfaces, and lines of symmetry. Figure 2.1 illustrates typical boundary surfaces for a transistor. The boundary conditions at each of these types of surface are as follows:

**Ohmic contacts:**

\[ V = V_A + \text{constant}, \tag{2.14} \]

\[ n = n_0, \tag{2.15} \]

\[ p = p_0. \tag{2.16} \]

Here, \( n_0 \) and \( p_0 \) are the equilibrium carrier concentrations and \( V_A \) represents some bias which is being applied to the contact. As an example, to determine common-emitter characteristics for a bipolar transistor operating at some
Figure 2.1 Boundary surface types for a typical transistor.
given $V_{BE}$ and $V_{CE}$, an arbitrary voltage can be placed on the emitter, and the base contact voltage boundary condition is $V = V_{BE} + \text{constant}$, and the collector contact boundary condition is similarly $V = V_{CE} + \text{constant}$.

Free surfaces:

\[ \nabla V \cdot \hat{n} = \frac{q N_{SS}}{\epsilon_s}, \quad (2.17) \]
\[ J_n \cdot \hat{n} = -q R_S, \quad (2.18) \]
\[ J_p \cdot \hat{n} = q R_S. \quad (2.19) \]

$\hat{n}$ is a unit vector normal to the surface in question, $\epsilon_s$ is the dielectric constant at the surface, and $N_{SS}$ is an effective surface charge density which is related to the actual surface charge density by:

\[ q N_{SS} = \rho_s - \overrightarrow{D}_{ext} \cdot \hat{n}, \quad (2.20) \]

where $\rho_s$ is the actual surface charge density and $\overrightarrow{D}_{ext}$ is the external displacement field. In PUPHS2D, $\overrightarrow{D}_{ext}$ is assumed to be zero and the user may specify $\rho_s$. $R_S$ in (2.18) and (2.19) is the surface recombination rate, which will be discussed in greater detail later in this chapter.

Lines of symmetry:

\[ \nabla V \cdot \hat{n} = \nabla p \cdot \hat{n} = \nabla n \cdot \hat{n} = 0. \quad (2.21) \]

With the semiconductor equations now subjected to the proper boundary conditions, the problem is fully specified and ready for solution.

2.3. Numerical Analysis Techniques

2.3.1. Discretisation

Except in the most simplified cases, closed-form solutions to (2.1-2.3) do not exist. The highly nonlinear nature of the equations to be solved suggests the use of approximate numerical methods. The well known finite difference method is chosen for simplicity and accuracy [63-65]. Implicit to
the choice of a difference method is the concept of a grid. Instead of finding a continuous solution for p, n, and V, the device domain is covered with a finite difference grid (mesh) and the equations are solved at each grid point. A two-dimensional finite difference grid is illustrated in Figure 2.2. Solution of the finite difference problem gives the values of p, n, and V only at the grid points. This is in contrast with the finite element method of solution [66], in which the device domain is broken into elements over which the variable is assumed to have some polynomial form and the polynomial coefficients are the solution variables. This is a somewhat immaterial difference, however, since for the final solution the polynomials must still be evaluated at discrete positions. A more important difference is the greater ease with which the

Figure 2.2 Two-dimensional finite difference grid for a transistor.
finite element method will treat nonplanar geometries. This ease is gained at
the high price of code complexity. Finite difference codes are easier to
understand and maintain and a procedure for modeling nonplanar geometries will be presented in Chapter 3.

The finite difference method itself involves writing the differential
equations at each node in terms of the variables at the node and neighboring
nodes. Derivatives in the differential equation are replaced by difference approximations, hence the name of the method. For optimal accuracy, center
differences are used. The derivative of a function is evaluated midway
between two adjacent nodes (hence the appellation center difference) by
finding the slope of a line passing through the function values at the nodes.
Mathematically, this is:

\[
\frac{df}{dx} \bigg|_{x_{i+1/2}} = \left[ \frac{df}{dx} \right]_h \bigg|_{x_{i+1/2}} = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i},
\]

where the brackets and subscript of \( h \) indicate the discretised derivative. Figure 2.3 illustrates the center difference approximation to a first order derivative. For the semiconductor equations, it is necessary to discretise the second derivative as well. This is done by simply reapplying the same process, and since the first derivatives are known midway between nodes, the center difference method will give the second derivative evaluated at the node itself. Thus:

\[
\left[ \frac{d^2f}{dx^2} \right]_h \bigg|_{x_i} = \frac{\left[ \frac{df}{dx} \right]_h \bigg|_{x_{i+1/2}} - \left[ \frac{df}{dx} \right]_h \bigg|_{x_{i-1/2}}}{x_{i+1/2} - x_{i-1/2}}
\]

\[
= \frac{2f(x_{i-1})}{h_L(h_L + h_R)} - \frac{2f(x_i)}{h_L h_R} + \frac{2f(x_{i+1})}{h_R(h_L + h_R)}.
\]

In this equation \( h_L \) and \( h_R \) are the node spacings, given by:

\[
h_L = x_i - x_{i-1},
\]

\[
h_R = x_{i+1} - x_i.
\]
It is easily seen that closely spaced grid points lead to a better approximation to the derivatives. If the function is nearly linear, the finite difference is fairly exact and excess grid points do not increase the accuracy of the solution very much. Since three equations are solved at each grid point, increasing the number of grid points greatly increases the execution time of the program. The optimal mesh is therefore one in which the mesh points are closely spaced where the variables are rapidly changing, and which does not waste nodes in uninteresting areas of the device. Choosing a good finite difference mesh is of paramount importance both to the speed and the accuracy of the computed solution.

Using (2.22) and (2.23), the semiconductor equations may now be discretised. In two dimensions and for interior nodes, (2.1–2.3) become:

![Graph](image)

**Figure 2.3** Graphic illustration of the center difference approximation to a first order derivative.
$f_v = \frac{\varepsilon_L V_L}{h_L(h_L + h_R)} + \frac{\varepsilon_R V_R}{h_R(h_L + h_R)} - \varepsilon_{ij} V_{ij} \left( \frac{1}{h_L h_R} + \frac{1}{h_B h_T} \right) + \frac{\varepsilon_B V_B}{h_B(h_T + h_B)}$

\[ + \frac{\varepsilon_T V_T}{h_T(h_T + h_B)} + \frac{\rho_{ij}}{2} = 0 \quad (2.26) \]

$f_p = \frac{2(j_{pR} - j_{pL})}{h_L + h_R} + \frac{2(j_{pT} - j_{pB})}{h_T + h_B} + q(R_{ij} - G_{ij}) = 0 \quad (2.27)$

$f_n = \frac{2(j_{nR} - j_{nL})}{h_L + h_R} + \frac{2(j_{nT} - j_{nB})}{h_T + h_B} + q(G_{ij} - R_{ij}) = 0 \quad (2.28)$

In (2.26), $\varepsilon_{ij}^*$ is a weighted dielectric constant given by:

$\varepsilon_{ij}^* = \frac{h_T h_B \left( \varepsilon_{R} h_L + \varepsilon_{L} h_R \right) + h_L h_R \left( \varepsilon_{T} h_B + \varepsilon_{B} h_T \right)}{h_L h_R + h_T h_B} \quad (2.29)$

Subscripts of $ij$ indicate that the quantity is evaluated at node $ij$, while $L$, $R$, $B$, and $T$ refer to the neighboring nodes as depicted in Figure 2.4. The current densities and dielectric constants, however, are evaluated midway between nodes and thus for these the subscript $R$ refers to the point midway between node $ij$ and its neighbor to the right, and similarly for the other subscripts.

For free surface boundary nodes, (2.26–2.28) can still be used, if proper definitions are made for node variables outside of the device domain. For instance, consider the top boundary as shown in Figure 2.5. In this case, the top node is outside of the device. If a phantom node (denoted by the PH subscript) is introduced, then the boundary condition equation (2.17) can be discretised:

\[ \frac{V_{PH} - V_B}{2h_B} = \frac{q N_{SSij}}{\varepsilon_{ij}} \quad (2.30) \]

Or, solving for $V_{PH}$:

\[ V_{PH} = V_B + \frac{2h_B q N_{SSij}}{\varepsilon_{ij}} \quad (2.31) \]
Wherever $V_T$ occurs in (2.26) it is replaced by $V_{PH}$, and all occurrences of $h_T$ are replaced by $h_B$. In (2.29), $h_T$ is also replaced by $h_B$, and $e_{ij}$ is substituted for $\varepsilon_T$.

The current boundary conditions can also be discretised in a similar fashion for the top free surface to produce:

$$f_p = \frac{2(J_{pR} - J_{pL})}{h_L + h_R} + \frac{2(q R_{Sij} - J_{pB})}{h_B} + q (R_{ij} - G_{ij}) = 0,$$

(2.32)

$$f_n = \frac{2(J_{nR} - J_{nL})}{h_L + h_R} + \frac{2(-q R_{Sij} - J_{nB})}{h_B} + q (G_{ij} - R_{ij}) = 0.$$

(2.33)

The left, bottom, and right surfaces have completely analogous boundary equations.

The drift-diffusion transport equations are all that remain to be discretised. These equations can be discretised using center differences, but this method has been shown to be conditionally unstable [67]. If the potential variation between nodes is greater than twice the thermal voltage, the difference equation produces negative coupling between adjacent carrier densities. This will almost certainly cause the solution to oscillate and may lead to clearly non-physical solutions with negative carrier densities. Scharfetter and Gummel [67] demonstrated that by multiplying the transport equations by an appropriate factor and integrating, positive coupling (and hence stability) can be achieved. The normalized Scharfetter–Gummel current equation for the $x$–directed hole current evaluated midway between node $ij$ and its neighbor to the right (refer to Fig. 2.4) is:

$$J_{pXR} = -\mu_{pR} \frac{\Delta V_R}{h_R} \left( \frac{P_R e^{\Delta V_R} - P_{ij}}{e^{\Delta V_R} - 1} \right),$$

(2.34)

where $\Delta V_R = V_R - V_{ij}$. Like potential, the carrier concentrations are evaluated at the node points, while the hole mobility is evaluated midway between nodes. The astute reader will have noticed that (2.34) is written in terms of the hole mobility, and that the hole diffusion coefficient is nowhere to be seen. This is because the Einstein relationship,
Figure 2.4  Node spacing, subscript definitions, and evaluation points for interior of a two-dimensional finite (center) difference mesh.

Figure 2.5  Node spacing and evaluation points for top boundary of a two-dimensional finite (center) difference mesh.
\[
D_p = \frac{kT}{q}\mu_p , \quad (2.35)
\]

has been assumed to hold true. PUPHS2D presently assumes Boltzmann statistics, so (2.35) is used rather than the generalized form shown in (2.13). Although it is customary to assume the Einstein relationship, the Scharfetter–Gummel equations can be written without this assumption and it will be necessary to do so later when high-field transport is addressed. Assumptions which are integral to this discretization method are that the current, mobility, and electric field are roughly constant between nodes. Thus, rather than requiring potential variations to be less than \(2V_T\), the restriction is only that the voltage vary linearly between nodes.

### 2.3.2. Solution

Once the problem is completely specified mathematically, the solution is sought. Given the highly nonlinear nature of the equations to be solved, the Newton iterative technique is chosen for linearizing the problem [68]. In matrix form, the Newton technique takes the form:

\[
\bar{J}(\bar{u}^k)\Delta\bar{u}^{k+1} = -\bar{F}(\bar{u}^k) . \quad (2.36)
\]

In (2.36), \(\bar{F}\) is the vector of the three equations at each node, and \(\bar{u}^k\) is the vector of unknowns at each node after the kth Newton iteration, or:

\[
\bar{F} = \begin{bmatrix}
\left[v_{1,1}\right] \\
\left[p_{1,1}\right] \\
\left[n_{1,1}\right] \\
\vdots \\
\left[v_{m,n}\right] \\
\left[p_{m,n}\right] \\
\left[n_{m,n}\right]
\end{bmatrix}, \quad \bar{u}^k = \begin{bmatrix}
\left[v_{1,1}\right]^k \\
\left[p_{1,1}\right] \\
\left[n_{1,1}\right] \\
\vdots \\
\left[v_{m,n}\right] \\
\left[p_{m,n}\right] \\
\left[n_{m,n}\right]
\end{bmatrix} . \quad (2.37)
\]

\(\bar{J}\) is the Jacobi matrix and contains the derivatives of each of the nodal equations with respect to each of the nodal variables. The Jacobian is a very large matrix, its dimensions being determined by the number of nodes in the \(x\) direction and the number of nodes in the \(y\) direction. For a mesh which is
50 nodes square, the full Jacobi matrix contains over 56 million elements. Since the variables at any given node only depend on the four adjacent nodes, the Jacobi matrix is sparse and block tridiagonal, so that most of the elements need not be stored.

At each Newton iteration, the problem is just a linear system of equations of the form $Ax = b$, where $A$ is the Jacobi matrix, $b$ is the vector of the equations to be solved evaluated at the solution to the last iteration, and $x$ is a vector of corrections to the variables. The direct inversion (i.e. brute force) method is applied for solution of the linear system. Inversion of the Jacobi matrix, even in banded form, requires a very large amount of memory space, but the direct technique is quite robust. For improved convergence [60], the actual matrix equation is formulated in terms of the hole and electron quasi-fermi levels $\phi_n$ and $\phi_p$. The carrier concentrations are then easily computed from [44]:

$$p = n_{\text{ref}} \exp \left( \frac{-q(V - V_p - \phi_p)}{kT} \right),$$  \hspace{1cm} (2.38)

$$n = n_{\text{ref}} \exp \left( \frac{q(V + V_n - \phi_n)}{kT} \right).$$  \hspace{1cm} (2.39)

Each iteration solves the linear problem (2.35) for the correction vector $\Delta \vec{u}^{k+1}$, which consists of corrections to the variables $V$, $\phi_n$, and $\phi_p$. The vector $\vec{u}^k$ is then updated prior to the next iteration. After several iterations, $\Delta \vec{u}^{k+1}$ approaches the zero vector and the solution has been reached.

2.4. Physical Models

2.4.1. Bulk Recombination

Accurate analysis of realistic semiconductor devices requires the use of a proper model for recombination, since recombination plays a primal role in determining device characteristics. Recombination in bulk semiconductors can be classified into five different processes [69]:

1) Radiative (band-to-band) recombination
2) Shockley–Read–Hall (R–G center) recombination
3) Auger recombination
4) Recombination involving excitons
5) Recombination via shallow level traps

Of these five processes, the last two are least likely to contribute significantly to the total recombination, becoming important only at low temperatures. We choose to model radiative, Auger, and Shockley–Read–Hall (SRH) recombination.

The total recombination rate due to these three processes may be written as [62]:

\[ R = \left[ B_r + A_n n + A_p p + \frac{1}{\tau_n(p + p_1) + \tau_p(n + n_1)} \right] (np - n_{i0}^2). \]  

(2.40)

In (2.40), \( B_r \) is the radiative recombination coefficient, \( A_n \) and \( A_p \) are Auger recombination coefficients and \( n_{i0} \) is the intrinsic carrier combination in the semiconductor without bandgap narrowing effects. In the SRH term, \( \tau_n \) and \( \tau_p \) are the SRH lifetimes and \( n_1 \) and \( p_1 \) are constants which depend on the energy of the deep-level traps.

For purposes of the simulation program, the default radiative recombination coefficient for n–type material is taken from the results of Hwang [70], while for p–type material it comes from Nelson and Sobers [71]. The default Auger recombination coefficients are from Haug [72]. These three coefficients can be over–ridden by inclusion of user–specified coefficients in the program input deck. The default SRH lifetimes are 1 nS for both electrons and holes, and the default trap energy level is the intrinsic energy level, \( E_i \). These parameters can also optionally be defined by the user of the program.

2.4.2. Surface Recombination

Recombination at the surface of devices is likely to play an important role due to the highly imperfect nature of the surface. In general, there will be a high number of surface states, caused by a high density of crystalline defects, which are distributed in energy. This will produce SRH–like recombination which will completely dominate the recombination processes.
Due to the similarity to the SRH recombination process, a similar mathematical description is expected. The difference is that the formulation for SRH recombination assumed recombination via one deep-level trap energy. For surface recombination, it will be desirable to integrate over the energy distribution. The surface recombination rate will then be given by [69]:

$$R_S = \int_{E_V}^{E_C} \frac{np - n_i^2}{c_p (n + n_{1S}) + c_n (p + p_{1S})} D_{IT}(E) \, dE,$$

(2.41)

where $D_{IT}$ is a function giving the distribution in energy of the surface states. $p_{1S}$ and $n_{1S}$ are the analogs of $p_1$ and $n_1$ in (2.40), and during integration must be evaluated at each $E$ since they depend on the trap energy level. $c_p$ and $c_n$ are the surface hole and electron capture coefficients and are a measure of how likely the carriers are to be captured when in the vicinity of a trap. For simplicity and to decrease execution time, a single-level trap energy is assumed. The trap distribution function is written as:

$$D_{IT} = N_{TS} \delta(E_{TS}),$$

(2.42)

where $N_{TS}$ is the total number of surface states per unit area, $E_{TS}$ is the energy level of the traps, and $\delta$ is Dirac's delta function. Integration of the delta function distribution of surface states in (2.41) gives:

$$R_S = \frac{np - n_i^2}{S_p (n + n_{1S}) + S_n (p + p_{1S})},$$

(2.43)

where $S_p = c_p N_{TS}$ and $S_n = c_n N_{TS}$ have units of cm/sec and are therefore termed surface recombination velocities. $n_{1S}$ and $p_{1S}$ are determined from $E_{TS}$.

The simulation program allows user input of the trap level $E_{TS}$ and the surface recombination velocities. For unpassivated GaAs, a surface recombination velocity as high as $10^7$ cm/sec may be reasonable [73].
default values in the model are surface recombination velocities of zero and traps located at the intrinsic energy level.

2.4.3. Carrier Mobilities

Proper modeling of carrier mobilities is of critical importance for the heterojunction bipolar transistor, since carrier drift across the base and collector may determine the high frequency performance of the transistor. The simplest mobility model is one in which the mobilities are viewed as being dependent only upon the material composition and doping. This is clearly inadequate for the simulation of transistors, since this model will never predict velocity saturation. Due to the high electric field present in the base-collector junction of a bipolar transistor operating in the forward active mode, velocity saturation will almost certainly occur. The use of a field-dependent mobility, in which the mobility varies inversely with electric field, is undoubtedly a step in the right direction, as velocity saturation can be modeled [38]. This model will still fail to predict velocity overshoot, however. A full energy and momentum balance equation solution is necessary to produce the best of all possible mobilities, which will include the effects of velocity overshoot. It is quite costly to solve the energy and momentum balance equations in addition to the basic semiconductor equations and for this reason simpler models are usually employed.

The original formulation of PUPHS2D uses the field-independent model, which is known to be inappropriate for the high fields found in HBT's. One of the objectives of this work is to extend the mobility models for accurate high-field description. To this end, a hybrid system for computing energy-dependent mobilities will be presented in Chapter 3.

The low-field hole mobility is computed using the method of Sutherland and Hauser [41]. From a very general standpoint, hole mobility may be written as:

$$\mu_p = \frac{q\langle \tau_p \rangle}{m_p^*},$$  

(2.44)
where \( m_p^* \) is the hole effective mass and \( \langle \tau_p \rangle \) is the average time between scattering events. Thus, given the mobility in some reference material, we can multiply by an appropriate ratio of effective masses and scattering times to find the mobility in an arbitrary material. The effective masses (hole and electron) in \( \text{Al}_x\text{Ga}_{1-x}\text{As} \) are specified as a function of the aluminum mole fraction by [74]:

\[
\frac{1}{m^*} = \frac{x}{m_{\text{AlAs}}^*} + \frac{1-x}{m_{\text{GaAs}}^*}.
\]  

(2.45)

Assuming that the hole mobility is controlled by polar optical phonon scattering, we have:

\[
\langle \tau_p \rangle = \frac{K}{\sqrt{m_p^* \left( \frac{1}{\kappa_h} - \frac{1}{\kappa_l} \right)}}.
\]  

(2.46)

In this equation \( \kappa_h \) and \( \kappa_l \) are the high and low frequency relative dielectric constants. Using (2.44–46), the hole mobility may now be related to the hole mobility in the reference material:

\[
\mu_p = \mu_{p, \text{GaAs}} \frac{m_p^* \frac{3/2}{m_{p, \text{GaAs}}^*} \left[ \frac{1}{\kappa_h, \text{GaAs}} - \frac{1}{\kappa_l, \text{GaAs}} \right]}{m_p^* \frac{3/2}{\left[ \frac{1}{\kappa_h} - \frac{1}{\kappa_l} \right]}}.
\]  

(2.47)

The GaAs subscripts indicate the parameter is to be evaluated in the reference material, GaAs, while no subscript indicates the parameter is to be evaluated in the material for which the mobility is desired. The doping-dependent hole mobility of GaAs is taken to have the form [61,75]:

\[
\mu_{p, \text{GaAs}} = \frac{380}{\left( 1 + 3.17 \times 10^{-7} (N_A + N_D) \right)^{0.266} \left[ \frac{300}{T} \right]^{2.7}}.
\]  

(2.48)

Modeling the field-independent electron mobility is somewhat more complex due to the existence of multiple conduction bands. Considering only electrons traveling in the \( \Gamma \) and \( X \) valleys, a direct and an indirect electron mobility may be defined by [41]:
\[
\mu_{n\Gamma} = \frac{m_{n\text{GaAs}}^{*} 3/2 \left( \frac{1}{k_h \text{GaAs}} - \frac{1}{k_l \text{GaAs}} \right)}{m_{n\Gamma} ^{*} 3/2 \left[ \frac{1}{k_h} - \frac{1}{k_l} \right]}, \quad (2.49)
\]

\[
\mu_{nX} = \frac{m_{nX \text{AlAs}}^{*} 3/2 \left( \frac{1}{k_h \text{AlAs}} - \frac{1}{k_l \text{AlAs}} \right)}{m_{nX} ^{*} 3/2 \left[ \frac{1}{k_h} - \frac{1}{k_l} \right]}, \quad (2.50)
\]

The doping-dependent electron mobilities in GaAs and AlAs are taken to be [61,74]:

\[
\mu_{n\text{GaAs}} = \frac{7200}{\left[ 1 + 5.51 \times 10^{-17} (N_A + N_D) \right]^{0.233} \left[ \frac{300}{T} \right]^{2.3}}, \quad (2.51)
\]

\[
\mu_{n\text{AlAs}} = \frac{165}{\left[ 1 + 8.1 \times 10^{-17} (N_A + N_D) \right]^{0.13} \left[ \frac{300}{T} \right]^{2.5}}, \quad (2.52)
\]

Having defined appropriate mobilities for the \( \Gamma \) and \( X \) valleys, it is now possible to compute the effective electron mobility [41]:

\[
\mu_n = \mu_{n\Gamma} R_\Gamma + \mu_{nX} (1 - R_\Gamma), \quad (2.53)
\]

where \( R_\Gamma \) is the fraction of electrons in the \( \Gamma \) valley and is given by

\[
R_\Gamma = \frac{1}{1 + \left[ \frac{m_{nX}^{*}}{m_{n\Gamma} ^{*}} \right]^{3/2} e^{(E_{g\Gamma} - E_{gX})/kT}}. \quad (2.54)
\]

Carrier mobility curves at room temperature for GaAs and Al\(_{0.3}\)Ga\(_{0.7}\)As are shown in Figure 2.6.

The mobility models used here are over a decade old and should probably be updated, or at the very least reviewed. Majority and minority carrier mobilities are assumed to be identical and the empirical forms for mobilities given by (2.48, 2.51, 2.52) are equations fit to old data, and are possibly not representative of the mobilities achievable with current growth techniques. In addition, the two-valley model must be questioned for
devices where intervalley scattering into the L-valley may limit high speed performance.

2.4.4. Bandgap and Bandgap Narrowing

The accuracy of the bandgap model in a heterostructure simulation program is clearly important. \( \Gamma, X, \) and L-valley bandgap models for GaAs were presented by Blakemore [76] and have been combined with the data of Casey and Panish for \( Al_xGa_{1-x}As \) [77] to form a compositional and temperature dependent bandgap model [61].

Bandgap narrowing in heavily doped silicon has been known to be substantial for several years [78]. Initial data for heavily-doped p-type GaAs based on optical absorption measurements showed the presence of bandgap narrowing [79], and recent data from electrical measurements in p-type GaAs show even larger bandgap narrowing [80]. The data of Klausmeier-Brown [81] have been used to implement a bandgap narrowing model for p-type GaAs.

The experimental procedure of Klausmeier-Brown [82] allows the determination of \( n_{ie}^2D_n \), where \( n_{ie} \) is an effective intrinsic carrier concentration and can be related to the effective bandgap shrinkage by [5]:

\[
n_{ie}^2 = n_{i0}^2 \exp \left\{ \frac{\Delta E_{cb}^{bn}}{kT} \right\}.
\] (2.55)

For purposes of modeling it is necessary to compute \( \Delta E_{cb}^{bn} \), so the diffusion constant must be known to extract \( n_{ie}^2 \). This is done by using the electron mobility model of the previous section and then applying the Einstein relation to find \( D_n \). After extraction of \( n_{ie}^2 \), (2.55) is used to compute \( \Delta E_{cb}^{bn} \).

A curve of the same form used by Slotboom and de Graaff in silicon [78] is then fit to the data for inclusion in the simulation program:

\[
\Delta E_{cb}^{bn} = 6.1 \left[ \ln \frac{N_A}{8.5 \times 10^{16}} + \sqrt{\ln^2 \frac{N_A}{8.5 \times 10^{16}} + 0.5} \right] \text{meV}.
\] (2.56)

A plot of (2.56) and the data it was fit to is shown in Figure 2.7.

It must be stressed that (2.56) is valid only when using the mobility model of section 2.4.3 and applying the Einstein relation. The data of
Electron Mobility in $\text{Al}_x\text{Ga}_{1-x}\text{As}$

Hole Mobility in $\text{Al}_x\text{Ga}_{1-x}\text{As}$

Figure 2.6 Majority carrier mobilities in $\text{Al}_x\text{Ga}_{1-x}\text{As}$. 
Klausmeier–Brown give only \( n_{ie}^2D_n \), and if a better diffusivity model exists it should be used in conjunction with the data to compute \( n_{ie}^2 \). Strictly speaking, (2.56) is also valid only at 300 K, although for lack of better data the use of (2.56) is probably better than assuming no bandgap narrowing.

2.4.5. Sundries

Models for a few other material parameters remain to be discussed. These are the electron affinity, dielectric constants, and effective masses. For each of these parameters, the interpolation schemes of Sutherland and Hauser [41] are used with the data of Casey and Panish [77] to produce models for arbitrary compositions of \( \text{Al}_x\text{Ga}_{1-x}\text{As} \) [61]. For carrier generation, the optical absorption coefficient of Aspnes et al. is used [83].

![Bandgap Narrowing in p–GaAs](image)

**Figure 2.7** Bandgap narrowing in p–GaAs.
2.5. Summary

In this chapter the formulation of a two-dimensional heterojunction device simulation tool, PUPHS2D, has been presented. The equations to be solved were introduced. The equations were shown in their discretised form for numerical analysis by the finite difference technique, which was briefly explained. The method used for solution of the difference equations was discussed. Finally, the material models used to define the physical properties of \( \text{Al}_x\text{Ga}_{1-x}\text{As} \) were given.
CHAPTER 3: MODEL DEVELOPMENT

3.1. Introduction

In order to make PUPHS2D an efficient and effective simulation program for heterojunction bipolar transistors, several changes were necessary. PUPHS2D was originally developed for the analysis of GaAs solar cells and, as previously mentioned, one of its parents is the silicon solar cell program SCAP2D. It is well developed, therefore, for solving solar cell problems, and is capable of computing solar cell characteristics, quantum efficiency versus incident wavelength, and current versus solar intensity. The basic control structure for treating bipolar transistors was in place, but was undeveloped and untested. This chapter discusses the evolution of PUPHS2D into a useful HBT design and evaluation tool for the solid state engineer.

3.2. Extensions for Transistor Analysis

The first course of action was to decide precisely what sort of information was desired from the program. A literature search was conducted to determine what characteristics were most important to device engineers. From a current–voltage characteristic point of view, it was deemed necessary to be able to produce:

1) Collector current as a function of collector–emitter voltage for a given base–emitter voltage (curve tracer plot)
2) Base and collector currents as a function of base–emitter voltage (Gummel plot)
3) Current gain as a function of base–emitter voltage.

To produce this information, it is necessary to have the capability of holding the base–emitter voltage constant while varying the collector–
emitter bias, and vice versa. For the case of $J_C$ versus $V_{CE}$, $V_{BE}$ is stepped up (in 0.05 volt increments) to the bias for which the $V_{CE}$ sweep is desired. $V_{CE}$ is then stepped out in user-specified increments, tracing the requested curve as it goes. For Gummel plots, the same algorithm is used, except that $V_{BE}$ is stepped to the initial (user-specified) bias and then, after the $V_{CE}$ sweep is complete, is again incremented, until the final base-emitter bias is reached. A plot of current gain versus $V_{BE}$ may then be directly computed from the Gummel plot data. The nonequilibrium algorithm of PUPHS2D for bipolar transistors is diagrammed in Figure 3.1. $V_{BASE}$ is the base-emitter bias and $V_{COLL}$ is the collector-emitter voltage. If the $J_C$ versus $V_{CE}$ analysis is to be performed, then VARY=COLL, whereas for a Gummel plot, VARY=BASE and $V_{BMAX}$ is the final base-emitter bias ($V_{BASE}$ is used for the initial base-emitter bias).

The stepping of bias voltages described above is necessary due to the need of the Newton method for an initial guess to the solution for the first iteration. If this initial guess is not close enough, the method may not converge to a solution. For this reason, small increments are chosen and after each step the present solution is used as the initial guess for the next. This is probably not an optimal method; an alternative is to make an educated guess as to the position of the quasi-Fermi levels after application of the next bias and use this as a first guess. This should allow larger steps to be taken without loss of convergence, and may significantly reduce computation time.

3.3. Treatment of Nonrectangular Geometries

One significant distinction between modeling solar cells and HBT's is the necessity of the capability of modeling nonplanar devices. Most fabricated $Al_xGa_{1-x}$As-GaAs HBT's are mesa-etched devices and need to be modeled as such. Surface recombination at the edge of the emitter mesa where an exposed base-emitter junction exists is an important effect. Current crowding at high current densities must also be realistically modeled.

First, one might ask why modeling a nonplanar device as a planar device is a poor compromise. Referring to Figure 3.2, several problems are apparent. The most obvious problem is the existence of lateral injection
Figure 3.1  Bipolar transistor nonequilibrium solution flowchart for PUPHS2D.
currents $J_{nL}$ and $J_{pL}$ across a vertical base–emitter junction which does not exist in a mesa-etched transistor. Under high current conditions, the vertical base–emitter junction will be more forward-biased than the horizontal base–emitter junction due to the finite conductivity of the base and emitter. This will cause current crowding and a substantially different current flow pattern than would exist in a mesa-etched transistor. Another obstacle is the location of the surface recombination current. In a mesa-etched transistor, the surface recombination current is separated from the collector only by the width of the base. A substantial portion of electrons injected from the emitter into the base may recombine at the exposed surface instead of being collected, thereby reducing the current gain. In the planar transistor, it is very unlikely that electrons injected from the emitter into the base will recombine at the surface, since the surface is much farther away. Admittedly, this will be offset to some degree by the fact that many of the laterally injected electrons in the planar transistor will recombine at the surface, thereby reducing the current gain of the planar transistor as well.

In order to model nonplanar devices using a rectangular domain, it is therefore essential to prevent lateral injection of carriers across the vertical base–emitter junction and to somehow move surface recombination into the interior of the device.

The prevention of lateral injection of carriers is accomplished by inserting an insulating material between the extrinsic base and the emitter, as shown in Figure 3.3. The insulator is specified to have a band gap of 3.4 eV, an electron affinity of 3.07, and a relative dielectric constant of 1. The insulator is left undoped. Basically, this is akin to actually modeling the air that would be present in the mesa-etched structure. The effectiveness of this approach will be illustrated in Chapter 4. The barrier presented to carriers in the base and emitter is seen in Figure 3.4, which shows energy band diagrams along the section marked A–A' in Figure 3.3.

In addition to preventing lateral carrier injection, the base contact must be “moved into” the device, where the contact would actually be in a mesa-etched structure. The contact is defined mathematically by the ohmic contact boundary conditions of (2.14–2.16), so it is only necessary to ensure that these boundary conditions are applied to the appropriate surface within the device.
Figure 3.2 Cross-sections of a) planar bipolar transistor, and b) mesa-etched bipolar transistor.
This can be done quite simply, without explicitly enforcing the boundary conditions. The carrier concentration boundary conditions of (2.15) and (2.16) state that the concentrations must remain at their equilibrium values, hence it is sufficient to set the minority carrier lifetimes in the etched region (indicated by diagonal shading in Figure 3.3) to some arbitrary small value. One femtosecond is typically chosen for convenience. This ensures that the carrier concentrations remain at their equilibrium values since any excess carriers will almost instantaneously recombine. Minority carrier lifetimes are a user-specified parameter, therefore this method is easily applied through the input deck.

The final task remaining to allow modeling of nonplanar devices is to address surface recombination. Surface recombination would normally be handled by setting a recombination velocity at the surface. Now, however, the surface has been effectively moved to the interior of the device. What is needed, then, is a way to relate surface recombination to a bulk recombination. Re-examining the similarity of the SRH portion of (2.40) and (2.43), it is immediately apparent that a close relationship exists between the surface recombination velocities and the minority carrier lifetimes. If an appropriate minority carrier lifetime can be specified, it should be possible to model a given surface recombination velocity. The surface recombination rate $R_S$, however, has units of cm$^{-2}$ s$^{-1}$, whereas the bulk recombination rate is in terms of cm$^{-3}$ s$^{-1}$. To compute the minority carrier lifetimes corresponding to given recombination velocities we use:

$$\tau_n = \frac{w}{S_n}, \quad \text{(3.1)}$$

$$\tau_p = \frac{w}{S_p}, \quad \text{(3.2)}$$

where $w$ is some incremental width over which the lifetime is defined to approximate a surface. Naturally, a small value of $w$ will produce a more surface-like effect than a larger value.
Figure 3.3 Placement of insulating material for prevention of lateral carrier injection across base–emitter sidewall.
Figure 3.4  p–I–N energy band diagrams along section A–A′ of Figure 3.3 under a) equilibrium, and b) 1.2 volts of forward bias.
3.4. Quasi-Static Frequency Evaluation

An obvious area of importance for transistor evaluation is high frequency performance. It is therefore desirable to be able to determine the frequency capabilities of the transistor being modeled. For use in high speed analog circuits, the unity current gain cutoff frequency is often quoted as a figure of merit. Approximating an AC signal as small perturbations in the base and collector biases and computing the DC solution at the perturbed biases, it is possible to compute the cutoff frequency of the transistor [84,85]. This method has been routinely used for computation of frequency characteristics; it requires only a DC analysis program. As has been pointed out by Laux [84], this technique is not rigorous and is difficult to apply to devices with more than two terminals. In any case, due to the treatment of the AC signal as a succession of DC states the quasi-static method is only valid for relatively low frequencies. A rigorous method for computation of the frequency characteristics of devices by Fourier analysis of a transient excitation is described by Laux in [84]. This technique is much more computationally intensive than the quasi-static method and requires a transient analysis program. While transient capability has been added to PUPHS2D as will be described later in this chapter, the Fourier analysis approach has not yet been attempted. A third manner for determining frequency characteristics is sinusoidal steady-state analysis, which requires significant changes to the model formulation. References [84] and [85] contain comparisons of these three techniques.

Using the quasi-static model as reported by Yokoyama et al [49], the cutoff frequency is computed from the difference between the charge stored in the device at the original and some perturbed bias. The total hole charge in the device, for example, is:

\[ Q_h = \int \int_{device} q (p - N_A^-) \, dx \, dy. \]  

(3.3)

The cutoff frequency is then computed using:

\[ f_T = \left. \frac{\Delta I_C}{2\pi \Delta Q_{BE}} \right|_{V_{CE} = \text{const}}. \]  

(3.4)
The solution is first obtained for the original bias point, and the hole or electron charge is computed. $V_{BE}$ is then incremented by a small amount, 0.005 volts, while $V_{CE}$ is held constant. The solution for this perturbed bias is computed and the incremental charge between the two base-emitter biases, $\Delta Q_{BE}$, is computed. (3.4) is then applied to find the cutoff frequency. In a similar manner it is possible to compute the element values of the small-signal hybrid-$\pi$ equivalent circuit model shown in Figure 3.5. The element values are given by:

$$g_m = \frac{\Delta I_C}{\Delta V_{BE}} \bigg|_{V_{CE} = \text{const}}$$  \hspace{1cm} (3.5)  

$$r_\pi = \frac{\Delta V_{BE}}{\Delta I_B} \bigg|_{V_{CE} = \text{const}}$$  \hspace{1cm} (3.6)  

$$r_0 = \frac{\Delta V_{CE}}{\Delta I_C} \bigg|_{V_{BE} = \text{const}}$$  \hspace{1cm} (3.7)  

$$c_\mu = \frac{\Delta Q_{CE}}{\Delta V_{CE}} \bigg|_{V_{BE} = \text{const}}$$  \hspace{1cm} (3.8)  

$$c_{\pi} = \frac{\Delta Q}{\Delta V_{BE}} \bigg|_{V_{CB} = \text{const}}$$  \hspace{1cm} (3.9)  

A simple flow chart of the solution technique necessary for solving (3.4-3.9) is shown in Figure 3.6. As can be seen from Figure 3.6, three solutions are needed for each point: the solution at the original bias, and the solution at two perturbed biases. Holding $V_{CE}$ constant while incrementing $V_{BE}$ gives $r_T$, $g_m$, and $r_\pi$. Holding $V_{BE}$ constant while incrementing $V_{CE}$ gives $r_0$ and $c_\mu$. In addition, if the same increment is used to perturb both the base and collector bias, then $V_{CB}$ can be held constant while both $V_{BE}$ and $V_{CE}$ are incremented, thus allowing the determination of $c_{\pi}$. 
Figure 3.5  Small-signal hybrid-$\pi$ equivalent circuit for bipolar transistors.

Figure 3.6  Quasi-static frequency analysis flowchart for bipolar transistors.
3.5. The Transient Problem

In order to evaluate transistors for use in high speed digital logic applications, the unity current gain cutoff frequency is no longer a very useful performance guide. In logic applications, the criterion used to assess the high speed capability of devices is more apt to be the switching time of the transistor. Determination of the switching characteristics of devices requires that a transient analysis be performed. After some switching action is performed, the terminal characteristics as a function of elapsed time are desired.

3.5.1. The Time-Dependent Semiconductor Equations

In writing the semiconductor equations (2.1–2.3), steady-state conditions were assumed to exist. In their time-dependent form, the basic semiconductor equations are [86]:

\[ \nabla \cdot \vec{D} = \rho , \]  \hspace{1cm} (3.10)

\[ \nabla \cdot \vec{j}_n = q (R - G + \frac{\partial n}{\partial t}) , \]  \hspace{1cm} (3.11)

\[ \nabla \cdot \vec{j}_p = q (G - R - \frac{\partial p}{\partial t}) . \]  \hspace{1cm} (3.12)

The Poisson equation has no explicit time dependence, but the divergence of the currents is modified by the time rate of change of the carrier concentrations. When evaluating the terminal currents, we have:

\[ \vec{J}_T = \vec{j}_n + \vec{j}_p + \vec{j}_D , \]  \hspace{1cm} (3.13)

where \( \vec{J}_T \) is the total current density and \( \vec{j}_D \) is the displacement current arising from time dependency of the displacement field. The displacement current is

\[ \vec{j}_D = \varepsilon \frac{\partial \vec{F}}{\partial t} . \]  \hspace{1cm} (3.14)

Since the transistor current is evaluated at the contacts, we assume the contacts are heavily doped and quasi-neutral, so that the electric field is zero.
and the displacement current is then zero. The appropriateness of this assumption is supported by the results of De Mari [87], which show that negligible displacement currents exist in the heavily doped side of n+–p germanium homojunctions after passage of the dielectric relaxation time. For GaAs doped in typical ranges, the dielectric relaxation time is on the order of 0.05 ps.

3.5.2. Discretisation and Solution

Several alternative methods for solving the time–dependent semiconductor equations have been devised. Due to the lack of explicit time dependence of the Poisson equation, several methods have been suggested to introduce such a dependence [63]. These techniques have been shown to be conditionally unstable, however, and full backward time differencing is preferred. While this method is more computationally intensive, it is stable for arbitrarily large time steps. If the carrier concentrations are assumed to vary linearly with time, then the backward time difference for electrons is

\[
\left[ \frac{\partial n}{\partial t} \right]^k = \frac{n_k - n_{k-1}}{\Delta t_k} ,
\]

where \( k \) indicates evaluation at the current time, \( k-1 \) refers to the previous time, and \( \Delta t \) is the time step used. An analogous equation exists for holes. If the carrier concentrations are assumed to have an exponential time dependence, the backward time difference is instead

\[
\left[ \frac{\partial n}{\partial t} \right]^k = \frac{n_k \ln \left( \frac{n_k}{n_{k-1}} \right)}{\Delta t_k} .
\]

The time–dependent analogs of (2.27) and (2.28) are then:

\[
f_p + \left[ \frac{\partial p}{\partial t} \right]^k = 0 ,
\]

\[
f_n - \left[ \frac{\partial n}{\partial t} \right]^k = 0 ,
\]
where \( f_p \) and \( f_n \) are the time independent functions of (2.27) and (2.28). PUPHS2D allows the user to choose the time discretisation model used, linear or exponential. The Poisson equation, (2.26), remains unchanged. To accommodate the derivatives of (3.17) and (3.18), the Jacobi matrix must be slightly modified.

The solution of (3.17) and (3.18) with (2.26) is accomplished in the same manner as the steady-state solution. The steady-state solution is found for some initial bias condition and the carrier concentrations are stored. The switching action is then performed and the solution is computed at time \( \Delta t_1 \). The initial conditions are used for computation of the backward time difference. The solution at \( \Delta t_1 \) is then used as the initial guess for time \( \Delta t_1 + \Delta t_2 \), and so on until a user-specified number of time steps have been taken. Presently PUPHS2D assumes a constant time step which is set by the user. An adaptive time step would allow faster computation of transients, since a small step must be used for the initial response and is not necessary as the transient dies out.

It should be mentioned that the transient analysis represents another method for producing faster steady-state results. As was discussed in Section 3.2, it is necessary to take many voltage steps to reach the desired bias point without loss of convergence. Instead of stepping in voltage, the problem could be viewed as a transient analysis. The transistor could be switched on to the desired bias and then by using an initially small time step which increases with each iteration, the steady-state value will eventually be reached. This method is potentially faster than voltage-stepping due to the inclusion of time-dependency directly in the Jacobi matrix. When stepping in voltage we can only hope that the steps are small enough for the last bias solution to serve as a good initial guess for the next.

3.6. Modeling Hot Electron Transport Phenomena

As previously stated, hot carrier effects such as velocity saturation and nonstationary phenomena such as velocity overshoot must be addressed by device modelers and designers alike. As device dimensions continue to shrink these properties are of increasing importance. While methods for including hot carrier velocity saturation in drift–diffusion models have been
in use for some time, velocity overshoot has proven to be much more difficult to treat. It is for this reason that Monte Carlo techniques and hydrodynamic models have become widely used and accepted [50-57].

The most common method used for treating velocity saturation has been the use of field-dependent mobilities, in which an empirical formula relating the mobility to the electric field is used. It will be seen that the mobility model presented in 2.4.3 contains no electric field dependency, nor was it necessary for the use of PUPHS2D as a solar cell model. For use as a transistor model, however, it is apparent that velocity saturation must be treated. We choose solution of a localized energy balance equation over a field–dependent mobility due to its modularity. The energy balance equation can be readily non-localized at a later date to produce an even more accurate model.

3.6.1. Generalized Drift–Diffusion Equations

The first step toward providing a true hot carrier model is to remove the Einstein relation from the drift–diffusion current equations. The Einstein relation was used to write the Scharfetter–Gummel current discretisation of (2.34). In regions of high electric field (or alternatively for high energy carriers), the Einstein relation is no longer valid. This is shown in Figure 3.7, which shows the correct diffusion coefficient and that computed from the Einstein relation, as a function of electric field in a homogeneous slab under steady–state conditions. The low–field diffusion coefficient is extrapolated for reference. It is clear that the use of the Einstein relation for high electric fields is not warranted, and in fact it is more accurate to use the low–field diffusion coefficient.

In order to remove the Einstein relation from the Scharfetter–Gummel current discretisation, a parameter $\xi$ is defined which relates the diffusion coefficient and the mobility [88]:

$$\mu_n \xi_n = D_n$$  \hspace{1cm} (3.19)

The $x$–directed electron current can then be written as

$$J_{nx} = q\mu_n \left( -n \frac{dV}{dx} + \xi_n \frac{dn}{dx} \right).$$  \hspace{1cm} (3.20)
where compositional uniformity has been assumed for ease of expression. The Scharfetter–Gummel current discretisation method [67] is now applied to (3.20), resulting in expressions for current density which are only slightly changed from their previous form. The normalized x-directed electron current midway between node ij and its nearest neighbor to the right is:

\[
J_{n\text{m}} = -\mu \frac{\Delta V_R}{h_R} \left( \frac{n_R \exp \left( \frac{-\Delta V_R}{\xi n_R} \right) - n_{ij}}{\exp \left( \frac{-\Delta V_R}{\xi n_R} \right) - 1} \right) \tag{3.21}
\]

It can be seen by comparison of (3.21) with (2.34) that the only change necessary when the Einstein relation is removed is the scaling of the exponentials by \( \xi \). For regions of small electric fields, the Einstein relation is expected to hold, so that \( \xi = kT/q \). The normalization factor for \( \xi \) is also \( kT/q \), so for low-field regions the normalized \( \xi \) is 1 and (3.21) becomes the proper expression. Like the mobilities, \( \xi \) is evaluated midway between nodes.

### 3.6.2 Localized Energy Balance Equations

The general energy balance equation comes from the second moment of the BTE and may be written as [89]:

\[
\frac{\partial W}{\partial t} = -\nabla \cdot \overrightarrow{F_W} + \overrightarrow{J_n} \cdot \overrightarrow{E_{\text{eff}}} - \frac{W - W^0}{\langle \tau_E \rangle}, \tag{3.22}
\]

where \( W \) is the carrier energy density, \( W^0 \) is the equilibrium carrier energy density, \( \overrightarrow{F_W} \) is the energy flux, \( \overrightarrow{E_{\text{eff}}} \) is an effective field which includes the effect of the band parameters \( V_p \) and \( V_n \), and \( \langle \tau_E \rangle \) is the energy relaxation time. The energy relaxation time is a measure of the decay of the average carrier energy toward its equilibrium value. At this point it is desirable to make some assumptions in order to relieve the computational burden. If steady-state transport is assumed, the partial derivative of the carrier energy density with respect to time becomes zero. The balance equation is then localized by assuming that the energy flux is not position-dependent, so its
Figure 3.7  Rigorous and Einstein relation diffusion coefficients versus electric field in an homogeneous GaAs slab.

The simplified energy balance equation therefore becomes:

$$\mathbf{J}_n \cdot \mathbf{E}_{\text{eff}} - \frac{W - W^0}{\langle \tau_E \rangle} = 0.$$  \hspace{1cm} (3.23)

Rearranging and dividing by $q_n$ to convert from energy density ($\text{J/cm}^3$) to energy in eV, we obtain:

$$u_n = \frac{\mathbf{J}_n \cdot \mathbf{E}_{\text{eff}}}{q_n} \langle \tau_E \rangle + u_n^0,$$  \hspace{1cm} (3.24)

where $u_n$ is the electron energy and $u_n^0$ is the equilibrium electron energy. The equilibrium energy is just the thermal energy of an ideal gas, and is given by:

$$u_n^0 = \frac{3}{2} \frac{kT}{q},$$  \hspace{1cm} (3.25)

where $u_n^0$ has units of eV.
3.6.3. Balance Equation Solution Method

A simultaneous solution of the basic semiconductor equations with the local energy balance equation of (3.24) is desired. For coding simplicity and to allow the later inclusion of a more complete balance equation, an iterative, uncoupled method of solution is used. The solution to the basic problem is first sought, and only after the initial problem converges to some predetermined tolerance is the energy balance equation solved. The electron energies computed from the energy balance solution are then used to update the mobilities (and if desired the diffusivities) in the device. The basic problem is then solved again for the new mobilities and the resulting currents and electric fields are used to resolve the energy balance equation, and so on until the solutions to the basic problem and the energy balance equation are self-consistent. A flow chart of the solution algorithm is shown in Figure 3.8.

The solution to (3.24) is found using the bisection method. This is necessary since the energy relaxation time is itself a function of the local energy. The energy relaxation time as a function of energy comes from the Monte Carlo simulations of Constant [90] and is shown in Figure 3.9. Once the electron energy is computed, the corresponding electron mobility is computed from the shaping function of Constant shown in Figure 3.10, where \( \mu_n = \mu_{n0} F_5 \). The data of Constant was computed using only one value of doping; the energy relaxation time is therefore assumed to be independent of doping and the mobility function is assumed to have the same shape for all doping levels. If the electron diffusion coefficient is to be considered energy-dependent then it can be computed from [89]:

\[
D_n = \frac{2}{3q} u_n \mu_n .
\] (3.26)

3.7. Summary

This chapter has described the changes made to PUPHS2D to produce an effective HBT analysis tool. The algorithm used for obtaining transistor characteristics was presented, as well as a method for treating nonrectangular
Figure 3.8  Local energy balance equation solution flowchart.
Figure 3.9  Energy relaxation time in GaAs as a function of electron energy [90].

Figure 3.10  Mobility shape function as a function of electron energy [90].
devices. The high speed and high frequency evaluations of HBT's were discussed. For analog circuits, the quasi–static method was offered as a technique for determining transistor cutoff frequencies. Digital applications were addressed by the inclusion of transient analysis to provide an indicator of switching performance. A local energy balance equation was introduced for the purpose of treating hot carrier effects. The local energy balance equation should be seen as a first step toward proper computation of high–field transport characteristics. PUPHS2D stands as an accurate tool for modeling the low–field characteristics of Al_xGa_{1-x}As–GaAs heterojunction bipolar transistors.
CHAPTER 4: MODEL VERIFICATION

4.1. Introduction

In order to gain confidence in any numerical model, it is necessary to examine some test cases for which an approximate analytical solution is known and verify that the model will produce similar results. In the instance of PUPHS2D, it is somewhat difficult to find appropriate test cases due to the lack of simple solutions to two-dimensional problems. This chapter presents simulation results from several example tests. Where possible the solutions are compared to simple analytical models. Where analytical solutions do not exist the simulation results are examined for qualitatively correct characteristics. In the case of hot electron effects, the PUPHS2D solution is compared to Monte Carlo results.

4.2. Treatment of Non-Planar Devices

In order to study the effectiveness of modeling nonrectangular devices as described in Section 3.3, a simple p+-n diode was simulated. A one-dimensional analytic solution (see Appendix A) was first computed and compared to the solution given by PUPHS2D for a “one-dimensional” diode. This diode was simulated in 2-D, but was specified such that there was no material variation in one direction. The current-voltage characteristics produced for the one-dimensional diode are plotted together in Figure 4.1. The numeric solution matches the analytic solution very well.

The diode was then specified as a two-dimensional mesa-etched device using the method outlined in Section 3.3. The current-voltage characteristics of the 1-D and 2-D diodes as computed by PUPHS2D are shown in Figure 4.2. The characteristics match very well until high biases are applied to the diode. This is due to the larger series resistance of the mesa-etched device, and is an effect which will exist in a real device. Figure 4.3
Figure 4.1 Analytic and numeric I–V characteristics of a 1-D homojunction diode.

Figure 4.2 I–V characteristics of 1-D and 2-D homojunction diodes.
illustrates the effectiveness of the technique for modeling nonrectangular devices. This figure is a plot of the electron concentration in the two-dimensional diode and contains several important tidbits of information. Most importantly, the suppression of lateral injection of electrons into the p-region is clearly shown. In addition it is readily seen that the boundary condition for the p-type contact is being satisfied; there are no minority carriers in the p-contact region and the contact has been effectively moved into the device interior. A side effect of the nonrectangular methodology is also seen: an accumulation layer of electrons exists on the n-type side of the insulator. This is easily explained by the fact that the insulator serves as the oxide layer of a parasitic lateral "MOS" capacitor, where the heavily-doped p-type region acts as the metal. For a positive bias on the p-type contact, the MOS capacitor is in the accumulation region and electrons pile up at the insulator boundary. In a real device, the existence of surface states would actually lead to a depleted surface. Such a small accumulation region is not expected to have a large impact on the total device characteristics.

4.3. Transient Analyses

4.3.1. Short-Circuit Current Decay

To determine the accuracy of the transient solution routines in PUPHS2D, the short-circuit current decay behavior of a solar cell was investigated, since a known approximate analytic solution exists for the one-dimensional case. The solar cell is subjected to some initial uniform generation rate and at time $t=0^+$ the generation of carriers is stopped. The current then decays with time as minority carriers recombine. The short circuit current as a function of time may be written as a series solution of an infinite number of modes [91]. After a short time, the higher order modes die out and the current decay is described by the first mode time constant:

$$J_{sc}(t) = J_{sc0} \exp \left( -\frac{t}{\tau_{J1}} \right) ,$$  \hspace{1cm} (4.1)

where:
Figure 4.3  Electron concentration in 2-D GaAs homojunction diode.
\[
\frac{1}{\tau_{II}} = \beta_{II}^2 D_p + \frac{1}{\tau_p},
\]
(4.2)
\[
\beta_{II} \cot (\beta_{II} W) = -\frac{S_p}{D_p}.
\]
(4.3)

Thus, the time constant of the current decay is close to the minority carrier lifetime, but may be a little less due to the fact that for a finite length device some carriers may transit the device without recombining. The time-dependent short circuit current computed by PUPHS2D for a test solar cell is shown in Figure 4.4.

The transient response shown in Figure 4.4 has precisely the expected shape. The higher-order modes are seen dying out after 0.25 nS, and the current decay is thereafter exponential, with a computed time constant of 1.05 nS. The expected time constant from the analytic solution is 0.94 nS. This

![Short-Circuit Current Decay](image)

**Figure 4.4** Short-circuit current decay in a GaAs solar cell.
close agreement was obtained without optimizing the finite difference mesh or the time step.

4.3.2. Transistor Switching

To check for qualitative agreement of a transistor transient response with theory, a transistor switching response was computed. For this test, an initial solution for an HBT with $V_{CB} = 3.0$ volts and $V_{BE} = 0.0$ volts was first computed. The base-emitter voltage was then instantaneously stepped to 1.5 volts and the transient response to this switch-on action was calculated. The transient collector current is shown in Figure 4.5. The transient collector current is initially negative due to capacitive feedthrough from the base to the collector. Initially, $V_{CB} = 3.0$ volts, but after the transient dies out $V_{CB}$ will be 1.5 volts and thus $dV_{CB}/dt < 0$. Remembering basic circuit theory, we know that $i = CdV/dt$, where $C$ is the junction capacitance of the collector-base junction. Since $dV/dt$ is negative, there will be a large initial negative collector current, as shown in Figure 4.5. After the capacitive feedthrough, the collector current slowly starts to rise toward its final value as the emitter-base junction becomes charged and the step of electrons injected from the base into the emitter finally start to be collected. Although the transient may appear to be rather long, it should be noted that in a circuit external resistors will play a role in determining the actual switching time. Additionally, the transistor would be subject to a much smaller voltage swing. Since the transistor won't begin to turn on until $V_{BE}$ is almost a volt, switching between $V_{BE} = 1.0$ volt and $V_{BE} = 1.5$ volts would be more typical and would produce a considerably shorter switching time. The transient shown compares favorably with the results of Tang for a one-micron silicon BJT [92], which are computed in a similar manner by ramping the base-emitter voltage from 0.5 volts to 0.85 volts. The $Al_xGa_{1-x}As$ HBT transient is about 4 times faster than the silicon transient.

4.4. Hot Electron Diode

The hot electron diode was chosen as the standard device for investigating the hot electron transport model. The one-dimensional $n^+-i-$
n+ structure analyzed by Constant [90] using the Monte Carlo method was simulated. Figure 4.6 shows the electron energy in the device computed by PUPHS2D by solving the local energy balance equation. High energy is imparted to the electrons by the large electric field in the intrinsic region of the diode ($V_A = 0.5$ volts). The Monte Carlo results of Constant show the same shape and similar magnitudes. PUPHS2D produces a maximum energy of about 0.37 eV, whereas the Monte Carlo model shows a maximum of around 0.33 eV. An effective electron velocity (which includes the effect of diffusion as well as drift velocity) defined by $V = J/qn$ is also plotted in Figure 4.6. The electron velocity from PUPHS2D is significantly different from the Monte Carlo calculation due to the inability of PUPHS2D to produce velocity overshoot. The Monte Carlo model shows velocity overshoot being maintained throughout the intrinsic region.

The double humps in the PUPHS2D velocity plot are interesting and easily explained. Electrons entering the high-field region are accelerated and gain energy. As they gain energy, however, their mobility starts to decrease,
Figure 4.6  Electron energy and effective velocity in a GaAs n⁺-i-n⁺ diode.

Figure 4.7  Electron concentration and mobility in the GaAs n⁺-i-n⁺ diode.
and they must slow down. This causes a pile-up of electrons in the intrinsic region where their energy is highest. Both of these features (the mobility behavior and accumulation of electrons) are shown in Figure 4.7. Since the effective velocity is inversely proportional to the electron concentration, the pile-up of electrons causes a momentary dip in their velocity.

4.5. Summary

In this chapter test computations were compared with expected solutions. PUPHS2D was shown to be capable of producing diode current-voltage characteristics which agreed very well with those predicted using an analytical model. The effectiveness of the method for modeling nonplanar geometries was verified. The accuracy of the transient solution technique was confirmed by comparison with analytical models for short-circuit current decay in solar cells. The transient response of a transistor was calculated and observed to have the correct general characteristics. The hot-electron model was tested on an n⁺-i-n⁺ diode and compared to Monte Carlo analysis of a similar device.
CHAPTER 5: MODEL APPLICATIONS

5.1. Introduction

Now that a working model for HBT's has been developed and its accuracy confirmed, it remains to apply the model to "real world" devices. The operating characteristics of an HBT as computed by PUPHS2D are presented and analyzed. Also included are studies of some fundamental questions which face the HBT designer, namely the issues of abrupt versus graded heterojunctions and the effect of the base contact placement on current gain. The results of these analyses will be compared to experimental studies.

5.2. HBT Characteristics

As an illustration of the ability of PUPHS2D to compute the internal and external (terminal) characteristics of an HBT the device shown in Figure 5.1 was simulated. This HBT is a mesa-etched structure and was modeled using the technique described in Section 3.3. A surface recombination velocity of $10^7$ cm/s was specified along the exposed areas of the emitter and base mesas. The solution was computed for a variety of base-emitter bias conditions, while the collector-emitter bias was held at 2.0 volts.

A log plot of the electron concentration in the transistor is shown in Figure 5.2. The base-emitter bias is 1.5 volts. Lateral diffusion of electrons in the quasi-neutral base and the depleted collector regions is a prominent feature. At this base-emitter bias the transistor is on the threshold of high level injection, which is evidenced by the large concentration of injected electrons which are present in the collector depletion region.

A linear plot of the recombination rate throughout the device of Figure 5.1 with $V_{BE} = 1.5$ volts is shown in Figure 5.3. Surface recombination at the exposed edge of the emitter mesa completely dominates the plot. This recombination behavior will be discussed in more detail in Section 5.3.
Figure 5.1 Typical heterojunction bipolar transistor.
Figure 5.2  Electron concentration in an HBT at $V_{BE} = 1.5$ V, $V_{CE} = 2.0$ V.
The terminal characteristics of the HBT in the form of a Gummel plot are shown in Figure 5.4. The high surface recombination velocity leads to base currents which are greater than the collector currents until the transistor is biased to around $V_{BE} = 1.1$ V, at which point the current gain finally surpasses 1. Series resistance causes the currents to "bend over" by 1.3 V. The collector current exhibits an ideality factor of 1.00. The base current ideality factor is around 1.8 for low $V_{BE}$ and reaches a minimum value of 1.4 for $V_{BE}$ between 1.2 and 1.3 volts. Surface recombination causes a severe degradation of the current gain, which is only 8.3 at $V_{BE} = 1.3$ V. The cutoff frequency performance of the HBT as calculated using the quasi-static method of Section 3.4 is displayed in Figure 5.5.

5.3. Placement of Base Contacts in HBT's

With the advent of self-aligned HBT fabrication processes, a subject of considerable interest has been the effect of placing the base contact in close proximity to the active device region. As was mentioned in Chapter 1, the primary reason for using a wide-gap emitter is the reduction in base resistance afforded by the use of an inverted base-emitter doping profile. The aim of self-aligned processes is to further reduce the base resistance by placing the base contact as close to the active region of the device as is possible. This may reasonably be expected to raise the base current (and lower the current gain) as injected minority carriers recombine at the base electrode rather than transit the base and contribute to the collector current.

This problem was studied using PUPHS2D by simulating the device shown in Figure 5.1: The emitter mesa to base contact spacing, $L$, was varied from 0.1 $\mu$m to 1.0 $\mu$m. Simulations were done for an unpassivated surface ($S = 10^7$ cm/s) and a perfectly passivated ($S = 0$ cm/s) surface. The results of the modeling are shown in Figure 5.6. It is clear that gain degradation occurs in the passivated device as the base contact is brought near the active region. No noticeable degradation occurs in the unpassivated device.

The current gain degradation in the passivated device is easily explained by examining the recombination rate in devices with different mesa-contact spacings. Figure 5.7 is linear plot of the recombination rate in a passivated HBT with $L = 1.0$ $\mu$m. Recombination in the base-emitter
Figure 5.3  Recombination rate in an HBT at $V_{BE} = 1.5$ V, $V_{CE} = 2.0$ V.
Figure 5.4  Base and collector current densities in an HBT.

Figure 5.5  Cutoff frequency behavior of an HBT.
Current Gain versus Emitter–Base Spacing

![Graph showing current gain dependence on emitter mesa–base contact spacing. Points shown are for $J_C = 10^4$ amps/cm².](image)

Figure 5.6  Current gain dependence on emitter mesa–base contact spacing. Points shown are for $J_C = 10^4$ amps/cm².

depletion region and in the quasi-neutral base constitute the vast majority of the base current. Figure 5.8 is the same plot for a passivated device with $L = 0.2 \mu$m. Recombination at the edge of the base electrode has become dominant, and the peak in recombination rate in the depletion region can only faintly be seen. When placed close to the active device region, the base electrode acts as a sink for injected minority carriers. A significant number of minority carriers injected from the emitter into the base do indeed recombine at the base electrode rather than reach the collector. The onset of this behavior occurs at a spacing of about 0.5 \mu m, and by a spacing of 0.2 \mu m (often quoted as the spacing for a self-aligned process), the current gain has dropped to half of its original value.

The absence of gain degradation for the unpassivated device may be understood by studying the plot of recombination rate previously shown in Figure 5.3. While this plot is for a mesa–contact spacing of 0.2 \mu m, the plot
Recombination Rate

Figure 5.7 Recombination rate in passivated HBT with $L = 1.0\, \mu m$, $V_{BE} = 1.35\, V$, and $V_{CE} = 2.0\, V$. 
Recombination Rate

Figure 5.8 Recombination rate in passivated HBT with $L = 0.2 \, \mu m$, $V_{BE} = 1.35 \, V$, and $V_{CE} = 2.0 \, V$. 
for \( L = 1.0 \mu m \) (not shown) is virtually identical. If the presence of a base electrode near the active device acts like a sink, then a high surface recombination velocity located at the junction behaves like a minority carrier black hole, sucking in all nearby minority carriers. Moving the base electrode has little or no effect because the region of high surface recombination is nearer (actually inside) the active region and completely overpowers any other effects.

The results of Figure 5.6 are in good qualitative agreement with the modeling results of Hiraoka and Yoshida, who performed a similar study using a two-dimensional model with a more rigorous treatment of surface recombination [93]. Hiraoka and Yoshida also see the onset of gain degradation at about 0.5 \( \mu m \) spacing, with considerable effect by \( L = 0.2 \mu m \). They report a slight degradation for their unpassivated HBT, however. This is because a much lower surface recombination velocity \( (S = 2 \times 10^5 \text{ cm/s}) \) is used, which doesn't overpower the base contact recombination completely.

While good agreement with other modeling results is obtained, a recent experimental paper by Lee et al. [94] shows a considerably larger degradation effect. Attenuation of the current gain by more than a factor of two is reported for \( L = 1.0 \mu m \). Lee does mention, however, that the emitter-base spacing might have 0.5 \( \mu m \) of alignment errors. The onset of degradation may therefore occur at a smaller spacing than reported. SEM determination of the actual spacings after fabrication would help in studying whether this is the case. A self-aligned process was used by Lee for the smallest base-emitter spacing, which was reported to be 0.2 \( \mu m \). It is difficult to say what the actual spacing for this device might have been, since it is determined by the properties of the wet etch procedure, and is not an explicit design parameter. A smaller spacing of 0.1 \( \mu m \) might easily have existed, which would help explain the drastic gain reduction observed by Lee. In any case, further investigation is required to reconcile the experimental and theoretical results.

5.4. Base Currents in Abrupt and Graded HBT's

Another area of interest to HBT designers is the use of graded base–emitter heterojunctions versus abrupt base–emitter heterojunctions. From a
process point of view, it is very difficult to make a good base-emitter junction because there are actually two junctions, a doping junction and a metallurgical junction, which must occur at the same position for best results. Grading the base-emitter junction helps to decrease the sensitivity to the metallurgical junction placement. The difficulty of producing both junctions at the same position is compounded by base dopant out-diffusion. The p-type dopants used to dope the base of an Npn HBT (typically Be or Zn) have high diffusivities and tend to out-diffuse into the emitter during subsequent processing. Frequently an undoped setback layer is placed between the emitter and the base in the hope that the p-type dopants will diffuse into the region and stop by the time they reach the emitter. The effect on current gain of a discontinuity in the conduction band profile of an abrupt base-emitter heterojunction was discussed in Section 1.1. Theoretically the current gain of an HBT can be raised by eliminating such discontinuities and for this reason, as well the above processing concerns, most contemporary HBT's incorporate some metallurgical grading scheme, usually linear grading over the last few hundred microns of the emitter.

In practice graded emitter transistors show a degradation of current gain due to increased base current. A recent study of this effect was reported by de Lyon et al. which showed base currents nearly two orders of magnitude greater for graded HBT's than for abrupt HBT's [95]. The authors explained this disparity by postulating the existence of a surface channel induced by Fermi level pinning at the surface. The energy barrier in the surface channel was deduced to be significantly smaller for graded junctions than for abrupt junctions, thereby increasing the base current in graded HBT's.

The Npn HBT's of de Lyon were simulated using PUPHS2D to determine whether a similar effect would be predicted. The results of this study are shown in Figure 5.9, which displays the abrupt and graded HBT base currents as predicted by PUPHS2D and as reported by de Lyon. To match the graded HBT current it was necessary to specify a surface recombination velocity of $2 \times 10^7$ cm/s. This is very high and is right at the thermal limit for recombination velocity. In the actual device it is possible that factors such as incomplete out-diffusion of the base dopants into the setback layer contributed to base current. With a surface recombination velocity of $2 \times 10^7$ cm/s the graded HBT base currents computed by the model match the
Figure 5.9 Base currents in HBT’s with abrupt and graded base–emitter heterojunctions. The experimental data are from [95].
experimental data of de Lyon very well. At high values of base–emitter bias, contact resistance in the real devices causes a more rapid bending over of the base current than is predicted by PUPHS2D. The ideality factor for the graded base current is constant at about 1.8.

The base current in the abrupt device as computed by PUPHS2D differs somewhat from the experimental result. This is not surprising, since the drift–diffusion model inadequately describes transport across a junction with a conduction band discontinuity, where effects such as tunneling and reflection may be significant. The proper treatment of injection across an abrupt heterojunction requires the use of thermionic emission theory and should include quantum mechanical effects (see, for example [96]). The uncertainty of junction placement introduced by the presence of out–diffused dopants further complicates transport. The same general characteristics are observed, however; the base current in the abrupt HBT is considerably less than in the graded HBT. Why is this the case? Fermi level pinning results from considering the charge present at surface trap sites when the Poisson equation is solved. Since this modification to the Poisson equation is not included in the basic surface recombination model of Section 2.4.2, PUPHS2D will not predict Fermi level pinning. The difference in base currents must therefore be explainable by some other mechanism. For a p–n junction a peak in recombination rate will occur somewhere within the depletion region. In the case of an HBT, where the base doping is greater than the emitter doping, this peak will be in the emitter. The peak in the abrupt HBT will occur in the wide–gap emitter, where for this example the aluminum mole fraction is 0.25. In the graded HBT, the peak recombination occurs in the grading region, where the aluminum mole fraction is less than 0.25. Peak recombination in the graded HBT therefore occurs in a region with a greater intrinsic carrier concentration than that of the abrupt HBT. Recombination current is proportional to intrinsic carrier concentration (see (A.3) in the Appendix); and therefore the graded HBT exhibits a larger base current than the abrupt HBT. While Fermi level pinning at the surface is almost certainly present, the different base currents for abrupt– and graded–junction HBT’s can be accounted for by the difference in $n_i$ at the recombination peaks.
5.5. Summary

In this chapter the two-dimensional drift-diffusion model was used to investigate some experimentally observed phenomena. The utility of the model was first demonstrated by finding the internal and external characteristics of a typical heterojunction bipolar transistor. Current gain degradation due to base electrode placement was studied and compared to other reported modeling and experimental assessments. The effect on recombination currents of using graded versus abrupt emitter-base heterojunctions was also examined and compared to experimental findings.
CHAPTER 6: CONCLUSION

6.1. Summary and Conclusions

The goal of this research was to develop an accurate model for studying heterojunction bipolar transistors. As has been described, this goal was achieved by modifying an existing solar cell analysis program, PUPHS2D, to treat HBT's. Changes in the control structure were necessary to add the capability of producing the desired current-voltage characteristics. Due to the highly nonplanar nature of modern mesa-etched transistors, a methodology for treating nonplanar geometries was devised. In order to evaluate the high-speed performance of HBT's for digital applications, transient capability was added by considering the time-dependent semiconductor equations. Provisions for analyzing the high-frequency performance of HBT's for analog applications were made by invoking quasi-static approximations, thereby allowing the computation of unity current gain cutoff frequencies. Self-consistent solution of a simplified energy balance equation with the conventional semiconductor equations was introduced in an effort to more accurately model high-field electron transport.

The modifications to the control structure necessary for transistor analysis were relatively minor. The bipolar transistor analysis routines allow the user to either sweep $V_{CE}$ as $V_{BE}$ is held constant, or to sweep $V_{BE}$ as $V_{CE}$ is held constant during computation of current-voltage characteristics.

The method used for treating nonrectangular device geometries was described in Section 3.3. While the method is admittedly somewhat crude, it is easily applied and produces very good results, as shown in Section 4.2. The alternative to this method is to allow termination of grid lines by using finite box discretisation. This would require a substantial rewrite of the code and would also significantly increase the code complexity; the method of Section 3.3 is deemed to be sufficient and therefore preferable. Development of a
general 3-D code which allows noncubic structures is being pursued at Purdue by J. L. Gray.

Transient capability was successfully added to PUPHS2D and shown to be accurate for the case of short circuit current decay in a solar cell, where an analytic solution is known. The transient behavior of an HBT to a turn–on transient was computed and shown to be qualitatively correct. The large number of time steps for which solutions are required to compute a transient response leads to a corresponding increase in computation time.

The newly–incorporated quasi–static analysis routines were exercised and produced reasonable results, as was shown in Figure 5.5. Theoretically, the cutoff frequency as computed by (3.4) should be the same whether computed using the incremental hole or the incremental electron charge. In practice, this was found to not necessarily be the case. This is probably due to the fact that a small perturbation in the base–emitter bias produces a very small change in the incremental charge. Given the very limited number of nodes available to the user, it is difficult to resolve the small change in incremental charge. This is especially true of the incremental hole charge in an Npn HBT, since the base is so heavily doped that the peak in incremental hole charge is very narrow. PUPHS2D computes cutoff frequencies using both the incremental electron and hole charges; if the frequencies differ by more than ten percent, the user is warned of possible accuracy problems. The actual cutoff frequency then reported in the tabular output is that computed from the incremental electron charge, since it will in general be more accurate.

A localized energy balance equation was implemented in the code as described in Section 3.6. While the balance equation was able to produce the results of Section 4.4, the solution was found to be reluctant to converge. The hot electron diode results were produced by heavily damping the mobility corrections computed from the energy balance equation. For the case of a heterojunction bipolar transistor, the energy balance equation method was unsuccessful due to the breakdown of the bisection solution to the balance equation at the interface of the base–emitter junction and the insulator used for nonplanar modeling. The balance equation as it stands must therefore be viewed as only a preliminary step toward accurate high–field computations.
6.2. Recommendations

While PUPHS2D is a functional and useful heterostructure device simulation tool, ongoing development is necessary to keep pace with the rapid advancement of semiconductor technology and experimentation. This continuing development falls into two main categories: optimization and device physics enhancements.

A small amount of optimization could make PUPHS2D considerably more tractable. The most significant change would involve reducing the number of voltage steps required to reach desired bias points. As was briefly mentioned in Section 3.2, an educated guess as to the position of the quasi-Fermi levels could be used to generate a much better initial guess to the solution than using the previous solution as the initial guess. If this method were to be perfected and implemented in the code, it could significantly decrease execution time by allowing much bigger voltage steps to be taken on the way to the desired bias points.

The transient routines should also be optimized. Presently the time step used for computation of transient responses is constant. This time step is usually very small in order to resolve the beginning of the transient, which may be quite fast. As the solution approaches steady state, however, the time step is much too small to be economical. An adaptive time step should be incorporated in order to increase computational efficiency. It should also be reiterated that the transient response represents a potentially faster method for obtaining steady-state results without voltage stepping (see Section 3.5.2). The advantage of this method over the quasi-Fermi level method is that in addition to the desired steady-state solution, the transient characteristic of the device is computed as a bonus. Inclusion of displacement current computation should be considered for improved generality of the transient response.

The most obvious topic for further investigation under the heading of device physics enhancement is the treatment of high-field electron transport. Before considering this topic, however, a more fundamental question should be raised, namely the appropriateness of the mobility models currently being used. As mentioned in Section 2.4.3, the mobility models in place are fit to data that are over a decade old and probably not representative of the
mobilities achieved by today's high quality growth techniques. The issue of
majority versus minority carrier mobility should also be investigated, as
PUPHS2D assumes both are equal.

The question of high-field transport may now be addressed. The
inclusion of a localized energy balance equation has proven to be a promising
first step, but is insufficient to accurately and efficiently model high-field
transport. It is therefore recommended that the balance equation be
delocalized and further developed. Since delocalization will introduce a term
containing the gradient of the energy, this will tend to damp changes in the
mobility and stabilize the problem.

Two more general issues should also be mentioned. The most
important of these is the extension of PUPHS2D to treat material systems
other than the AlGaAs–GaAs system. As more interest is focused on systems
such as InAlAs–InGaAs and Si–SiGe, it is appropriate to consider adding
material subroutines to allow modeling of devices in these systems. Since the
material parameters used by PUPHS2D are contained in a very small number
of subroutines, it is expected that the changes would be fairly minor and
straightforward.

The other general issue of concern involves the matrix solution
technique currently being employed. The large matrix equation (2.36) is
solved by direct inversion of the Jacobi matrix. Since this is a very large
matrix, inversion takes a substantial amount of memory. The number of
nodes allowed by the LINPACK solver routines is limited to 2000. This is
marginal, and almost certainly not enough to obtain accurate results for very
complex structures. In the future it may be necessary to use a technique
which allows more nodes.
LIST OF REFERENCES
LIST OF REFERENCES


[81] M. E. Klausmeier-Brown, M. R. Melloch, and M. S. Lundstrom, "Transistor-based measurements of electron injection currents in p-type GaAs doped $10^{18}$ to $10^{20}$ cm$^{-3}$," to be published.


APPENDIX
APPENDIX: ANALYTIC SOLUTION TO THE ONE-DIMENSIONAL P⁺–N DIODE UNDER FORWARD BIAS

The current in a forward–biased p–n junction diode may be written as the sum of diffusion current and current due to recombination in the space-charge region:

$$J_D = J_{01} \exp \left( \frac{qV}{kT} \right) + J_{02} \exp \left( \frac{qV}{2kT} \right), \quad (A.1)$$

where $J_{01}$ is the diffusion saturation current density and $J_{02}$ is the recombination saturation current density. The first term is simply the current given by the well–known ideal diode equation [1]. The diffusion saturation current density for a p⁺–n diode is:

$$J_{01} = q \left[ \frac{D_p n_{i0}^2}{L_p N_D} \right], \quad (A.2)$$

where $D_p$ is the hole diffusion constant, $L_p$ is the hole diffusion length, and $N_D$ is the doping density in the n–type region. The recombination saturation current density in the depletion region can be written as [2]:

$$J_{02} = \frac{q n_{i0} kT}{\sqrt{t_n t_p} \frac{kT}{qE}}, \quad (A.3)$$

where $t_n$ and $t_p$ are the electron and hole minority carrier lifetimes and $E$ is the electric field at the junction. $kT/qE$ has units of length and is therefore viewed as the characteristic width $W_{\text{eff}}$ over which the recombination takes place. $W_{\text{eff}}$ for a p⁺–n junction may be written as:

$$W_{\text{eff}} = W_{\text{SCR}} \frac{\pi}{4} \frac{kT}{q} \left( (V_{\text{bi}} - V) \left( V_{\text{bi}} - \frac{V}{2} - \frac{kT}{q} \ln \left( \frac{N_A}{n_{i0}} \right) \right) \right)^{1/2}. \quad (A.4)$$
In this equation, $W_{SCR}$ is the width of the space-charge region, $V_{bi}$ is the built-in junction potential, and $N_A$ is the doping density in the p+ region.

LIST OF APPENDIX REFERENCES
