INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Applications of resonant tunnel diodes and effects of a novel cathode structure

Lear, Kevin L., Ph.D.
Stanford University, 1990

Copyright ©1990 by Lear, Kevin L. All rights reserved.
APPLICATIONS OF RESONANT TUNNEL DIODES
AND EFFECTS OF A NOVEL CATHODE STRUCTURE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Kevin L. Lear
January 1990
I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

James S. Harris, Jr.
(Principal Adviser)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Malcolm Beasley
(Applied Physics)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Thomas Kailath

Approved for the University Committee on Graduate Studies:

Elizabeth Close Franpott
Dean of Graduate Studies
Abstract

The double barrier resonant tunnel diode (DBRTD) is a device based on wave-function interference in nanometer sized layers. Its characteristics resemble those of the junction tunnel diode, but it can be faster, more reliable, and more flexible in design. These improvements warrant an examination of the use of DBRTDs in existing or proposed applications using junction tunnel diodes as well as entirely new applications. The DBRTD is an excellent research vehicle for studying hot electron and quantum effects which will dominate future nanoelectronic devices. Some aspects of transport in these devices are not yet well understood, and more adequate models need to be developed.

Fabrication and measurement techniques for the characterization and application of DBRTDs have been developed. Known semiconductor technology has been applied to the microfabrication of DBRTDs using etched mesa and ion implantation isolation. The effects of parasitic oscillations on DBRTD characteristics have been analyzed and suppressed. Stabilized DC and microwave measurements including the negative differential conductance region have been made.

The use of DBRTDs as loads in tunnel diode FET logic has been investigated. Optimized tunnel diode load characteristics have been determined as a function of the speed vs. power tradeoff. Monolithic inverters, fabricated from composite molecular beam epitaxy layers for DBRTDs and MESFETs, have been characterized. The inverters offer significant savings in chip area and can function as static memory elements with a single transistor.

The full role of the cathode region and particularly the accumulation layer has not been incorporated into either the design or theory of DBRTDs. To gain a better understanding of the role of the cathode region in transport in DBRTDs, device structures with pseudomorphic quantum wells in their cathodes have been fabricated and characterized. The devices show up to a 3.5 times increase in peak current, achieving a peak current density of $1 \times 10^8$ A/cm$^2$ and simultaneously a peak-to-valley ratio of 5.2 at room temperature. Possible mechanisms to explain the effects are offered, and future experiments are suggested.
Acknowledgements

I welcome the opportunity to express my gratitude to a number of people who have helped me during my time at Stanford. At the top of the list is my advisor, Professor James Harris, whose diligent efforts have created an environment that made this work possible. He maintains a group of excellent students and has always been able to provide us with adequate resources and funds for our work. My exposure to the many areas of research in our group is a direct consequence of the enormous breadth of his experience and knowledge of semiconductor materials and devices. Perhaps most importantly, he has taught me about individually directed research by allowing me to determine my own directions without pressure, while still encouraging me when I doubted my work.

In addition to Professor Harris, I thank Professors Beasley and Kailath for reading my dissertation and Professors Bloom and Tobagi for serving on my orals committee. I also appreciate the advice and encouragement of Professor Pease. I benefited from frequent interaction with Professor John Owens of Santa Clara University regarding the microwave properties and applications of DBRTDs. One of his colleagues, Dave Halchin, performed several of the initial measurements on the devices.

This work could not have been done without the assistance of a number of students. Foremost, I wish to thank those who have grown MBE samples for me, especially Won-Seong Lee who grew the majority. Won has also been a valuable source of processing information for myself and others. Kanji Yoh was responsible for helping me get started on fabricating circuits. A large number of Professor Harris's past and present students are to be credited with establishing, maintaining, and improving the GaAs lab in which my work has been done. The patient answers and help I have received from several of Professor Bloom's students are greatly appreciated. I have had a number of useful discussions and collaborations with others who have worked on DBRTDs or other quantum devices including Peng Cheng, Steve Chou, Byung Park, Ken Shepard, Dominique Thomas, and Ed Wolak of the Harris Group and Scott Diamond and Eckmel Ozbay from the Bloom Group.

The technical staff of the CIS and Ginzton Microfabrication Laboratories have performed fabrication steps and made masks on a number of occasions. I will miss...
the helpful service of Gail Chun-Creech, Susie Burns, Sherrie Harvey, and Lachen Pence who seemed to always get their work done despite being readily available for friendly conversation with dozens of graduate students.

Finally, I thank my many good friends from school and church, and my family for their support.
to God
and his people here,
especially Monna
Contents

Abstract iv

Acknowledgements v

List of Tables xii

List of Figures xiii

1 Introduction 1

2 Background Material 8

2.1 Compound Semiconductors ........................................ 8

2.1.1 Band Energies .................................................... 9

2.1.2 Heterojunctions .................................................. 9

2.1.3 Effects of Strain ................................................. 11

2.2 Molecular Beam Epitaxy ............................................ 12

2.2.1 MBE System ..................................................... 13

2.2.2 Growth Considerations .......................................... 14

2.2.3 RHEED ............................................................ 15

2.2.4 Migration Enhanced Epitaxy .................................... 16

2.3 Resonant Tunnel Diodes ........................................... 16

2.3.1 DBRTD Structure ............................................... 19

2.3.2 DBRTD Operation ............................................... 20

2.3.3 Design Parameters .............................................. 23

2.3.4 Peak Parameter Expressions ................................... 28

3 Device Fabrication and Measurement 32

3.1 Device Fabrication ............................................... 32

3.1.1 Alloyed Contacts .............................................. 32

3.1.2 Non-alloyed Contacts ......................................... 34

3.1.3 Mesa Isolation .................................................. 34
3.1.4 Implant Isolation ................................................. 36
3.2 Stabilized Measurements ............................................. 38
  3.2.1 Amplification by NDR ......................................... 38
  3.2.2 Parasitic Oscillation ......................................... 40
  3.2.3 Stabilization Techniques ...................................... 40
3.3 Microwave Measurements .............................................. 45
  3.3.1 Equivalent Circuit ........................................... 45
  3.3.2 Measurement Technique ....................................... 47
  3.3.3 Measurement Results .......................................... 47

4 Tunnel Diode Applications ........................................... 52
  4.1 Tunnel Diode Comparison ......................................... 53
    4.1.1 Junction Tunnel Diode Operation ............................ 53
    4.1.2 Flexibility .................................................. 54
    4.1.3 Speed ....................................................... 55
    4.1.4 Reliability ................................................ 57
    4.1.5 Uniformity ............................................... 58
  4.2 Tunnel Diode Applications .................................... 60
    4.2.1 Amplifiers and Oscillators .................................. 60
    4.2.2 Detectors and Mixers ....................................... 61
    4.2.3 Switching and Pulse Circuits ................................. 62
    4.2.4 Digital Proposals .......................................... 62
    4.2.5 Multi-level Applications .................................... 63
    4.2.6 Resonant Tunneling Transistors ............................... 64

5 Tunnel Diode FET Logic ............................................. 67
  5.1 Static Operation ................................................. 67
    5.1.1 Operating Points .......................................... 68
    5.1.2 Bistability ............................................... 69
  5.2 Dynamic Operation ............................................... 71
    5.2.1 Node Equations ........................................... 71
    5.2.2 Delay Minimization ......................................... 73
    5.2.3 Origin of Power vs. Speed Tradeoff .......................... 73
  5.3 Modeling of Optimized Loads ................................... 74
# List of Tables

2.1 Lattice constants of InAs, GaAs, and AlAs. ........................................... 12
2.2 Average wavelengths of thermal Γ-point electrons and electron scattering lengths for undoped AlAs, GaAs, and InAs at room temperature. 18

5.1 Combined MBE layers for MESFET and RTD ............................................. 80

6.1 Layer structure of DBRTD cathode well samples ..................................... 91
6.2 Thicknesses of undoped layers in anode and cathode spacers of cathode well DBRTD samples ................................................................. 91
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Bandgaps of InAs, GaAs, AlAs, and intermediate InGaAs and AlGaAs alloys at T=300K</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Molecular beam epitaxy system’s growth chamber</td>
<td>13</td>
</tr>
<tr>
<td>2.3</td>
<td>Band extrema as a function of position in a typical AlAs/GaAs double barrier resonant tunnel diode structure</td>
<td>20</td>
</tr>
<tr>
<td>2.4</td>
<td>Typical current-voltage characteristics of a DBRTD</td>
<td>21</td>
</tr>
<tr>
<td>2.5</td>
<td>Resonant energies referred to GaAs electrodes for In(<em>y)Ga(</em>{1-y})As wells as a function of well size</td>
<td>25</td>
</tr>
<tr>
<td>2.6</td>
<td>The transmissivity of single monolayers of GaAs, Al(<em>{0.3})Ga(</em>{0.7})As, and AlAs as a function of energy below the (\Gamma)-points of those materials</td>
<td>27</td>
</tr>
<tr>
<td>2.7</td>
<td>Current voltage characteristics of a DBRTD with three monolayer barriers at liquid nitrogen temperature</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>Cross-section of a mesa isolated DBRTD</td>
<td>36</td>
</tr>
<tr>
<td>3.2</td>
<td>Cross-section of an implant isolated DBRTD</td>
<td>37</td>
</tr>
<tr>
<td>3.3</td>
<td>Schematics of mismatched transmission lines</td>
<td>39</td>
</tr>
<tr>
<td>3.4</td>
<td>Items used to package and fixture DBRTDs for connection to 3.5 mm 50 (\Omega) line</td>
<td>42</td>
</tr>
<tr>
<td>3.5</td>
<td>Wirebonded device inside a pill package</td>
<td>43</td>
</tr>
<tr>
<td>3.6</td>
<td>DC current-voltage characteristics of a DBRTD with and without suppression of parasitic oscillations</td>
<td>44</td>
</tr>
<tr>
<td>3.7</td>
<td>Small signal equivalent circuit for tunnel diodes</td>
<td>45</td>
</tr>
<tr>
<td>3.8</td>
<td>Measured and modeled data for the reflection gain of a DBRTD</td>
<td>48</td>
</tr>
<tr>
<td>3.9</td>
<td>Capacitance of different sized DBRTDs devices on semi-insulating substrates</td>
<td>49</td>
</tr>
<tr>
<td>3.10</td>
<td>DC characteristics and junction capacitance for DBRTD measured at 130 MHz</td>
<td>51</td>
</tr>
<tr>
<td>4.1</td>
<td>Band diagram of a junction tunnel diode under small forward bias</td>
<td>53</td>
</tr>
<tr>
<td>4.2</td>
<td>Difference between peak and valley current density vs. peak voltage for various devices measured in connection with this work.</td>
<td>56</td>
</tr>
</tbody>
</table>
4.3 DBRTD peak current and peak voltage as a function of radial position on 2” wafers .................................................. 59

5.1 Schematic diagrams and characteristics of enhancement/depletion FET and tunnel diode FET logic inverters ........................................ 68
5.2 Transfer characteristics of a TDFL gate exhibiting hysteresis ........... 70
5.3 Functional diagram of an inverter composed of a driver and a load ...... 72
5.4 Partitioning of available driver currents into two regions by the load line 74
5.5 Calculated waveforms for subsequent gates of a chain of inverters .... 76
5.6 Average static current as a function of the maximum gate delay for RTD loads and D-FET loads ................................................. 77
5.7 Specific optimum models for TD load characteristics .................... 78
5.8 A schematic cross-section of an integrated MESFET and RTD ........ 80
5.9 Photograph of a TDFL gate made from a monolithically integrated MESFET and RTD ......................................................... 81
5.10 Current-voltage characteristics of RTD loads and MESFET drivers and the resulting inverter transfer characteristics ............................................... 82

6.1 Conduction band diagrams for calculated resonant biases ................. 87
6.2 MEE shutter timing cycles ...................................................... 90
6.3 Photoluminescence spectra of the DBRTD samples with cathode wells 92
6.4 Room temperature, forward current-voltage characteristics of diodes with and without cathode wells .............................................. 94
6.5 Boiling nitrogen temperature characteristics of diodes with and without cathode wells .......................................................... 95
6.6 Room temperature current-voltage characteristics of diodes with different sizes of cathode wells .............................................. 96

A.1 Potential profile of a hypothetical barrier used for comparing methods of calculating tunneling spectra .............................................. 130
A.2 Spectra calculated for the tunneling barrier in Figure A.1 using numerical integration and transfer matrix methods .................................. 132
A.3 Root-mean-square error of calculated transmission probabilities and reflection phase shifts ..................................................... 133
Chapter 1
Introduction

The shrinking dimensions of microelectronics have pushed into the new regime of nanoelectronics, aided by the advent of new technology for lithography and crystal growth. Electron beam, ultraviolet, and X-ray lithographic tools permit the patterning of features that are on the order of 10 nanometers in size. Molecular beam epitaxy and metal-organic chemical vapor deposition can create crystal layers with a resolution on the order of the crystal’s lattice constant, a fraction of a nanometer. At these scales, not only is the flow of single electrons in devices like transistors no longer subject to statistical averaging, but also the effective wavelength of thermal electrons can be greater than the device size. Quantum interference phenomena can thus play a direct role in determining the properties of a device.

Nearly all aspects of contemporary electron devices can be successfully analyzed in terms of semiclassical theories. These relatively large devices contain many particles that interact with the crystal lattice many times within the device, so only the average and not the individual particle’s dynamic evolution is apparent at the terminals. The underlying quantum mechanics can be wrapped up in terms of phenomenological bulk material parameters such as band gaps, effective masses, and mobilities. Clearly, these devices will function differently if scaled to nanometer dimensions where single carriers can pass through the device without equilibrating with the lattice.

Some problems due to non-equilibrium transport, such as hot electrons, have already been encountered by devices on the micrometer scale. Further and more dramatic complications can be expected for nanometer-sized versions of contemporary devices due to quantization effects. While the nanometer regime holds new problems, it also provides great promise for revolutionary systems that exploit these new phenomena. One proposal for quantum coupled integrated circuits amounts to interacting single-electron boxes[1]. Other ideas for “wavefunction computation”, perhaps analogous to optical computation, have begun to be voiced.

Although today’s nanofabrication technology has capabilities well beyond those
required for the manufacture of VLSI ICs, these capabilities are not sufficient for the fabrication of the proposed new quantum systems. Before these can be made, lithography, processing, and epitaxial regrowth must advance so that true three-dimensional structures can be made with the resolution and crystal perfection molecular beam epitaxy (MBE) now affords in one-dimension. Even then, the current perception of these concepts seems somewhat dubious, or at least ill-defined, and it is almost certain that we do not yet know the best use of this hypothetical, future technology.

In the nearer term, some advanced devices are being developed as potential replacements or extensions for contemporary devices in known applications. Nanoelectronics technology benefits these devices in two ways. First, it can produce the nanometer features required by devices that utilize hot-electron or quantum effects. Second, and perhaps more importantly, it provides immense freedom to tailor devices and particularly materials for specific applications. For example, MBE can juxtapose certain materials with varying properties in a nearly arbitrary fashion. Artificial crystals or superlattices can be created that have electrical and optical properties that cannot be found in any natural crystals. The resulting devices are faster and more flexible than their conventional counterparts, and they may provide entirely new functional levels. To the circuit designer, these are the properties that count. Today, nanofabrication can produce electronic devices which are revolutionary on the inside but still appear evolutionary on the outside. Hopefully, they will be stepping stones to future devices which are revolutionary inside and outside.

The double barrier resonant tunnel diode (DBRTD), the subject of this thesis, is prototypical of these intermediate devices. It requires epitaxial layers that are 10 nanometers or less and operates on the basis of electron wavefunction interference. Its terminal characteristics resemble an older tunneling device, the junction tunnel diode, although it is faster and has far greater design and optimization possibilities. The junction tunnel diode is also called the Esaki diode, after its inventor, Leo Esaki, who discovered it while doing experiments on heavily doped p-n junctions in 1957[2]. It was also Esaki, who along with colleagues Tsu and Chang at IBM, did the seminal work on the DBRTD more than 15 years later[3, 4]. In 1973, Esaki received the Nobel prize for his development of tunneling devices[5, 6]. Exemplary of the design freedom
afforded by molecular beam epitaxy, the tunneling structure in the DBRTD can be engineered to produce a broad range of characteristics, while tunneling in the junction tunnel diode is mainly controlled by a fundamental bulk semiconductor parameter, the bandgap.

Because of its improved speed and design flexibility, the DBRTD is expected to be a successful replacement for the junction tunnel diode in those applications using tunnel diodes. Unfortunately modern uses of the junction tunnel diode are limited to a few niche applications, despite early enthusiasm about a pervasive role for its use in electronics. Numerous proposals for circuits were published in the early 1960s, many of which are collected in handbooks[7, 8]. Part of the optimism derived from the precept that tunneling devices are inherently fast. In the early 1960s, junction tunnel diodes were faster than competing devices like the transistor, created more than 10 years earlier.

However, between 1965 and 1980, the junction tunnel diode’s reputation declined, and it has failed to ever be used to any significant extent in digital circuit applications. In the absence of any well known, clear postmortem on the device, common lore has arisen among semiconductor device physicists that its failure in digital circuits was due to having one too few terminals[9, 10]. Because the transistor uses separate terminals for the input and output signals with respect to a common ground, the input and output are isolated, a luxury not afforded with just two terminals. Schemes, sometimes elaborate, were invented to solve the problem. Two common ones were to use either rectifying diodes or transistors between tunnel diodes.

The junction tunnel diode has endured better in some special communications applications[11, 12]. By far, the largest among these is its use as a zero bias detector or mixer for microwaves. Here it effectively functions as a zero volt zener diode and is called a back diode. It is notable that the trademark negative differential conductance of the tunnel diode is not important for this application. The reflective gain provided by the negative differential conductance in tunnel diodes is, however, important for their use in amplifiers or oscillators. The extent of the tunnel diode’s application in these applications appears to be quite small. Other devices such as Gunn and IMPATT diodes are capable of providing higher power output. Finally, the tunnel
diode is used in fast switching circuits, such as those in sampling oscilloscopes.

Because these present applications of the junction tunnel diode are limited and they occur outside the realm of devices for digital integrated circuits, many have questioned the need to reincarnate the tunnel diode, complete with its two-terminal tragic flaw. In response to these criticisms, great emphasis has been placed on trying to make the DBRTD into a three-terminal resonant tunneling transistor. This requires contacting the thin tunneling layers to form a base, or control electrode, which would have high parasitic resistance. Other intermediate device composites of DBRTDs and transistors are easier to make, but rely on classical drift and diffusion transport mechanisms in the transistor, in addition to wavefunction interference in the DBRTD.

Because the fabrication of true three-terminal resonant tunneling devices is so difficult, much work has continued to be done on the resonant tunnel diode as a two-terminal device. There are different justifications offered to support research on DBRTDs in response to the pragmatic criticisms it receives. One is that the improvements of DBRTDs over junction tunnel diodes are great enough to warrant renewed investigation of tunnel diode circuits. The improved speed of DBRTDs should allow them to surpass the performance of junction tunnel diodes or other negative differential conductance devices in existing microwave and switching applications. The additional flexibility of DBRTDs compared to tunnel diodes could make them appropriate for new two-terminal applications or older unsuccessful propositions. It may also be possible that it was not the two-terminal nature of the junction tunnel diode, but other frailties that kept it from competing with the transistor. It is interesting to note that even after Shockley invented the transistor, he was interested in two-terminal negative resistance devices for circuit applications[13, 14].

A second motivation for continued DBRTD research is that knowledge gained about DBRTDs may benefit other tunneling or hot-electron devices. This is certainly a defendable position. Much has been discovered about materials and structural configurations in the context of DBRTD research that can readily be applied to more sophisticated devices. DBRTDs, an extreme example of finite superlattices, have been an excellent vehicle for studying fundamental aspects of quantum transport in semiconductor heterojunction systems as well as for pushing materials quality.
Today, mostly relatively simple theories exist to describe DBRTD operation, and the differences between the predictions of these simple theories and the observations of real devices have pointed out important phenomena that must be considered in the fabrication of nanoelectronic devices. This exploratory work on DBRTDs can also uncover new applications or design parameters that can be used to enhance or optimize DBRTD performance.

The above motivations encompass the two foci of DBRTD research presented in this dissertation. The first is a specific application of resonant tunnel diodes used as two-terminal devices. The circuit configuration is called tunnel diode FET logic, and it uses tunnel diodes as loads for enhancement-mode MESFET drivers to form logic gates. The combination is topologically similar to one of the composite three-terminal devices mentioned earlier. Here however, there is an internal connection between the FET and the diode; so the diode functions as a two-terminal element rather than the entire combination appearing as a three-terminal device. Tunnel diode FET logic can be designed to have low power dissipation, and single transistor gates can be used as static memory cells.

The second focus is an investigation of the effects of the cathode structure on the characteristics of the standard DBRTD structure. The experiment was designed to emphasize attributes of transport not accounted for in most models of DBRTDs by systematically changing the accumulated carrier concentration in the cathode next to the barriers. Various sizes of quantum wells placed in the cathodes of DBRTDs produced dramatic effects. The peak current of the devices was increased by 3.5 times to current densities of $1 \times 10^5$ A/cm$^2$, concurrent with peak-to-valley ratio increases to 5.2 at room temperature. The results have both practical and theoretical implications.

Before the details of these two items are presented, chapter 2 summarizes the necessary background information. It contains descriptions of the compound semiconductor materials used in DBRTDs, the molecular beam epitaxy technology capable of combining these materials into structures with sub-nanometer features, and a qualitative description of the operating fundamentals of resonant tunnel diodes. Description of the work completed for the thesis begins in chapter 3, which covers experimental
CHAPTER 1. INTRODUCTION

technical items pertinent to the remainder of the work. The first part describes fabrication techniques used to make devices from epitaxial layers on both semi-insulating and conducting substrates. The second part of Chapter 3 deals with measurement issues. It covers the difficulties of measuring negative differential conductance devices, and results of DC and microwave measurements on DBRTDs.

Prerequisite to application of DBRTDs is establishing that they can be produced with attributes appropriate for modern integrated circuits. This is especially important in view of the junction tunnel diode’s failure. The first half of chapter 4 accomplishes this by comparing the salient points of flexibility, speed, reliability, and uniformity of the two devices. The second half of chapter 4 surveys a wide variety of tunnel diode applications.

With the appropriate foundation laid, chapter 5 describes work on tunnel diode FET logic using DBRTDs as load elements, the major application discussed in this dissertation. It reviews the proposal of tunnel diode FET logic using junction tunnel diodes, details its benefits, and recalls the comparison of the two types of diodes made in chapter 4 to motivate the use of DBRTDs. Because a wide range of characteristics are available using DBRTDs, the question of an optimum load characteristic is studied along the speed vs. power curve, and a comparison between DBRTD and other types of loads is presented. The presentation of the experimental applications work starts with the epitaxial layer design, a brief mention of the fabrication, and the resulting monolithic inverter structure. After the measured characteristics are shown, the chapter concludes with a discussion of the difficulties encountered and possible solutions.

The last main part, chapter 6, turns from applications to experimental work on the effects of cathode wells in DBRTD structures. It begins by outlining the space charge effects, transport issues, modeling assumptions, and previous studies that formed the motivation for the experiments. The structures used in the experiments are described and modeled. The results are presented along with evidence that the observed changes are due to the intended modification. Possible mechanisms to explain the effects are offered, and additional experiments proposed to the effects of the individual mechanisms. The chapter closes with a discussion of the experiment’s
implications for DBRTD design and potential applications. Chapter 7 concludes this dissertation with a summary, ideas for future work, and an opinion about the outlook for DBRTD applications.
Chapter 2
Background Material

The work described in this thesis was done on double barrier resonant tunnel diodes (DBRTDs) made from the binary III-V compound semiconductors GaAs, AlAs, and InAs and their ternary alloys. The small electron effective masses and variability of electron potential energy make this an excellent materials system for DBRTDs. Molecular beam epitaxy is not only capable of growing high quality layers of these materials, but of doing so with resolution approaching that of the crystal lattice. One can make reasonably arbitrary structures of layers with differing attributes and relatively defect free interfaces. The combination of high quality layers and interfaces, varying material properties, small effective masses, and sub-nanometer features allows the creation of devices, like the DBRTD, based on quantum interference effects. The high degree of control and multiplicity of layer arrangements are the source of the great flexibility in device design.

This chapter surveys background material necessary to understand the operation, fabrication, and applications of DBRTDs. It opens with the properties of the arsenide compound semiconductors pertinent to DBRTDs. The fundamentals of molecular beam epitaxy are overviewed next. Having covered the materials and how structures are made with molecular beam epitaxy, the last section begins with the specific structure used in DBRTDs. A qualitative description of DBRTD operation, a discussion of design parameters, and formulae for estimating their important characteristic parameters complete the section.

2.1 Compound Semiconductors

The III-V compound semiconductors that are most widely used for devices at present have arsenic as the group V element, combined with indium, gallium, or aluminum from group III. The resulting binary compounds InAs, GaAs, and AlAs along with ternary mixtures or alloys of the binaries form a materials system appropriate for making heterojunction devices. Most DBRTD experiments have been done with
CHAPTER 2. BACKGROUND MATERIAL

electrons as the majority carriers because their effective mass is much smaller than that for holes. Silicon is the usual donor dopant used to provide conduction electrons. The small effective masses of electrons in the arsenide materials allow them to have Bloch state wavelengths greater than 10 nm at thermal energies. The materials can be made with high crystalline quality so that they possess large mobilities and long scattering lengths. The semiconducting properties of the binaries are different, and intermediate properties can be obtained with alloys whose properties vary linearly with composition, according to Vegard's Law[15]. This allows interfaces with controllable differences which alter the electron wavefunction. In addition to changes in effective mass, this is accomplished with differences in conduction band energies.

2.1.1 Band Energies

The III-V arsenide compounds have bandgap energies that differ from semiconductor to semiconductor. The direct, $\Gamma$-point bandgaps of InAs, GaAs, AlAs, and intermediate InGaAs and AlGaAs alloys are shown in Figure 2.1 by the solid line. The bandgap can be adjusted to values in between those for the binaries by an appropriate choice of the ternary alloy composition. The indirect $X$-point bandgap is smaller than the direct band gap for AlGaAs alloys of more than about 45% AlAs. The $X$-point bandgap of the compound semiconductors in Figure 2.1 is plotted with a dashed line.

2.1.2 Heterojunctions

When two different semiconductor materials are juxtaposed, a heterojunction is formed which has different properties on each side. If the connection of the materials does not modify their physical structure, it is reasonable to assume that the properties of the materials a long way from the interface are unperturbed. The details of how the properties change near the interface is more involved and can depend on the nature and perfection of the interface[16]. Of particular interest to DBRTDs are the band parameters that control the dispersion relation for electrons, namely the band energy minima and maxima and effective masses.

Because carrier transport is largely determined by the profile of the individual
band energies rather than the bandgap, it is necessary to determine how the difference in bandgaps at heterojunctions is accommodated separately by changes in the conduction and valence bands. This problem was one of the first to be addressed in heterojunction device research, but adequate results are still not available for all systems. The work in this paper will assume that the change in conduction band Γ-point is 70% percent of the bandgap difference for all arsenide compound semiconductors unless noted otherwise.

Both the concepts of band energies and effective masses are rigorous only for truly
periodic materials. Since the periodic symmetry sensed by a localized electron is broken near an interface, these concepts become poorly defined near heterojunctions. If we wish to retain these concepts, we must assume a spatial variation of the properties near the heterojunction. The simplest assumption is that the properties change discontinuously from one bulk value to another at the interface. This is called the envelop function approximation and has been enormously successful[17]. It will be used in this work.

2.1.3 Effects of Strain

It is possible for the connection of two materials at a heterojunction to modify the physical properties of one or both materials. If the lattice constants of the two separate materials are not identical, then the materials must be strained to make them commensurate, otherwise crystal dislocations will be generated at the interface, degrading its electrical properties. Thin epitaxial layers can be grown commensurately on bulk semiconductor substrates or thick layers with different lattice constants. However, at some critical thickness, it will be energetically favorable for the thin layer to form dislocations and relax rather than remain strained to the thick layer's lattice constant. This critical thickness depends on the difference in natural lattice constants between the two materials[18].

Because the lattice constants of AlAs and GaAs, seen in Table 2.1, are nearly the same they are called lattice matched. The critical thickness for layers of either of these on the other is so large that practically infinite thicknesses can be grown without detrimental effects. The lattice constant of InAs, on the other hand, is 7% larger than that of AlAs and GaAs. Only thin layers of InGaAs alloys can be grown on GaAs before they exceed the critical thickness. For example, the limit for In$_{0.15}$Ga$_{0.85}$As is about 100 Å[18].

The strains in a layer to accommodate the lattice constants of another layer cause a modification of the strained layer's electronic properties. The biaxial strain of the layer in the plane of the heterojunction changes the material's bandgap, including splitting the normally degenerate light and heavy hole valence band maxima. In the case of a thin InGaAs layer compressively strained to match the lattice constant of
Table 2.1: Lattice constants of InAs, GaAs, and AlAs. Lattice constants of intermediate ternary alloys obey Vegard's Law, so they can be linearly interpolated from those for the binary compounds.

GaAs, the bandgap increases and the light hole valence band maximum is greater than heavy hole maximum. Biaxial tensile strain narrows the bandgap and pushes the heavy hole band above the light hole band. Given the values of elastic constants and deformation potentials, the resulting bandgaps can be calculated in terms of the strain[19]. The results of these calculations for the materials in Figure 2.1 strained to match GaAs are shown there.

2.2 Molecular Beam Epitaxy

Molecular beam epitaxy (MBE) is the growth of crystals on a substrate by evaporation in an ultra-high vacuum (UHV) environment[20, 21]. It uses thermal beams from elemental sources which are evaporated onto a heated substrate. The flux of the beam is controlled by varying the source temperature, and the flux can be abruptly initiated and terminated by opening and closing a shutter in front of each of the sources. MBE thus offers precise control of epitaxial layer growth conditions and has been used for both compound and elemental semiconductors. MBE easily combines various materials into complicated structures of layers only a few atomic layers thick with abrupt interfaces. This can be done under the control of a computer that manipulates both the source temperatures and shutters. The UHV environment prevents background contamination and, when combined with clean sources, can produce very high quality material. These virtues have allowed epitaxial material with a wide variety of properties for both electronic and optical devices to be grown by MBE. The UHV environment also allows several analytical instruments to be used during growth,
including reflection high energy electron diffraction (RHEED) which can determine surface quality, structure layer thicknesses, and composition during growth.

![Diagram of molecular beam epitaxy system's growth chamber]

Figure 2.2: Molecular beam epitaxy system's growth chamber. See text for details.

2.2.1 MBE System

Figure 2.2 is a schematic illustration of the growth chamber of a MBE system. The growth chamber is connected to a load lock where samples can be introduced and prepared. The system is kept at high vacuum by a combination of cryo-, ion, turbo, and sublimation pumps. The elemental constituents of the desired compounds are placed in the effusion cells pictured at the left of the diagram. Appropriate dopant sources are also put into these cells. The cells point toward the substrate holder and each has a shutter that can be opened or closed. Individual furnaces heat the
cells to establish appropriate vapor pressures or fluxes of the various elements. The fluxes can be measured with an ionization gauge that can be positioned where the substrate holder normally resides during growth. A refrigerated shroud surrounds the cells to help thermally isolate them. The substrate on which the epitaxial layers will be grown is prepared and placed on the substrate heater facing the furnaces. During growth, the wafer in many systems, including the Varian Gen IIIs at Stanford, is rotated to improve the epitaxial layer uniformity across the wafer. The wafer may either be directly mounted with indium on a heated block, the common method for early systems, or it may be radiatively heated as in the Stanford system. The RHEED gun produces an electron beam that is incident on the wafer surface at a glancing angle and is diffracted by the surface on to a screen opposite the gun.

2.2.2 Growth Considerations

During growth, the elemental atoms or molecular complexes arrive at the surface, migrate on the surface, and either become reduced to single atoms and incorporated in the crystal or desorb. When growing III-V compounds, fluxes of the the group V elements are usually several times those of group III elements. However, the group V atoms tend to incorporate only with group III atoms, preserving the stoichiometry of the crystal. Thus the group III fluxes control the growth rate of the material which is usually on the order of 1 μm/hr. Faster growth rates tend to increase the number of defects in the material. If ternary compounds are desired, the relative flux ratios between the elements from the same group must be carefully controlled to achieve the right alloy composition.

If an abrupt interface between different materials is required, shutters can be opened and closed to change which elemental fluxes are allowed to reach the wafer. More gradual or graded interfaces can be obtained by changing the temperature of a cell with its shutter open, or by creating an artificial grading by pulsing a shutter open for different amounts of time. The interactions between the different mechanisms involved in growth change with temperature, and different growth temperatures are optimal for different materials. The best temperature range for GaAs growth is between 550 and 650 °C, with lower temperatures being better for InGaAs growth,
and higher temperatures being better for AlGaAs growth.

2.2.3 RHEED

RHEED is a crucial analytical instrument found on all MBE systems to monitor surface conditions of the wafer before and during the growth. Initially, it is used to monitor the desorption of the wafer’s chemical oxide developed during preparation. The temperature is increased until a streaked pattern develops on the RHEED screen indicating that the oxide has been removed. The specific pattern can indicate whether the surface is terminated in group III or group V elements. It is usually necessary to maintain the surface in a group V rich condition at elevated temperatures. The most important use of RHEED is to calibrate the growth rate. Under the proper conditions, the specular reflection of the RHEED pattern oscillates once in intensity for every atomic layer grown. RHEED oscillations occur because the epitaxy is formed monolayer by monolayer.

Initially, the beam strikes an atomically smooth surface giving an intense reflection. As a new monolayer begins to grow, islands form on the initial surface, scattering the beam and diminishing the reflection. As additional atoms arrive at the wafer, they migrate to the edges of the islands where they are held in place by multiple bonds. When the surface is 50% covered with monolayer thick islands, the reflection is at a minimum. As the islands grow further, they begin to coalesce. Arriving atoms migrate to fill in the missing areas of the monolayer. As the monolayer is completed, scattering is eliminated, and the reflection intensity is again maximized. Usually some atoms have begun to form in new islands on the next monolayer before the first layer is completed. Thus the maxima in the RHEED reflection tend to diminish with time. If RHEED oscillations can be monitored throughout the growth, the exact number of atomic layers grown can be determined. Unfortunately, the electron beam must be aligned to specific crystallographic orientations to observe the RHEED pattern. Thus RHEED cannot be used when the wafer is rotating. The normal practice is to utilize a stationary wafer to calibrate growth conditions, and assume that they remain constant during growth of the desired structures, which are rotated to achieve areal uniformity.
2.2.4 Migration Enhanced Epitaxy

The decay of RHEED oscillations is symptomatic of a general roughening of the growth front. If the material composition is suddenly changed in this condition, the resulting heterojunction interface will not be atomically smooth. Interface roughness can degrade electron transport both parallel and perpendicular to the interface. One method used to smooth interfaces is to interrupt the growth to allow the atoms in small islands an extended time period to migrate to step sites on a lower monolayer. Because impurities can accumulate during the time of the interruption, it is sometimes done several monolayers prior to any critical interface, in the hope that the growth front will remain rather planar until the interface.

Another option is to interrupt the growth for a short time after every monolayer in order to allow the atoms additional time to migrate to step sites of the incomplete previous layers. This is essentially the motivation behind a technique called migration enhanced epitaxy (MEE)[22]. In MEE the migration of atoms is improved still further by depositing the group III elements in the absence of the group V flux. This greatly enhances the surface mobility of the group III atoms, increasing the likelihood that they will incorporate at step sites in the lattice. After enough group III atoms have been deposited, that flux is shuttered off and the group V atoms are provided to satisfy the bonds of the group III atoms on the surface. Any additional group V atoms will not incorporate and will be pumped away. It is important to balance the flux and shutter times of the group V elements to supply enough atoms for a monolayer while having a diminished group V pressure during the group III deposition. The resulting growth surface remains planar and strong RHEED oscillations can be observed throughout the growth. MEE also allows high quality layers to be grown at temperatures well below those for optimum GaAs growth.

2.3 Resonant Tunnel Diodes

Several texts on quantum mechanics or tunneling phenomena have addressed tunneling in a double barrier system, often using a transfer matrix formalism[23, 24]. Tsu and Esaki first discussed tunneling in a double barrier system as an extreme
example of a finite superlattice with a three-dimensional supply function[3]. Chang, Esaki, and Tsu later published the first observations of structure in DBRTD characteristics using a GaAs/AlGaAs diode grown by MBE[4]. The following description is only qualitative. More formal calculations of DBRTD characteristics can be found in other sources[25, 26, 27].

The interference effects of electron waves in a double barrier resonant tunnel diode are analogous to those of light waves in a Fabry-Perot etalon made of two partially transmitting mirrors. The barriers act to reflect most of the electron wave except for a small portion that is transmitted by tunneling through the thin barrier. Since the electron encounters two barriers, there are multiple reflections between the barriers. Usually these reflections don't reinforce each other, and the net effect is that only a doubly attenuated portion of the impinging electron wave is transmitted with the rest being reflected. If however, the multiple reflections constructively interfere, there is no reflection and the entire wave is transmitted. Only electrons with wavelengths that match the resonance determined by the cavity between the barriers are transmitted; the rest are mostly reflected. The resonant wavelengths of both the DBRTD and etalon are determined by the relation for constructive interference. This requires that the roundtrip in the cavity bring the wave back in phase,

$$2L = \left(n - \frac{\phi_0}{2\pi}\right)\lambda$$

(2.1)

where $\lambda$ is the wavelength, $L$ is the cavity length, $n$ is a positive integer indexing the resonances, and $\phi_0$ is the cumulative phase shift due to reflections at the barriers.

Thus the double barrier structure filters out electrons in the same way that an etalon only passes certain wavelengths of light. It is clear that in both cases the coherence of the wave must be preserved across the cavity or else there will be no constructive interference at the resonance. This points out the first dimensional restriction on the double barrier structure: the cavity length must be much shorter than the electron coherence length for strong interference phenomena. The scattering lengths for electrons in undoped arsenide semiconductors at room temperature are listed in Table 2.2.

Furthering this analogy, the transmission resonances of both the DBRTD and
CHAPTER 2. BACKGROUND MATERIAL

<table>
<thead>
<tr>
<th>Material</th>
<th>$\lambda_{kT}$ [nm]</th>
<th>$L_{e-\phi}$ [nm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlAs</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>GaAs</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>InAs</td>
<td>260</td>
<td>330</td>
</tr>
</tbody>
</table>

Table 2.2: Average wavelengths of thermal $\Gamma$-point electrons, $\lambda_{kT}$, and electron scattering lengths, $L_{e-\phi}$, for undoped AlAs, GaAs, and InAs at room temperature.

The etalon have a Lorentzian line shape with the width or finesse determined by the reflectivity of the barriers or mirrors and cavity length. Very reflective mirrors will give very narrow transmission resonances. Usually, a moderately wide transmission width is desired for DBRTDs to increase their maximum current. This points out a second dimensional restriction on the double barrier structure: thin or low barriers are required to obtain high peak currents. It will be shown later why thin, high barriers are preferable to thick, low barriers. If the barriers or mirrors have unequal reflectivities, the maximum transmission is less than unity[25].

There are some differences between the optical etalon and the DBRTD. One is that the typical etalon operates at resonances with a relatively high index, corresponding to large $n$ in equation 2.1. This is necessary because a macroscopic physical length is orders of magnitude larger than the wavelength of visible light, 0.5 $\mu$m. However, it is desirable in DBRTDs to have electrons interact with only one resonance, and because it is difficult to inject electrons at only high energies, the lowest lying resonances must be used. In order for a thermal electron distribution to interact with only one resonance at a time, the energy separation between these low lying resonances needs to be much greater than the thermal energy. This points out the third dimensional restriction on the double barrier structure: the cavity must be much smaller than the thermal wavelengths listed in Table 2.2 in order to exhibit strong quantization effects. In practical DBRTD designs, this is a more stringent criterion than electron coherence through the structure. The other differences between the DBRTD and optical etalon are concerned with the fact that the light’s wavelength is the same everywhere in free space, while the electron’s wavelength changes for different potential energies.
CHAPTER 2. BACKGROUND MATERIAL

encountered throughout the structure. Also, the electron barriers' transmissivity is a function of energy, while the etalon mirrors' transmissivity is relatively constant.

2.3.1 DBRTD Structure

DBRTD structures most typically consist of two thin layers of large bandgap material separated by a thin region of narrow bandgap material imbedded in a narrow bandgap bulk crystal. Since the electron wave must be attenuated in the barriers and propagating in the cavity, it is necessary that the Fermi energy of electrons in the surrounding material's conduction band be smaller than the barrier material's conduction band minimum and larger than the conduction band minimum of the material between the barriers. In the case of the arsenide compounds on a GaAs substrate, the barriers are often $\text{Al}_x\text{Ga}_{1-x}\text{As}$, the intervening layer $\text{In}_y\text{Ga}_{1-y}\text{As}$, and the outside material is GaAs. Quite frequently, $x = 1$ and $y = 0$. Because this structure resembles a quantum well with finite width barriers, the region between the barriers is usually referred to as the well.

Figure 2.3 shows the valence band maximum and conduction band minima as a function of distance for a typical DBRTD structure with 20 Å AlAs barriers and a 50 Å GaAs well. Note that these distances meet the dimensional requirements mentioned earlier. The figure shows that the conduction band's X-point minimum is actually lower than the Γ-point minimum in the AlAs barriers. Experiments have shown that electrons which start in the Γ-point minimum of the conduction band mainly experience the Γ-point potential in thin AlAs barriers rather than the smaller X-point potential. Also, only conduction band electrons are involved in the tunneling processes in the n-type DBRTDs used in this work, so further band diagrams will only show the Γ-point minimum of the conduction band.

In order to supply electrons for tunneling through the double barrier structure, at least a portion of the device must be doped with donor impurities, silicon in the case of GaAs. If these ionized impurities are near the double barriers, they can scatter electrons that are interacting with the barriers. When the scattering changes the magnitude of the electron's longitudinal momentum it will alter the resonance condition for transmission, usually degrading the device's tunneling characteristics[28,
Figure 2.3: Band extrema as a function of position in a typical AlAs/GaAs double barrier resonant tunnel diode structure. The solid lines indicate the Γ-point valence band maximum and conduction band minimum and the dashed line indicates the X-point conduction band minimum.

To avoid these problems, doping is typically confined to electrode regions physically separated from the barriers on each side of the tunneling structure. The barriers and well are left undoped as well as these additional setback or spacer regions surrounding the barriers. Studies have been done on the effects of varying the size of the spacer layers[29]. The electrodes act as electron reservoirs that are internally in equilibrium. Under the right conditions electrons can resonantly tunnel between the reservoirs. The heavily doped electrodes also serve as contact regions.

2.3.2  DBRTD Operation

The current-voltage characteristics of DBRTDs result from the convolution of the transmission spectrum, determined by the double barriers, with the electrons supplied
by the reservoir in the electrode. The general steps of the process are illustrated in Figure 2.4. The different band diagrams correspond to different applied biases that cause different amounts of current to flow as shown on the accompanying current-voltage curve. The dark band between the barriers represents the resonant energy.

Figure 2.4: Typical current-voltage characteristics of a DBRTD. The conduction band diagrams at the left correspond to zero, peak, and valley biases. The dark band indicates the resonance energy. See the text for a detailed discussion.

Initially there is no bias across the device. The few electrons which can tunnel from left to right are exactly balanced by those flowing the opposite way, making the total electron flux and associated current zero. As bias is applied, the bands begin to bend so that the energy of the electrons in the reservoir on the left and the resonant energy get closer. At a bias corresponding to the threshold on the current-voltage curve, the resonant energy comes into line with the top of the reservoir and substantial amounts of current begin to flow.
CHAPTER 2. BACKGROUND MATERIAL

The current reaches a maximum when the resonance is in line with the bottom of the reservoir, as depicted by the second band diagram. At this point there is a peak current, \( I_p \), and the corresponding bias is the peak voltage, \( V_p \). A further increase in bias will carry the resonance below the electron reservoir and the current will decrease rapidly. As the tunneling current is decreasing, other non-ideal currents are increasing.

When the magnitude of the rate of change of these two components is equal, the total current will be minimized. The current at this point is the valley current, \( I_v \), and the voltage is the valley voltage, \( V_v \). The third band diagram shows that the resonance is completely below the lefthand electron reservoir at this bias. Further increases in bias will increase the non-resonant currents until they dominate. Between the peak and valley voltages, incremental changes in voltage cause incremental changes in current of the opposite sign. Hence, \( dI/dV < 0 \), and this region is called the negative differential conductance (NDC) region.

One of the most commonly quoted figures of merit for tunnel diodes is the peak-to-valley current ratio (PVCR), \( I_p/I_v \). This has some basis in both theory and applications. Most models do not account for incoherent current components which are the major portion of the valley current in practice. Accordingly, they predict PVCRs much larger than those measured in the laboratory. Devices with high PVCRs are considered more ideal because they more closely match the coherent model. In applications, the amount that the current can change within the vicinity of the NDC region is \( I_p - I_v \), and the amount of current required to bias into this region is about \( (I_p + I_v)/2 \). A ratio of these two quantities gives a measure of how much useful NDC is available compared to the required DC bias:

\[
\frac{I_p - I_v}{(I_p + I_v)/2} = \frac{1}{2} \left( \frac{\text{PVCR} - 1}{\text{PVCR} + 1} \right).
\]  

(2.2)

This ratio indicates that a PCVR of 3 is 50% better than a PVCR of 2, but a PVCR of 100 is only 20% better than a PVCR of 10. PVCRs of 3 or 4 are probably adequate for most applications.

A more important number for tunnel diode applications is the speed index[7, 30]. This is a ratio of the peak current, \( I_p \), to junction capacitance, \( C_j \). The speed of
tunnel diodes in many applications is proportional to this ratio. Generally, currents in a tunnel diode tend to scale with the peak current, whether the change is due to a change in area or a change in structure. These currents must charge and discharge the junction capacitance as the diode bias point changes. The speed of a tunnel diode can be increased by increasing the peak current density or decreasing the junction capacitance per unit area. For DBRTDs, the peak current density is largely a function of the electrode doping and the resonance width, which in turn is determined by the barrier and well configuration. The junction capacitance is largely determined by the combined thickness of the barriers, well, and depleted anode spacer. It can be estimated relatively well with a simple parallel plate capacitor model. Approximate analytical expressions for peak voltage, peak current density and junction capacitance are given in Section 2.3.4.

A final figure of merit is the voltage width of the NDC region, \( \Delta V_{\text{NDC}} = V_v - V_p \). The desired width depends on the particular application. For switching applications, it is desirable to have the transition from peak to valley current occur as rapidly as possible[30, 7]. For microwave applications, it is desirable to make the NDC region as broad as possible. This allows larger amplitude signals which correspond to more power. A rough number for comparison of tunnel diodes in microwave applications can be formed from the ratio of the maximum AC power in the NDC region, assuming it is exactly linear, to the required DC bias power.

\[
\frac{P_{\text{AC,NDC}}}{P_{\text{DC,BIAS}}} = \frac{\frac{1}{2}(I_p - I_v)(V_v - V_p)}{\frac{1}{2}(I_p + I_v)(V_v + V_p)} = 2 \left( \frac{\text{PVCR} - 1}{\text{PVCR} + 1} \right) \frac{1}{1 + 2V_p/\Delta V_{\text{NDC}}}.
\]

This number clearly increases with broadening NDC regions and decreasing peak voltages. Unfortunately, it is difficult to make DBRTDs with high \( I_p \) and low \( V_p \)[31].

2.3.3 Design Parameters

The key parameters of peak current density and peak voltage are related to the width of the transmission resonance and resonant energy, respectively. The details of the transmission spectrum are determined by the materials and thicknesses of the barriers and well. In realistic devices, the resonant energy is almost independent of the barrier thickness, as long as it is only a small fraction of the barrier height. The
energy of the resonance above the conduction band minimum of the well material can be reasonably approximated using the condition for constructive wave interference, equation (2.1), and assuming a constant electrostatic potential in the well and equal height barriers. The resulting resonant energy is given implicitly by the relation

\[ k_w L_w = n\pi - \arctan \left( \frac{2k_w m^*_b}{\alpha_b m^*_w} \right), \]  

(2.4)

where \( n \) is the resonance index, \( L_w \) is the length of the well between the barriers, \( m^*_w \) and \( m^*_b \) are the effective masses, and \( k_w \) and \( \alpha_b \) are the propagation constants in the well and barriers. The propagation constants are a function of the relative energy compared to the potential in the well and barriers. Under the effective mass approximation, they can be written as

\[ k_w = \sqrt{\frac{2m^*_w (E - V_w)}{\hbar^2}} \quad \text{and} \quad \alpha_b = \sqrt{\frac{2m^*_b (V_b - E)}{\hbar^2}}, \]  

(2.5)

where \( \hbar \) is Planck's reduced constant, and \( V_w \) and \( V_b \) are the constant potentials in the well and barriers. In reality, the well experiences a large electric field which Stark shifts the resonance slightly, and charge accumulation in the well makes the well potential spatially nonlinear. For most analytical calculations, these effects can be neglected and the resonant energy referred to the middle of the well. Figure 2.5 shows the approximate resonant energy as a function of well size for various well material compositions.

As the well narrows, the resonance is pushed to higher energies and more voltage must be applied to a DBRTD structure to reach the peak current. Low peak voltages can be realized with wide wells or wells with conduction band minima lower than those of the surrounding electrodes. Wells of InGaAs are often used with GaAs electrodes to lower the peak voltage\[32\]. If the well is too wide the resonances become closely spaced. If the spacing is less than the electrons' Fermi energy and thermal distribution, or if the well approaches the scattering length, the resonant effects will become washed out.

Both the well length and the barrier transmissivity affect the width of the Lorentzian-shaped transmission resonance. The barrier transmissivity is roughly the square of
Figure 2.5: Resonant energies referred to GaAs electrodes for In$_y$Ga$_{1-y}$As wells as a function of well size. The energy of the first (solid line) and second (dashed line) resonances were calculated from equation 2.4 for InGaAs wells strained to match the lattice of GaAs. The decreasing resonant energies correspond to increasing InAs fractions of $y = 0.0, 0.1, 0.2, \text{and} 0.3$.

the wavefunction attenuation factor, $\alpha_b$, neglecting electric fields and reflection due to effective mass mismatches in the materials on each side of the barrier:

$$T = \left(e^{-\alpha_b L_b}\right)^2.$$ (2.6)

Thinner barriers increase the transmissivity, widening the the resonance, and allowing more electrons from the electrode reservoir to flow through. Thus thinning the barriers increases the peak current density up to the point where the transmission resonance is as wide as the Fermi energy in the electrode. The transmission’s full width at half
the maximum (FWHM) is
\[
\Delta E_{\text{FWHM}} = \frac{2T}{\sqrt{1 - T}} \left( \frac{\hbar^2((2n - 1)\pi - \phi_b)}{4m^*_wL_w^2} \right),
\]
(2.7)
where \(\phi_b\) is the previously mentioned phase shift for reflection from the barriers,
and \(T\) is the average of the transmissions for the left and right barriers, \(T_L\) and \(T_R\)
which have both been assumed to be much less than unity. The current density is
proportional to the area under the transmission versus energy curve. For a Lorentzian
resonance, this is the product of the transmission maximum and the FWHM, which is
\[
T_T \Delta E_{\text{FWHM}} = \frac{2T_LT_R}{T\sqrt{1 - T}} \left( \frac{\hbar^2((2n - 1)\pi - \phi_b)}{4m^*_wL_w^2} \right).
\]
(2.8)
To first order, the peak current density is proportional to the barrier transmissivity
and inversely proportional to the square of the well size.

The parabolic form of the attenuation constant, \(\alpha_b\), in equation 2.5 is actually only
accurate for electron energies near the conduction band minimum of the barrier where
the effective mass approximation is good. Further away, the attenuation constant
must eventually begin to diminish and become zero again at the valence band maxima
of matching symmetry. One approach to the problem is to use an elliptic rather
than parabolic dispersion relation for the attenuation constant. The details of this
treatment are available elsewhere[33, 26]. The effect is that the maximum attenuation
occurs somewhere in the bandgap; for AlAs, it is about 1.5 eV below the conduction
band \(\Gamma\)-point energy. Near this maximum, the attenuation is relatively constant, and
the transmissivity is only a weak function of electron energy.

The transmission probability of hypothetical one monolayer barriers of GaAs,
Al\(_{0.3}\)Ga\(_{0.7}\)As, and AlAs are plotted in Figure 2.6 as a function of energy below the
\(\Gamma\)-points of those materials. The transmissivities for barriers of realistic width can
be calculated from the data in the figure by raising the transmissivity of a single
monolayer to the power of the number of monolayers in the barrier. It is highly
unlikely that a single monolayer barrier could be accurately treated with the envelop
function approximation. However, barriers as thin as three monolayers were grown
in conjunction with the work described in this thesis. Liquid nitrogen characteristics
exhibiting NDC for this structure are shown in Figure 2.7.
Figure 2.6: The transmissivity of single monolayers of GaAs, Al\textsubscript{0.3}Ga\textsubscript{0.7}As, and AlAs as a function of energy below the Γ-points of those materials.

There are four motivations for using thin AlAs barriers rather than wider AlGaAs barriers to accomplish similar attenuations. The first can be seen from the data in Figure 2.6. Electrons from a GaAs electrode will have energies about 0.25 eV below a Al\textsubscript{0.3}Ga\textsubscript{0.7}As barrier and about 1.0 eV below the Γ-point of an AlAs barrier. The figure shows that the variation of transmissivity with energy is smaller for the AlAs barrier than for the Al\textsubscript{0.3}Ga\textsubscript{0.7}As barrier. Second, the relative change in electron energy as a fraction of barrier height due to electric fields is greater for the AlGaAs barrier both because it is wider, and lower in potential. The third reason involves the fact that variations in growth rate cause variations in barrier width proportional to the total width. However, the fractional change in tunneling current is exponentially related to the absolute change in barrier width. Thus, it is best to minimize the barrier width in order to minimize the changes in tunneling current[34]. Finally, AlGaAs barriers
Figure 2.7: Current voltage characteristics of a DBRTD with three monolayer barriers at liquid nitrogen temperature. The device had 10 Å AlAs barriers, a 50 Å GaAs well, and GaAs electrodes doped $1 \times 10^{18}$ cm$^{-3}$ with silicon beyond 500 Å spacer layers. Three regions of NDC are observed.

are a source of alloy scattering and cannot be grown with the well defined binary interface that occurs between AlAs and GaAs. The one detriment to AlAs barriers is that they likely increase the probability of scattering to the lower energy X-valley over that for similar scattering in direct gap AlGaAs.

2.3.4 Peak Parameter Expressions

The peak voltage is expected to occur when the resonant energy level is approximately equal to the conduction band minimum of the electrode supplying electrons.
Figure 2.4 indicates that there will be voltage drops across the barrier and well structure as well as in both electrodes, particularly in the undoped regions just outside the barriers. It is possible to make an analytical estimate of the peak voltage by adding up the potential drop in each region based on a few assumptions.

The first assumption is that the net change in the cathode potential is the thermal voltage, $kT/q$. This is based on the fact that little further change in potential is needed to produce far more net negative charge once an accumulation layer is formed. The second assumption is that the barriers, well, and anode spacer are depleted of charge, making the E-field constant in this region. This essentially amounts to these regions being in equilibrium with the anode Fermi level. Finally, the Stark shift and other effects due to the electric field across the well are ignored. The resonant energy is taken as that given by equation 2.4 above the middle of the well. Based on an accumulation potential of $kT/q$ and constant electric field of $\mathcal{E}$ across the active structure, the total voltage drop is

$$V_{\text{tot}} = \frac{kT}{q} + (L_{\text{dbs}} + L_{\text{as}})\mathcal{E} + V_{\text{ad}}(\mathcal{E}),$$

(2.9)

where $L_{\text{dbs}} = 2L_{b} + L_{w}$ is the total width of the double barrier structure, $L_{\text{as}}$ is the width of the anode spacer region, and $V_{\text{ad}}$ is the potential drop across the anode depletion layer needed to support the electric field. The specific dependence of $V_{\text{ad}}$ on $\mathcal{E}$ is obtained by integrating the Poisson equation. This gives the potential drop as

$$V_{\text{ad}} = \frac{kT}{q} \left( \frac{1}{2} \left( \frac{qL_{D}\mathcal{E}}{kT} \right)^2 + 1 \right),$$

(2.10)

where $L_{D} = \sqrt{\frac{kT}{qN_{d}}}$ is the extrinsic Debye length determined by the anode doping concentration, $N_{d}$. The total voltage drop depends quadratically on the electric field.

The constant electric field at the peak voltage can be calculated by making the voltage drop in the accumulation layer and across half of the barrier structure align the resonant level with the bottom of the conduction band. This gives

$$\mathcal{E}_{p} = \frac{E_{\text{res}} - kT}{q\frac{1}{2}L_{\text{dbs}}}. $$

(2.11)

This expression can be substituted for $\mathcal{E}$ in equations 2.9 and 2.10 to obtain the peak
voltage,

\[ V_p = \frac{2kT}{q} \left[ \left( \frac{L_D}{L_{\text{dbs}}} (E_{\text{res}}/kT - 1) \right)^2 + \left( 1 + \frac{L_{\text{as}}}{L_{\text{dbs}}} \right) (E_{\text{res}}/kT - 1) + 1 \right]. \]  

(2.12)

For example, a 50 Å GaAs well gives a resonant energy of 0.126 eV. The room temperature extrinsic Debye length for a $1 \times 10^{18}$ cm$^{-3}$ doped anode would be 43 Å. If a structure with these parameters had 25 Å barriers and a 100 Å anode spacer, the peak voltage would be 0.6 volts.

The peak current can be estimated rather simply if the transmission resonance width is much smaller than the Fermi energy. In this case, the transmission coefficient, $T(E)$, in Tsu and Esaki's integration over the three dimensional supply function\[3,\]

\[ J = \frac{q m^* kT}{2\pi^2 \hbar^3} \int_{0}^{\infty} T(E) \ln \left( \frac{1 + e^{(E_F - E)/kT}}{1 + e^{(E_F - E + qV)/kT}} \right) dE, \]  

(2.13)

is effectively a delta function of area $T_{\text{res}} \Delta E_{\text{FWHM}}$. The maximum value of the integral or peak current density is the largest value of the remaining part of the integrand within the range of the integral. Since this occurs when $E = 0$, the peak current density is

\[ J_p = \frac{q m^* kT}{2\pi^2 \hbar^3} T_{\text{res}} \Delta E_{\text{FWHM}} \ln \left( \frac{1 + e^{E_F/kT}}{1 + e^{(E_F - qV_p)/kT}} \right). \]  

(2.14)

Typically, the electrode will be degenerately doped so the last term is approximately $qV_p/kT$ for $qV_p$ at least a few $kT$ less than $E_F$, and approximately $E_F/kT$ for $qV_p$ at least a few $kT$ greater than $E_F$. Note that the assumptions used to derive equation 2.12 don't hold for the first set of conditions. Numerically, $(qm^* kT)/(2\pi^2 \hbar^3) = 2.8 \times 10^7$ A/cm$^2$-eV at room temperature with the electron effective mass of GaAs. Thus a device with an electrode doping that gave $E_F/kT = 3$ and a 1 meV wide resonance would give $J_p = 8.4 \times 10^4$ A/cm$^2$.

The junction capacitance per unit area can be calculated from $C^{-1} = \frac{1}{\varepsilon} dV/d\varepsilon$. Differentiating equation 2.9 gives

\[ C = \frac{\varepsilon}{L_{\text{dbs}} + L_{\text{as}} + L_D \frac{qL_{\text{dbs}}}{kT}}. \]  

(2.15)

The last term in the denominator corresponds to the anode depletion width, so the expression is just that for the parallel plate model. Equations 2.9 and 2.10 can be
inverted to find the normalized electric field,

\[ \frac{qL_D\mathcal{E}}{kT} = \sqrt{\left(\frac{L_{\text{dbs}} + L_{\text{as}}}{L_D}\right)^2 + 2\left(\frac{qV_{\text{tot}}}{kT} - 2\right) - \left(\frac{L_{\text{dbs}} + L_{\text{as}}}{L_D}\right)} \]. \hspace{1cm} (2.16)

As an example, a structure with \( L_{\text{dbs}} = 100 \) Å, \( L_{\text{as}} = 100 \) Å, and an anode doping of \( N_d = 1 \times 10^{18} \) cm\(^{-3} \) with an applied bias of \( V_{\text{tot}} = 0.6 \) volts would have a 140 Å depletion layer and a junction capacitance of about \( 3.4 \times 10^{-7} \) F/cm\(^2\).
Chapter 3
Device Fabrication and Measurement

Several techniques used to prepare and measure DBRTDs were common to all the experimental work that was done. The experimental portion of the research cycle included MBE growth of epitaxial layers, fabricating devices from the layers, and measuring the devices. The last step often included special packaging of the devices in order to test their microwave or cryogenic properties. The devices were fabricated using standard microelectronics processing including photolithography, metal and dielectric deposition, etching, and ion implantation. The general attributes of MBE have already been described in chapter 2. This chapter covers the remaining experimental steps involved in device fabrication and measurement. The techniques described here were used for the rest of the thesis work described in this dissertation. General measurement results are presented here.

3.1 Device Fabrication

Once the MBE layers for a structure like the DBRTD are grown, it is necessary to process them into individual devices that can be measured. Fabrication of devices based on epitaxial layers mainly consists of sculpting individual devices by effectively removing unwanted material and then making electrical contacts to the separate devices. There are some heterojunction devices which are made by altering the doping profiles or to some extent even the band properties of the epitaxial layers laterally, but many research devices rely solely on the structure grown into the layers. The parts of this section describe different procedures for isolating and contacting devices.

3.1.1 Aligned Contacts

Devices are commonly contacted by evaporating metallic combinations onto the semiconductor that are then alloyed. To contact the n-type GaAs electrodes of DBRTDs, layers of gold, germanium, nickel, and gold were successively evaporated onto heavily doped regions using an electron beam evaporator. The bottom layers of
gold, germanium, and nickel were 400 Å, 100 Å, and 125 Å, respectively. The ratio of gold to germanium is chosen to provide a low melting temperature eutectic, and the nickel helps to break up surface oxidation[35]. The uppermost gold layer acts as the body of the contact and is usually at least 1500 Å thick.

The contact is alloyed at 450 °C in a mixture of hydrogen and nitrogen. Alloying causes the eutectic mixture to form a liquid and melt back a thin region of the GaAs. This allows the germanium to move into the GaAs, doping it and decreasing the contact resistance. During the alloying process, the molten metal and germanium alloy can penetrate irregularly into the semiconductor. If the DBRTD barriers are too close to the surface, this irregular meltback can locally penetrate or “spike” through the barriers. The DBRTD structures that had alloyed contacts always had at least 2000 Å of material above the barriers and no spiking problems were observed.

The evaporated metals were patterned with liftoff techniques. If there was no deposited insulating layer on the wafer, the photolithography was done with AZ 1370SF resist spun on at 3500 rpm to produce a layer thick enough to liftoff up to 5000 Å of metal. After 15 minutes of prebake, the resist was soaked in chlorobenzene for 5 minutes to help establish a proper profile for liftoff. Then, after another 5 minute prebake, the resist was exposed and developed. No postbake was done unless the resist was also to serve as an etch mask for a gate recess during FET fabrication. The metal was evaporated and then lifted off in the places were it was on top of resist by removing the resist in acetone.

Often insulating oxide or nitride layers are deposited on exposed semiconductor surfaces to passivate and protect them and to insulate them from any overlying metallization. If an insulating layer had been deposited on a wafer that was to be contacted, it was used to do a dielectric assisted liftoff. In this case the photolithography was done with the normal resist thickness, without the chlorobenzene soak, and with a prebake. The exposed oxide was then etched away before the metal deposition and liftoff. This technique leaves most of the area inside the oxide opening metalized.
3.1.2 Non-alloyed Contacts

Alloyed ohmic contacts are sufficient for most device processing requirements, but it is possible to make n-type contacts with lower resistances which require no alloying and thus don't expose the device to contact spiking or other degradation due to high temperature processing. If GaAs metal contacts are not alloyed, they form rectifying Schottky contacts because the Fermi level of the GaAs surface is pinned midgap. The surface states in InAs are such that they pin the Fermi level in the conduction band instead of midgap, thus making it easy to contact InAs with a variety of metals without alloying. However, if an InAs layer thick enough to pin in the conduction band is placed directly onto the GaAs, it will be highly dislocated because of the lattice mismatch described in section 2.1.3. These dislocations will affect the bulk and interface electrical properties of the InAs.

In order to gradually accommodate the difference in lattice constant between the top GaAs electrode and the InAs contact, the DBRTDs grown with InAs contacts used a chirped superlattice grading technique developed by W.-S. Lee at Stanford[36]. This structure, grown at 480 °C, consists of a heavily doped 10 period superlattice of GaAs and In_{0.5}Ga_{0.5}As where each period is 30 Å long. The width of the GaAs layer in each period is linearly decreased from 27 Å to 0 Å just below the InAs layer with the balance of the period being the In_{0.5}Ga_{0.5}As layer, which increases in thickness closer to the InAs. Finally, a heavily doped, 200 Å layer of InAs is deposited. This seemingly complicated structure is easily grown with computer control, and is an excellent example of the beneficial capabilities of MBE. A titanium/gold layer was usually used to contact the structure, although other metals, as well as tungsten-silicide could be used. The thin 200 Å titanium layer was used to improve the surface adhesion of the thicker gold body of contact metal.

3.1.3 Mesa Isolation

The simplest way to define isolated devices is to etch off the epitaxial material between them, leaving individual device mesas. The most common etchant used on arsenide compound semiconductors for this purpose is a mixture of phosphoric acid and hydrogen peroxide diluted in water. In order to etch specific amounts of
material, the etch rate, which is a function of etchant mixture and temperature, must be known. The mixture and temperature used to prepare the devices discussed in this dissertation were $\text{H}_3\text{PO}_4: \text{H}_2\text{O}_2: \text{H}_2\text{O} = 1:1:30$ at $15^\circ\text{C}$, which etched GaAs at approximately 800 Å/min. Before the wafer is subjected to this etchant, a resist mask is deposited and defined using photolithography. The unmasked or exposed portions of the wafer are then etched to a depth determined by the etch rate and time.

For DBRTD structures, the first etch should be deeper than the level of the barriers so that the top electrodes of the diodes are separated. If the etch is not as deep as all of the doped layers, which includes the substrate if it is conducting, then the bottom electrodes are still connected by the doped material. Often, only single diodes need to be measured and the bottom electrodes can be left connected together. If the substrate is conducting, a single ohmic contact can be made to the entire backside of the wafer which serves as a common connection to the bottom electrodes of all the devices. If the substrate is insulating, the bottom electrode layers must be contacted from the top side. Unless the bottom electrode layer is very thick, it is usually advantageous to place the patterned bottom ohmic contacts as close to the mesas as possible to reduce spreading resistances.

If the devices will be used for high frequency measurements, care must be taken to eliminate parasitic capacitances that can dominate the device speed. These capacitances result from placing metal bond pads or interconnect lines too close to conducting material. Either metal pads and lines should not be used at all, or they must be placed on insulating material such as semi-insulating GaAs substrates. In this later case, another etch is done which leaves some conducting material around the bottom of the mesa for contacting the bottom electrode, but removes the rest. A cross-section of a completed mesa isolated device on a insulating substrate is illustrated in Figure 3.1.

Sometimes, it is important to precisely control etch depths to uncover a thin epitaxial layer under the surface of the wafer without etching through it. This would be required, for example, to contact the well of a DBRTD. Selective etches which etch the material above the layer but not the layer itself, or etch it much more slowly, can be used in these circumstances. In GaAs processing the superoxyl etch,
Figure 3.1: Cross-section of a mesa isolated DBRTD. The dark bands indicate the location of the double barriers. The top and bottom electrodes are contacted through openings in an insulating oxide. Outside of the bottom contact area, the conducting layers have been etched off to reduce the parasitic capacitance of interconnects and contact pads.

a pH balanced mixture of hydrogen peroxide and ammonium hydroxide, is used to selectively etch GaAs but not AlGaAs[37, 38]. This etch was shown to be quite successful for DBRTD processing, stopping on 20 Å AlAs layers. It was also discovered in the course of etching devices with nonalloyed InGaAs superlattice contacts that superoxyl will selectively etch GaAs from InGaAs[39].

3.1.4 Implant Isolation

One difficulty with the mesa isolation method is that it creates nonplanar structures which may complicate further processing. Another is that alignment tolerances place limits on the minimum size of a mesa that can still be contacted. If the alignment tolerance is 1 μm, the mesa must be at least 2 μm larger in diameter. This means that although a 1 μm diameter contact can be defined, the smallest contacted mesa has an area nine times larger. In order to make high impedance, high peak current density DBRTDs, it is necessary to make the device size as small as possible.
One way to accomplish this is to self-align the isolation to the contact using ion implantation. The implantation can isolate devices without etching so that the surface remains planar.

Figure 3.2: Cross-section of an implant isolated DBRTD. The dark bands indicate the location of the double barriers. The gold metal contacts the top electrode via the non-alloyed InGaAs contact structure and serves to mask the implant and the etch of the InGaAs surface layer. This top layer is lightly damaged by the implant and may contribute to surface leakage if not removed. This step is not necessary for alloyed contacts.

The cross-section of an implant isolated device is illustrated in Figure 3.2. First, either an alloyed or non-alloyed ohmic contact is defined on the surface of the wafer. Then the entire wafer is implanted with ions to damage the semiconductor surface layers, making them insulating. The contact metal is made thick enough to stop the implanted ions from penetrating the contact. Thus the material under the contact remains undamaged, but is isolated. The implant needs to be done with enough energy to make the crystal down to the barriers insulating. Doses of $1 \times 10^{12}$ boron
atoms/cm\(^2\) at 100 keV were typically used for the DBRTD layers. This was enough to reach barriers 2500 Å below the surface\[^{40}\].

This process was done with conducting substrates so that the bottom electrode could be contacted through the back of the wafer, making the entire process a relatively quick way to fabricate small devices. Similar self aligned isolation could be combined with deep etches to contact the bottom electrode to make small devices on insulating substrates. In order to contact very small devices, larger contact pads can be placed over the devices. However, this adds a significant parasitic capacitance to the devices since the contact pads extend over the implanted region above conducting layers.

### 3.2 Stabilized Measurements

One of the difficulties in measuring devices like the DBRTD that have NDC is that their characteristics are often obscured by parasitic oscillations. These result from amplification of random noise by the NDC element. In order to suppress these oscillations, proper care must be taken with the measurement setup. The mechanisms causing the oscillations and their prevention are discussed below.

#### 3.2.1 Amplification by NDR

Quasi-static DC measurements do not usually require much attention to the details of signal propagation in the setup. However, it is possible for devices biased into NDC to amplify spurious noise, and even oscillate if the device is not properly matched to the cabling. Figure 3.3a shows a transmission line with impedance \( \hat{Z}_0 \) terminated in a load of impedance \( \hat{Z}_L \). Waves traveling through the transmission line and impinging on the load will be reflected from the load with a magnitude and phase determined by the match between the line’s impedance and the load’s impedance\[^{41}\]. The ratio of the reflected wave’s amplitude to that of the incident wave is the reflection coefficient,

\[
\hat{\Gamma} = \frac{\hat{Z}_L - \hat{Z}_0}{\hat{Z}_L + \hat{Z}_0}.
\]  

(3.1)

If the real part of \( \hat{Z}_L \) is negative, as it is in the NDC region, then \(|\hat{\Gamma}| > 1\), and the pulse is amplified. If \( \hat{Z}_L \approx -\hat{Z}_0 \) the reflection coefficient is very large, and large
amplification of the incident signal is possible. If $|\hat{Z}_L| \gg |\hat{Z}_0|$, the reflection coefficient is only slightly larger than unity for $\text{Re}(\hat{Z}_L) < 0$. Thus high load impedances, even negative ones, result in a reflection about the size of the incident one. As will be discussed below, large reflection coefficients can result in unwanted oscillations. The reflection coefficient can obviously be changed by adjusting either the load impedance, the line impedance, or both.
3.2.2 Parasitic Oscillation

Measurement setups often have multiple impedance mismatches at interfaces between instruments, cabling, fixtures, and devices. For example, devices are often probed at stations whose wiring is effectively a high impedance connected to 50 Ω cables, which are connected to high impedance curve tracers. These systems can be schematically represented by a device serving as the load for a mismatched transmission line as shown in Figure 3.3b. Noise pulses from various sources can be reflected multiple times between the mismatched interfaces. Usually, the pulse is diminished at each interface since the reflection coefficient is less than one. However, the reflection coefficient can be greater than one at the interface between a NDC device and the line whose impedance is $Z_3$ in the diagram. If enough of the amplified reflection is returned by the subsequent reflection at the interface between $Z_2$ and $Z_3$, the noise pulse can grow. The large pulse bouncing back and forth between the two interfaces keeps growing until it is large enough to bias the device out of the NDC region. This is also called self-detection.

The mismatched line acts as a resonant structure for the NDC device. If the Q of the cavity is high enough, the device will oscillate. The frequency of oscillation is determined by the length of the mismatched line. High frequency oscillations occur over short distances, like the interconnect metal or wire bond in a device package, and lower frequency oscillations occur over longer pieces of coaxial line. Curve tracers or other low frequency measurement instruments usually display average values of the oscillating current at bias voltages in the NDC region. Since the current oscillations are clipped at the peak and valley values, the measured current values tend to be somewhat constant throughout the NDC region. This leads to the formation of "chairs" in the observed DC characteristics[42]. Some of these can be quite complicated, even showing bifurcation[43].

3.2.3 Stabilization Techniques

In order to accurately measure the NDC region characteristics, it is necessary to prevent parasitic oscillations by decreasing the reflection gain at the mismatched interfaces. The most crucial interface is the one at the device. The gain at this
interface can be reduced to almost unity by using devices that have an impedance much larger than the first section of transmission line. This can be accomplished by altering the device characteristics, decreasing the line impedance, or most simply by decreasing the device area. If the on-chip line impedance can be adjusted, it is preferable to match it to the external probe and cabling impedance.

In practice, it was difficult to stabilize devices with peak currents much greater than 1 mA in 50 Ω systems. For typical voltage scales, this corresponded to a minimum negative impedance magnitude of about 100 Ω. Based on empirical observations, the minimum magnitude of negative impedance must be at least 2 to 3 times larger than the line impedance in order to eliminate parasitic oscillations. Since devices with current densities of $1 \times 10^5$ A/cm$^2$ must be less than 1 μm$^2$ to have such small currents or large impedances, it was difficult to stabilize high current density devices. Very small devices also tend to have poorer characteristics than larger ones, presumably because depletion effects alter the resonant energy near the edges of the device. Surface leakage and other currents that are proportional to the device perimeter also become proportionately larger for small area devices.

Some high frequency oscillations can be suppressed with capacitive bypasses. It is important to place these capacitors as close to the device as possible. In the case of the implant isolated DBRTDs, the 50 μm diameter contact pad acted as a bypass with a capacitance of 1 pF. Since the impedance of a capacitor at the end of a length of transmission line can be phase shifted to appear as an open or short at the right frequencies, it is important to put a resistor in series with the capacitor to dissipate energy and damp the reflection. The resistor can be empirically adjusted to help suppress the oscillations. Lossy inductive elements such as ferrite beads may also diminish the oscillations.

It is also beneficial to reduce reflections at other points by measuring the device in an environment with a uniform impedance. Unless the device has been fabricated with impedance matched pads that permit microwave probing, it is unlikely that the device leads will be the proper impedance. If the leads are kept short enough so that their resonant frequency is higher than the device's cutoff frequency, they won't contribute to oscillations. A fixture should be used that transitions from the device
impedance to that of the rest of the environment as close to the device as possible. External matched loads can be placed in parallel with high impedance measurement equipment in order to match them to cabling, if the associated shunt current can be accurately subtracted from the measured device characteristics.

![Image](image.jpg)

**Figure 3.4:** Items used to package and fixture DBRTDs for connection to 3.5 mm 50 Ω line. The devices are epoxied into the cavity of the small pill package with a threaded post and covered with the circular lid after being wirebonded. The pill packages are then threaded into the machined fixture, and the bulkhead connector screwed into the fixture until the center conductor touches the pill package lid.

The device measurements done for this thesis were most successfully stabilized using the fixtures pictured in Figure 3.4. The devices were fabricated from samples with conducting substrates using the ion-implantation isolation process described previously. They were then cleaved into 800 μm × 800 μm squares and silver epoxied
Figure 3.5: Wirebonded device inside a pill package. The ribbon wire is stitched from the rim of the package on one side to the contact pad for the device and back to the rim of the package on the other side. The device contact pads are 50 μm in diameter.

into threaded S4 pill packages. Ribbon wire bonds were stitched from the rim of the package to the contact pad and to the opposite rim before the caps were soldered on with indium solder. Figure 3.5 shows a wire bonded device in a pill package without a cap. The capped devices were then threaded into the inside of the pictured, custom machined fixture, and a 3.5 mm bulkhead connector was threaded in above it until the center conductor contacted the pill package cap. This provided a rapid transition from the device to a 50 Ω coaxial line. Figure 3.6 shows the characteristics of a diode measured with and without the suppression of parasitic oscillations.
Figure 3.6: DC current-voltage characteristics of a DBRTD with (solid line) and without (dashed line) suppression of parasitic oscillations. The "chairs" seen in the dashed curve have been misinterpreted in terms of DBRTD structure rather than a circuit effect.
3.3 Microwave Measurements

In addition to the DC current-voltage characteristics of a DBRTD, the frequency dependence of the characteristics must be known to model the performance of the device in a high speed circuit. This information can be obtained from scattering parameter measurements which are then fit to an equivalent circuit model. The parameters of the components in the equivalent circuit are chosen to give a good fit to the data. Microwave characterization of DBRTDs is presented here, including data for the NDC region which is usually not reported in the literature because of stabilization problems[44].

![Diagram of equivalent circuit](image)

Figure 3.7: Small signal equivalent circuit for packaged tunnel diodes. Intrinsic elements of the diode are a junction resistance, $R_j$, junction capacitance, $C_j$, and a series resistance $R_s$. The package contributes an additional series inductance and resistance, $L_p$ and $R_p$, and a parallel capacitance, $C_p$.

3.3.1 Equivalent Circuit

A simple small signal equivalent circuit for tunnel diodes is shown in Figure 3.7[8]. It consists of a parallel junction resistance, $R_j$, and junction capacitance, $C_j$, in series with a contact or series resistance, $R_s$. More complete models exist to describe the
effects of spacer layer transit times and quantum mechanical evolution times[45], but these are not required for the range of frequencies investigated here. In addition to the intrinsic components of the DBRTD, there are external parasitics due to packaging, $R_p$, $L_p$, and $C_p$.

At low frequencies, the capacitances and inductances can be neglected. The measured DC characteristics correspond to $R_s$ and $R_j$, neglecting the packaging or measurement setup resistances. At any point, the derivative of the DC characteristic is equal to the conductance of $R_s$ and $R_j$ in series, or

$$\frac{dV}{dI} \bigg|_{f=0} = \frac{1}{R_s + R_j}, \quad (3.2)$$

which clearly changes as a function of voltage.

Physically, $R_s$ corresponds to the resistance associated with the contacts and bulk regions which should be ohmic, so $R_s$ is constant. The variation in the slope of the DC characteristic is all attributed to the changes in $R_j$. It is negative between the peak and valley voltages. The junction capacitance, $C_j$ corresponds to the capacitance between the electrodes and can be modeled as a parallel plate capacitor with variable spacing between the plates. As the voltage across the device increases the depletion region extends further into the anode and $C_j$ is expected to decrease slightly.

At high frequencies $C_j$ will effectively begin to short $R_j$. If the tunnel diode is biased into the NDC region, when the negative real part of the parallel combination's impedance equals $R_s$, the overall real impedance is zero and the device no longer appears to have a NDC. This happens at the cutoff frequency given by

$$f_{co} = \frac{1}{2\pi R_j C_j \sqrt{\frac{|R_j|}{R_s} - 1}} \quad (3.3)$$

and places a limit on the useful frequency range of the DBRTD as a NDC device. Usually, the negative junction resistance is much greater than the series resistance, making the cutoff frequency inversely proportional to the geometric mean of $R_s$ and $R_j$. Thus, decreasing the contact resistance is just as important as increasing current density for high speed operation.
3.3.2 Measurement Technique

Values for the equivalent circuit elements of the tunnel diode can be obtained by making one port scattering parameter measurements of the device on a HP 8510 or similar vector network analyzer. This amounts to obtaining the phase and magnitude of the reflection coefficient given in equation 3.1 as a function of frequency for a known $\hat{Z}_0$ with the device being tested as the load. The reflection coefficient has unity magnitude when $f = f_{co}$. Typically, a microwave modeling program like Libra is used to adjust the values of the equivalent circuit elements until the resulting modeled data matches the measured data. An example of matching measured and modeled data for a DBRTD is shown in Figure 3.8. The values of the elements used in the model are $R_s = 1.7 \, \Omega$, $R_j = -200 \, \Omega$, and $C_j = 1.8 \, \text{pF}$.

3.3.3 Measurement Results

Initially, microwave measurements were made to determine the junction capacitance of devices in the NDC region. Since the characteristics needed to be stabilized without using bypass capacitors, which would change the measured device capacitance, low current devices that were naturally stable were used. They were mounted on microstrip fixtures and measured in a compatible jig. The measurements were done in the NDC region for both polarities of different sized devices fabricated on semi-insulating substrates. The results are shown in Figure 3.9.

The device layer structure, which is listed in Table 5.1, had a 60 Å GaAs well, 30 Å AlAs barriers, and 100 Å undoped spacers. The electrodes were doped differently, one at $1 \times 10^{17} \, \text{cm}^{-3}$ and the other at $1 \times 10^{18} \, \text{cm}^{-3}$, making the characteristics asymmetric. Under positive bias, electrons were injected from the more lightly doped side. Because a small depletion length in the heavily doped electrode was sufficient to support the electric field, the capacitance is larger than in the opposite polarity. Measuring different sized devices allowed the true capacitance per unit area to be separated from other fixed capacitance contributions which were about 0.1 pF. For forward bias of 0.25 volts, the measured capacitance was 7.2 fF/μm², and the value calculated from equation 2.15 was 4.2 fF/μm². For reverse bias of -0.5 volts, the measured capacitance was 3.0 fF/μm² compared to a calculated value of 1.4 fF/μm².
Figure 3.8: Measured (crosses) and modeled (boxes) data for the reflection gain of a 5 μm diameter DBRTD. The model uses values of $R_s = 1.7 \, \Omega$, $R_j = -200 \, \Omega$, and $C_j = 1.8 \, \text{pF}$.

It is unlikely that the underestimation comes from neglecting the charges associated with current flow in the well and spacer since the peak current density of the devices was only about $2 \times 10^3 \, \text{A/cm}^2$. Even if this current is attributed to electrons at a saturated velocity of $1 \times 10^7 \, \text{cm/sec}$, switching the associated charge in the depleted 300 Å structure with 50 mV gives a capacitance of only about 0.1 fF/μm². More probable sources for error include an underestimation of the voltage drop in the cathode and failure of the parallel plate model because charges are being modulated in the well or accumulation layer.
Figure 3.9: Capacitance of different sized DBRTDs on semi-insulating substrates determined from scattering parameter measurements. The capacitance was measured versus device area for a forward bias of $V = 0.25$ volts (squares fit with the solid line) and for a reverse bias of $V = -0.5$ volts. (circles fit with the dashed line).

More complete measurements were made on a variety of samples with conducting substrates and ion implant isolated devices. These structures were measured for a large range of biases, but $C_j$ could not be extracted because it was obscured by the contact pad capacitance. In general, the series resistance was less than 50 $\Omega\cdot \mu m^2$ and $R_j$ was accurately given by equation 3.2. Based on the measured $R_j$ and $R_s$ for devices such as those presented in Figure 3.8, and relatively conservative values of 4.0 fF/$\mu m^2$, cutoff frequencies in excess of 100 GHz can be expected if the same devices were fabricated on semi-insulating substrates.

Anomalous low frequency behavior was observed in a number of these samples.
Large capacitance spikes were observed for some biases. Typical low frequency capacitance vs. bias data and DC characteristics are shown in Figure 3.10 for one of these devices. At some biases, the measured capacitances were ten times the expected value of 2 pF, which was also the minimum measured value. The anomalously large capacitances rolled off very quickly at higher frequencies. The capacitance data was usually flat with bias at 2 pF by 1 GHz. This observation is consistent with trapping phenomena. Low frequency noise due to trapping phenomena in DBRTDs has been recently reported[46]. More work is required to tell if this is related to the measured changes in capacitance, but seems most probable for these devices.

Another possibility is that the large capacitance variations are artifacts of measurement errors. Since the devices had a large impedance, the reflection coefficient was near unity with a small imaginary part. Small variations in measured phase get transformed into large changes in capacitance in this low frequency region. Thus the capacitance structure might be attributed to measurement errors, although the original data does show small but distinct changes in phase in these regions. More careful measurements with greater attention to phase angle calibration and variation need to be made to confirm these initial measurements.
Figure 3.10: (a) DC characteristics and (b) junction capacitance for DBRTD measured at 130 MHz.
Chapter 4
Tunnel Diode Applications

Today, three-terminal transistors are strongly preferred over two-terminal diodes as building blocks for modern integrated circuits. The tunnel diode was born, however, in an era prior to integrated circuits or even small geometry discrete devices. At that time, the reduced number of terminals was considered a virtue because it allowed smaller devices. Attaching a third terminal required a nearly macroscopic contact in another dimension, greatly increasing the device size[14]. In the early 1960s, the tunnel diode was the leading candidate among two-terminal devices for circuit applications because of its speed and unique current-voltage characteristics. However, the advent of better fabrication technology decreased the size of the transistor and increased its speed.

Microlithography not only removed the size advantages of the tunnel diode, but brought about integrated circuits which required multiple devices on a chip, all working reliably. The junction tunnel diode was not a high-yield or high-reliability device[47], and was sometimes manufactured with techniques inappropriate for integrated circuits[48, 49]. The DBRTD appears to have overcome these problems, having only the two-terminal flaw in common with the junction tunnel diode. These reliability issues along with speed and flexibility will be discussed in comparing the junction and resonant tunnel diode in the first section of this chapter.

The excitement over the junction tunnel diode as a general purpose semiconductor device was brief but produced a large number of proposed applications. Since reliable performance and integrability are not barriers to the DBRTD, it is useful to survey these proposals as well as new potential applications of tunnel diodes. This is done in the second section of the chapter. Present three terminal resonant tunneling devices are also discussed and categorized.
4.1 Tunnel Diode Comparison

The current-voltage curves of junction tunnel diodes (JTDs) and DBRTDs have similar shapes, but there are also many differences between the devices. Several of the differences hinge on the fact that the JTD is a bipolar device where tunneling is controlled by the bandgap[11]. In contrast, the DBRTD is a unipolar device, usually n-type, and the tunneling is determined by the conduction band structure.

![Diagram of a junction tunnel diode]

Figure 4.1: Band diagram of a junction tunnel diode under small forward bias. Both sides are degenerately doped, allowing electrons from the conduction band of the n-side to tunnel into hole states in the valence band on the p-side.

4.1.1 Junction Tunnel Diode Operation

Junction tunnel diodes are formed at abrupt junctions between heavily doped n-type and p-type regions. Under small forward biases, electrons can tunnel through the bandgap from the conduction band of the n-type side to the valence band of the p-type side. This condition is illustrated in Figure 4.1. As more bias is applied, the n-type side's conduction band minimum rises above the p-type side's valence band.
maximum. Since there are no longer available states in the bandgap opposite the tunneling electrons from the conduction band, the tunneling current goes to zero. With further bias, currents due to minority carrier injection increase. At biases in between those for tunneling and the p-n junction diode threshold, excess currents flow due to non-ideal mechanisms, such as recombination at traps in the junction. The level of this excess current determines the minimum or valley current[11].

4.1.2 Flexibility

The doping levels on each side of the junction as well as the host material determine the current-voltage characteristic of the JTD. Since the doping degeneracy is limited to a few tenths of an eV at most, the peak voltage and width are limited to this range. The peak voltage is reasonably well approximated[50] as a third of the total degeneracy, \( qV_p = ((E_{F,n} - E_{c,n}) + (E_{v,p} - E_{F,p}))/3 \). For example, if a GaAs p-n junction is grown by MBE with acceptor and donor concentrations of \( 2 \times 10^{20} \text{ cm}^{-3} \) and \( 2 \times 10^{19} \text{ cm}^{-3} \) respectively, which are the near the highest possible using beryllium and silicon[20, 51], the degenerate Fermi levels calculated with the Joyce-Dixon approximation[52] are \( E_{F,n} - E_{c,n} = 0.485 \text{ eV} \) and \( E_{v,p} - E_{F,p} = 0.348 \text{ eV} \). Although both of these degeneracies are overestimated due to the Joyce-Dixon approximation and nonparabolicity of the bands, the calculated peak voltage is only 0.278 volts. Experimentally, MBE grown GaAs JTDs with \( 5 \times 10^{19} \text{ cm}^{-3} \) beryllium and \( 1.3 \times 10^{19} \text{ cm}^{-3} \) silicon doping concentrations have been reported to give peak voltages of 0.140 volts[51]. The JTD's valley voltage is largely a function of the bandgap energy. Thus it is difficult to change the voltage scale or general shape of the junction tunnel diode characteristic for a particular material. Only the doping can be adjusted over a limited range, and this also changes the device current and capacitance.

In contrast, the DBRTD offers a multidimensional design parameter space. As discussed in section 2.3.2, the peak voltage can be adjusted with the well size or composition, and the peak current density can be adjusted with the well size, barrier thickness, or cathode doping. The entire scale of the characteristic can be variably adjusted by modifying the relationship between the electric field across the well and
the total applied voltage with different doping profiles in the anode. Additionally, the work presented in chapter 6 shows that changing the band profiles with heterojunctions in the cathode region near the barrier can have dramatic and beneficial effects. The controlled variation of these parameters can produce a wide variety of current-voltage characteristics.

Although experimental structures can be repeatably grown with certain properties, present models do not adequately predict the NDC region's shape. Better models must be developed in order to take advantage of the many design freedoms permitted by the DBRTD. Also, it is empirically difficult to get both low peak voltages and high peak currents and valley voltages concurrently[31]. JTDs typically have a lower ratio of peak voltage to NDC region width, $V_p/\Delta V_{ND}$, than is easily obtained with DBRTDs. This gives the JTD better power efficiency for oscillators as discussed in connection with equation 2.3. There seems to be a general trend of reduced peak current density as the peak voltage is reduced. The reasons for this are not entirely clear, but it likely involves the blocking of incident electrons by undoped spacer layers. Figure 4.2 plots the difference in peak and valley current density against peak voltage for devices made in this work.

### 4.1.3 Speed

As noted before, tunneling at a bipolar junction under forward bias requires degenerate doping on both sides of a very abrupt junction. The two sides must be close to give sufficient tunneling probability, but this also results in very large junction capacitances. The junction capacitance can be at least an order of magnitude smaller in DBRTDs because the doped regions are separated by an undoped region surrounding the barriers. DBRTD capacitances can be reasonably approximated by a parallel plate model where the charges reside in the cathode next to the barrier and at the edge of the depletion layer in the anode. Large spacer layers can be used to increase the distance between these two effective plates. Charge separations of 1000 Å are easily accomplished and give junction capacitances of about $1 \times 10^{-7}$ F/cm². Calculated junction capacitance values for high current density germanium junction tunnel diodes are above $1 \times 10^{-6}$ F/cm² [11]. The validity of the parallel plate model
Figure 4.2: Difference between peak and valley current density \((I_p - I_v)\) vs. peak voltage for various devices measured in connection with this work. Open symbols denote room temperature measurements and closed symbols denote boiling nitrogen temperature measurements.

should be examined for devices with high current densities. Assuming a saturated electron velocity of \(1 \times 10^7\) cm/s, a current density of \(1 \times 10^5\) A/cm² corresponds to a charge concentration of \(1.6 \times 10^{17}\) electrons/cm\(^3\) which might be comparable to the electrode doping concentration.

Thus, for similar peak current densities the speed index, which was discussed in section 2.3.2, is in general higher for DBRTDs than JTDs. The best currently published results for this index are \(I_p/C_j = 1000\) mA/pF[53] and \(I_p/C_j = 70\) mA/pF [48] for resonant and junction tunnel diodes, respectively. However, the JTD in this case was formed by a welding technique inappropriate for integrated circuit applications. Typical best numbers for GaAs, planar JTDs suitable for integration are
\( I_p/C_j \approx 10 \text{ mA/pF}[54]. \)

### 4.1.4 Reliability

The junction tunnel diode suffers reliability problems that would cause very poor manufacturing yields for integrated circuits using multiple devices[9]. This problem, in part, can also be attributed to the bipolar nature of the JTD. When the diode is biased into forward conduction, thermal minority currents flow which recombine in the vicinity of the junction. A significant portion of the energy released by the recombination goes into producing phonons, which drive recombination enhanced processes that are detrimental to the junction[55]. For example, the dopant impurities may diffuse, reducing the abruptness of the junction[56], or defects can increase or migrate. The processes are similar to those involved in dark line defects in semiconductor lasers.

The problem is exacerbated by any defects near the junction. These same defects also increase the excess current. To avoid future degradation junction tunnel diodes are often screened for excessive valley current by rejecting devices where the peak-to-valley current ratio is too small[47]. Modern manufacturers of planar JTDs reject about 20% of their discrete components in this way[57]. Such a failure rate would make integrated applications requiring tunnel diodes infeasible. In contrast, DBRTDs appear to be quite stable and robust, although quantifying work has not yet been done in this area. Some of the DBRTDs fabricated for other purposes as a part of this work were subjected to intense gamma radiation with little effect. Only those devices with \( 1 \times 10^{17} \text{ cm}^{-3} \) doping or less showed increased contact resistance.

One of the causes leading to junction defects in JTDs is mechanical stress. Before planar fabrication techniques were used, many methods of manufacture required the relatively fragile junction to serve as a structural support for the contact. One method of manufacture formed the junction by applying electrical pulses to a metallic whisker in contact with the semiconductor bulk until the desired characteristics were obtained[48]. Either the metal itself or a coating served to dope the welded junction to the opposite type of the bulk. The inevitable stress placed on the whisker in packaging was applied directly on the junction.

Prior to planar devices, the most common fabrication method formed the junction
by alloying a small, dopant coated ball on the surface of the semiconductor[49]. Since
the ball formed a much larger junction than desired, the sample was etched down
until only a small neck supported the ball above the semiconductor. The stress of the
ball and a contacting wire neck were again placed directly on the junction formed in
the small pillar.

The most modern and reliable method of manufacturing JTDs is to diffuse dopants
or make alloy contacts through small openings in an oxide over the semiconductor[58].
Even in this case, it is necessary to move the bond pads away from the junctions to
avoid degrading them[59].

4.1.5 Uniformity

In addition to yield, device uniformity is required for integrated applications. To
check the uniformity of DBRTDs, the peak voltage and peak current were measured
for a number of devices throughout half of 2" wafers. The radial profile of these
peak parameters are shown for two wafers in Figure 4.3. The wafers had nominally
identical DBRTD epitaxial layers that were grown in the two different chambers of
the Stanford MBE system. There was almost no correlation between the diodes’
parameters and their angular position.

The peak current varies by almost a factor of four across the wafer from system #1
but only 20% across the wafer from system #2. This later variation would be accep-
table for most applications. The differences in peak current can be attributed to
changes in the AlAs barrier thickness. The system #2 wafer’s uniformity is rather
remarkable since the peak current depends exponentially on the barrier thickness.
The total of the barrier thicknesses must not vary by more than a quarter of a mono-
layer to achieve the observed peak current uniformity. The difference between the two
wafers points out the potential for DBRTDs to be used as a very sensitive measure
of uniformity. Proposals for their use in this fashion have been published[60].
Figure 4.3: DBRTD (a) peak current and (b) peak voltage as a function of radial position measured from the center of a 2" wafer. The data are for nominally identical structures grown on Stanford MBE system #1 (+) and system #2 (x).
4.2 Tunnel Diode Applications

Junction tunnel diode applications in general fall into two categories: microwave and pulse circuit applications using single discrete diodes and digital circuit applications using multiple diodes. In general, the former met with some success while the later met with none. The applications originating from the junction tunnel diode era are discussed before going on to new ones for DBRTDs. Among these newer applications are ones that make use of multiple NDC regions obtained with stacked or other special RTD structures. It should be noted that all of these applications can benefit from reliable manufacture of DBRTDs discussed above. This includes those applications using single discrete devices because superior performance and economy may result from monolithic integration with microwave structures.

4.2.1 Amplifiers and Oscillators

As discussed in section 3.2.1, microwaves incident on tunnel diodes biased into NDC can generate larger amplitude reflections. Tunnel diodes used in this fashion constitute one port amplifiers[49]. Isolation between input and output signals is usually accomplished with directional couplers that deflect the amplified reflection of the input into another waveguiding structure. The amplification is broadband, from DC to the cutoff frequency determined by the diode equivalent circuit parameters. Reflected power gain from DBRTDs has been demonstrated as high as 420 GHz[61].

Since the NDC characteristic of tunnel diodes are curved, the use of tunnel diodes as linear amplifiers is restricted to applications with small input signals. The maximum power for non-linear amplification is determined by the width of the NDC region, $\Delta V_{\text{NDC}}$. Making this region broader and more linear are of great interest in DBRTD studies.

It is evident from the problems encountered with parasitic oscillations during DC measurements that devices biased into NDC can readily serve as local oscillators to generate microwave signals. In this case, the nonlinear amplification characteristics are not important, although the power output is again limited by the width of the NDC region. This is particularly important since other NDC devices, such as the Gunn and IMPATT diodes, can achieve much higher powers, on the order of watts.
CHAPTER 4. TUNNEL DIODE APPLICATIONS

However, the Gunn diode is limited to operating frequencies less than 100 GHz, and the IMPATT diode is noisy and it cannot be tuned since its oscillation frequency is determined by the length of the transit region[11]. Present DBRTDs can generate 30 milliwatts of power at 100 GHz[61].

The broadband device is forced to oscillate at a particular frequency by placing it in a resonant structure that provides properly phased positive feedback at the desired frequency. The exact oscillation frequency can be tuned by adjustments to the bias voltage or other perturbations. Both cavity and hybrid microstrip resonator structures have been investigated for DBRTDs.

4.2.2 Detectors and Mixers

The overwhelming current use of the junction tunnel diode is as a zero bias detector or mixer[12]. The critical aspect of the characteristic for this application is not NDC, but the large non-linearity at the origin. The conductance is large at reverse and small forward biases up to the peak voltage, but beyond the peak, the conductance is small until significant minority current injection occurs. The overall characteristic near the origin resembles that about the turn-on threshold of a normal diode, except with the opposite polarity of current and voltage, hence the name "back diode". Other devices, including Schottky diodes and resonant tunnel diodes, have strong nonlinearities, but not at the origin. In contrast to the back diode, these devices would need to be biased into the vicinity of their nonlinearity.

Most resonant tunnel diodes use the same material for electrodes on each side of the barrier structure. As a consequence it is extremely difficult to design DBRTDs whose conductance changes rapidly at the origin. In view of the large market for back diode detectors, it seems worthwhile to investigate unipolar tunneling structures for this application. The electrodes for such a device will likely need to be materials with different band structures. One possibility is to use a tunnel barrier between two superlattices designed so that their minibands are not aligned. Applying bias in one direction would improve the alignment, increasing conduction, and applying bias in the opposite direction would worsen the alignment, decreasing conduction.
4.2.3 Switching and Pulse Circuits

The tunnel diode's speed makes it appropriate for use in switching circuits that can tolerate the lack of isolation[62, 30]. Usually, an input signal is applied as a voltage to a series combination of a tunnel diode and a resistor. The conductance of the resistor is chosen smaller than the maximum magnitude of NDC, making the combination bistable for some range of applied voltages. The circuit is biased near the edge of this bistability and the output is taken as the portion of the applied voltage dropped across just one of the elements. Two tunnel diodes in series, called tunnel diode pairs, could also be used.

Since the circuit is biased near the edge of the bistability, a small pulse applied to the input can trigger the combination to switch from one stable state to the other. This creates an output pulse that is on the order of the bias, which can be made much larger than the input signal. Thus the tunnel diode can generate fast, large amplitude pulse edges from small, slowly varying signals. This scheme is used in triggering circuits and pulse generating circuits. Once the circuit is switched, it must be reset by either reducing the bias or by applying a large input with the opposite sense.

4.2.4 Digital Proposals

Most of the proposals for digital circuits involve tunnel diodes in bistable combinations like the one described in the last section. By combining them with other elements, a wide variety of circuits can be implemented. A natural application for bistable components is latches or flip-flops. Proposals for a variety of different flip-flop circuits were made in the literature, including ones that were isolated with rectifying diodes[63] and ones that would alternately set and reset, based on dynamical interactions with inductive circuits[64, 65].

Memories using tunnel diodes were the most frequently proposed digital circuits, and they still attract some interest today. One early publication described 15 nanosecond access time memories based on tunnel and back diodes. Bistability is again important because it allows single tunnel diodes to function as very compact static memory cells. Memories based on combinations of JTDs and transistors, either
CHAPTER 4. TUNNEL DIODE APPLICATIONS

GaAs MESFETs[66] or silicon MOSFETs[67, 68], have been proposed repeatedly this decade.

Several more elaborate digital circuits were also proposed, and many demonstrated as prototypes. These include shift registers[69] and binary full adders made with tunnel diodes as their only active component[70]. The undulating nature of the tunnel diode characteristic was employed in the adders, and was often expanded to larger voltage scales using resistors in series with the tunnel diodes. By using different values of resistors, the peak voltages could be shifted to different places. The currents form these shifted characteristics would then be recombined to perform logic functions. Transistors were sometimes used to isolate and amplify signals between stages for adders[71] or ring counters[72] or general sequential circuits[73].

JTD based analog-to-digital conversion[74] was also possible by using series resistors to effectively scale the inputs that were applied to tunnel diodes. If the scaled input was adequate, it would switch the associated diode. Since properly loaded tunnel diodes switch very abruptly between two stable states, the output was essentially discrete. A circuit was also demonstrated that converted analog video signals to pulse width modulated binary signals[75].

Tunnel diode FET logic was also proposed, based on the unique nonmonotonic characteristics of JTDs. The tunnel diodes were used as loads for inverters to derive certain benefits that will be described in detail in chapter 5.

4.2.5 Multi-level Applications

Virtually all of the new applications that have been proposed in connection with DBRTDs rely on characteristics with more than one peak. Several means of accomplishing such characteristics have been demonstrated. The first and easiest is to use DBRTDs with more than one resonance[76]. However, most applications require nearly equal peak currents, and excited state peak currents are usually much larger than ones associated with lower energy resonances. Another option is to use identical parallel DBRTDs, but shift their characteristics so the peaks occur at different voltages. This may be accomplished by biasing them with external voltage sources[77] or series resistors of different values[78] similar to the scheme used for JTD based
analog-to-digital converters.

Currently, the most popular method is to combine DBRTDs in series[79, 80, 81]. The number of useful peaks is equal to the number of stacked devices. One drawback to this method is that the characteristics are hysteretic except for small stacks under special circumstances. These stacks have also been integrated with heterojunction bipolar transistors[82].

The applications of characteristics with multiple peaks are mostly the same as existing proposals from the JTD period, except the multiple peaks allow additional signal levels. The characteristics of stacked structures could also be generated from series combinations of JTDs. Multi-state memories was one of the first applications mentioned for the multi-peaked characteristics[77, 78, 79]. Other applications include frequency multiplication, waveform scrambling, parity generation, and analog-to-digital conversion[80, 82]

4.2.6 Resonant Tunneling Transistors

Devices with the speed and flexibility of DBRTDs but an additional, isolated input to control them would be a desirable replacement for transistors in many applications. In fact, great emphasis has been placed on adding another terminal to DBRTDs. This has been attempted in three general ways. The approaches differ, based on the largest size active feature of the device, which in turn determines the nature of the transport in the device and is also the easiest point to attach a third terminal. Smaller dimensions are desirable from a speed perspective but greatly complicate attempts to contact the region.

The easiest devices to fabricate are ones which are essential hybrids with a resonant tunnel diode placed in series with one terminal of a transistor. The connecting region is greater than or on the order of micrometers. Except for the effects of interconnection parasitics, their performance is identical to discrete devices connected with wires. Transport in the connecting region is by the drift and diffusion mechanisms found in contemporary devices. In this case, the third or controlling terminal is already connected to the transistor and no contact is made to the region between the DBRTD and the transistor. Examples of this type are hybrids of field-effect[83, 84], or bipolar
transistors[85, 86] with DBRTDs.

The next class of devices are ones which have active regions up to 100 nm, small enough to support hot-electron transport. They typically use resonant tunneling barriers to filter out relatively hot electrons which will traverse the active region rapidly. Contact is made to the active region, often called a base, to provide a third terminal which can turn the injection of hot electrons from the tunneling structure on and off. The DBRTD is still used as a diode, but the rest of the device is close enough to it to have its transport mechanisms modified, although it does not make use of quantum interference effects. The predominant example of this type of device is the resonant hot-electron transistor[87], although some versions of the vertical FET and DBRTD hybrid may come close[84].

The most aggressive approach is to actually contact the active region of the DBRTD itself, which must be on the order of 10 nm or less in order to preserve the coherence of the electron required for interference effects. Then by modifying the potential energy within the device, electrons can be brought in and out of resonance. Devices of this type have only recently been realized in practice. A couple of workers have fabricated lateral FETs using multiple, very small gates on HEMT layers. The high electron mobility increases the coherence length of the electron so that the tunneling structure can have larger dimensions[88, 89, 90]. Other work has used more traditional vertical MBE structures, but with some modifications to make contacting the internal layers easier[91]. Because the contact layers are so thin, parasitic resistances which will significantly reduce the useful speed of the device are a very serious concern. Another version of a vertical resonant-tunneling transistor places the DBRTD structure in the base of a heterojunction bipolar transistor[92]. In this case, a p-type contact is made surrounding the DBRTD structure. It is not clear that there are any advantages to just contacting the well when the structure is bipolar.

The applications proposed for these devices have been very similar to those for stacked devices. One of the additional advantages of the resonant tunneling transistors over DBRTDs is they offer both negative conductance and negative transconductance. The measured speeds of the relatively crude resonant tunneling transistors
fabricated to date have been well below those for heterojunction bipolar transistors.
Chapter 5
Tunnel Diode FET Logic

Tunnel diode FET logic was one of the last proposals for junction tunnel diode applications prior to the advent of the DBRTD, coming well after the early 1960s when most others were made. The use of junction tunnel diodes as load elements for FETs, tunnel diode FET logic (TDFL), was originally proposed by Lehovec [54] in 1979 to reduce the power consumption and enhance the noise margins of GaAs MESFET based logic gates. He also indicated that a simultaneous speed advantage could be obtained because of increased drive current, even though the relatively high capacitance of the JTDs dominated the node capacitance. The superiority of the DBRTD over the JTD in terms of speed and reliability for integrated circuits that require many tunnel diodes offered a natural motivation to investigate TDFL using DBRTDs[93].

The additional flexibility of the DBRTD design also raised the question of optimal load characteristics for inverters. To answer the question, inverter performance was simulated for various load characteristics as a function of the speed vs. power tradeoff. The different characteristics would, in practice, be obtained by varying the design parameters of the diode. The effects of these design parameters on the overall transfer characteristics were both modeled and demonstrated experimentally. Finally, it was desirable to show that TDFL gates could in fact be fabricated by monolithically integrating DBRTDs and MESFETs. This section describes these aspects of the thesis after a general introduction to the operation of TDFL.

5.1 Static Operation

Tunnel diode FET logic gates are topologically identical to the common enhancement/depletion (E/D) FET logic gate, but use a tunnel diode as the load element in place of the depletion mode load. The schematic diagram for inverters in both E/D logic and TDFL are shown in Figure 5.1, along with the characteristic load
lines for the different types of elements superimposed on the characteristics for an enhancement-mode MESFET (EFET) driver.

Figure 5.1: Schematic diagrams for (a) an enhancement/depletion FET inverter and (b) a tunnel diode FET logic inverter using JTDs and (c) DBRTDs. The characteristic load line for each of the loads is superimposed on the characteristics for an enhancement-mode MESFET driver.

5.1.1 Operating Points

The inverter's static output voltage and current dissipation are determined by the intersection of the EFET characteristics and the load line. When the input or gate voltage is high, the output voltage is low. Note that the static current and output voltage in this state are both lower for the tunnel diode (TD) load line than for the
depletion-mode FET (DFET) load line. This occurs because the valley voltage of
the tunnel diode has been chosen to be similar to the supply voltage. When the
input voltage is low, the output voltage is high. In this case both load lines give
approximately the same current and output voltage. For sufficiently small input
voltages, the current in this state is negligible.

On average, the inverter is assumed to be in each state the same amount of time,
so the average static current is one-half the value for the low output state. Clearly,
the TD loaded inverter dissipates less static current, or equivalently, less static power
than the DFET loaded inverter. Since the the output voltage is also lower, the total
voltage swing with the TD load is larger than the voltage swing with the DFET load.
The difference will be increased by the application of the output voltages to a chain
of several gates with similar loads.

5.1.2 Bistability

Figure 5.1 also illustrates the fact that for certain intermediate gate voltages, the
TD load line intersects the EFET driver characteristic at three points, one or two of
which are stable. The intersection is stable if the slopes of the intersecting character-
istics have opposite signs. Thus the EFET and TD load inverter can be bistable for
some input voltages, similar to the bistability of the series tunnel diode and resistor
combination used for switching as discussed in section 4.2.3. The initial conditions
and dynamics of the inverter determine at which of the stable operating points the
inverter settles. Note that some of the stable points can never be reached if the
TD load's peak current exceeds the EFET's maximum current at the correspond-
ing voltage. The inverter's bistability produces a hysteretic transfer characteristic as
seen in Figure 5.2. The DFET loaded inverter exhibits no bistability since there is
always only a single intersection between the monotonic DFET load and the EFET
characteristic for any gate voltage.

One of the attractive features of a bistable gate is its potential as a compact
memory element. With reference to Figure 5.2, if the input voltage is a midrange
value such as $V_Q$, the output voltage can be in one of two stable states. In order
to force the output voltage high, the input voltage can be pulled below $V_I$ and then
Figure 5.2: Transfer characteristics of a TDFL gate exhibiting hysteresis that leads to bistable operation. The stable operating points for normal logic application are indicated with circles, and the associated noise margins are shown in the figure.

allowed to return to $V_Q$. The clockwise transversal of the hysteretic loop requires the inverter to come back in the high output state. Similarly, a momentary positive voltage pulse greater than $V_2$ on the input puts the gate in the low output state. Thus the simple TD loaded inverter can function as a static memory cell. The one disadvantage of this implementation is that bipolar pulses are required to write to it, so the scheme is not directly compatible with existing binary logic.

The other benefit of the bistable transfer characteristic is that it improves the noise margins of the inverter. Noise margins can be defined as the difference in voltage from the stable operating points to the nearest unity gain point on the transfer characteristic. The stable operating points for a transfer function $v_{out} = f(v_{in})$ are $v_{in} = V_H, v_{out} = V_L$ and $v_{in} = V_L, v_{out} = V_H$ where $V_L$ and $V_H$ are the solutions to $v = f(f(v))$. The noise margins for the transfer function in Figure 5.2 are illustrated
there. Note that it would be impossible to have noise margins whose sum was greater than the supply voltage unless the transfer curve was hysteretic. Large noise margins help to prevent logic gates from functioning improperly due to power supply or electromagnetic noise, device variability, or other environmental changes. Essentially, the inverter functions as a very compact Schmitt trigger because of the bistability.

5.2 Dynamic Operation

The inverter's current and voltage in the low output state could be made identical for the DFET and TD loads by scaling down the DFET load characteristic so that it intersects the low output operation point of the TD load. While this would give identical behavior at the two static operating points, it also results in the inverter taking a longer time to switch from the low output state to the high output state. Since the entire load line would be scaled equally, the current through the load for all voltages would decrease. During a transition of the output from low to high, the load would source less current to charge the capacitances associated with the node, increasing the time required to charge the node capacitance as the output rises. The advantage of the TD load is that it can source a larger current for charging the node capacitance as the output voltage rises than the static current it sources in the low output state. This is possible because the nonmonotonic tunnel diode characteristic has a low valley current at high voltages and a high peak current at low voltages.

5.2.1 Node Equations

The inverter can be most simply modeled as a two-terminal load, a three-terminal driver, and a constant node capacitance, \( C \), which accounts for the capacitance of the load and driver as well as the input capacitance of the next stage. Figure 5.3 shows a schematic with these components and labels for the currents and voltages. The current in the load, \( i_L(V_{dd} - v_{out}) \), depends only on the difference between the supply and output voltages, \( V_{dd} \) and \( v_{out} \). The driver current is \( i_d(v_{in}, v_{out}) \) and depends on the gate or input voltage, \( v_{in} \), and the drain or output voltage, \( v_{out} \). Changes in the output voltage depend on the amount of net current flowing into the node.
CHAPTER 5. TUNNEL DIODE FET LOGIC

Figure 5.3: Functional diagram of an inverter composed of a driver and a load. The node capacitance represents the effective capacitance of the inverter elements and the input capacitance of following stages.

capacitance,

\[
\frac{dv_{out}}{dt} = \frac{1}{C} i_{net},
\]

(5.1)

where \( i_{net} = (i_L(V_{dd} - v_{out}) - i_d(v_{in}, v_{out})) \), the difference between the current sourced by the load and that sunk by the driver. For an enhancement-mode MESFET driver not in the velocity-saturation regime,

\[
i_d(v_{in}, v_{out}) = \begin{cases} 
Kv_{out}(2v_{in} - v_{out}) & \text{if } v_{out} < v_{in} \\
Kv_{in}^2 & \text{if } v_{out} > v_{in},
\end{cases}
\]

(5.2)

with \( K = (\mu_\epsilon W)/(2aL) \), where \( \mu \) is the electron mobility, \( \epsilon \) is the dielectric permittivity, \( a \) is the channel depth, and \( W \) and \( L \) are the gate width and length, respectively. Assuming that the input voltage suddenly changes from \( V_H \) to \( V_L \) when the output is stable in the low output state, \( v_{out} = V_L \), the time required for the output to charge up to another voltage \( v \) is

\[
T_{\text{rise}}(v) = C \int_{v_{out}=V_L}^{v_{out}=v} \frac{dv_{out}}{i_L(V_{dd} - v_{out}) - i_d(V_L, v_{out})}.
\]

(5.3)
The time required for the output to make the opposite transition, starting at \( v_{\text{out}} = V_H \), in response to the input changing from \( V_L \) to \( V_H \), is
\[
T_{\text{fall}}(v) = C \int_{v_{\text{out}} = V_H}^{v_{\text{out}} = V_L} \frac{dv_{\text{out}}}{i_d(V_H, v_{\text{out}}) - i_L(V_{dd} - v_{\text{out}})}.
\]
(5.4)

5.2.2 Delay Minimization

To quicken the rising and falling transitions, the magnitudes of the net currents in the denominators of the integrals should be made large. However, this implies making \( i_L \) large for rising transitions and small for falling transitions. Clearly, both cannot be done simultaneously unless the load is an active device as it is for complementary logic. Usually, digital circuits cannot be designed to take advantages of differences in the speed of rising and falling transitions. Designers must design circuits based on the slower of the two. Equations 5.3 and 5.4 show that small changes in \( i_L \) will produce changes in \( T_{\text{rise}} \) and \( T_{\text{fall}} \) of opposite sign. Since \( T_{\text{rise}} \) and \( T_{\text{fall}} \) change in opposition, the maximum of the two will be minimized when \( T_{\text{rise}} = T_{\text{fall}} \).

These concepts are displayed graphically in Figure 5.4. The two types of transitions are represented by the different slants of the hatching. The right leaning hatch marks denote the region where the driver current is greater than the load current. This is the case when the node capacitance is discharging and the output voltage is decreasing. The left leaning hatch marks denote the region where the driver current is less than the load current. This corresponds to the excess load current charging up the node capacitance. The load line serves to divide the EFET driver characteristic into these two regions. If the load line moves up, the discharging region shrinks and the charging region grows, indicating that the rising transition is being hastened at the expense of the falling transition. This would benefit the overall speed if the rising transition was slower than the falling one.

5.2.3 Origin of Power vs. Speed Tradeoff

It is important to note that the the integrals in equations 5.3 and 5.4 involve the inverse of the net currents. The value of the delays will be dominated by regions where the load and driver currents are nearly the same. For example, if the peak current of the diode load is too close to the maximum driver current, the falling transition will
Figure 5.4: Partitioning of available driver currents into two regions by the load line. The current in the upper hatched section of the characteristic corresponds to the discharging of the node capacitance as seen at the left. The charging current for rising output transitions is represented by the hatched region below the load line.

slow down when the output voltage goes through this region. One point at which the currents are necessarily small for TD loads is in the low output state where only the valley current flows through the load. While small valley currents decrease static current or power dissipation, they also decrease the inverter’s speed, as would be expected for a standard power vs. speed tradeoff.

5.3 Modeling of Optimized Loads

Since the parameters of DBRTDs can be easily varied, it is desirable to know the optimum load characteristics for inverters. These were determined by simulating inverter performance for different load characteristics. To simplify the problem, only
piecewise linear models for tunnel diode characteristics of the form shown in Figure 5.7 were used. Also, only the shape of the load characteristic is important, so the calculations were done in a normalized fashion using the EFET driver characteristic of equation 5.2. Performance and power dissipation were expected to depend on the valley current, so the optimization was done for various valley currents to determine a power vs. speed tradeoff curve. The valley voltage was fixed at $0.9V_H$. Thus only peak voltage and peak current were allowed to vary. The current was considered linear from the origin to the peak, and from the peak to the valley. Beyond the valley voltage the current was taken as constant at the valley current. The power vs. speed tradeoff was also calculated for DFET loads that had the same value of $W/L$ as the EFET drivers. Larger $W/L$ ratios improve the performance of the DFET loads by effectively making their scaled characteristics less rounded, but it is a reasonable constraint that the physical size of the DFET loads not be larger than the EFET drivers. Different static current dissipations were obtained by varying the threshold voltage of the DFETs used as loads.

5.3.1 Gate Delay Calculation for Inverter Chains

Realistically, the inputs to the gates will not instantaneously change as was assumed for equations 5.3 and 5.4. The inputs to the gates will most likely be derived from similar gates. For this reason, the optimization was based on calculated delay times for chains of inverters with a particular load line. The node equation 5.1 was solved iteratively by using one $v_{\text{out}}(t)$ as the input to the next stage until waveforms had the same shape, i.e. $v_{\text{out},i+2}(t) = v_{\text{out},i}(t + \tau)$, where $i$ indexes the inverter or iteration. The difference in the times between when the input and output of an inverter crossed the midpoint, $(V_H + V_L)/2$, was taken as the gate delay time. If the output was rising, it is considered the rising delay time, $T_{\text{rise}}$, and if the output is falling, it is considered the falling delay time, $T_{\text{fall}}$. Calculated waveforms and an indication of these delay times are shown in Figure 5.5.
Figure 5.5: Calculated waveforms for subsequent gates of a chain of inverters. Tunnel diode characteristics were optimized to minimize the gate delay times measured between the midpoints of the characteristics as shown.

5.3.2 Optimum Performance and Load Characteristics

The TD load was considered optimum when \( T_{d,\text{max}} = \max(T_{d,\text{rise}}, T_{d,\text{fall}}) \) was minimized. In most cases this resulted in equal rising and falling delay times. For the DFET load the rising delay time always exceeded the falling delay time because the DFET characteristic did not partition the EFET driver current equally. This unequal partition was partially a consequence of the assumed restriction that the DFET gate width was less than the EFET gate width. If this restriction were removed, the DFET load would perform slightly better, mostly for load configurations with large current or power dissipation. The RTD load could partition the driver current equally because of the additional freedom of adjusting the peak position. The results are plotted in Figure 5.6 as the normalized static power dissipation versus the normalized maximum...
gate delay for both the DBRTD and the DFET load.

Figure 5.6: Average static current is plotted as a function of the maximum gate delay in a chain of inverters using enhancement MESFET drivers and RTD loads (dark box) or D-FET loads (open box). Different points were obtained by varying the load characteristic.

The power vs. speed tradeoff for inverters with TD loads is slightly better than that for inverters with DFET loads, especially at low currents and correspondingly longer gate delays. In this regime, the nonmonotonic shape of the TD load line is most beneficial, since the peak current can be several times higher than the DFET load current at the same voltage. For shorter gate delays and larger static currents, the performance curves of the two loads converge. Here the peak-to-valley current ratio is decreasing, as indicated by the labels in the figure, so the TD characteristic is approaching the shape of the DFET characteristic. Typically, E/D FET logic is designed to work in this shorter gate delay region where there is only a slight benefit due to the TD load shape in normalized terms. However, the RTD can have a smaller capacitance and area than a D-FET with the same current and might be easier to
Figure 5.7: Specific optimum models for TD load characteristics at different points along the static current vs. gate delay tradeoff curve represented in Figure 5.6. The characteristics correspond to normalized gate delays of (a) 3.1, (b) 2.0, (c) 1.5, and (d) 1.2.

fabricate. Specific optimum TD characteristics for different points along the power vs. speed curve are shown in Figure 5.7.

5.4 Inverter Design and Fabrication

The transfer characteristic of a TDFL inverter gate depends on the current-voltage characteristic of the tunnel diode, which in turn depends on the layers of the RTD. The RTD's characteristic must be designed on a voltage scale that is compatible with the logic swing of the inverter. The logic swing for enhancement MESFET drivers is limited to the turn on voltage of the MESFET gate, approximately 0.7 volts for GaAs devices. A small supply voltage, $V_{dd} = 0.5$ volts, was chosen to prevent large gate currents and to decrease power dissipation. Accordingly, the RTD's valley voltage was
limited to 0.5 volts, and while the peak voltage needed to be a small fraction of this to maximize the voltage swing. The RTD layers were designed to have 100Å spacers, 30Å AlAs barriers, and a 60Å well that was GaAs for one sample and In$_{0.1}$Ga$_{0.9}$As for the other sample. The InGaAs well was used to lower the resonant level in the well and thus reduce the peak voltage. The anode was heavily doped to reduce the depletion voltage required to support the E-field for resonance, reducing the overall voltage.

The combined layers for the MESFET followed by those for the RTD were grown by molecular beam epitaxy. The layers, as listed in Table 5.1, consisted of a confinement layer for the backside of the MESFET, an enhancement MESFET channel, a source/drain contact layer for the MESFET which also served to contact the cathode or bottom of the RTD mesa, the RTD layers, and finally a RTD anode contact layer. Figure 5.8 is a schematic cross-section of the integrated RTD load and MESFET driver. The processing of the upper DBRTD layers was similar to the general mesa isolation procedure described in section 3.1.3. A second mesa isolated the MESFET on the lower layers, and ohmic contacts were made to source and drain regions. A gate was made between the source and drain contacts by opening a window in the oxide, etching a recess down to the enhancement MESFET channel, and depositing Ti/Pt/Au gate metal. A photograph of one of the inverters is reproduced in Figure 5.9. It shows the small diode mesa in the middle of the drain contact metallization above the gate and source of the MESFET driver. Current flows vertically through the DBRTD, so it occupies a much smaller area than the MESFET driver, or any other planar devices that could be used as a load, such as a DFET. In functioning as an inverter, the source would be grounded and the metal line attached to the top of the diode mesa would be connected to the supply voltage. Inputs would be applied to the gate and the output would be taken from the drain metallization and applied to subsequent stages.

5.5 Measured Results

Figure 5.10 shows measured RTD characteristics as load lines superimposed on MESFET characteristics and the resulting transfer characteristics of the inverter.
<table>
<thead>
<tr>
<th>Layer</th>
<th>Composition ($N_d$ [cm$^{-3}$])</th>
<th>Thickness [Å]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact</td>
<td>n+ - GaAs ($5 \times 10^{18}$)</td>
<td>2000</td>
</tr>
<tr>
<td>Anode</td>
<td>n - GaAs ($1 \times 10^{18}$)</td>
<td>2000</td>
</tr>
<tr>
<td>Spacer</td>
<td>$\nu$ - GaAs</td>
<td>100</td>
</tr>
<tr>
<td>Barrier</td>
<td>$\nu$ - AlAs</td>
<td>30</td>
</tr>
<tr>
<td>Well</td>
<td>$\nu$ - GaAs or In$<em>{0.1}$Ga$</em>{0.9}$As</td>
<td>60</td>
</tr>
<tr>
<td>Barrier</td>
<td>$\nu$ - AlAs</td>
<td>30</td>
</tr>
<tr>
<td>Cathode</td>
<td>n - GaAs ($1 \times 10^{17}$)</td>
<td>1000</td>
</tr>
<tr>
<td>S/D contact</td>
<td>n+ - GaAs ($1 \times 10^{18}$)</td>
<td>1000</td>
</tr>
<tr>
<td>E-FET channel</td>
<td>n - GaAs ($1 \times 10^{17}$)</td>
<td>1000</td>
</tr>
<tr>
<td>Confinement layer</td>
<td>$\nu$ - Al$<em>{0.3}$Ga$</em>{0.7}$As</td>
<td>2000</td>
</tr>
<tr>
<td>Growth buffer</td>
<td>$\nu$ - GaAs</td>
<td>$\approx$ 2000</td>
</tr>
<tr>
<td>Substrate</td>
<td>Semi-insulating GaAs</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Combined MBE layers for MESFET and RTD.

Figure 5.8: A schematic cross-section of an integrated MESFET and RTD.
The monolithic combination in Figure 5.10a used the GaAs well DBRTD at room temperature. The transfer characteristic has a hysteretic loop centered about an input voltage of 0.3 volts. The measured current-voltage plot shows that the driver current characteristic with this gate voltage intersects the DBRTD load line three times. The drain voltages at the two stable intersections, $V_{ds} = 1.9$ and $3.3$ volts correspond to the output voltages in the middle of the hysteretic loop. The sloped DBRTD characteristic for small voltages across the load results in the approximately linear decrease in the high output voltage with increasing input voltage. If the peak load current occurred at higher drain voltages, the high state output voltage would be more constant. The valley current of this DBRTD is also rather high making the
Figure 5.10: Current-voltage characteristics of RTD loads and MESFET drivers and the resulting inverter transfer characteristics for (a) GaAs well RTD at room temperature, (b) GaAs well RTD immersed in liquid nitrogen, and (c) InGaAs well RTD immersed in liquid nitrogen. The MESFET characteristics were displayed for gate voltages from 0 to +0.5 volts in steps of 0.1 volt.
hysteresis relatively small and the low state output voltage higher than it would be otherwise.

Better DBRTD characteristics and correspondingly improved inverter transfer characteristics were obtained by operating the DBRTDs in liquid nitrogen. The characteristics of such a combination again using the GaAs well DBRTD, are shown in Figure 5.10b. The resulting transfer characteristic is appropriate for proper logic operation and there is a large hysteresis that improves the noise margins. The high and low state output voltages are about 0.41 volts and 0.07 volts with noise margins of 0.24 volts and 0.26 volts, respectively. The bistability could also be used for memory as discussed in section 5.1.2. An appropriate quiescent voltage would be 0.22 volts, and the bipolar writing pulses could then be as little as ±0.10 volts. The steep slope of the load line near the peak voltage and beyond the valley region make the high and low state output voltages more constant than they were for the first inverter where the load line had less steep features.

One problem with the GaAs well DBRTD is that its threshold of about 0.08 volts is a significant fraction of the supply voltage. As a result, the high output state voltage of the corresponding transfer curve is depressed. The InGaAs well RTD has a very small threshold and can be used to rectify this situation. The load and driver characteristics and inverter transfer function using an InGaAs well DBRTD are shown in Figure 5.10c. The RTD was again immersed in liquid nitrogen to obtain a better characteristic. Since the InGaAs well RTD has a linear tunneling current component at the origin, the high output level voltage has been increased to nearly $V_{dd}$. This particular load current is actually too large for the driver current; the hysteresis could be positioned closer to the middle by scaling it down. Since the peak current is only slightly less than the maximum MESFET driver current, a large gate or input voltage must be applied to switch the inverter to the low output state. The relatively high valley current also displaces the hysteresis to the right, allowing the inverter to switch from the lower to the higher bistable state with almost 0.3 volts on the gate. The large, flat valley current results in a sloped low output voltage for the inverter.

In general, steep load line characteristics result in transfer characteristics with flat output voltages that are more immune to small changes in the input voltage.
CHAPTER 5. TUNNEL DIODE FET LOGIC

The position of the hysteresis depends on the relative magnitude of the load and driver currents. Large peak and valley currents shift the hysteresis towards higher input voltages while smaller currents have the opposite tendency. The output voltages can be adjusted by changing the peak and valley voltage. DBRTDs with small peak voltages are desirable for maintaining high output voltages that are a large fraction of the supply voltage. Large, broad, and low DBRTD valley characteristics give small low state output voltages and correspondingly reduce static power dissipation. DBRTD valleys that end in a sharp upturn in the current at slightly less than the supply voltage give more constant low state output voltages but also increase static power dissipation. Thus the optimum shape of DBRTD valleys depends somewhat on whether large, square hysteretic transfer characteristics or low static current dissipation is more important.

5.6 Conclusions

The possibility of monolithically integrating MESFETs and DBRTDs, as well as the experimental dependence of the transfer characteristics on well composition were demonstrated by these experiments. To some extent, the optimum load characteristics depend on the desired transfer characteristics as well as the relative importance of speed and power dissipation. The particular experimental diode characteristics studied here had shortcomings, and it is possible to greatly improve them with better layer designs. One problem was that the peak current densities were not large enough. Both the peak current density and the peak-to-valley ratio could be improved by using thinner AlAs barriers.

A more serious problem is controlling the characteristics of the EFET drivers in such a nonplanar process. The gate recess was effectively controlled by a timed etch through more than 6000 Å of material and needed to be accurate to within ±50 Å for reasonable threshold voltage control. Because of variations in the etch depth, the threshold voltage varied widely and only isolated chains of inverters functioned properly. These difficulties could be diminished by using selective etches and reducing the diode mesa height or employing planarization layers. Another possibility being pursued at Honeywell is growing the DBRTD structure on undoped material. The
MESFETs are then made with implantations after a diode mesa isolation has exposed the undoped layers. They have shown that DBRTD structures can withstand high temperature anneals required for implant activation[94].

While MESFETs were used in this study, MODFETs could be used with a similar process. The AlGaAs gate layer could be used to stop a selective planarization etch that would improve recess uniformity. Other devices might form useful integrated combinations with other types of transistors, especially vertical devices like the heterojunction bipolar transistor. Vertical devices with similar current densities to the DBRTD would allow their sizes to be more closely matched. Combinations of DBRTDs and lasers have been investigated for use as bistable optical components.
Chapter 6
Effects of Cathode Wells

Derivation and calculation of the transmission spectrum for the double barrier structure are relatively straightforward; however, the details of the interaction of the resonance with the electrons supplied from the electrodes that generates the current-voltage characteristics are complicated and subtle. The use of large spacer layers and other modifications to the electrode doping make the band profile far different than the simple model assumed by Tsu and Esaki[3]. The literature that has been published on effects of varying the spacer layer doping profiles has attributed the differences to reduced impurity scattering[29, 95]. These conclusions are suspicious in view of recent experiments which show little degradation due to intentional doping of the wells[28] or barriers[96, 97].

It seems more plausible that the electrode doping profile has modified the electrode potential profile. Of particular interest is the formation of a two-dimensional electron gas in the cathode accumulation layer. Prior to this study, there was no known work on the effects of heterostructure in the cathode. Instead, attention focused on the effects of different barrier and well configurations[98, 32, 99, 100]. The motivation of this work was to determine if electrode heterostructure, particularly cathode structure, affected the current-voltage characteristics, and if so, could it be used as a beneficial design parameter. The answer to both questions, based on these experiments, is emphatically yes.

6.1 Band Profiles

The models of band profiles and effective supply functions have evolved to more sophisticated forms since the first calculation of DBRTD current by Tsu and Esaki[3]. They assumed that all of the potential difference of the electrodes was dropped across the barrier and well structure, giving the band profile shown in Figure 6.1 by the dot-dashed line. The problems of a discontinuous E-field in that model are clear. The electric field across the barriers must be screened out over some finite distance by
net charges in the electrodes. In DBRTDs with n-type electrodes, this is done by the space charge of donor ions in the depleted portion of the anode and by accumulated electrons in the cathode[101].

![Graph showing conduction band diagrams](image)

Figure 6.1: Conduction band diagrams for calculated resonant biases. The dot-dashed curve is the profile used for Tsu and Esaki's original analysis. It has been offset for clarity. More realistically computed band profiles are shown for diodes with (solid line) and without (dashed line) a quantum well in the cathode.

The electrons accumulate in the spacer region and the notch next to the cathode barrier. The details of the cathode potential profile depend on the electric field and the cathode doping profile. If the spacer is large, it can actually deplete the edge of the doped region of the cathode [102]. In this case, the band will bend upwards in the spacer before coming back down and forming the notch next to the barrier as shown by the dashed curve in Figure 6.1. This bulge in the spacer can block lower energy electrons, effectively forming a third barrier. Electrons may either coherently
tunnel through all three barriers or thermalize into the notch next to the barriers. For sufficiently large spacers and E-fields, the lowest states in the notch are quantized.

The motivation for including cathode quantum wells described in this chapter is to enhance the two-dimensional nature of the accumulated electrons in the cathode next to the barriers. This is expected to emphasize phenomena involving the accumulation layer as an alternate transport mechanism. A profile of the conduction band with a quantum well in it is shown by the solid curve in Figure 6.1.

6.2 Sample Design and Fabrication

Six similar DBRTD epitaxial structures were made by molecular beam epitaxy in a modified Varian Gen II system. The layers grown for each structure on Si-doped GaAs substrates were a doped growth buffer, a doped GaAs cathode, an undoped GaAs spacer, an optional undoped In$_{0.15}$Ga$_{0.85}$As quantum well, a double barrier structure, an undoped GaAs spacer, a doped GaAs anode, and the nonalloyed InGaAs contact described in section 3.1.2. All samples were designed to have identical 20 Å AlAs barriers separated by a 45 Å GaAs well. The cathode and anode silicon doping concentrations were $2 \times 10^{18}$ cm$^{-3}$ except for the last 100 Å of the cathode next to the undoped cathode spacer, where the doping was reduced to $1 \times 10^{17}$ cm$^{-3}$. The alloy composition of the strained InGaAs quantum wells was determined to be 15% InAs based on microprobe analysis of a 1 μm thick InGaAs layer grown under the same conditions.

Because impurities would accumulate during interruptions long enough to change the substrate temperature, the entire structure except for the nonalloyed contact was grown continuously with a substrate thermocouple reading of 600 °C which was thought to correspond to an actual temperature of about 520 °C[103, 36]. This temperature is a compromise between the optimal temperature for AlGaAs and InGaAs growth. The temperature falls in a range reported to give planar interfaces between layers of AlAs and InGaAs of this composition[104]. The final InGaAs contact layer was grown at at 450 °C.

Special migration enhanced epitaxy (MEE) techniques were used at critical interfaces and within the AlAs barriers to improve the interface quality. Prior to the
growth of the cathode quantum well and the barriers, approximately one monolayer of GaAs was grown with the Ga and As fluxes separated. Figure 6.2a shows the specific shutter timing used. The entire AlAs barriers were grown with MEE, and Figure 6.2b shows the shutter timing that was used there. The quality of the interfaces generated by this technique could not be monitored in-situ with RHEED because of the rotating substrate on the Gen II machines. The As/Ga flux ratio of 3 may have been too high to allow improved surface mobility of the group III atoms. It is suspected that the technique was neither significantly beneficial nor detrimental to the interface quality based on an additional experiment. The experiment consisted of growing three identical DBRTD and quantum well photoluminescence structures (1) without any interruption, (2) with 30 second interruptions at the interfaces, and (3) using the partial MEE techniques described above. The measured characteristics and spectra were indistinguishable. While all of the DBRTD characteristics were rather poor, all of the GaAs well photoluminescence peaks were quite narrow, having 4 meV FWHM linewidths.

The size of the undoped spacer layers and strained InGaAs quantum well were different for each of the six samples. Two of the samples had a large 500 Å anode spacer layer and a 25 Å cathode spacer layer. The first, sample A, had a 45 Å InGaAs quantum well, while the second, sample B, was used as a control sample that had no InGaAs quantum well. The remaining four samples all had 100 Å anode spacers, while the total thickness of the cathode spacer and InGaAs quantum well was kept constant at 70 Å. The four samples differed in the division of the 70 Å of undoped material between the GaAs spacer and InGaAs quantum well. The respective thicknesses of the GaAs and InGaAs were 70 Å and 0 Å for sample C-0, 45 Å and 25 Å for sample C-25, 25 Å and 45 Å for sample C-45, and 5 Å and 65 Å for sample C-65. The layers for each device are summarized in Tables 6.1 and 6.2. Various size devices were fabricated using the ion implantation isolation process described in section 3.1.4. Circular dots of 5000 Å thick titanium/gold were used to contact the nonalloyed InGaAs contact structure and mask the implant. Other 50 µm diameter titanium/gold dots were then evaporated over the smaller devices to serve as bonding and probing pads.

Photoluminescence (PL) measurements were used to check the InGaAs quantum
Figure 6.2: MEE shutter timing cycles used to grow (a) single smoothing monolayers of GaAs and (b) multiple monolayers of AlAs barriers.
CHAPTER 6. EFFECTS OF CATHODE WELLS

<table>
<thead>
<tr>
<th>Layer</th>
<th>Composition (N_d [cm^{-3}])</th>
<th>Thickness [Å]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact</td>
<td>n+ - InGaAs (≈ 1 \times 10^{19})</td>
<td>500</td>
</tr>
<tr>
<td>Anode</td>
<td>n - GaAs (2 \times 10^{18})</td>
<td>2000</td>
</tr>
<tr>
<td>Spacer</td>
<td>ν - GaAs</td>
<td>(see Table 6.2)</td>
</tr>
<tr>
<td>Barrier</td>
<td>ν - AlAs</td>
<td>20</td>
</tr>
<tr>
<td>Well</td>
<td>ν - GaAs</td>
<td>45</td>
</tr>
<tr>
<td>Barrier</td>
<td>ν - AlAs</td>
<td>20</td>
</tr>
<tr>
<td>Spacer</td>
<td>(see Table 6.2)</td>
<td></td>
</tr>
<tr>
<td>Inner Cathode</td>
<td>n - GaAs (20 to 1 \times 10^{17})</td>
<td>1000</td>
</tr>
<tr>
<td>Outer Cathode</td>
<td>n - GaAs (2 \times 10^{18})</td>
<td>1000</td>
</tr>
<tr>
<td>Growth buffer</td>
<td>n - GaAs (2 \times 10^{18}\text{?})</td>
<td>≈ 2000</td>
</tr>
<tr>
<td>Substrate</td>
<td>n+ - GaAs</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: Layer structure of DBRTD cathode well samples. Table 6.2 contains the details of the anode and cathode spacer layer for each sample.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Spacer Layer Thickness [Å]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cathode</td>
</tr>
<tr>
<td>A</td>
<td>45</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>C-0</td>
<td>0</td>
</tr>
<tr>
<td>C-25</td>
<td>25</td>
</tr>
<tr>
<td>C-45</td>
<td>45</td>
</tr>
<tr>
<td>C-65</td>
<td>65</td>
</tr>
</tbody>
</table>

Table 6.2: Thicknesses of undoped layers in anode and cathode spacers of DBRTD cathode well samples. The anode spacer is entirely GaAs while in the cathode spacer the In_{0.1}Ga_{0.9}As layer is adjacent to the AlAs cathode barrier and the GaAs layer is adjacent to the doped inner cathode.
wells. The photoluminescence spectra of the DBRTD samples C-25, C-45, and C-65 were measured near 77 K using approximately 500 W/cm² of excitation from an Ar-ion laser. The PL spectra are presented in Figure 6.3. Calculated peak photoluminescence energies for square quantum wells of strained In₀.₁Ga₀.₉As on GaAs are indicated by arrows in the figure. The bandgap of the strained In₀.₁Ga₀.₉As was computed using Asai and Oe's method[19]. The agreement is worst for the 65 Å well. The difference between the computed and measured values might be due to a broadening of the well or a relaxation of its strain. Since the well width is still well below the critical limit calculated by Matthews and Blakeslee[18], it is unlikely that the well layer is relaxed.
CHAPTER 6. EFFECTS OF CATHODE WELLS

6.3 Measured Results

The current-voltage characteristics of diodes formed in this manner from sample A, with a cathode quantum well, and sample B, the control without a cathode well, were measured at room temperature using a HP4145A. Small area \( \approx 3\mu m \) diameter, high impedance devices were packaged in S4 pill packages and stabilized as discussed in section 3.2.3.

Typical characteristics of both samples are shown in Figure 6.4a for forward bias and Figure 6.4b for reverse bias. Forward bias is defined as the top or epitaxial side having a more positive voltage than the substrate side. In the forward direction, where electrons are injected from the cathode, the diode with the cathode quantum well has a much higher peak current and a higher peak voltage than the control sample. While the cathode well diode's valley current is also higher, it has increased by a much smaller ratio than the peak current. The general shape of the characteristics' valley regions are similar but the peak shapes, and particularly the NDC regions are quite different. The control sample's characteristic is smooth and single valued in contrast to the cathode well sample's current, which switches abruptly at the peak with approximately 7 mV of hysteresis. This extrinsic switching is consistent with an intrinsic negative differential resistance of smaller magnitude than the parasitic series resistance of the device and measurement setup.

To confirm that the differences in the samples' characteristics are associated with the quantum well in the cathode rather than a modification of the double AlAs barriers due to the preceding growth of a strained InGaAs layer, the reverse characteristics of diodes from samples A and B were compared. Under reverse bias, electrons are injected from the anode layers grown after the barriers and the undoped region between the doped cathode and AlAs barriers is depleted. Since, in both samples, the electrons are accumulated in identical structures and injected at energies much greater than the conduction band energy of the depleted InGaAs quantum well, it is expected that the presence of the InGaAs well in sample A will make little difference. Thus samples A and B are expected to have similar reverse characteristics. This is indeed the case, as shown in Figure 6.4b. The reverse peak currents of the samples with and without a cathode well are the same. The peak voltage is slightly larger with the
Figure 6.4: Room temperature, stabilized current-voltage characteristics of $\approx 3\mu m$ diameter diodes with a cathode well (sample A) and without it (sample B) in (a) forward bias and (b) reverse bias.
cathode well because of the increased length of the undoped spacer layer.

![Graph](image)

Figure 6.5: Boiling nitrogen temperature characteristics of $\approx 4\mu$m diameter diodes from sample A with a cathode well (solid line) and from sample B without it (dashed line).

Current-voltage characteristics were also measured with samples A and B immersed in liquid nitrogen and are shown in Figure 6.5. The characteristics have the same general features as the ones measured at room temperature; the forward peak current of sample A is almost three times as large as that for sample B in the forward direction, and the peak currents and voltages are similar in the reverse direction. The initial increase in current prior to the peak voltage is somewhat nonlinear, having a small bump that is often seen in low temperature characteristics of DBRTDs. The bump does occur at approximately the same current where sample B’s current peaks. The correspondence of these features might be interpreted by attributing the main peak in sample A’s characteristic to tunneling from some additional, lower
level induced by the cathode well not present in the other sample. The bump would then be attributed to tunneling alignments similar to those that produce the peak in sample B's current.

![Graph showing current-voltage characteristics](image_url)

Figure 6.6: Room temperature current-voltage characteristics of ≈ 5μm diameter diodes with different sizes of cathode well. The characteristics of these lower impedance devices were not stabilized.

The other four samples, C-0, C-25, C-45, and C-65, were used to measure the variation of the current-voltage characteristic with cathode well size. The room temperature characteristics of 5 μm diameter diodes made from these samples are shown in Figure 6.6. In the forward direction, the peak voltage increases as the cathode well size increases. While the peak current increases for quantum well sizes of 25 Å and 45 Å, there is no further increase in peak current as the cathode well is broadened to 65 Å. Since these last two configurations have equal peak currents, the shift in the peak voltage cannot be attributed to series voltage drop. Because the valley current
increases with cathode quantum well size for all samples, the sample with the 45 Å cathode well, C-45, has the largest peak-to-valley current ratio of any of the samples. Diodes from this sample had peak to valley ratios of up to 5.2 at room temperature. Both the devices with 45 Å and 65 Å cathode wells had peak current densities of approximately $10^5$ A/cm². Again, the similarity of the reverse characteristics of the samples with different cathode wells indicates that the differences in forward characteristics can be attributed to the cathode structure instead of differences in the double barrier structure.

6.4 Analysis: Possible Mechanisms

While our understanding of the mechanisms responsible for changes in the characteristics is not complete, some general comments can be made. The changes in the tunneling current can be attributed to a combination of factors, including changes in material quality, changes in the electron supply function, or changes in the transmission spectrum.

Growth of the InGaAs quantum well may affect the quality of the subsequent barriers. However, because of the lattice mismatch, it would be expected to degrade the quality of the AlAs barriers and decrease the peak-to-valley current ratio in contrast to the measurements.

If the electrons conserve their transverse wavevector across the entire structure, including the spacer regions, then the supply function remains three-dimensional, and the transmission spectrum is modified by the cathode well and any changes in potential induced in the surrounding region by the accumulated charge in the well. Analogous to a HEMT channel, the cathode well depletes carriers from the adjacent doped cathode region. The increased potential of the region just outside the cathode well causes it to act as another barrier, effectively forming a triple barrier resonant tunnel diode. However, the resulting transmission peak with the cathode well should then be narrower and the peak current smaller than when the well is not present. This result is in opposition to the experimental observations.

The other factor controlling the tunneling current, in addition to the transmission spectrum, is the supply function. If there are quantized levels in the cathode at
CHAPTER 6. EFFECTS OF CATHODE WELLS

energies lower than the continuum conduction band states, the supply function due to these states is two-dimensional. The filling of the quantized levels from the higher energy continuum states requires that the transverse wavevector of electrons from the bulk or three-dimensional cathode be broken by either elastic or inelastic scattering. If the electrons did not rapidly tunnel out of the quantized states, they would form a two-dimensional electron gas that could supply electrons to the remainder of the tunneling structure with a resonant transmission spectrum determined by the double barriers.

The current would be expected to quickly increase when enough bias was applied to align the resonant level of the double barriers with the quantized level in the cathode. Placing InGaAs quantum wells in the cathode reduces the energy of the quantized levels and separates them from the continuum. As the well deepens or widens, the lowest confinement energy decreases, and higher electric fields across the double barrier structure are required to align the resonant level with the quantized level in the cathode well. Eventually, widening the well will decrease the separation of the quantized levels below the thermal energy, making them again appear as a continuum.

Further analysis is required to determine if the observed changes in the DBRTD characteristics are consistent with tunneling from quantized levels. It is not clear whether the current due to quantized levels should be greater or smaller than that due to continuum states. If quantizing the electron supply function does result in increased tunneling current, then widening the well too much would eventually begin to decrease the tunneling current as the quantized levels merged into a continuum. In the limit of an infinitely wide well the supply function would again become threedimensional. It is interesting to note that the tunneling current given by equation 2.13 is proportional to the effective mass, so the electron flux supplied by an InGaAs electrode is smaller than that supplied by a GaAs electrode.

The increase in peak voltage agrees with a decrease in the energy of the quantized level, but it can also be attributed to differences in electrostatics between the structures. Since the negative charge creating the E-field across the barriers and undoped spacer is centered closer to the barriers when the cathode well is present, less voltage
is dropped in front of the barriers for a given E-field. The E-field and total voltage drop must increase to align the transmission resonance with the conduction band minimum in the bulk cathode. The increased E-field also increases the transmission asymmetry for barriers of equal thickness and composition.

While these arguments are far from definitive, they indicate that the effects of the cathode well on material quality or the transmission spectrum should reduce the tunneling current or degrade the peak-to-valley ratio. Since these conclusions are in opposition to the experimental results, the increased peak current is most probably a consequence of quantization effects in the cathode. The cathode well enhances the quantization of the accumulation layer, and the two-dimensional injection gives higher tunneling currents based on the experimental observations and this interpretation.

6.5 Conclusions

The increased peak current density of the devices with cathode wells compared to those without would decrease their charging time and increase their speed if their capacitance did not increase proportionally with the peak current. Since the capacitance is largely controlled by the configuration of the anode spacer, it is expected to be unchanged by the introduction of the cathode well. The shape of the characteristics of the diodes with cathode wells reported here are more appropriate for use in switching or pulse forming circuits than for microwave applications that require a broad region of negative differential conductance to obtain high power. The abrupt transition from the peak current to near the valley current is beneficial for applications where a diode with a resistive load is being switched by a small voltage [30]. The increased peak voltage of the cathode well devices increases their power dissipation which is disadvantageous for all applications.

This work has demonstrated the dependence of resonant tunnel diode characteristics on the band structure outside the barriers, particularly the cathode region. Thus multiple alloy and doping layers in the cathode are another means of producing devices tailored for particular purposes. They might, for example, be used to engineer a quantized electron supply function that would give controllable negative differential conductance characteristics.
A full understanding of the experimental results requires additional theoretical treatment and analysis. Experiments that could differentiate between tunneling from two-dimensional and three-dimensional states would be immensely useful. Magneto-tunneling studies in strong magnetic fields parallel to the tunneling currents are one such experiment that should be performed on these samples.
Chapter 7
Conclusions

7.1 Summary

The work described in this dissertation has addressed multiple aspects of research on double barrier resonant tunnel diodes. The primary areas of contribution concerning DBRTD applications are associated with the evaluation of DBRTDs for integrated circuit logic applications in general, and tunnel diode FET logic specifically and more extensively. Experimental work on DBRTDs with cathode wells has contributed data on the important effects of cathode profiles, as well as demonstrated that cathode heterostructures are an additional beneficial design freedom that has produced dramatic improvements. In the course of this work, knowledge about fabrication and measurement techniques has been generated that is useful outside of the scope of the work presented here.

7.1.1 Comparison of DBRTDs and JTDs

In attempts to affirm that DBRTDs may be beneficially applied in digital circuits, their superiority over JTDs with regard to integrated circuit applications has been ascertained. JTDs suffer from degradation due to recombination enhanced reactions. These problems were worse for early methods of fabrication, but the 80% yield of JTDs made with modern planar techniques still precludes their use in even small scale integrated circuits. In contrast, DBRTDs appear quite robust. Demonstrated uniformities of 20% variation in DBRTD peak current across a 2" wafer are appropriate for integrated circuits. DBRTDs have also been made that have an order of magnitude higher peak current density and an order of magnitude lower capacitance than JTDs giving a factor of 100 improvement in tunnel diode speed index. DBRTDs are recognized as having characteristics that can be controllably varied much more readily than JTDs. It is difficult to simultaneously achieve high peak current and low peak voltage with DBRTDs which would be of particular advantage in TDFL.
Also, broader NDC voltage ranges would increase the available power in microwave applications.

7.1.2 Tunnel Diode FET Logic

Motivated by this generally improved performance, tunnel diode FET logic has been investigated as an application for DBRTDs. TDFL offers the possibility of a more advantageous speed vs. power tradeoff compared to enhancement/depletion-mode FET logic, reduced load size, improved noise margins, and operation of single inverter stages as memory elements. Optimized tunnel diode load lines were determined for various points along the speed vs. power curve, based on a normalized analysis of inverter chains. Because of the additional freedom in tailoring the tunnel diode characteristics to equally partition the available driver current, the resulting inverter performance is superior to that for DFET loads, especially for relatively long gate delays or correspondingly low static power.

Monolithic TDFL inverters based on integrated DBRTDs and MESFETs were fabricated. The inverter transfer characteristics were appropriate for stable logic operation. They exhibited bistability which served to increase noise margins and would allow a single inverter to function as a static memory cell if the input was controlled appropriately. The fabrication highlighted problems of threshold voltage control of the EFET drivers due to the nonplanarity of the structures and deep recess etch.

7.1.3 DBRTDs with Cathode Wells

DBRTDs were fabricated with quantum wells in the cathode adjacent to the barriers to enhance the accumulation layer and thus accentuate the role it played in transport. The characteristics for devices with various sizes of these cathode wells were measured. The peak and valley current and peak voltage all increased with cathode well size. Initially the peak current increased proportionately faster than the valley current, increasing the peak-to-valley current ratio. However, for the largest wells, the peak current did not increase but remained constant. The level of peak current saturation level was 3.5 times greater than the peak current of the device.
without a cathode well. The intermediate sized quantum well device had a peak current density of $10^5$ A/cm$^2$ and peak-to-valley current ratios of up to 5.2 at room temperature.

In addition to producing devices with excellent characteristics for switching applications, the experiment provides important data for analyzing transport mechanisms in DBRTDs other than the common coherent models. Solutions of the band profile for the structure indicate that the quantum well depletes the adjacent region, causing it to appear as a third barrier to electrons. However, it is unlikely that the device is acting as a fully coherent triple barrier RTD because the observed increase in current is inconsistent with a decreased transmission resonance width that would result from additional total barrier width. The remaining most plausible explanation is that electrons are tunneling from quantized levels in the cathode well. These wells would be filled by processes not considered in coherent models.

### 7.1.4 Measurement and Fabrication Techniques

Some of the fabrication and measurement techniques developed in conjunction with this work are appropriate for DBRTD research in general. Standard microlithography techniques have been applied to DBRTD fabrication using both mesa and ion implantation isolation. The stabilization techniques used to prevent oscillations in DC measurements are important for measuring NDC devices.

### 7.2 Suggestions for Future Work

There are a number of promising investigations that are suggested by this work. Ones mentioned here represent the different aspects of using new devices in practical applications. Beneficial steps in this development process include improving the device's performance, modifying applications to take advantage of the special attributes of the devices, and producing accurate, predictive models to move device engineering out of the strictly experimental realm.
7.2.1 Device Engineering

DBRTDs can be further improved, both in general and by tuning their characteristics for particular applications. Obtaining satisfactory current densities at low peak voltages is the key to efficient operation of DBRTDs in most applications and has been elusive until very recently[31]. This requirement is necessitated by the limitations on power dissipation in dense integrated circuits and is desirable for improved power efficiencies in analog circuits.

Microwave power amplification efficiencies are also increased by widening the NDC region. This must be done to make the DBRTD produce power-bandwidth products that are competitive with those produced by Gunn and IMPATT devices. For switching applications, the opposite extreme of NDC region breadth is required. It is desirable for the current to abruptly jump between a peak and valley value. Since the DBRTDs with cathode wells studied in this dissertation have this property, they would be excellent for switching circuits although their peak voltage should be reduced. If large voltage changes are required in switching circuits, then the valley current should be low over an extended voltage range.

Multi-valued logic and memory and other applications based on characteristics with multiple peaks are most likely to be implemented with stacked structures. It is quite difficult to stack more than a few devices without producing undesirable hysteresis in the characteristics, so ways of preventing this should be investigated. It may be possible to diminish the hysteresis by allowing small interactions between separate devices via the electric field in depletion layers.

The issues of reliable and repeatable manufacture of appropriate DBRTDs need to be more thoroughly and quantitatively explored. Accelerated aging tests would give important data about expected lifetimes as well as critical processes causing diode degradation. Susceptibility to radiation should also be examined because of the probable environments for DBRTDs in microwave applications. Run to run variations as well as individual wafer uniformities should be characterized based on statistically accurate sample sizes. Comparative studies of these properties in MBE versus MOCVD grown wafers would be interesting because of the different advantages and disadvantages of each method of crystal growth.
CHAPTER 7. CONCLUSIONS

7.2.2 DBRTD Applications

Tunnel diode FET logic needs to be further explored with more sophisticated processing technology to make devices that are competitive with other state-of-the-art logic families. Problems of nonplanarity and threshold variation could be overcome by using very short electrodes and non-alloyed contact layers. HEMTs or other heterojunction transistors would perform better than MESFETs as drivers. There would be little additional cost for such devices as compared to the MESFET because the DBRTD already requires precise epitaxial growth.

High density static memory is one application that could benefit from the bistable characteristics of single inverters that are relatively compact. TDFL inverters might prove to be an attractive means of implementing memories based on DBRTDs. The memory elements would be quite compact with only a single FET and a small DBRTD required for each memory cell. The small power dissipation and large noise margins of TDFL inverters would also be beneficial for dense static memories using small supply voltages. While the output of TDFL inverters used as memory elements is compatible with binary logic, appropriate circuits need to be developed that could generate the bipolar pulses required to write to the memory cells. It might be possible to reduce the memory cell size further by fabricating the DBRTDs underneath the FETs, or by using vertical FETs or bipolar transistors stacked on top of the DBRTDs, although these schemes would complicate circuit fabrication and could add parasitics.

If the DBRTD is ever to make contributions outside of small niche markets, research efforts should be aimed at important applications. Microwave detection is the largest market for junction tunnel diodes, yet research on applying DBRTDs to this application remains minimal, at best. Most existing DBRTD characteristics are relatively symmetric about the origin, making them inappropriate for use as detectors. Ideas for detector structures based on tunnel barriers between dissimilar electrode superlattices have already been presented in this dissertation. Alternatively, electrodes of different bulk semiconductors might suffice. It seems likely that material systems with more diverse properties than those of just AlGaAs and GaAs would be beneficial. The critical issues to address are the frequency response of a detector and its nonlinearity at zero bias.
CHAPTER 7. CONCLUSIONS

Much room remains for imaginative new analog and digital applications built on the DBRTD's strengths of speed and potentially high functional density. Based on Shockley's observation that two-terminal devices can be built more compactly than three-terminal devices, it might be possible to obtain higher circuit densities with diodes than with transistors. Proposed superconducting circuits, excluding SQUIDs, bear examination because of the functional similarity of Josephson junctions and tunnel diodes, both of which are two-terminal devices that are bistable when appropriately loaded. Rectifying Schottky diodes seem good candidates for providing input isolation for tunnel diode circuits. If it were desirable to use unipolar logic with bistable elements, photogenerated carriers created by a broadcast optical clock signal might prove an efficient way to reset latched circuits before each cycle.

7.2.3 Cathode Well Research and Applications

Existing DBRTD models based solely on coherent tunneling are clearly inadequate for predicting realistic device characteristics. Inclusion of dissipative mechanisms is an obvious first step towards improving these models. DBRTDs with modified cathode structures, such as cathode wells, that accentuate transport via these mechanisms appear to be good vehicles for gathering the data and insight necessary for better models. There are several additional experiments described below that could be done on DBRTDs with cathode wells.

The dependence of current-voltage characteristics on a full range of experimental parameters should be established. Both structural and environmental variations are possible. It would be quite interesting to examine larger cathode wells. This would help to determine whether strong quantization was related to the observed changes in the characteristics. Varying the composition of the InGaAs cathode well would also provide additional useful data. For deeper cathode wells, well defined multiple quantized states could exist that might cause multiple peaks in the characteristic. The difference in energy levels can be measured by varying the temperature to obtain an activation energy. Since the resonant energy can be accurately calculated with this technique, it could be determined whether electrons were being supplied from the bulk cathode or the lower energy quantum well. Strong magnetic fields produce
further quantization of two-dimensional electrons to one dimension. Thus magneto-
tunneling experiments could also shed light on the role of reduced dimensionality of
the supplied electrons.

The structure of DBRTDs with cathode wells also suggests several interesting de-
vice possibilities. The unique feature of the structure is the smaller bandgap quantum
well in the immediate vicinity of the tunneling structure. This well can be accessed
independently from the rest of the structure by optical and perhaps electrical means.
Electrical access would be relatively difficult; it is similar to contacting the well of
resonant tunneling transistors. Optically, carriers could be excited from the quan-
tum well to interact resonantly or non-resonantly with the barriers, depending on
the optical wavelength. The wavelength for resonant interaction depends on the bias
across the device. As the bias increases, the quantized levels of the cathode well and
the central well between the barriers move closer. This description indicates that the
device might be used as a tunable longwavelength detector.

7.3 Outlook for Resonant Tunneling Devices

It seems appropriate to end this dissertation with an opinion about the future of
resonant tunneling device applications. In some specific applications, mostly involving
microwave communications, the success of RTDs is imminent. This statement is based
on their excellent performance at present, and the knowledge that they can be further
improved. There is also a high probability that they will be used in non-cryogenic
pulse and sampling measurement circuits, because they will be the fastest switching
devices available at room temperature. These uses of a few RTDs at a time look
promising, but the significant use of RTDs as components in large integrated circuits
is dubious.

RTDs will have to perform significantly better than transistors to overcome their
inertia, but heterojunction bipolar transistors are only about a factor of two slower
than RTDs, and will likely remain so. It is also probable that tunneling transistors,
with a large parasitic base resistance and capacitance, will not outperform bipolar or
hot electron transistors. It seems the only circumstances where the RTD would be
preferred over more traditional transistors is one in which RTDs would be functionally
more complex or denser than transistors.

It is probable that power dissipation restrictions would not permit individually isolated RTDs to be spatially denser than transistors, so individual RTDs must function on a higher order than single transistors. The terminal characteristics of each device will have to represent a more complex function than simple linear amplification or binary logic. It has been pointed out that single RTDs can be used to implement exclusive-or gates, and similarly that stacks of devices can function as parity generators. However, it is hard to imagine very many general classes of problems that are more compactly expressed in exclusive-ors than nands and nors.

There are general problems with implementing complex functions with individual devices. Independently adjusting the function of different devices by internal changes would be almost impossible, so either all devices would have the same complicated function, or they would require external biasing components. These components, interconnections, and contacts to the device would probably nullify any original size advantage. Also maintaining multiple logic levels and other overhead associated with complex logic schemes is much more difficult than maintaining their binary counterparts.

These arguments can be summarized by stating that either the entire application should be done in one complicated device or it should be built up out of just a few types of simple devices. The knowledge and tools associated with transistor device and circuit design are massive, giving the traditional approach a great deal of momentum. The other approach, of accomplishing the complete function with one device, makes it ironically similar to the other microwave and pulse applications that are likely to succeed; they are all effectively based on discrete devices. The conclusion is that integrated RTD circuits will not be widely applied. Quantum interference devices will not likely displace the transistor as an electronic building block until the devices are linked with light or tunneling electrons instead of wires.
Appendix A
Transfer Matrix Calculations

A.1 Introduction

Transfer matrix methods are used in many problems in electromagnetics and quantum mechanics where waves propagate perpendicular to slabs of different homogeneous materials. The method uses matrices to relate the complex amplitude of the forward and reverse propagating waves in one region to those in another region, and is applicable when the wave equation is linear. It is an especially convenient method for analyzing propagation in structures composed of multiple elements with individual transfer matrices. The matrices of each element can be multiplied in order to find the overall transfer matrix.

The method is readily applied to quantum mechanical tunneling structures, like the double barrier resonant tunnel diode (DBRTD). The mathematics is most simple when each section of the structure has a constant potential so that there is a single pair of forward and reverse plane waves with a constant wavevector in that section. Then only the matrices for connecting wavefunction solutions at interfaces are required to calculate the wave propagation, and thus the tunneling probability, through the entire structure.

This appendix discusses quantum mechanical transfer matrix method calculations for rectangular potential profiles with the aim of obtaining useful approximate analytic forms for DBRTD tunneling spectra. Initially, the matrices for flat potential regions and their interfaces are derived and then used to calculate the tunneling probability for a single tunnel barrier. With this background, the transfer matrix for a double barrier structure is constructed from those for single barriers. Based on the composite transfer matrix, the resonant transmission spectra for DBRTDs is evaluated and discussed.

Analytical tunneling expressions that are derived for rectangular double barriers are useful for gaining insight but are typically inadequate for accurate quantitative results on DBRTD characteristics. Electric fields due to doping gradients and applied
bias result in sloped potential profiles both in the double barrier structure and the electrodes. Profiles that more closely match the real ones give better results but make analytic calculations intractable, forcing numerical computations. Many such computations are based on transfer matrix methods using matrices with numerical elements.

One method for computing transmission spectra of arbitrary potential profiles uses many small rectangular steps to approximate the true profile. The transfer matrices have the same form as those derived in the first part of this appendix. The origins of this method have not been completely researched, but at Stanford this method was first used on DBRTDs by Ed Wolak[26] who extended a program written by Alex Harwit[105] who refers to similar computations by Miller[106]. Ed Wolak and Byung Park have also found that it is important to use non-parabolic dispersion relations to calculate the attenuation of tunneling barriers[26], and this feature is included in a tunneling program used at Stanford. The form of the attenuation constant is based on work by Schulman and McGill[33] and effectively results in an elliptic rather than parabolic dispersion relationship. This formulation will be reviewed here.

In this first approach to numerically computing tunneling spectra, many small steps are required to accurately represent a complicated potential profile, and complex exponentials need to be evaluated to propagate the solution across each step. Since this must be done for a wide range of energies based on a new potential profile at each bias, computing the current-voltage characteristics of a DBRTD is somewhat numerically expensive. It is useful to investigate more efficient means of computing transmission spectra for DBRTDs.

Another approach is to estimate the profile with linear rather than constant potential segments. This permits the use of fewer regions but complicates individual transfer matrices. The solutions of the Schrödinger equation for linear potentials are Airy functions. Some workers have implemented transfer matrix computations based on numerically evaluating these Airy functions. However, only certain combinations of Airy functions appear in the transfer matrices. When these are expanded appropriately, the transfer matrix elements reduce to polynomial series that are more easily evaluated than Airy functions. Computation of tunneling spectra based on
piecewise-linear approximations is discussed as a second method. The calculation of the simplified transfer matrix elements is an original contribution that is presented in detail.

An alternative to transfer matrix methods for calculating wavefunctions is to solve a discrete or finite difference form of the Schrödinger equation. Bill Frensley has endorsed this approach as being more efficient than transfer matrix methods. To evaluate this claim, this appendix concludes with a comparison of tunneling spectra computations using transfer matrix and finite difference methods. Numerical accuracy and efficiency are the criteria upon which the different methods are evaluated.

While much of the mathematical expression presented here is original, many of the underlying concepts have been derived or influenced by the published work of others. Tsu and Esaki are perhaps responsible for popularizing the use of transfer matrix methods for calculating the transmission spectra of finite superlattices as it was used in their paper[3]. Prior to their work, Kane had described double-barrier tunneling using transfer matrices[23]. One popular paper that interprets tunneling spectra of DBRTDs based on transfer matrix calculations is by Ricco and Azbel[25]. Some of the work here resembles that published by Price[27], although his papers were read after most of this work was completed.

A.2 Transfer Matrices for Rectangular Potentials

To the extent that electron motion in solids can be decomposed into propagating plane waves with piecewise-constant wavevectors perpendicular to the layers in the solid, coherent electron transport can be analyzed with traditional transfer matrix methods. If the left and right traveling plane waves on one side of a structure can be linked to the outgoing plane wave on the other side of the structure, the probability for tunneling through the structure is readily calculated. Transfer matrices are a convenient means of linking the solutions in different regions.
APPENDIX A. TRANSFER MATRIX CALCULATIONS

A.2.1 Wavefunction Solutions

This analysis uses the effective mass and envelop function approximations, and both the effective mass, \( m^* \), and the potential, \( V \), are assumed to be piecewise constant:

\[
m^*(x) = \begin{cases} 
  m_0^* & \text{for } x < x_1 \\
  m_j^* & \text{for } x_j < x < x_{j+1} \\
  m_n^* & \text{for } x_n < x
\end{cases}
\]  

(A.1)

and

\[
V(x) = \begin{cases} 
  V_0 & \text{for } x < x_1 \\
  V_i & \text{for } x_j < x < x_{j+1} \\
  V_n & \text{for } x_n < x
\end{cases} 
\]  

(A.2)

The translational invariance of the structure in the directions perpendicular to the \( x \)-axis allows the wavefunction to be factored. Only the portion that depends on \( x \) will be discussed below. The wavefunction, \( \psi(x) \), for an electron with energy \( E \) is governed by the time-independent Schrodinger equation,

\[
\left( -\frac{\hbar^2}{2m^*} \frac{d^2}{dx^2} + V(x) \right) \psi(x) = E \psi(x),
\]  

(A.3)

where \( \hbar \) is Planck's reduced constant. The solution for the region \( j \) between \( x_j \) and \( x_{j+1} \) is

\[
\psi = C_j^+ e^{+\gamma_j x} + C_j^- e^{-\gamma_j x},
\]  

(A.4)

where the leading coefficients, \( C_j^+ \) and \( C_j^- \), are determined from boundary conditions, and the plane wave propagation constant in the region is

\[
\gamma_j = \sqrt{\frac{2m^*_j(V_j - E)}{\hbar^2}}.
\]  

(A.5)

A.2.2 Boundary Matching Conditions

The coefficients, which are the amplitudes of plane waves if \( E > V \), of two adjoining regions can be related by requiring that the wavefunction and the probability flux, \( (\hbar/m^*) \psi d\psi/dx \), be continuous across the interface between the two regions. Since
the wavefunction is continuous, continuity of the flux can be reduced to continuity of
the probability flux normalized to the wavefunction,

$$\left( \frac{\hbar}{im^*} \right) \frac{d\psi}{dx} = \nu \left( C^+ e^{+\gamma x} - C^- e^{-\gamma x} \right), \quad (A.6)$$

where equation A.4 has been explicitly substituted for $\psi$, and $\nu \equiv \hbar \gamma / im^*$. In the case
that $E > V$ so that the wave is propagating, $\gamma = ik$, and $\nu = \hbar k / m^*$ is the effective
velocity of the wavefunction. If $E < V$, then $\gamma = \alpha$, the attenuation constant, and the
effective velocity, $\nu = \hbar \alpha / im^*$, is imaginary. Relationships between the wavefunction
amplitudes in adjoining regions are simply expressed by forming a column vector of
the wavefunction and normalized probability flux, which are both linear in the
wavefunction amplitudes:

$$\begin{bmatrix} \psi(x) \\ \frac{\hbar}{im^*} \psi'(x) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \nu & -\nu \end{bmatrix} \begin{bmatrix} e^{\gamma x} & 0 \\ 0 & e^{-\gamma x} \end{bmatrix} \begin{bmatrix} C^+ \\ C^- \end{bmatrix}. \quad (A.7)$$

Since the quantities in the vector on the left hand side must be identical on each side
of an interface between two regions, the right hand side of the equation evaluated in
each of the two regions at the interface must also be equal.

### A.2.3 Transfer Matrices

If the interface is at $x_j$ so that the coefficients in the two regions are indexed by
$j - 1$ and $j$, the resulting matrix equation can be manipulated into the form

$$\begin{bmatrix} C^{j-1}_+ \\ C^{j-1}_- \end{bmatrix} = X_j \begin{bmatrix} C^+_j \\ C^-_j \end{bmatrix}, \quad (A.8)$$

where $X_j$ is a transfer matrix. It transforms the vector of coefficients for region $j$
into a vector of coefficients for region $j - 1$. An equation relating the coefficients in
regions $j$ and $j - 2$ can be obtained by left multiplying equation A.8 by $X_{j-1}$. The
relationship between the coefficients of the plane waves in the unbounded end regions
can be expressed in terms of a product of transfer matrices across each region:

$$\begin{bmatrix} C^+_0 \\ C^-_0 \end{bmatrix} = \prod_{j=1}^{n} X_j \begin{bmatrix} C^+_n \\ C^-_n \end{bmatrix}. \quad (A.9)$$
Thus if the amplitude of the two plane waves in one region are known, the others can be determined. The product of all the transfer matrices is called the global transfer matrix, \( X_G = \prod_{j=1}^{n} X_j \), because it is the transfer matrix for the entire structure.

The direction of operation of the transfer matrix defined by equation A.8 is to transform coefficients for regions of larger \( x \) values into coefficients for regions of smaller \( x \) values. The transformation could equally as well have been defined to operate in the opposite direction. If \( x \) is considered to increase to the right, then the choice made here is to express the coefficients on the left hand side region in terms of those for the right hand side region. While a left to right transformation may seem more natural for waves incident from the left, it will be seen that the right to left definition simplifies the calculation of transmission probabilities.

### A.2.4 Transfer Matrices for Abrupt Steps and Constant Potentials

An explicit form for \( X_j \) is found by equating the right hand side of equation A.7 for regions \( j \) and \( j - 1 \) evaluated at \( x = x_j \). The manipulation of this equality into the form of equation A.8 gives the transfer matrix as

\[
X_j = \begin{bmatrix}
  e^{\gamma_{j-1} x_j} & 0 \\
  0 & e^{-\gamma_{j-1} x_j}
\end{bmatrix}^{-1}
\begin{bmatrix}
  1 & 1 \\
  v_{j-1} & -v_{j-1}
\end{bmatrix}^{-1}
\begin{bmatrix}
  1 & 1 \\
  v_j & -v_j
\end{bmatrix}
\begin{bmatrix}
  e^{\gamma_j x_j} & 0 \\
  0 & e^{-\gamma_j x_j}
\end{bmatrix}
\]

\[
= \frac{1}{2}
\begin{bmatrix}
  e^{-\gamma_{j-1} x_j} & 0 \\
  0 & e^{\gamma_{j-1} x_j}
\end{bmatrix}
\begin{bmatrix}
  1 + \Delta_j & 1 - \Delta_j \\
  v_{j-1} & -v_{j-1}
\end{bmatrix}
\begin{bmatrix}
  1 + \Delta_j & 1 - \Delta_j \\
  v_j & -v_j
\end{bmatrix}
\begin{bmatrix}
  1 & 1 \\
  0 & e^{-\gamma_j x_j}
\end{bmatrix}
\]

\[
X_j = \Phi(-\gamma_{j-1} x_j) \Theta(\Delta_j) \Phi(\gamma_j x_j).
\] (A.10)

where \( \Delta_j \equiv v_{j-1} / v_j \), \( \Theta(\Delta) \equiv \frac{1}{2} \begin{bmatrix}
  1 + \Delta & 1 - \Delta \\
  1 - \Delta & 1 + \Delta
\end{bmatrix} \), and \( \Phi(\phi) \equiv \begin{bmatrix}
  e^\phi & 0 \\
  0 & e^{-\phi}
\end{bmatrix} \). If \( x_j = 0 \), then \( X_j = \Theta(\Delta_j) \). Thus, \( \Theta(\Delta_j) \) is the transfer matrix for a potential step at the origin. For potential steps at \( x_j \neq 0 \), \( \Phi(-\gamma_{j-1} x_j) \) and \( \Phi(\gamma_j x_j) \) serve to transform \( \Theta(\Delta_j) \) to a location away from the origin by accounting for changes in phase of propagating waves or changes in amplitude of attenuated waves. In either case, these matrices will be referred to as phase shift matrices. Note that if \( v_j = v_{j-1}, \Delta_j = 1 \) and \( \Theta(\Delta_j) \) is the identity matrix.
The symbol for the ratio of the effective velocities is chosen as $\Delta$ because it is the determinant of the $\Theta(\Delta)$ matrix. Since $\det(\Phi(\phi)) = 1$ for arbitrary $\phi$, $\Delta_j$ is also the determinant for $X_j$. If the waves in both regions linked by the transfer matrix are propagating, the $\Delta$ will be real, as it will be if the waves are attenuated in both regions. However, if the waves are propagating in one region but attenuated in the other region, then $\Delta$ will be imaginary.

To find the explicit relationship between waves in the end regions of the structure, equation A.10 can be substituted for $X_j$ in equation A.9. The result is

$$
\begin{bmatrix}
C_0^+ \\
C_0^-
\end{bmatrix} = \left( \prod_{j=1}^{n} \Phi(-\gamma_{j-1} x_j) \Theta(\Delta_j) \Phi(\gamma_j x_j) \right) \begin{bmatrix}
C_n^+ \\
C_n^-
\end{bmatrix}.
$$

(A.11)

The global transmission matrix is $X_G = \prod_{j=1}^{n} \Phi(-\gamma_{j-1} x_j) \Theta(\Delta_j) \Phi(\gamma_j x_j)$, and its determinant is

$$
\Delta_G = \prod_{j=1}^{n} \Delta_j = \frac{v_n}{v_0}.
$$

(A.12)

It is useful to manipulate equation A.11 further by left multiplying by $\Phi(\gamma_0 x_0)$, where $x_0$ can be chosen arbitrarily. Removing $\Phi(\gamma_n x_n)$ from the product, and regrouping the phase shift matrices gives

$$
\Phi(\gamma_0 x_0) \begin{bmatrix}
C_0^+ \\
C_0^-
\end{bmatrix} = \left( \prod_{j=1}^{n} \Phi(\gamma_{j-1} x_{j-1}) \Phi(-\gamma_{j-1} x_j) \Theta(\Delta_j) \right) \Phi(\gamma_n x_n) \begin{bmatrix}
C_n^+ \\
C_n^-
\end{bmatrix}
$$

$$
= \left( \prod_{j=1}^{n} \Phi(-\gamma_{j-1} L_{j-1}) \Theta(\Delta_j) \right) \Phi(\gamma_n x_n) \begin{bmatrix}
C_n^+ \\
C_n^-
\end{bmatrix},
$$

(A.13)

where $L_j \equiv x_{j+1} - x_j$ and products of phase shift matrices obey exponentiation rules so that $\Phi(\phi_1) \Phi(\phi_2) = \Phi(\phi_1 + \phi_2)$. The last form of equation A.13 makes it clear that the complex amplitudes of wavefunctions are changed by reflection and transmission at steps according to $\Theta(\Delta)$ and that they are modified by propagation to the left across constant potential layers by $\Phi(-\gamma L)$.

### A.2.5 Transmission Coefficients

The transmission and reflection of a plane wave by a structure are readily calculated from the transfer matrix linking the plane wave solutions in the unbounded
regions of constant potential at each end of the structure. If the coefficients of the wave in the unbounded left hand region are denoted with an \( L \) subscript, and similarly \( R \) for the right hand region, a transfer matrix equation linking them is

\[
\begin{bmatrix}
C_L^+ \\
C_L^-
\end{bmatrix} = X \begin{bmatrix}
C_R^+ \\
C_R^-
\end{bmatrix}.
\]

Initially there is a right going wave with an amplitude \( C_L^+ \) in the left hand region. A portion of it is transmitted to a positive going wave on the right with an amplitude \( C_R^+ \) and a portion is reflected by the structure as a left going wave with amplitude \( C_L^- \). There is no cause for the wave impinging from the left to generate a left going wave in the right hand region, so \( C_R^- = 0 \). The fraction of the wave amplitude that is transmitted from left to right is the ratio of \( C_R^+ \) and \( C_L^+ \). From equation A.14,

\[
C_L^+ = X_{11} C_R^+ + X_{12} C_R^- \\
= X_{11} C_R^+,
\]

so the transmitted fraction is

\[
t \equiv 
\frac{C_R^+}{C_L^+} = \frac{1}{X_{11}}.
\]

If the transmission matrix had been defined in the opposite sense, i.e. transferring coefficients from left to right, then some algebra would have led to \( t = 1/X_{11}^{-1} \). Since the probability flux is \( (\hbar/i m^*) \psi d\psi/dx \), the transmission ratio of probability flux is the ratio of the wavefunction probabilities weighted by their effective velocities,

\[
T \equiv \frac{v_R(C_R^+ C_R^+)}{v_L(C_L^+ C_L^+)} = \Delta t^* t = \frac{\Delta}{|X_{11}|^2},
\]

where \( \Delta \) is for the entire structure: \( \Delta = \det(X) = v_R/v_L \).

The reflection coefficient can also be calculated from elements of the transfer matrix. The amplitude of the reflected wave in the left hand region is

\[
C_L^- = X_{21} C_R^+ + X_{22} C_R^- \\
= X_{21} C_R^+.
\]
Since $C^+_{R}$ is related to the amplitude of the wave impinging on the structure, $C^+_{L}$, by equation A.15, equation A.18 can be solved for the fraction of the wave amplitude that is reflected. The wave amplitude reflection ratio is

$$ r \equiv \frac{C^-_{L}}{C^+_{L}} = \frac{X_{21}}{X_{11}}, \quad (A.19) $$

which gives a probability flux reflection coefficient of

$$ R \equiv r^* r = \frac{|X_{21}|^2}{|X_{11}|^2}. \quad (A.20) $$

This relation could have also been deduced from the requirement that $T + R = 1$ and the fact that

$$ \Delta = \text{det}(X) = |X_{11}|^2 - |X_{21}|^2. \quad (A.21) $$

The last equality of this equation implies that $X_{11} = X^*_{22}$ and $X_{12} = X^*_{21}$.

Since the transmission probabilities are based on the magnitude of the wavefunction coefficients as written in equation A.17, the relative phase of these coefficients is not important. Thus, if the electron energy is greater than the potential in the end regions so that the wavefunction is propagating there, the phase shift matrices that are adjacent to the coefficient vectors in equation A.13 can be dropped for purposes of calculating the transmission probability. The resulting global transfer matrix between allowed regions is

$$ X_G = \Theta(\Delta_1) \left( \prod_{j=2}^{n} \Theta(-\gamma_{j-1} L_{j-1}) \Theta(\Delta_j) \right), \quad (A.22) $$

which includes only phase shift matrices for regions between potential steps.

### A.3 Elliptic Dispersion Formula for Tunneling in Bandgaps

Electron energies in semiconductors are often close enough to the band extrema to make parabolic approximations of the dispersion relations adequate for many purposes. This amounts to effective mass theory, with the curvature of the bands being given by the value of the effective mass. In this case the wave propagation constants can be calculated with the normal parabolic relation in equation A.5. In tunneling experiments, however, it is possible to inject electrons into layers at energies a
substantial fraction of the bandgap below the conduction band minimum. The imaginary dispersion relation in bandgaps cannot continue to grow parabolically since it must eventually bend back and become zero at the edge of the band of matching symmetry. Schulman and McGill[33] have proposed a dispersion relation that matches the expected effective mass approximations near the band extrema and has a maximum attenuation in the middle of the bandgap. The resulting dispersion relation effectively has two elliptic branches and is reviewed here.

The proposed dispersion relation for attenuation factors in the bandgap is

\[ E = E_0 \pm \frac{\hbar^2 \alpha_0}{m^*_\pm} \sqrt{\alpha_0^2 - \alpha^2}, \] (A.23)

where \( \alpha \) is an attenuation factor that takes on a maximum value of \( \alpha_0 \) at an energy \( E_0 \) in the bandgap, \( m^*_+ = m^*_c \) is the conduction band effective mass which is used for \( E > E_0 \), and \( m^*_- = m^*_v \) is the effective mass for the valence band of the same symmetry as the conduction band minimum. In III-V semiconductors, this is the light-hole band. When \( E < E_0 \), the valence band effective mass, \( m^*_- \), should be used. The form of equation A.23 guarantees that the dispersion relation is of the expected parabolic form near the band edges. The effective mass is

\[ \frac{1}{m^*} = \mp \frac{1}{\hbar^2} \frac{dE}{d\alpha^2} = \frac{1}{m^*_\pm} \frac{\alpha_0}{\sqrt{\alpha_0^2 - \alpha^2}} \left(1 + \frac{\alpha^2}{\alpha_0^2 - \alpha^2}\right), \] (A.24)

so that \( m^* = m^*_\pm \) at the band edge where \( \alpha = 0 \).

The maximum attenuation factor is chosen so that the two branches of the dispersion relation join smoothly at \( E_0 \) where \( d\alpha/dE = 0 \) for both branches. Equation A.23 gives \( E_\pm = E_0 \pm \hbar^2 \alpha_0^2/m^*_c \) for the upper branch when \( \alpha = 0 \), and \( E_v = E_0 - \hbar^2 \alpha_0^2/m^*_v \) for the lower branch when \( \alpha = 0 \). Thus the bandgap is \( E_g = E_+ - E_v = \hbar^2 \alpha_0/(1/m^*_c + 1/m^*_v) \), or alternatively, the maximum attenuation factor is

\[ \alpha_0 = \sqrt{\frac{E_g}{\frac{1}{m^*_c} + \frac{1}{m^*_v}}}. \] (A.25)

\( E_0 \) can be calculated in terms of \( \alpha_0 \) from equation A.23 at either \( E = E_+ \) or \( E = E_v \) where \( \alpha = 0 \). With reference to the conduction band,

\[ E_0 = E_+ - \frac{\hbar^2 \alpha_0^2}{m^*_v}. \] (A.26)
Usually, \( m_c^* < m_v^* \), so \( E_0 \) lies closer to the valence band than to the conduction band.

When rearranged and squared, equation A.23 becomes

\[
\alpha^2 + \left( \frac{m^*}{\hbar^2 \alpha_0} \right)^2 (E - E_0)^2 = \alpha_0^2. \tag{A.27}
\]

In this form, the elliptic nature of the dispersion relation in the \( \alpha-E \) plane is apparent. Normally, it is desirable to compute \( \alpha \) from a known \( E \). When equation A.23 is solved for \( \alpha \) it can be expressed in a particularly convenient form in terms of the attenuation factor calculated from the parabolic dispersion relation. If the parabolic attenuation factor is

\[
\alpha_p = \begin{cases} 
\sqrt{2m_c^*(E_0 - E)/\hbar^2} & \text{for } E_0 > E > E_0 \\
\sqrt{2m_v^*(E - E_0)/\hbar^2} & \text{for } E_0 > E > E_v
\end{cases}, \tag{A.28}
\]

then the elliptic attenuation factor is

\[
\alpha(E) = \alpha_p \sqrt{1 - \left( \frac{\alpha_p}{2\alpha_0} \right)^2}. \tag{A.29}
\]

The elliptic attenuation factor is always smaller than the parabolic attenuation factor. At \( E = E_0, \alpha = \alpha_p / \sqrt{2} \). The attenuation factor calculated from the elliptic dispersion relation also changes less in response to changes in energy than does the attenuation factor calculated from the parabolic relation.

### A.4 Tunneling Through Single Barriers

One of the simplest tunneling structures to analyze is a single rectangular barrier. Even in this modest problem, transfer matrices are an apt tool for managing the simultaneous equations that must be solved at the two interfaces at \( x_1 \) and \( x_2 \) between the three regions. The example is instructive and is also one of the components of the double barrier structure to be analyzed later.

The global transfer matrix for the single barrier structure neglecting the end region phase shift matrices as written out in equation A.22 is

\[
\mathbf{X} = \Theta(\Delta_1) \Phi(-\gamma_b L_b) \Theta(\Delta_2)
\]

\[
= \frac{1}{2} \begin{bmatrix} 1 + \Delta_1 & 1 - \Delta_1 \\ 1 - \Delta_1 & 1 + \Delta_1 \end{bmatrix} \begin{bmatrix} e^{-\gamma_b L_b} & 0 \\ 0 & e^{\gamma_b L_b} \end{bmatrix} \frac{1}{2} \begin{bmatrix} 1 + \Delta_2 & 1 - \Delta_2 \\ 1 - \Delta_2 & 1 + \Delta_2 \end{bmatrix}
\]
where $\gamma_b$ is the propagation constant in the barrier region which is of length $L_b$. $F^+$ and $F^-$ are the ratios of the amplitudes of the waves going in the same direction on each side of the barrier. They are respectively for positive and negative going waves. $G^-$ is the ratio between waves that are moving away from the barrier on each side, and $G^+$ is the ratio between the waves that are converging on the barrier on each side. These elements can be written for the general case as

$$F^+ = (1 + \Delta_1 \Delta_2) \cosh(\gamma_b L_b) - (\Delta_1 + \Delta_2) \sinh(\gamma_b L_b), \quad (A.31)$$

$$G^- = (1 - \Delta_1 \Delta_2) \cosh(\gamma_b L_b) + (\Delta_1 - \Delta_2) \sinh(\gamma_b L_b), \quad (A.32)$$

$$F^- = (1 + \Delta_1 \Delta_2) \cosh(\gamma_b L_b) + (\Delta_1 + \Delta_2) \sinh(\gamma_b L_b), \quad (A.33)$$

$$G^+ = (1 - \Delta_1 \Delta_2) \cosh(\gamma_b L_b) - (\Delta_1 - \Delta_2) \sinh(\gamma_b L_b). \quad (A.34)$$

Note that the sign associated with $\Delta_1$ changes between $F^+$ and $G^-$, and that the signs associated with both $\Delta_1$ and $\Delta_2$ change between $F^+$ and $F^-$ and again between $G^+$ and $G^-$. The case of propagating states on each side of a highly attenuating barrier is particularly important and allows the single barrier transfer matrix to be simplified. Since at both interfaces the states are propagating on one side and attenuated on the other, both $\Delta_1$ and $\Delta_2$ are imaginary. Thus the sign reversals discussed in the previous paragraph amount to conjugations of complex numbers: $(F^-)^* = F^+$ and $(G^-)^* = G^+$. The form of $X$ becomes

$$X = \begin{bmatrix} F & G^* \\ G & F^* \end{bmatrix}, \quad (A.35)$$

where $F = F^+ = (F^-)^*$ and $G = G^- = (G^+)^*$. If the barrier is highly attenuating, $e^{\gamma_b L_b} \gg 1 \gg e^{-\gamma_b L_b}$, so that the terms multiplied by $e^{-\gamma_b L_b}$ in equation A.30 can be dropped. The remaining terms approximate the matrix elements as

$$F = (1 - \Delta_1)(1 - \Delta_2) \frac{e^{\gamma_b L_b}}{4}, \quad (A.36)$$

$$G = (1 + \Delta_1)(1 - \Delta_2) \frac{e^{\gamma_b L_b}}{4}. \quad (A.37)$$
The other matrix elements are the complex conjugates of these quantities.

The amplitude reflection ratio for a wave impinging on the highly attenuating barrier from the left is

\[ r = \frac{G}{F} = \frac{1 - \Delta_1}{1 + \Delta_1}, \quad (A.38) \]

and the ratio of the transmitted to impinging amplitude is

\[ t = \frac{1}{F} = \frac{4e^{-\gamma_b L_b}}{(1 - \Delta_1)(1 - \Delta_2)}. \quad (A.39) \]

If the energy of the electron, \( E \), is not much greater than the potential in the regions on each side of the barrier but is much smaller than the potential of the barrier, \( V_b \), then

\[ |\Delta_1| = \left| \sqrt{\frac{(V_b - E)/m_b^*}{(V_0 - E)/m_0^*}} \right| \gg 1, \quad (A.40) \]

and

\[ |\Delta_2| = \left| \sqrt{\frac{(V_2 - E)/m_2^*}{(V_b - E)/m_b^*}} \right| \ll 1. \quad (A.41) \]

Thus equation A.39 gives the approximate transmission ratio as \( t = -4e^{-\gamma_b L_b}/\Delta_1 \) which includes a phase shift of \(-\pi/2\) since \( \Delta_1 \) is imaginary. The reflection coefficient in equation A.38 gives a phase shift of approximately \( \pi \) on reflection from the highly attenuating barrier. In order to get an estimate of \( r \) that doesn’t have unity magnitude, terms of order \( e^{-\gamma_b L_b} \) in equation A.30 need to be retained, although the magnitude of \( r \) can be easily calculated from \( |r| = \sqrt{1 - |t|^2} \).

A more accurate estimate of the phase shift due to reflection is, however, needed for later calculations of resonant energies in double barrier structures. The phase shift of the reflection coefficient is \( \arg(r) = \arg(G/F) = -\arg(FG^*) \), with

\[ \arg(FG^*) = \arg((1 - \Delta_1)(1 - \Delta_2)(1 + \Delta_1^*)(1 - \Delta_2^*)) \]

\[ = \arg((1 - 2\Delta_1 + \Delta_1^2)(1 - \Delta_2^2)) \]

\[ = \arctan \left( \frac{2|\Delta_1|}{1 + \Delta_1^2} \right) \]

\[ \approx \arctan \left( \frac{2}{-|\Delta_1|} \right) \]

\[ \approx \pi - \arctan \left( \frac{2}{|\Delta_1|} \right). \quad (A.42) \]
Note that even the exact form for the reflection phase shift is dependent only on the discontinuity at the near side of the barrier via $\Delta_1$. This is a consequence of the large attenuation of the barrier. It can be shown, by taking the inverse of $X$, that reflection from the opposite side of the barrier results in a phase shift related to $\arg(FG)$, which is worked out for the single rectangular barrier as

$$\begin{align*}
\arg(FG) &= \arg((1 - \Delta_1)(1 - \Delta_2)(1 + \Delta_1)(1 - \Delta_2)) \\
&= \arg((1 - 2\Delta_2 + \Delta_2^2)(1 - \Delta_2^2)) \\
&= \arctan \left( \frac{-2|\Delta_2|}{1 + \Delta_2^2} \right) \\
&\approx \arctan (-2|\Delta_2|) \\
&\approx -\arctan (2|\Delta_2|).
\end{align*}$$

(A.43)

Both $\arg(FG^*)$ and $\arg(FG)$ turn out to be small angles modulo $\pi$ if $|\Delta_1|$ and $|\Delta_2|$ are very small or very large. However, if the effective velocities and masses in the barrier and surrounding regions are similar, as they can be in semiconductor heterojunction structures, $|\Delta_1|$ and $|\Delta_2|$ can be $\sim 1$, making the phase shifts on reflection differ significantly from 0 or $\pi$.

### A.5 Tunneling Through Double Barriers

The transmission through a double barrier can be calculated by multiplying the appropriate transfer matrices for each potential step and intervening region in sequence. However, it is more constructive to group the matrices for barriers, using them to left and right multiply the phase shift matrix for the constant potential region between the barriers. If $\gamma = ik_w$ in the well of length $L_w$, the overall or global matrix can be written,

$$\begin{bmatrix}
F_G & G_G^* \\
G_G & F_G^*
\end{bmatrix} = \begin{bmatrix}
F_L & G_L^* \\
G_L & F_L^*
\end{bmatrix} \begin{bmatrix}
e^{-ik_wL_w} & 0 \\
0 & e^{ik_wL_w}\end{bmatrix} \begin{bmatrix}
F_R & G_R^* \\
G_R & F_R^*
\end{bmatrix},$$

(A.44)

where the outside phase shift matrices have been neglected. Multiplying out the matrices gives

$$F_G = F_L F_R e^{-ik_wL_w} + G_L G_R e^{ik_wL_w}.$$

(A.45)
Then, using \( T = \Delta / |F|^2 \) and \( R = |G|^2 / |F|^2 \), some algebra leads to the global transmission as

\[
T_G = \Delta_G \left\{ \frac{1 + R_L R_R}{T_L T_R / \Delta_L \Delta_R} + |F_L||G_L||F_R||G_R|e^{i\beta} + \text{c.c.} \right\}^{-1}
\]

\[
= \frac{T_L T_R}{2 - T_L - T_R + T_L T_R + 2 \sqrt{(1 - T_R)(1 - T_L)} \cos \beta},
\]

(A.46)

where c.c. denotes the complex conjugate of the previous term and \( \beta = 2k_w L_w + \phi_b \).

The sum of the phase shifts due to reflection from the barriers, \( \phi_b \), can be written using equations A.42 and A.43 as

\[
\phi_b \equiv -\arg(F_L G_L F_R G_R) = \arctan(2|\Delta_{w,L}|) + \arctan(2|\Delta_{w,R}|) - \pi,
\]

(A.47)

where \( \Delta_{w,L} = \sqrt{(v_w - E)/m_{s,L}} \) and \( \Delta_{w,R} = \sqrt{(v_{w,L} - E)/m_{s,R}} \).

In the limit that the transmission, \( T \), of each barrier is much less than one, \( T \ll 1 \), the square root may be expanded. Then

\[
T_G \approx \frac{T_L T_R}{(2 - T_L - T_R + T_L T_R)(1 + \cos(\beta)) - \frac{1}{4}(T_R + T_L)^2 \cos(\beta)}.
\]

(A.48)

Since the \( 2(1 + \cos(\beta)) \) term in the denominator is much larger than any of the other terms which all involve small transmissions, \( T_G \) is a maximum or at resonance when \( \cos(\beta) = -1 \). The resonance energy can be solved for using this condition which can be rewritten as

\[
2k_w L_w = (2n - 1)\pi - \phi_b
\]

\[
= 2n\pi - \arctan(2|\Delta_{w,L}|) + \arctan(2|\Delta_{w,R}|).
\]

(A.49)

Under the resonance condition,

\[
T_{G,\text{res}} \approx \frac{4T_L T_R}{(T_R + T_L)^2}
\]

\[
= \frac{4}{\left(\sqrt{T_R/T_L} + \sqrt{T_L/T_R}\right)^2}
\]

\[
= \cosh^{-2}(\tau),
\]

(A.50)

where \( \tau = \frac{1}{2} \ln(T_R/T_L) \). If \( T_R = T_L \) then \( T_{G,\text{res}} = 1 \). Note that for square barriers of equal height, \( T \sim e^{-2\alpha L_b} \) so that \( \tau \approx \alpha(L_{b,L} - L_{b,R}) \), where \( L_{b,L} \) and \( L_{b,R} \) are the
widths of the left and right barriers. In this case, it is not the relative amount of barrier width fluctuation, but the absolute amount that determines the reduction of \( T_{G,\text{res}} \) below unity.

The shape of the resonance can be examined by expanding \( \beta \) about its resonant value, \( \beta = (2n - 1)\pi + \delta \beta \), so \( \cos(\beta) = -1 + \frac{1}{2} \delta \beta^2 \). Using this value in equation (A.48) gives

\[
T_G \approx \left\{ \frac{2 - T_L - T_R + T_LT_R}{T_LT_R} \left( \frac{1}{2} \delta \beta^2 \right) - \frac{(T_R + T_L)^2}{4(T_LT_R)} \left( \frac{1}{2} \delta \beta^2 - 1 \right) \right\}^{-1}
\]

\[
\approx \left\{ \frac{1 - \bar{T}}{T_LT_R} \delta \beta^2 + \cosh^2(\tau) \right\}^{-1}
\]

(A.51)

as the transmission near the resonance, having neglected factors on the order of \( T^2 \) and using the average barrier transmission \( \bar{T} = (T_R + T_L)/2 \). Equation (A.51) shows that the transmission spectrum has a Lorentzian shape about the resonance. In order to calculate the width of the peak in the transmission spectrum in units of energy, \( \delta \beta \) must be related to small changes in energy by differentiating the expression for \( \beta \):

\[
\delta \beta = \frac{d}{dE}(2k_w L_w + \phi_b) \delta E = \frac{2m^*_w L_w}{\hbar^2 k_w} \delta E,
\]

(A.52)

where \( d\phi_b/dE \) has been assumed to be negligible compared to \( dk_w/dE \). Substituting \( k_w \) at resonance as given in equation (A.49) into this expression yields

\[
\delta \beta = \frac{4m^*_w L_w^2}{\hbar^2 ((2n - 1)\pi - \phi_b)} \delta E.
\]

(A.53)

A Lorentzian function of the general form \( f(x) = (a + bx^2)^{-1} \), has a full width at half the maximum (FWHM) of \( 2\sqrt{a/b} \) and a total integrated area of \( 2/\sqrt{ab} \) which is equal to the FWHM times its height, \( 1/a \). Thus, noting the parts of the expression in equation (A.51) corresponding to the \( a \) and \( b \) parameters of the general Lorentzian function, the FWHM of a double barrier structure's transmission resonances is

\[
\Delta E_{\text{FWHM}} = 2 \frac{\cosh(\tau)}{\sqrt{\frac{1 - \bar{T}}{T_LT_R}}} \left( \frac{\hbar^2 ((2n - 1)\pi - \phi_b)}{4m^*_w L_w^2} \right) = \frac{2\bar{T}}{\sqrt{1 - \bar{T}}} \left( \frac{\hbar^2 ((2n - 1)\pi - \phi_b)}{4m^*_w L_w^2} \right).
\]

(A.54)

The resonance width is inversely proportional to the square of the well width, and increases with increasing resonance index, \( n \). If \( \Delta E_{\text{FWHM}} \) is much narrower than
the supply function, the Lorentian peak can be treated as a delta function with a fixed area. The resonant flux through a double barrier structure is proportional to the area of the resonance. Multiplying the FWHM from equation A.54 by the peak transmission factor in equation A.50 gives the area of the transmission resonance peak as

$$A_{res} = \frac{2T_L T_R}{T \sqrt{1 - \frac{T}{T_L T_R}}} \left( \frac{\hbar^2 ((2n - 1) \pi - \phi_b)}{4m_w^* L_w^2} \right). \quad (A.55)$$

Approximately 80% of the area under a Lorentzian curve lies in a region twice as wide as the FWHM centered about the peak.

**A.6 Transfer Matrices for Piecewise Linear Potentials**

In the preceding sections, transfer matrices were constructed that accounted for potential steps and constant potential regions, namely $\Theta(\Delta)$ and $\Phi(\phi)$. These are sufficient for the rectangular potentials considered previously, but not for regions of varying potential. Since electric fields, and hence non-constant potential regions, are unavoidable in most electronic devices, it is desirable to generate transfer matrices that are appropriate for such regions. The simplest logical extension to the rectangular potential profiles considered before is to introduce regions of linear potential with constant, nonzero slopes or electric fields. Numerically expedient formulations for such transfer matrices are developed in this section.

**A.6.1 Airy Function Transfer Matrices**

As before, the wavefunction and normalized flux at the boundaries between adjacent regions must be equated, and a column vector of these quantities is convenient for handling the resulting simultaneous equations. In a linear potential region, say $V(x) = V_1 - Fx$ for $x_1 < x < x_2$, the wavefunctions are solutions of the Schrodinger equation with this potential energy, \{-\frac{\hbar^2}{2m^*} \frac{d^2}{dx^2} + (V_1 - Fx) - E\}\psi = 0. The equation is simplified by making a transformation to non-dimensional coordinates with $\xi = -\lambda(x + \epsilon)$ where $\lambda = (\frac{2mF}{\hbar^2})^{1/3}$ and $\epsilon = \frac{E - V_1}{F\lambda}$. Then the Schrodinger equation becomes $\frac{d^2\psi}{d\xi^2} - \xi \psi = 0$ with the solution

$$\psi = C^A Ai(\xi) + C^B Bi(\xi), \quad (A.56)$$
i.e. a linear combination of Airy functions. The column vector to be conserved across
interfaces can still be written as a linear combination of the coefficients of the two
types of solutions,
\[
\begin{bmatrix}
\psi(\xi(x)) \\
\frac{\lambda}{im} \psi'(\xi) \xi'(x)
\end{bmatrix} = A(\xi) \begin{bmatrix} C^A \\ C^B \end{bmatrix},
\]  
(A.57)

where
\[
A(\xi) \equiv \begin{bmatrix} Ai(\xi) & Bi(\xi) \\ \frac{\lambda}{im} Ai'(\xi) & \frac{\lambda}{im} Bi'(\xi) \end{bmatrix}.
\]  
(A.58)

Initially, consider a potential profile that is constant at \(V_0\) for \(x < x_1\) and constant
at \(V_2\) for \(x > x_2\). If \(V_1 - Fx_1 \neq V_0\) or \(V_1 - Fx_2 \neq V_2\), then there are steps in
the potential at the appropriate interface. Expressions of the form of equations A.7
and A.57 can be equated at both interfaces. At the left interface where \(x = x_1\) and
\(\xi = \xi_1 \equiv -\lambda(x_1 + \epsilon)\) the conservation conditions give
\[
\begin{bmatrix} 1 & 1 \\ v_1 & -v_1 \end{bmatrix} \begin{bmatrix} e^{ik_1 x_1} & 0 \\ 0 & e^{-ik_1 x_1} \end{bmatrix} \begin{bmatrix} C^+_1 \\ C^-_1 \end{bmatrix} = A(\xi_1) \begin{bmatrix} C^A \\ C^B \end{bmatrix}.
\]  
(A.59)

At the right interface there is a similar equation except that the variables are sub-
scribed with 2 rather than 1. Solving this equation at the second interface for the
column vector of Airy function coefficients and substituting this in the equation above
for the left interface after solving for the plane wave coefficients generates a transfer
matrix equation for the two plane wave regions,
\[
\begin{bmatrix} C^+_1 \\ C^-_1 \end{bmatrix} = X \begin{bmatrix} C^+_2 \\ C^-_2 \end{bmatrix},
\]
where
\[
X = \begin{bmatrix} e^{ik_1 x_1} & 0 \\ 0 & e^{-ik_1 x_1} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 \\ v_1 & -v_1 \end{bmatrix}^{-1} A(\xi_1) \]  
(A.60)
\[
\times A(\xi_2)^{-1} \begin{bmatrix} 1 & 1 \\ v_2 & -v_2 \end{bmatrix} \begin{bmatrix} e^{ik_2 x_2} & 0 \\ 0 & e^{-ik_2 x_2} \end{bmatrix}.
\]  
(A.61)

The determinate of the matrix of Airy functions is
\[
|A(\xi)| = \frac{-\lambda h}{im} [Ai(\xi)Bi'(\xi) - Ai'(\xi)Bi(\xi)]
\]
\[
= \frac{-\lambda h}{im} W\{Ai, Bi\}(\xi)
\]
\[
= \frac{-\lambda h}{im} \pi,
\]  
(A.62)
(A.63)
(A.64)
where \( W\{Ai, Bi\} \) is the Wronskian of the two types of Airy functions which equals \( \pi^{-1} \). The inverse of the Airy function matrix is

\[
A(\xi)^{-1} = \pi \begin{bmatrix}
Bi'(\xi) & \frac{-im}{-\lambda h}Bi(\xi) \\
-Ai'(\xi) & \frac{im}{-\lambda h}Ai(\xi)
\end{bmatrix}.
\] (A.65)

Thus the product of the Airy function matrix at \( \xi_1 \) and its inverse at \( \xi_2 \) is a matrix \( M^A(\xi_1, \xi_2) \equiv A(\xi_1)A(\xi_1)^{-1} \) with elements:

\[
\begin{align*}
M_{11}^A(\xi_1, \xi_2) &= \pi[ Ai(\xi_1)Bi'(\xi_2) - Bi(\xi_1)Ai'(\xi_2) ], \\
M_{12}^A(\xi_1, \xi_2) &= -\pi \frac{im}{-\lambda h} [ Ai(\xi_1)Bi(\xi_2) - Bi(\xi_1)Ai(\xi_2) ], \\
M_{21}^A(\xi_1, \xi_2) &= \pi \frac{-\lambda h}{im} [ Ai'(\xi_1)Bi'(\xi_2) - Bi'(\xi_1)Ai'(\xi_2) ], \\
M_{22}^A(\xi_1, \xi_2) &= -\pi [ Ai'(\xi_1)Bi(\xi_2) - Bi'(\xi_1)Ai(\xi_2) ].
\end{align*}
\] (A.66)

### A.6.2 Analysis of Airy Function Matrix Elements

The diagonal elements of this product matrix would involve Wronskians if \( \xi_1 = \xi_2 \). This provides a motivation to express the Airy functions and derivatives evaluated at one normalized coordinate in terms of an expansion about the other normalized coordinate. If \( Xi(\xi) \) is either \( Bi(\xi) \) or \( Ai(\xi) \), its Taylor series expansion is

\[
Xi(\xi + \delta) = \sum_{n=0}^{\infty} Xi^{(n)}(\xi) \frac{\delta^n}{n!}.
\] (A.67)

Since \( Xi(\xi) \) is a solution of the Airy equation, \( Xi^{(n)}(\xi) = \frac{d^n}{d\xi^n} Xi(\xi) \) can be expressed in terms of lower derivatives. It can be shown by construction that the \( n \)th derivative of \( Xi(\xi) \) can be written in the form,

\[
Xi^{(n)}(\xi) = P_n(\xi)Xi(\xi) + Q_n(\xi)Xi'(\xi),
\] (A.68)

where \( P_n(\xi) \) and \( Q_n(\xi) \) are polynomials in \( \xi \). For example, if \( n = 1 \) then \( P_1(\xi) = 0 \) and \( Q_1(\xi) = 1 \) and if \( n = 2 \) then \( P_2(\xi) = \xi \) and \( Q_2(\xi) = 0 \) because this gives the Airy equation, \( Xi''(\xi) = \xi Xi(\xi) \). An iterative formula for \( P_{n+1} \) and \( Q_{n+1} \) given \( P_n \) and \( Q_n \) can be determined by differentiating equation A.68 once to find

\[
Xi^{(n+1)}(\xi) = P_n(\xi)Xi(\xi) + P_n(\xi)Xi'(\xi) + Q_n(\xi)Xi'(\xi) + Q_n(\xi)Xi''(\xi).
\] (A.69)
But \( Xi''(\xi) = \xi Xi(\xi) \), so the right hand side can be reduced to terms proportional to \( Xi(\xi) \) and \( Xi'(\xi) \) without any higher order derivatives. Substituting \( \xi Xi(\xi) \) for \( Xi''(\xi) \) in equation A.69 and grouping the coefficients of \( Xi(\xi) \) and \( Xi'(\xi) \) yields

\[
Xi^{(n+1)}(\xi) = [P_n'(\xi) + \xi Q_n(\xi)] Xi(\xi) + [P_n(\xi) + Q_n'(\xi)] Xi'(\xi). \tag{A.70}
\]

The desired iterative equations for \( P_n \) and \( Q_n \) then come from realizing that the coefficients of \( Xi(\xi) \) and \( Xi'(\xi) \) in equation A.70 are respectively \( P_{n+1} \) and \( Q_{n+1} \) by the definitions in equation A.68:

\[
\begin{align*}
P_n(\xi) &= P_{n-1}(\xi) + \xi Q_{n-1}(\xi) & n > 0 \tag{A.71} \\
Q_n(\xi) &= P_{n-1}(\xi) + Q_{n-1}'(\xi) \\
P_n(\xi) &= 1 & n = 0. \tag{A.72}
\end{align*}
\]

Having developed a method of calculating \( P_n(\xi) \) and \( Q_n(\xi) \), and thus the \( n \)th derivative of \( Xi(\xi) \), the Taylor series expansion in equation A.67 and its derivative can now be expressed in terms of sums over these polynomials by substituting equation A.68 into equation A.67 to find

\[
Xi(\xi + \delta) = Xi(\xi) \sum_{n=0}^{\infty} P_n(\xi) \frac{\delta^n}{n!} + Xi'(\xi) \sum_{n=0}^{\infty} Q_n(\xi) \frac{\delta^n}{n!} \tag{A.73}
\]

and

\[
Xi'(\xi + \delta) = \sum_{n=0}^{\infty} Xi^{(n+1)}(\xi) \frac{\delta^n}{n!} = Xi(\xi) \sum_{n=0}^{\infty} P_{n+1}(\xi) \frac{\delta^n}{n!} + Xi'(\xi) \sum_{n=0}^{\infty} Q_{n+1}(\xi) \frac{\delta^n}{n!} \tag{A.74}
\]

As a notational convenience, further symbols are defined for the sums of \( P_n \) and \( Q_n \):

\[
\begin{align*}
P_0(\xi, \delta) &= \sum_{n=0}^{\infty} P_n(\xi) \frac{\delta^n}{n!} \\
Q_0(\xi, \delta) &= \sum_{n=0}^{\infty} Q_n(\xi) \frac{\delta^n}{n!} \\
P_1(\xi, \delta) &= \sum_{n=0}^{\infty} P_{n+1}(\xi) \frac{\delta^n}{n!} \\
Q_1(\xi, \delta) &= \sum_{n=0}^{\infty} Q_{n+1}(\xi) \frac{\delta^n}{n!}. \tag{A.75}
\end{align*}
\]
These quantities are, in fact, the primary portions of the elements of $M^4$ after the simplification of the expanded Airy function products. Each term in the sums requires a polynomial that can be generated from the polynomials for the previous terms.

The expansions of Airy functions written in equations A.73 and A.74 are used to simplify the elements of $M^4(\xi_1, \xi_2)$ by taking $\xi = \xi_1$ and $\xi_2 = \xi_1 + \delta$. With much algebra, and recognition of the Wronskian combinations in the elements, it can be shown that

$$M^4(\xi_1, \xi_2) \equiv A(\xi_1)A(\xi_2)^{-1} = \begin{bmatrix} Q_1(\xi, \delta) & \text{im}^* \frac{i\hbar}{m^*} Q_0(\xi, \delta) \\ -\text{im}^* \mathcal{P}_1(\xi, \delta) & \mathcal{P}_0(\xi, \delta) \end{bmatrix}. \quad (A.76)$$

**A.7 Comparison of Methods for Computing Tunneling Spectra**

An alternative to using transfer matrices is to numerically solve the Schrodinger equation in the region of interest. A variety of methods exist for numerically solving differential equations[107]; a rather simple one is used here. If the Schrodinger equation is discretized on a mesh with constant spacing, $d$, a difference equation analogous to the second order differential equation is

$$\frac{\psi_{j-1} - 2\psi_j + \psi_{j+1}}{d^2} = -\frac{2m^*(E - V(x_j))}{\hbar^2} \psi_j, \quad (A.77)$$

where $\psi_j$ is the value of $\psi$ at the $j$-th node, located at $x_j$. The value of the wavefunction at the leftmost of the three nodes, $\psi_{j-1}$, can be calculated from the values of the wavefunction at the other two nodes. For tunneling structures, boundary conditions on normalization of wavefunction magnitude and flux are sufficient to determine the values of $\psi$ at the two rightmost nodes. Then, the value of $\psi$ at the next node can be calculated from equation A.77. The solution can be propagated across the structure by solving for the wavefunction at the next node using the values at the two nodes to the right of the node with the unknown value of $\psi$. This amounts to numeric integration of the second order equation, starting from initial values determined by choices of the wavefunction amplitude and flux. Once the values of the wavefunction and its derivative at the left of the structure are known, coefficients for plane waves...
moving toward and away from the structure, and hence tunneling and reflection coefficients, can be calculated. One benefit of the numerical integration approach is that it implicitly generates the wavefunction for problems where it is needed.

In order to examine the relative merits of the numerical integration and transfer matrix approaches to calculating tunneling spectra, both were used to calculate the transmission probability and reflection phase shift from a potential barrier shown in Figure A.1. The barrier was chosen to be trapezoidal so that any rectangular profile

![Potential profile graph](image)

Figure A.1: Potential profile of a hypothetical barrier used for comparing methods of calculating tunneling spectra. The effective mass was taken to be a tenth of the free electron mass throughout the structure.

would only be an approximation to the true profile. The approximation of the true profile by a stepped potential can be improved by increasing the number of segments in the approximation. Decreasing the mesh spacing and increasing the number of points in the numerical integration likewise improve its accuracy. However increasing
the number of segments or points increases the time required for the computation, so it is desirable to compare the error between the computed and true spectra for various numbers of intermediate points.

The computed spectra for the structure using a 154-point numerical integration and a transfer matrix calculation with 24 segments are shown in Figure A.2. It is clear that the values computed with transfer matrices are in closer agreement with the precise values shown by the solid line than values from the numerical integration. The curves for the two transfer matrix calculations are indistinguishable in the figure. However, even with a large number of points, the spectra calculated from the numerical integration deviate noticeably from the more accurate transfer matrix calculation. A comparison of the root-mean-square error for the two methods is presented in Figure A.3 as a function of the number of segments or points used.

For an equivalent approximation of the potential profile, the transfer matrix method is more accurate. However, the numerical integration method is faster than the transfer matrix method for structures of equal complexity. The transfer matrix method is slow because it requires the evaluation of transcendental functions at each potential step. Using reasonably efficient C programs on an 4.77 MHz IBM-PC equivalent computer with a math coprocessor, it took 0.10 seconds to calculate each point in the spectrum of a structure with 150 potential points using numerical integration, but 0.45 seconds using transfer matrices for the same potential profile.

Although numerical integration was approximately five times faster than equivalent transfer matrix computations in this comparison, this is not sufficient to overcome the superior accuracy of the transfer matrix methods. Figure A.3 shows that the transfer matrix errors with 19 segments are more than an order of magnitude smaller than errors from numerical integration using 154 points. Since these computations take a similar amount of time, the transfer matrix is preferrable for calculating accurate transmission spectra. Numerical solutions of the Schrodinger equation are useful if the wavefunction is needed; however, more accurate methods for solving differential equations should be employed.
Figure A.2: Spectra calculated for the tunneling barrier in Figure A.1 using a 154-point numerical integration (long-dashed line), a 24-segment transfer matrix calculation (short-dashed line), and a very accurate 194-segment transfer matrix calculation (solid line). The (a) tunneling probability and (b) reflection phase shift are shown as a function of incident electron energy.
Figure A.3: Root-mean-square error of transmission probability (squares) and reflection phase shift in radians (circles) for calculations based on transfer matrices (open symbols) and numerical integration (filled symbols) as a function of the number of points or segments used to approximate the potential.
References


134
REFERENCES


REFERENCES

[34] H. Sakaki. Personal communication.


[57] FEI Microwave. Personal communication.


[59] Impulse Microwave. Personal communication.


REFERENCES


REFERENCES


REFERENCES


