QUANTUM ELECTRONIC DEVICE SIMULATION

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

Bryan A. Biegel

March 1997

Abstract

An accurate understanding of quantum wave effects in electronic devices is important for several reasons. In the short term, this understanding will enable the suppression of these increasingly significant parasitic effects in ever-smaller conventional devices. In the medium term, this understanding will enable the control of these effects, possibly extending down-scaling closer to the quantum realm with hybrid conventional-quantum electronic devices. In the longer term, an understanding of quantum electronic effects is necessary for the possible development of a true quantum device technology, with the potential for much greater functionality per unit cost, size, and power. To build this understanding, a numerical quantum device simulator called SQUADS (Stanford QUAntum Device Simulator) was developed. This dissertation describes the implementation, capabilities, and some illustrative simulation results of SQUADS.

The design of SQUADS was directed by two goals: the study of quantum device operation, and the study of quantum device simulation. In pursuing these goals, a comprehensive 1-dimensional simulation tool was developed for modeling quantum-effect electronic systems of arbitrary structure. Two independent formulations of quantum mechanics were implemented in SQUADS. The first is the widely-employed transfer-matrix method of quantum system simulation, which provides a source of quick initial simulation results, and is especially useful in detailing the energy spectrum of carriers in the device. The second method uses the Wigner function formulation of quantum mechanics, which is more computationally intensive, but which allows a more intuitive and complete description of real quantum electronic systems, especially including transient response and energy dissipation.

In addition to describing the basic implementation and simulation results of these sim-

ulation methods in SQUADS, this thesis also describes three detailed investigations of quantum device simulation and operation, using SQUADS as the simulator and the resonant tunneling diode as the test device. An investigation of self-consistency in quantum device simulation found that both the efficient steady-state and the more accurate transient self-consistency iteration methods have important roles to play, and that a Gummel (as opposed to full-Newton) iteration method is almost always quite adequate. An investigation of the effect of slew rate variation in transient RTD simulation showed that the use of an appropriate applied bias slew rate is necessary for accurate simulations and to prevent the misinterpretation of simulation results. Finally, the detailed simulation investigation of the physics of an RTD produced a better understanding of this device, corrected several errors in previous interpretation of simulation and experimental results, and resulted in improved agreement between simulation and experiment for this device.

In general, this work found that quantum device simulation is still in a formative stage, although significant advances have been made in this work and elsewhere. Quantum device simulation is not yet at a point where it can reliably reproduce or predict quantitative experimental results, whether because of non-idealities in experiment or inaccuracy of the simulator. Nevertheless, quantum device simulation in this work and elsewhere has already contributed to the debates surrounding significant unresolved issues of quantum device physics and operation.

Acknowledgments

I am happy to have this opportunity to acknowledge the assistance and support of the many people who contributed to the completion of this work and to my enjoyable and rewarding experience at Stanford. First, I would like to thank my research advisor, Professor James Plummer, who always supported and encouraged my research efforts, in spite of the long duration of this project. His broad understanding of the state-of-the-art and future of electronics made it possible for me to pursue my rather esoteric interests in quantum electronic devices while still obtaining useful guidance from him. At the same time, he allowed me to develop my own research abilities, make my own contributions, and learn self-sufficiency. I believe all Ph.D. graduates from Stanford should have this experience.

I would also like to thank my associate advisor, Professor Walter Harrison, for his generosity of time and candid discussions. Professor Harrison posed very thought-provoking questions about this work, prompting worthwhile examinations of subtle underlying assumptions. I appreciate his deep understanding of quantum physics, and his unique ability to make that knowledge accessible to others. Dr. Zhiping Yu has also been very helpful in sharing his knowledge of semiconductor device physics and simulation. Along with Professor Plummer, Professor Harrison, and Dr. Yu, I would like to acknowledge Professor Lambertus Hesselink as the final member of my reading committee. They have given a long and much-appreciated effort in the review and correction of this dissertation.

The time and assistance of many other people in the completion of this work should also be acknowledged. My orals committee; including Professors Plummer, Hesselink, Robert Dutton, and James S. Harris, as well as Dr. Zhiping Yu; were very gracious in accommodating my schedule and subject, and their insightful questions and comments were much appreciated. Dr. Kevin Jensen was always willing to discuss my latest discov-

eries and challenges. He provided key insights that led to some of the significant contributions of this work. Dr. Kiran Gullapalli and Dr. William Frensley also provided valuable technical input during this research. Dr. Daniel Scales was my main programming guru; able to help slay the hardiest of software bugs. Many people assisted during the brief experimental portion of this work, including Chan-Hong Chern and Vincent Arbet at UCLA, who grew SiGe MBE samples for me *pro-bono*; and all of the Integrated Circuits Laboratory technicians, who make useful work therein possible.

The information systems personnel at the Center for Integrated Systems, including Charlie Orgish, Laura Schrager, and Dr. Ernest Wood, were always willing and able to address computer-related problems and questions. Robert Taft, Walter Snoeys, William Wong, and Amit Paul also generously gave their time to make computing resources in the Plummer group as productive as possible. Vital to the success of this project were the impressive computing resources available to Stanford students - a reflection of Stanford's foresight in procuring and managing these resources. The Plummer administrative assistants over the years, including Joyce Pelzl, Susan Stout, and Jane Edwards, have been uniformly friendly and knowledgeable about how to get things done.

Finally, I would like to mention some of the many people who helped to make my time at Stanford enjoyable and rewarding. I am happy to call many Plummer students my friends, including Eric Perozziello, Walter Snoeys, Robert Taft, Mary Weybright, and Tiemin Zhou. Along with Scott Gallert, Jim Roche, Dan Scales, and many others, they made me wish my time at Stanford would never end. The support and encouragement of my long-time friend, Margaret O'Hare, has always been appreciated, though not often acknowledged. And, of course, I appreciate the love and support of my family: my parents Peter and JoAnn, and my siblings Denise, David, Mark, Glen, Andrea, Miriam, Alex, and Josh. My parents-in-law, Mel and Fran Holdener, gave me a home away from home, and have never failed to provide support (nutritional and emotional) on a regular basis (especially at holidays). Last and certainly not least, I want to recognize my wife, Teresa, for her superhuman endurance and love during one of the longest doctoral programs known to man (or woman). I'm sure she sometimes wondered whether I *did* still recognize her, but having her picture in my wallet helped.

Support for this research was provided in part by the Joint Services Electronics Program and by the Computational Prototyping Program at Stanford University.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter provides an introduction and overview of the research described in this dissertation. Section 1.1 describes the motivation for studying quantum effects in semiconductor devices, Section 1.2 presents the rationale for the specific approach and objectives of the research, and Section 1.3 describes the organization of the remainder of this dissertation.

1.1 Motivation

1.1.1 The Quantum Challenge

Tremendous advances have been seen in digital electronics technology in the past three decades due to a strong market demand for greater system speed and functionality. The amazing and apparently tireless advance of digital electronics technology provides us with empowering technological innovations, enables us to address new challenges in our world, and allows us to tackle ever more complex questions about our universe. With continuing efforts to improve the speed and functionality of integrated circuits, higher integration densities are forcing device dimensions to decrease to the scale of the quantum wavelength of the charge carriers[1] used in device operation. The transition between classical (particle-like) and quantum (wave-particle) behavior of carriers begins at device dimensions of around 0.1 μm (100 nm) [1]. With continued device scaling, the reliable

1. Hereafter, charge carriers, meaning electrons or holes, will often be referred to simply as carriers.

operation of ultra-large-scale integrated (ULSI) electronic devices, which depends on classical (particle-based) carrier transport, will be increasingly antagonized by “parasitic” quantum (wave-based) transport phenomena. Figure 1.1 shows three examples of carrier quantum “mis-behavior” that already occur in conventional electronic devices.

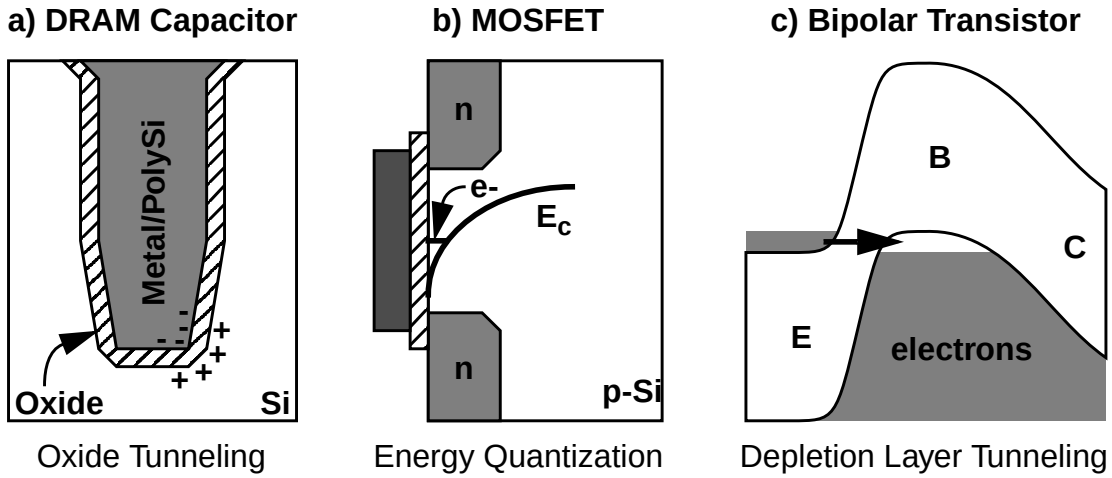


Figure 1.1: Parasitic quantum effects in conventional devices

Effects indicated include (a) charge carriers tunneling through a DRAM capacitor oxide at sharp corners, (b) energy quantization of carriers in the inversion layer of a MOSFET, and (c) electrons tunneling though the band gap at the base-emitter junction of a degenerately doped bipolar transistor.

The effort to maintain reliable operation of electronic devices as their dimensions inevitably shrink towards the quantum realm is herein called the “quantum challenge”. The usual response to the quantum challenge is to constrain or modify conventional device designs such that quantum effects are *avoided* entirely if possible, *ignored* if they can at least be made negligible, *suppressed* as much as possible even if they are not negligible, or *overwhelmed* if all of the above approaches fail. What happens when these compromises and constraints leave no room to advance ULSI electronics? If the conventional, *evolutionary* solution to the quantum challenge no longer works, and of course the demand for greater digital system functionality via device scaling will not diminish, a new, *revolutionary* response to the quantum challenge must be found.

1.1.2 The Quantum Solution

A revolutionary approach to addressing the quantum challenge, which has more long-

term possibility though at the same time less certainty, is to find some means by which quantum phenomena can be used as *enabling*, rather than *disabling*, mechanisms in the operation of electronic devices. Devices whose operation is fundamentally based on quantum wave phenomena are called quantum devices. The concept of producing useful computing (analog or digital signal processing) with quantum devices is herein called quantum electronics [2, 3].[2] The path to quantum electronics may in fact be evolutionary, with the development of hybrid conventional-quantum devices. Such devices would operate essentially as conventional devices, but would use quantum effects in a controlled but subordinate manner to achieve down-scaling or functionality beyond that attainable by a pure conventional electronic device. Three well-known examples of such hybrid devices, the quantum well laser diode, hot electron transistor, and EPROM, are shown in Figure 1.2. The quantum effect used to enhance the operation of each is also described.

The idea of using, rather than avoiding, quantum effects in electronic device operation has several significant benefits. First, it would finally allow (in fact *require*) at least one device dimension to be scaled into the quantum realm, whereas conventional electronics requires all device dimensions to be greater than quantum scale. Scaling devices to quantum dimensions, say 50 nm or less, would allow integration densities well beyond even ULSI, into a regime that can best be called quantum scale integration. Higher integration levels lead directly to greater system functionality. The additional benefits of device scaling are well known: faster device operation and lower device power. The use of quantum devices and quantum scale integration allow all of these improvements to continue into the quantum realm, solving the quantum challenge. Secondly, using quantum effects in the operation of electronic devices would allow the control electronic device operation with any of the phenomena that are parasitic in conventional devices, and thus presents the possibility of attaining much higher functional efficiency. Finally, quantum effects include a

2. The term “quantum electronics” is used with some reservations, since this term has been applied in the literature to semiconductor laser optoelectronics. However, “quantum electronics” aptly describes the electronic systems discussed in this dissertation, which use quantum phenomena in their operation, but which act externally much like conventional electronic systems. Alternative terms were also considered. “Quantum computing” has been applied narrowly to quantum systems whose operation is based on the physics of *discrete* quanta, which systems are described in Section 1.1.4. This dissertation mainly considers quantum systems whose operation is based on the physics of a *continuous distribution* of quanta. Some researchers use “nanoelectronics” to describe this type of quantum device, but this moniker is not adequately descriptive, and is more appropriately applied to deep submicron conventional electronics, as the natural successor to “microelectronics”.

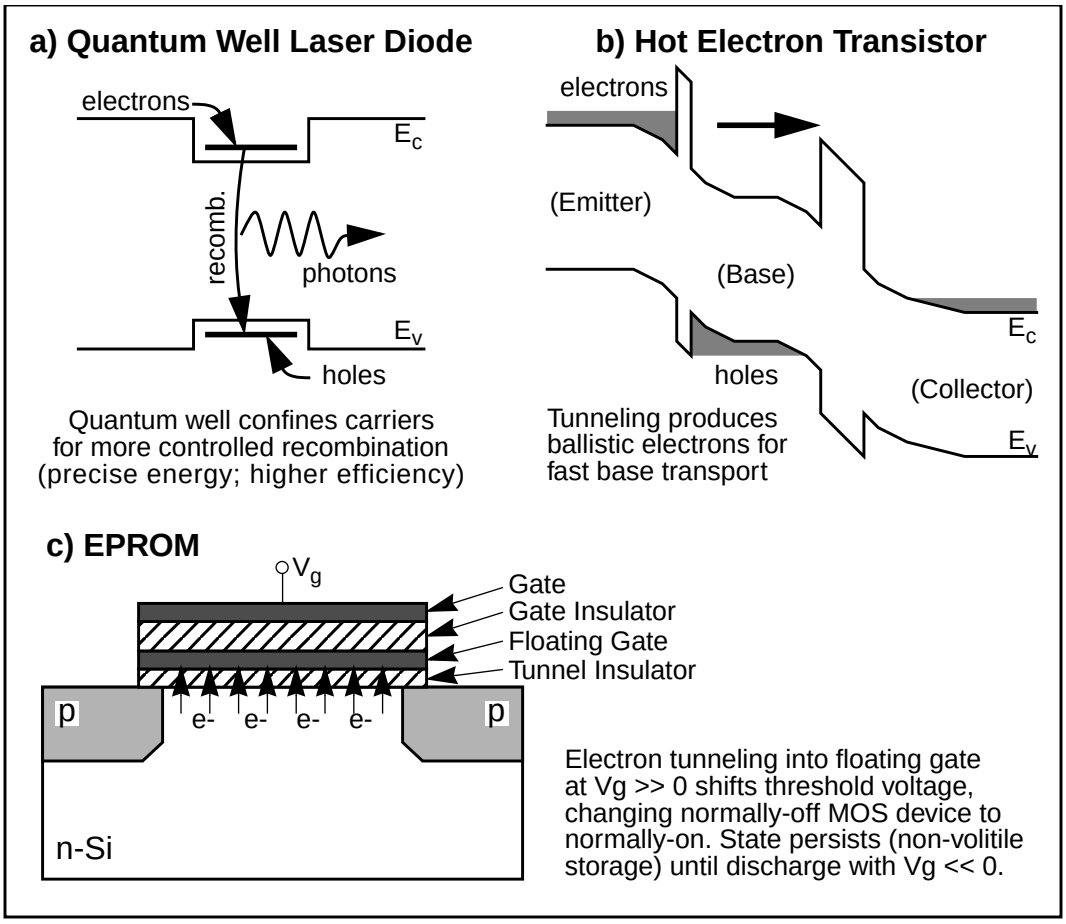


Figure 1.2: Hybrid conventional-quantum electronic devices

Three hybrid devices are shown: a) the quantum well laser diode, b) a hot electron transistor, and c) an EPROM (erasable, programmable ROM) device. The quantum effect used to enhance the operation of each device is also described.

diverse set of both analog (wave-like) and discrete (particle-like) phenomena, and may thereby allow the design of quantum effect-based computing systems which naturally produce both analog and digital functions.

1.1.3 More Challenges

While quantum electronics theoretically allows electronic system capabilities (efficiency and functionality) to advance perhaps orders of magnitude beyond that of conventional ULSI, several new challenges must be faced in the development quantum electronics, and these will require ingenuity to surmount. These new challenges include not only the need to develop new devices based on different physics, but also new fabrication technologies and circuit architectures will be needed for these quantum-scale devices,

and new computation paradigms will probably be required. These issues are discussed in more detail in Chapter 2.

Because of these challenges, it is not clear whether quantum electronics will ever become a viable substitute for, much less a successor to, conventional electronics. In fact, some of the strongest advocates of pursuing quantum electronics research are also its best debunkers [2, 4]. In spite of the touted potential of quantum electronics as the future and savior of electronics, research in this field may seem an academic exercise, considering all of the challenges in the way of its realization. Further, conventional electronics continues to improve almost faster than consumers can tolerate, with no end in sight. The wisdom of embarking on the costly and potentially fruitless endeavor of quantum electronics research seems foolish and wasteful. However, discoveries about quantum systems in just the past few years may have rendered this viewpoint invalid, as described below.

1.1.4 Further Possibilities

It turns out that quantum electronics may itself be a stepping-stone to an even more futuristic computing paradigm which has co-opted the name “quantum computing”. In this realm, discrete electrons and photons produce desired computing functions using coherent quantum waves. In late 1994, Peter Shor [5] described the “killer application” for quantum computing. This is none other than the potential to defeat public key encryption schemes, which are considered the best current hope for inexpensive and pervasive secure communication. Rather than the exponential computational requirements of a classical computer to decrypt such messages, the computation time using Shor’s quantum algorithm rises only as the square of the encryption key size [6]. Effectively, if Shor’s scheme were implemented, any message encrypted with the public key scheme, whether sent 10 years ago or tomorrow, would be immediately readable by the owner of the quantum decrypter. For better or worse, it appears that it will be extremely challenging to build a quantum decrypter that can handle an encryption key larger than a classical computer can (say 1000 bits). However, it is equally difficult to think of a device more coveted by governments (who are best able to fund the development of such a device). For this reason alone, if a quantum decrypter can be built, it certainly will be, regardless of the difficulty.

The quantum decrypter is just a specific (but important) example of the power of quantum computing. Quantum computing takes advantage of the inherent parallel nature of

processing coherent waves for the fast solution of recursive problems, essentially checking all possible solutions simultaneously. Only the correct solution will interfere constructively and produce a positive result. Other computations which are infeasible to solve on a classical computer may also be proven accessible using a quantum algorithm. Also, smaller quantum computers, perhaps handling only a few bits at a time, could solve other seemingly intractable problems. For example, provably secure quantum communication schemes (e.g., for passing secret encryption keys) have been proposed [7, 8] and tested [9]. These implementations are as yet impractical for general use. Several accessible reviews of these and other quantum computing possibilities have been written [6, 9-14].

Clearly, there are tremendous incentives driving the research and development of quantum computing, and equally momentous challenges to be overcome. Of course, the eventual success or failure of quantum computing can not be determined unless and until the concept is thoroughly investigated. And the ultimate in quantum computing will almost certainly never be achieved without passing through quantum electronics first. Further, it seems preferable to face the quantum challenge now, rather than when it becomes a quantum *crisis* at the limit of conventional ULSI advancement. Both the inevitability of the quantum crisis and the potential benefits of quantum electronics and quantum computing were the motivations for this research, the broad goal of which was the investigation of quantum electronics. Research on the ultimate quantum computing systems, the heir-apparent to quantum electronics, is left for future researchers.

1.2 Approach and Objectives

In general, there are three approaches to pursuing a scientific field of study: conceptual, computational, and experimental. The choice of approach for this research project was a fairly straight-forward decision. As argued above, quantum electronics truly is a revolutionary concept. It's realization will require new physics, new fabrication technologies, new circuit architectures, and new computing paradigms. Thus, in order to illuminate potentially useful directions to take with either simulation or experiment, an initial theoretical (*i.e.*, conceptual) analysis of quantum electronics and its future in quantum computing is absolutely essential. Many researchers have undertaken such an analysis, and the body of this dissertation begins with a summary of this theoretical work and its main conclusions (Chapter 2). The central focus of this investigation of quantum computing, as

described in this dissertation, takes the numerical simulation (*i.e.*, computational) approach. As in other fields, simulation fills the information gap between idealized theory and "exact" but expensive experiment. Conceptual research can only provide a general idea of how a real quantum system would behave, beyond which simulation and experiment are required.

In quantum electronics research, simulation will largely direct experiment, just as theory directs simulation. The most compelling reason for this is feasibility: the cost of simulation is many orders of magnitude less than experimental trial-and-error. In fact, experimental work with all but the simplest quantum-scale devices and quantum circuits is not feasible at all with existing fabrication technologies. Even with conventional ULSI research and development, simulation makes possible the investigation of a much wider range of device structure and operating condition alternatives than would be feasible with experiment alone. With the revolutionary concept of quantum electronics, where the eventual device "winners" are not even known, simulation represents a much more cost-efficient means of investigating new device concepts and operation phenomena.

There are other advantages besides cost of simulation over experiment in quantum electronics research. For example, a quantum system simulation capability is more widely applicable than fabricating and testing quantum devices, since the former can be useful for any quantum device for any application. Further, simulation can give detailed information about device operation which is not provided by experiment (or even theory), such as an internal view of device operation (*e.g.*, internal carrier concentrations or current flow lines). An accurate device simulator can thereby resolve any mysteries of device operation. Thus, in quantum electronics research, simulation will likely serve to illuminate promising directions for (future/expensive) experimental research.

Another issue which falls under the heading of "investigative approach" is to choose the time-range where the research is expected to have impact. In other words, is the research meant to be of short-term or long-term practical importance? Although quantum electronics inherently has a relatively long-term potential pay-off, this research attempts to make its impact more immediate where possible by taking a short-term focus.[3] For example, some simple quantum devices are being fabricated today, and being able to simulate

3. The short-term focus was, in fact, an important reason for choosing to research quantum electronics rather than quantum computing systems.

these devices and thus contribute to this work was considered essential. This focus has implications for the type of simulator developed, as discussed in Chapters 2 and 3.

Based on the above discussion, the objective of this research has been narrowed from a general investigation of quantum electronics to the development of a simulation tool for the investigation of quantum electronic devices. The resulting simulator is called SQUADS (Stanford QUAntum Device Simulator). Two goals were envisioned in the use of this tool: the investigation of quantum device *operation*, and the investigation of quantum device *simulation*. The general approach taken in developing SQUADS was therefore to include in it as many capabilities as possible. Progress toward the research objective was gauged by simulating various devices with generally known behavior to see whether the simulation results agreed with expectations. This dissertation contains many illustrative examples of such simulations. Defining the research approach for this work in any more detail requires a significant amount of additional background information, which is provided in the next two chapters. Before undertaking this discussion, an overview and outline of this dissertation is given below.

1.3 Organization

This dissertation is organized into nine chapters, of which this introduction is the first. Note that the bibliography for each chapter is given at its end, rather than as a combined bibliography at the end of the dissertation, since the references of different chapters have minimal overlap.

To give direction and focus to the numerical simulator development in this work, Chapter 2 presents an overview of the current understanding of quantum electronics. This begins with a description of the quantum phenomena and basic structures that will be building blocks for quantum devices, and then turns to the general characteristics of quantum effect devices. The discussion also analyzes the merits of some specific quantum devices. Finally, this chapter considers what quantum electronic circuits may look like, and the likely nature of computing with these circuits. The important results from this analysis are its implications for the implementation of SQUADS.

Chapter 3 completes the specification of the approach taken in developing SQUADS by selecting its underlying basis from the many mathematical formulations of quantum mechanics. To accomplish this, the capabilities and features required of a useful quantum

device simulator are first discussed. Based on this, the various quantum mechanics formulations are evaluated in terms of their ability to produce a quantum system simulator with the best combination of capabilities, accuracy, and computational efficiency. The conclusion is that a dual-formulation approach would serve as the best basis for SQUADS. The resulting implementation of SQUADS and its simulation results are the subject of Chapters 4 through 8.

Chapters 4 and 5 describe the basic implementation of the two formulations of quantum mechanics used in SQUADS. The implementation of the transfer-matrix method, the de-facto standard in quantum device simulation, is described in Chapter 4. This method is based on solving the time-independent Schrödinger equation. The transfer-matrix method is suitable for quick, reasonably accurate simulations of a wide range of quantum device structures to determine which merit more detailed study. The implementation of the Wigner function method in SQUADS is then described in Chapter 5. This method is analogous to solving the Boltzmann transport equation for conventional device simulation. As such, it is suitable for accurate, but computationally expensive, simulations. Basic simulation results for each method are given at the end of their respective chapters.

Chapters 6 and 7 describe some of the more advanced capabilities of SQUADS, and present simulation results showing the importance of these features. Chapter 6 covers the implementation in SQUADS of full quantum self-consistency, which requires that the charge density profile produced by the quantum transport equation is consistent with the energy band profile in the simulated device. Chapter 7 describes the transient simulation capabilities of SQUADS in more detail through an analysis of the effect of applied bias slew rate variation on the operation of a quantum device.

To bring closure to the body of this work, Chapter 8 presents a detailed and comprehensive quantum simulation investigation of a single quantum device. As in chapters 4-7, a resonant tunneling diode is used as the test device. The physics of this device is analyzed in as much detail as necessary to present a complete picture of its operation. Also discussed are the implications this simulated operation has for the resolution of several significant controversies regarding resonant tunneling diode operation. Finally, an attempt is made to assess the current accuracy and reliability of quantum device simulation tools, including SQUADS.

Finally, Chapter 9 summarizes this work, presents its contributions to the field of quan-

tum device simulation, and gives suggestions for future work in this field.

References

[1] K. Hess and G. J. Iafrate. “Approaching the quantum limit.” *IEEE Spectrum*, pages 44–49, July 1992.

[2] R. T. Bate. “Nanoelectronics.” *Nanotechnology,* 1(1):1–7, 1990.

[3] F. A. Buot. “Mesoscopic physics and nanoelectronics: Nanoscience and nanotechnology.” *Physics Reports,* 234(2-3):73–174, 1993.

[4] R. Landauer. “Can we switch by control of quantum mechanical transmission?” *Physics Today,* pages 119–121, Oct. 1989.

[5] P. W. Shor. “Algorithms for quantum computation: Discrete logarithms and factoring.” In S. Goldwasser, editor, *Proceedings 35th Annual Symposium on Foundations of Computer Science,* pages 124–134, Santa Fe, NM, Nov. 20-22 1994. IEEE Computer Society.

[6] D. P. DiVincenzo. “Quantum computation.” *Science*, 270:255–261, Oct. 13 1995.

[7] A. K. Ekert. “Quantum cryptography based on bell’s theorem.” *Physical Review Letters*, 67(6):661–663, 1991.

[8] C. H. Bennett, G. Brassard, and N. D. Mermin. “Quantum cryptography without bell’s theorem.” *Physical Review Letters*, 68(5):557–559, 1992.

[9] C. H. Bennett, G. Brassard, and A. K. Ekert. “Quantum cryptography.” *Scientific American*, 267(4), Oct. 1992.

[10] C. H. Bennett. “Quantum information and computation.” *Physics Today,* 48(10):24–30, 1995.

[11] D. Deutsch. “Quantum computation.” *Physics World*, pages 57–61, June 1992.

[12] C. H. Bennett. “Quantum cryptography: Uncertainty in the service of privacy.” *Science*, 257(5071), Aug. 7 1992.

[13] D. Deutsch and A. Akert. “Quantum communication moves into the unknown.” *Physics World*, pages 22–23, June 1993.

[14] S. Lloyd. “Quantum-mechanical computers.” *Scientific American*, 273(4):140–145, 1995.

Chapter 2

Overview of Quantum Electronics

This chapter provides a summary what is known and what can be reasonably deduced about the possibility of practical quantum electronics. The discussion covers the quantum effects, basic structures, devices, systems, and computing models, which might produce orders of magnitude denser, faster, and more efficient computing systems. Although this chapter is not intended to be a comprehensive review of quantum electronics, it will serve to show where this dissertation fits in the larger field of quantum electronics research. The more specific goal of this chapter is to discuss the characteristics of computationally useful quantum devices, and thus, what capabilities and features should be implemented in a practical quantum device simulator.

The summary of quantum electronics in this chapter is accomplished in four steps. In Section 2.1, some guidelines for the identification of useful quantum devices are established by discussing briefly the past, present, and ultimate future of quantum electronics. The summary of quantum electronics then proceeds from its basic elements to complete systems. In Section 2.2, the quantum phenomena and basic structures that will be building blocks for quantum electronic systems are enumerated. Section 2.3 then discusses the general characteristics of quantum effect devices designed with these structures and phenomena, and considers the merits of some specific quantum devices. This section concludes with the selection of a prototype quantum device for use in developing a practical quantum device simulator. Finally, Section 2.4 considers the likely characteristics of complete quantum electronic systems, and how they might accomplish useful computing. The

resulting conclusions about quantum electronics, especially those relevant to the development of a quantum device simulator, are summarized in Section 2.5.

2.1 Past, Present, and Future

In this section, the basic outline of the quantum electronics concept is defined by showing where quantum electronics has come from, how existing technologies can inform us about the nature of quantum electronics, and where quantum electronics may ultimately be headed.

2.1.1 The Genealogy of Quantum Electronics

As has been observed, quantum electronics is a revolutionary idea, and few of the rules of conventional electronics may apply to it. However, nothing is completely new, and relevant existing knowledge should certainly be used, where possible, to guide the successful development of quantum electronics. So what knowledge *is* relevant? On one hand, at the very smallest scale, charge carriers in quantum devices will obey wave mechanics. Therefore, a wave-based system should be sought from which to draw knowledge and insight. The closest cousin of quantum wave systems is optical systems. In fact, optical computing, which seeks to produce useful digital functionality through electromagnetic wave manipulation, is a well-established field of research [1-4]. On the other hand, at the terminals of the quantum circuit, the input and output signals will be voltages and currents, as in conventional electronics. Quantum electronics viewed from the macroscopic level will probably be very similar to conventional computing. Based on these arguments, Figure 2.1 shows the types of knowledge that should be applicable from optical computing and electronics in the development of a useful quantum electronics technology. Relevant knowledge will be drawn from these sources throughout this chapter.

2.1.2 The Optical Analogy

The basic premise of quantum electronics is to build electronic devices that are small enough that their operation is dominated by the wave nature of charge carriers. Several researchers have explored the strong similarity between such quantum electronic systems and electromagnetic wave systems [5-7], which relationship is often called the optical analogy.[1] The optical analogy is important in the development of quantum electronics

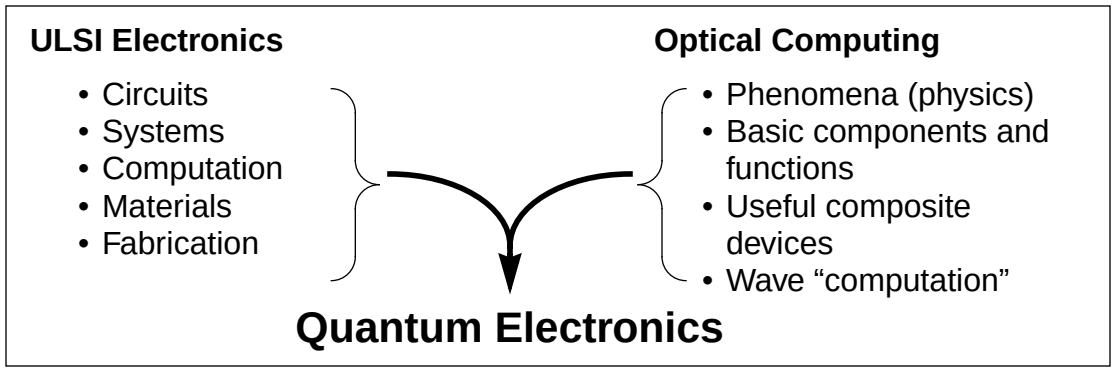


Figure 2.1: Progenitors of quantum electronics

Quantum electronics, despite having many revolutionary characteristics, can make use of relevant knowledge from ULSI electronics (which it seeks to replace), and from optical computing (which it seeks to emulate).

because it links common experience and intuition to the seemingly strange quantum phenomena and systems. The optical analogy indicates that in a quantum electronic system, charge carrier quantum waves will be guided with waveguides, reflected with mirrors, refracted with lenses, split, recombined, etc., to purposefully transform some input wave into a new output wave and thereby, in some manner, perform a desired computation. Clearly, quantum computing at this level is unrecognizable from the standpoint of conventional computing.

From the above arguments, the *function* of useful components in quantum computing systems are expected to be similar to those in optical systems, based on the optical analogy. Further, all basic wave effects can be produced in any system if refraction (wavelength change with position) and reflection can be produced. In optical systems, material changes result in refraction, and conductors produce reflection. Of course, a quantum electronic system is chiefly an *electron* system, not a photon system (although photons may be involved). The materials, structures, and material parameters of electronic systems must be used to create the desired wave effects. In particular, material parameters in semiconductors must be engineered at quantum dimensions to produce reflection and refraction. As it turns out, two material parameters, the energy band minimum[2] and the carrier effec-

1. “Electromagnetic” (EM) might be more appropriate than “optical”, since the latter implies only visible light - a very small part of the EM spectrum. However, optical systems are the most familiar and highly developed EM wave systems in common experience, especially for the purpose of computation, so the term will be used in this dissertation.

tive mass, fit this requirement in an analogous manner to the refractive index in optical systems. The refraction effect (due to wavelength modulation) can be calculated from:

$$\lambda = \frac{h}{\sqrt{2m^*(x)\,[E - U(x)]}} \,. \tag{2.1}$$

In (2.1), λ is the carrier wavelength, h is Planck's constant, and E is the (constant) total carrier energy. Also, m^* and U are the position dependent carrier effective mass and energy band minimum (a.k.a. potential energy), respectively. Just as with refractive index changes in optical systems, if U or m^* change gradually, refraction occurs according to (2.1), while there is little reflection. When the change is abrupt, both refraction and reflection occur. Finally, where the energy band minimum U rises above the total carrier energy E, the wave reflects completely. Thus, both basic wave effects (reflection and refraction) *can* be produced in quantum electronic systems. These effects will be depicted shortly.

Technologically, it is easier to manipulate the energy band minimum than the effective mass to produce reflection and refraction. This reality has resulted in a device design technique called band gap engineering [8]. With band gap engineering, a device is designed by simply determining the energy band structure required to produce a desired device function, and then putting together the necessary materials to create (as near as possible) this energy band structure. In the case of quantum devices, it is the (quantum-scale) energy band structure that the propagating quantum wave interacts with to produce quantum effects.[3] Note that because quantum device operation is based on wave interactions, the fabrication of these devices requires much more accuracy than does conventional device fabrication. Both the size and placement of structures in quantum devices determine the function produced.

Two methods are used to create static energy band offsets in electronic devices: *p-n* junctions and heterojunctions[4]. However, *p-n* junctions are far too large and have too much statistical variability when quantum scale interface control is essential, and they result in avalanche breakdown strength electric fields when doping is made high enough to reduce the first two problems. Thus, isotype heterojunctions[5] must be used to create band

---

2. The conduction band minimum (or edge) for electrons and the valence band minimum for holes.
3. The effective mass affects the relative *size* of quantum structures, but is a secondary consideration during initial quantum device design.
4. An interface between two different materials, such as GaAs and $Al_xGa_{1-x}As$, where differing electron affinities and band gaps result in an inherent conduction or valence band minima offset.
5. A heterojunction where the two materials have the same doping type, *p* or *n*.

offsets for quantum devices. The conclusion that there will be no *p-n* junctions in quantum devices leads to another conclusion: quantum devices will be majority carrier (unipolar) devices, since all local regions will be of the same conductivity type. In fact, not only device structures, but also inter-device isolation, must be produced by heterojunctions if integration densities much beyond that of conventional electronics are to be realized [9].

2.1.3 The Future

In the effort to apply quantum effects to computing, researchers have taken two fairly distinct approaches. The first approach is to evolve quantum electronics from existing computing technologies (ULSI electronics and optical computing), as described in Sections 2.1.1 and 2.1.2. By this approach, band gap engineering is used on semiconductors to create quantum systems whose elements function like optical (i.e., wave) components, but whose gross function is that of a conventional ULSI circuit. The resulting "distribution function" quantum devices use a continuous distribution of a large number of (indistinguishable) quanta in their operation, just like conventional electronic devices. These devices typically rely on quantum transport dominance only in the relatively small active region of the device, with classical transport (*i.e.*, scattering) allowable elsewhere. The result is that distribution function quantum devices are actually hybrid classical-quantum devices. Due to their external similarity, these quantum devices could conceivably enhance or replace conventional electronic devices *in-situ*. In fact, this is the reason this approach to using quantum effects in electronics is sometimes called "nanoelectronics" [10, 11]. Several other reviews of quantum electronics are available, including [9, 12-16].

The second approach to applying quantum effects to computation, called quantum computing, largely ignores existing computing technologies and goals. Instead, quantum computing researchers endeavor to develop a truly revolutionary and fully quantum technology which takes advantage of the unique features of quantum physics to accomplish feats not otherwise possible. The quantum decryptor described briefly in Section 1.1.4 is an example of such a uniquely quantum system. Quantum computing requires the development and use of "discrete-quanta" devices, which would operate on the quantum wavefunctions of individual quanta (*e.g.*, electrons, photons, or atoms). These devices rely on the relatively long-range and long-term coherence of each quanta involved (*i.e.*, scattering destroys the result), and the state of each quanta is significant to the computation. Accessi-

ble reviews of quantum computing include [17-21].

Many physical systems have been discussed [18, 22] as possible discrete-quanta devices, including the quantum dot, a nuclear spin, a localized electronic state in a polymer, a hydrogen atom, an ion trap, and even molecules in a salt crystal. For example, the basic operation of a hydrogen atom "bit" is indicated in Figure 2.2a. Building quantum computers requires forming a 1-D, 2-D, or 3-D array of these devices, "programming" them with a desired quantum state, and performing the desired computation through interactions with and among the devices. A comprehensive scheme for accomplishing this has been described [18]. For illustration, Figure 2.2b shows a 2-D array of closely-spaced, and therefore interacting, quantum dots.

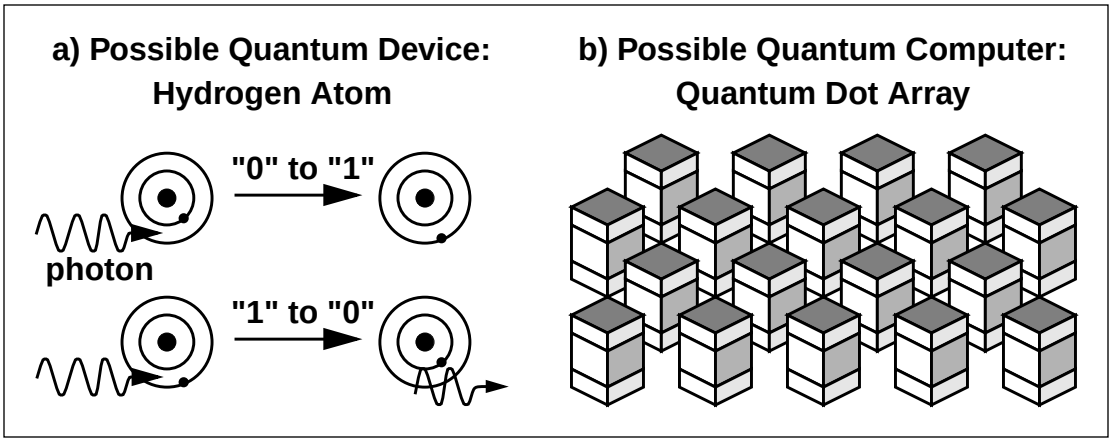


Figure 2.2: Possible quantum computing device and system

(a) shows a two energy states of a hydrogen atom, where a "probe" photon of the correct energy will shift the electron to the other energy level, causing the emission of a photon if the electron was is in the high energy state. (b) shows a 2-D array of interacting quantum dot structures, which might serve as a prototype quantum computer.

Discrete-quanta devices and the resulting quantum computing systems appear to represent the final frontier of computing, combining the ultimate in down-scaling and minimal energy consumption with the ability to perform amazing computational feats. However, the time frame and challenges for constructing a general-purpose computer are much greater for discrete-quanta devices than for distribution-function, quantum electronic devices. Therefore, since this research pursued a more short-term realization of applying quantum effects to computing, the focus in this research, and the discussion hereafter in this dissertation, will be on quantum electronic (*i.e.*, distribution-function) devices. This

focus has important implications for the type of quantum device simulator developed, and for the prototype device used to test it, as discussed in Section 2.3. Hereafter, any reference to quantum devices assumes the distribution function type, unless stated otherwise.

2.2 Phenomena and Structures for Quantum Devices

Given a general picture of quantum electronics from Section 2.1, and having decided in Section 2.1.3 on the type of quantum devices that will be considered, deducing the details of quantum electronics with those devices now begins, starting at the lowest (and least speculative) level. Section 2.1.2 reached the general (but important) conclusions that the relevant quantum effects are wave phenomena, and that the structures used to produce them can be created by band-gap engineering. This section inventories the available range of quantum phenomena and associated structural elements for quantum electronics. To facilitate this, complete and unabashed advantage is taken of the optical analogy.

2.2.1 Basic Quantum Wave Phenomena

In deducing the nature of operation of quantum electronic systems, the first step is to list the phenomena that can be used to develop a quantum wave processing technology. The basic wave phenomena that are familiar from optical systems include wave propagation, refraction, reflection, diffraction, interference, and evanescent wave penetration. Based on the optical analogy, and using band-gap engineering, it should be possible to produce all of these effects in quantum electronic systems as well. Indeed, Table 2.1 describes these wave phenomena from a quantum system perspective and gives sketches of simple energy band structures which have been used to produce each effect. See [5-7] for the theory behind these wave phenomena and structures.

2.2.2 Basic Wave Components and Quantum Structures

The next step towards describing complete quantum effect devices is to use the phenomena and energy band structures in Table 2.1 to design some of the basic wave processing elements for quantum electronic systems. Again drawing on the optical analogy, and using band-gap engineering, Table 2.2 shows six such elements: the refractor (lens) [23], reflector (mirror) [24], beam splitter [25], waveguide [26], partial reflector [27], and impedance matcher [28, 29]. Note here that the optical analogy should be seen as a source

Table 2.1: Fundamental quantum phenomena and associated structures

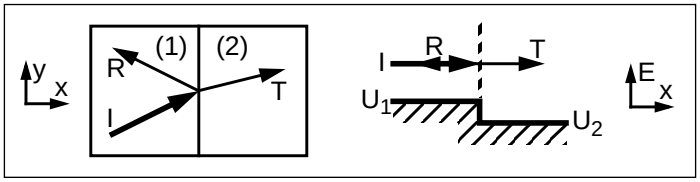
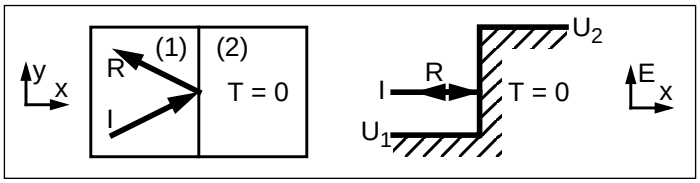
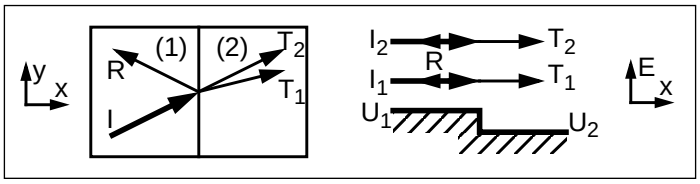
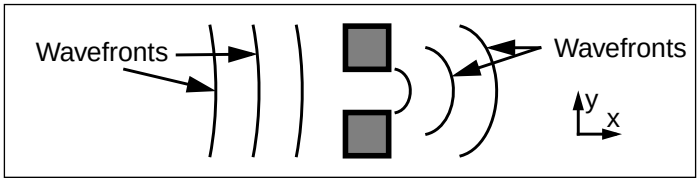
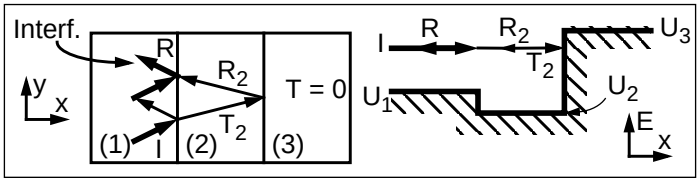
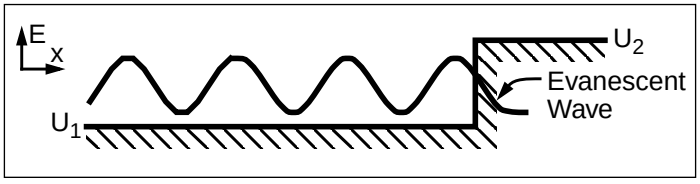
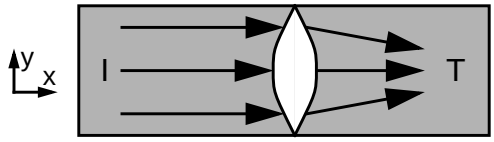
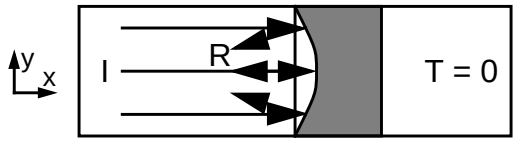
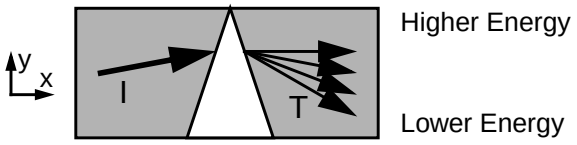
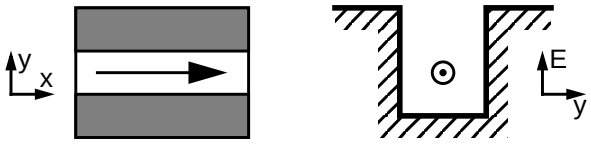
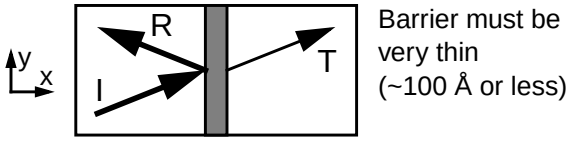
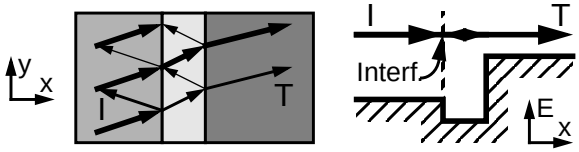
Quantum (Wave) Effect	Quantum System Component Structure (Example)
Refraction: Wave "trajectory" bending due to wavelength changes	 Low band-offset material interface
Reflection: Wave "trajectory" reversal at an interface	 High band-offset material interface
Dispersion: Refractive index varies with wavelength (and thus energy)	 Semiconductors are inherently dispersive
Diffraction: Wave bending around the edges of an object	 Small aperture defined by "opaque" material
Interference: Merging of two or more waves. Constructive interference: in-phase waves add. Destructive interference: out-of-phase waves subtract.	 Double partial-reflector interface
Evanescent Wave: Exponentially decaying wave penetration into an energetically "forbidden region"	 Moderate band-offset reflecting interface

Table 2.2: Basic wave processing elements in quantum electronic systems

Quantum Electronic Structure (Optical System "Equivalent") Basic Quantum Effects (BQEs)	Structure/Energy Band Diagram Example (Note: darker region = higher band edge) ☐ Low-band-edge; ■ High-bend-edge
Refractor (Lens) BQE: Refraction	
Reflector (Mirror) BQE: Reflection	
Beam Splitter/Analyzer (Prism) BQEs: Refraction, Dispersion	
Waveguide (Waveguide) BQEs: Reflection (abrupt waveguide); Refraction (graded waveguide)	
Tunnel Barrier (Partial Reflector, Beam Splitter) BQEs: Evanescent Penetration	
Impedance Matcher (Anti-Reflective Coating, Quarter Wave Plate) BQEs: Reflection, Refraction, Interference	

of ideas and understanding. It is not necessary to reproduce every optical component in a quantum system, although this could probably be done. The analysis to this point, and especially the optical analogy, should be sufficient to provide a visual image of what quantum phenomena and quantum devices will be like.

2.3 Devices for Quantum Electronics

The purpose of this section is to show how the phenomena and functional elements discussed in Section 2.2 can be combined to produce potentially useful digital quantum electronic devices and systems. It is not possible, of course, to discuss every potentially useful quantum electronic device — all such devices are certainly not known. Instead, some general concepts about quantum devices are considered. Then some prominent examples of quantum devices are considered, drawing from the two classes of (distribution function) quantum devices: quasi-equilibrium devices and far-from-equilibrium devices. Finally, conclusions are drawn about the characteristics of potentially useful quantum electronic devices in the near term, and thereby, a prototype quantum device is chosen as a test-case for the quantum device simulator developed in this work.

2.3.1 General Concepts of Quantum Devices

The motivation for investigating quantum devices was to solve the quantum challenge: to find a way to continue down-scaling digital electronic devices in spite of the quantum barrier. Given a short-term approach to answering this challenge, the solution sought in this section is a simple quantum replacement for the conventional transistor. First, the characteristics necessary for a direct replacement of a digital computing device are discussed. Then the general understanding of quantum phenomena and device structures discussed in Section 2.2 is used to predict ways in which such devices might be achieved with quantum wave phenomena and quantum wave processing elements.

Beginning with conventional electronics then, Figure 2.3 shows a block diagram of a generic switching device. A basic conventional digital electronic device functions by using an input voltage to modulate the height of a potential barrier to current flow along the output path. The “necessary” characteristics of conventional (digital) electronic computing devices can be described as follows [10, 30]:[6]

- gain — small input change produces large output change,

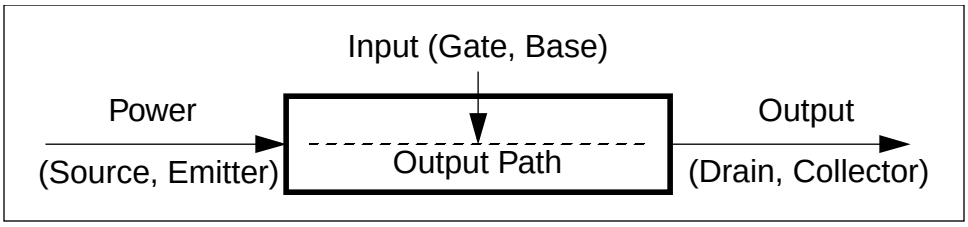


Figure 2.3: Conventional electronic switch (FET, BJT)

The basic digital device building block consists of a low-power controlling (Input) electrode which modulates the resistance of a high-power conduction path between the other two electrodes (Power and Output). A quantum replacement device must have an equivalent function.

- fan-out — output can supply sufficient current for multiple inputs,
- isolation of input from output — output voltage does not affect device operation,
- inversion — output varies oppositely to input, and
- well-defined logic (voltage) levels for I/O signals (strong device non-linearity).

Given these general characteristics, two classes of quantum devices are considered below in terms their potential to replace a conventional digital switching device.

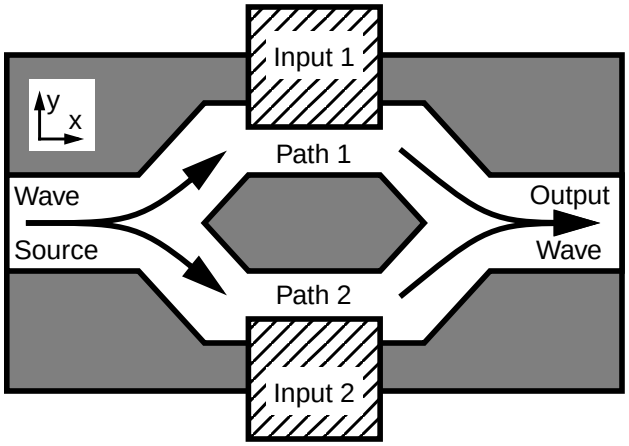
2.3.2 Quasi-Equilibrium Devices

Quasi-equilibrium quantum devices use low resistance waveguides and interference to implement switching. These devices work best (or only) with very low biases along the output waveguide path [9], which is the origin of the designation “quasi-equilibrium”. The prototype quasi-equilibrium device is the quantum interference transistor (QUIT) [31]. The QUIT (see Figure 2.4) has an analogous device in optical systems called the Mach-Zender interferometer [32]. In the QUIT, a quantum wavefunction splits, travels losslessly along two (or more) paths, and then rejoins to interfere constructively or destructively. Along one or both paths, the carrier wavelength can be purposely altered such that the waves recombine either in-phase (constructive interference), which results in high transmission, or out-of-phase (destructive interference), which results in low transmission.

From Equation (2.1), note that wavelength is related to the position-dependent potential energy U. Further, since propagation along either path of the QUIT is to be lossless,

---

6. It must be emphasized that these characteristics are relevant only for the hybrid classical-quantum devices, in which dissipation is allowed. In the ultimate, discrete-quanta devices, some of these characteristics are either not necessary or not desirable.



QUIT Gate Truth Table

Input 1	Input 2	Output
0	0	1
0	1	0
1	0	0
1	1	1

Figure 2.4: Quantum interference transistor XNOR gate

A symmetric QUIT exhibits constructive interference (high conductivity) for the same bias on both waveguides, but can be designed to produce destructive interference (low conductivity) if the inputs differ.

the total carrier energy E is constant. Therefore, a potential energy difference between the two paths (e.g., produced by applying a voltage between them) will result in a difference in wavelength and total phase change along the paths. Figure 2.4 shows a possible (XNOR) logic gate based on the QUIT [9]. Here, the device is designed to produce a 180˚ phase shift on the carrier wavefunction between a logic 0 and a logic 1 input (gate) voltage.

As a discrete device, the QUIT has long been touted as having an extremely low power-delay product [31], a good measure of switching efficiency. However, the issue is whether such a quasi-equilibrium quantum device can replace conventional digital logic in a densely integrated circuit. As their name indicates, quasi-equilibrium devices require very low applied biases to operate properly. In the case of the QUIT, carrier heating resulting from a potential drop along the waveguide paths increases inelastic scattering, so that interference effects begin to fade. Device function is also significantly affected by any *change* in the potential at the output. If the output potential changes (e.g., due to a load change), the phase difference between the two paths, and thus the interference result, also changes. Other problems of quasi-equilibrium devices are that they do not exhibit gain and require cryogenic operation [10]. Thus, quasi-equilibrium quantum devices do not appear to be suitable as a direct quantum replacement for conventional transistors.

2.3.3 Far-From-Equilibrium Devices

If there is to be a direct quantum enhancement of, or substitute for, conventional electronic switches, it will come from the far-from-equilibrium class of quantum devices. These devices use tunneling, quantum wells, and superlattices to achieve higher operating voltages, albeit with resulting energy dissipation. The prototype far-from-equilibrium quantum device is the resonant tunneling diode (RTD) [33]. Its basic energy band diagram and I-V characteristic are shown in Figure 2.5. Note that this device is essentially a quantum well defined by tunnel barriers. The RTD has high current (resonance) when an allowed energy state in the quantum well lines up with the band minimum in the emitter electrode. For details on RTD operation, see [34].

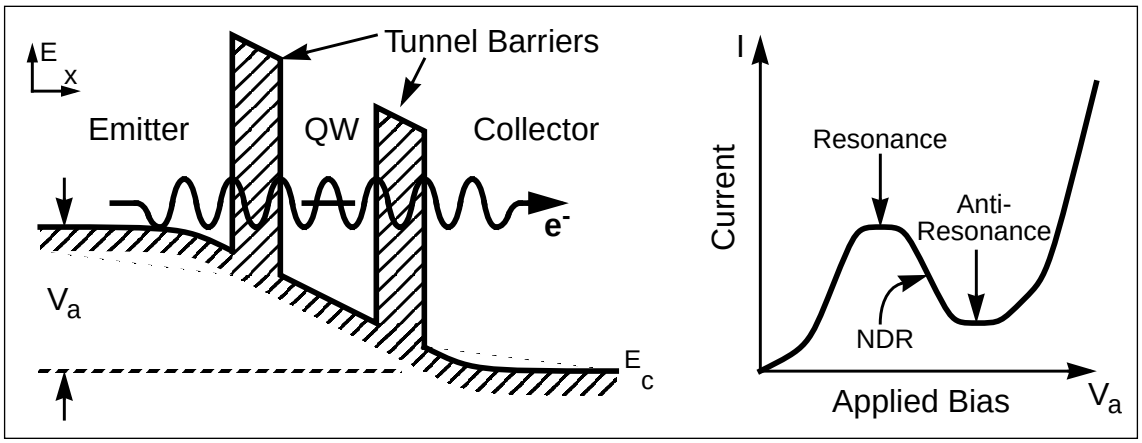


Figure 2.5: Resonant tunneling diode structure and I-V curve

Current is at a maximum (resonance) when the quantum well state lines up with the emitter minimum, so that electrons in the emitter (which are at energies near the minimum) can tunnel through the quantum well state.

Again the question is whether far-from-equilibrium quantum devices can potentially enhance or replace conventional transistors. This time, the answer appears to be “yes”. Resonant tunneling transistors have been proposed by contacting (either directly or through an insulator) the RTD quantum well. In fact, resonant tunneling transistors of various designs have actually been fabricated for proposed use as frequency multipliers, parity generators, multi-state memory, and A-to-D converters [35]. There are still difficulties with the successful implementation of this technology [9], though they appear less fundamental than the limitations of quasi-equilibrium quantum devices.

2.3.4 Prototype Quantum Electronic Device

Based on the above analysis, the RTD was used as the prototype quantum device for directing the development, testing the features, and benchmarking the performance of SQUADS, as described in the remainder of this dissertation. The RTD was chosen mainly in deference to the short-term focus: RTD-based devices are not totally different from conventional electronic devices - they support relatively large applied biases, they have I/O signals which are currents and voltages (rather than quantum waves which must somehow be converted), and quite useful room temperature operation has even been demonstrated [10]. In fact, in comparison to the discrete-quanta devices discussed in Section 2.1.3, RTD-based devices are just a small first step from conventional electronic devices towards the ultimate quantum devices and true quantum computing systems.

In addition to the RTD's similarity to conventional electronic devices as described above, the RTD has much to recommend it for its prominent role in the development of SQUADS. The general difficulty or impossibility of conducting experimental quantum electronics research slows both theory and simulation efforts, since each approach builds on the results, both successes and failures, of the others. Therefore, the second most important feature of RTDs is simply the fact that they can be fabricated with existing technology. The result is that RTD-based devices have been widely studied experimentally [33, 35]. The availability of experimental measurements has been very beneficial in the development of SQUADS, enabling the development of a more accurate simulation tool. The symbiotic relationship between simulation and experiment in this work is clear. On one hand, simulations indicate which devices, materials, dimensions, doping, etc., are promising, and they describe the "ideal" operation to which real devices should aspire. On the other hand, experiments indicate invalid simplifying assumptions in the simulation models, the numerical accuracy required, and secondary effects (e.g., scattering) that are (or are not) important for an accurate model. When an effect can be safely neglected, the simulation will be more efficient, but effects which are inappropriately ignored in a simulation or not implemented in the simulator result in an inaccurate prediction. Later chapters of this dissertation include comparisons between simulations and experimental measurements.

Another key feature of the RTD is the fact that it is the simplest structure in which both quantum tunneling and spatial quantization (due to quantum self-interference) can domi-

nate device operation, even at room temperature. As a result, the RTD has been widely studied not only experimentally, but also analytically and numerically, so that its physics is now well understood. Finally, quantum devices are intended for very high speed applications, and the RTD is a very fast device [36]. All of these facts made very easy the choice of the RTD as the prototype device for the development of SQUADS.

For the development of a general quantum device simulator, the device characteristics that the simulator must handle should be enumerated independently of the description of any test device. This requires essentially a summary of what has been deduced above about potentially useful quantum devices, based on the short-term approach in this work. Thus, SQUADS must simulate devices which are:

- unipolar (no bipolar effects such as recombination-generation),
- heterojunction-based (abrupt band offsets),
- far-from-equilibrium (non-linear, self-consistent band-bending),
- irreversible (scattering), and
- very high-speed (transient).

A reasonable case can be made for the implementation of two-carrier simulation in SQUADS, since some proposed resonant tunneling transistors and most quantum well laser diodes are bipolar. In these devices, both carrier types, both the conduction and valence bands, recombination-generation, and photon effects (laser diodes only) must be treated. Such bipolar effects are not currently handled by SQUADS, but their implementation in SQUADS is certainly feasible.

2.4 Quantum Electronic and Quantum Computing Systems

This section discusses the challenges and possibilities of integrating quantum devices into highly-functional quantum systems, and how the ultimate quantum computer might operate. These aspects of quantum-effect computing go beyond what is strictly necessary for the purposes of quantum electronic device simulator development. A prototype quantum device has already been selected for testing SQUADS, and the device characteristics which are essential for SQUADS to handle, and those which are *not*, have already been determined. The purpose in analyzing the challenges and possibilities of complete quantum-effect (including quantum electronic and true quantum) computing systems is to show where quantum electronic device simulation research fits in the progress towards the real-

ization of quantum-effect computing.

2.4.1 Architecture Challenges and Conclusions

In Section 1.1, the conclusion was that the only way to scale ULSI electronic devices beyond the quantum challenge (the inevitable increase of quantum effects) was to adopt quantum phenomena as the operating mechanisms of the smaller devices. Quantum-effect computing was advocated as a revolutionary approach to advancing electronics into the quantum regime. But in previous sections of this chapter, a short-term, *evolutionary* approach to using quantum phenomena in electronics was actually adopted, by looking for quantum devices that could directly replace conventional electronic devices (in the same circuit architectures). Note that the quantum challenge is not the only problem facing the advancement of ULSI as it pushes toward higher integration densities. Other serious problems include interconnect-dominated delay and scaling limits, increasing cross-talk, increasing hard and soft-errors, and inefficient architectures. If quantum integrated circuits don't remove or at least mitigate *each* of these problems, the potential benefits of using quantum devices over conventional devices will be small or non-existent. In other words, to significantly improve on ULSI, it will not be sufficient to simply replace conventional devices with quantum devices.

The following is a list of other problems and challenges faced by ULSI, and the resulting characteristics that computing systems must have to achieve quantum scale integration [12, 37, 38].

- Interconnect scaling and delay limits are serious and increasing.
  - Limited-interconnect architectures will be mandatory.
  - Function per interconnect must increase.
  - There will be many basic logic units, rather than a single universal switch.
- Device and interconnect cross-talk increases as device spacing decreases.
  - Isolated, functionally independent devices will be impossible.
  - Inter-device coupling must be used for communication.
  - Circuit operation will be governed by distributed computation and collective modes of behavior.
- Hard and soft error probability will tend to increase.
  - Fault-tolerant architectures will be necessary.

- Low parallelism limits speed in conventional architectures.
- High levels of computational parallelism are required.

In discussing quantum-effect computing architectures and operation, it is difficult to separate any of these issues and their implications from the others. Therefore, the discussion is presented in reverse, by simply describing a viable quantum electronic architecture and how it might operate, and then presenting the reasoning behind these speculations. The results are as follows. The quantum-effect computer will have a hierarchical architecture. On the large scale, it may look much like a conventional ULSI circuit, with relatively conventional interconnections between what appear to be single devices. But these "devices" are actually quantum "sub-circuits" in themselves, each producing a very complex function compared to the simple switch that they essentially replace from conventional circuits. The quantum sub-circuits must use a limited interconnect architecture, for which two options have been proposed: cellular automaton arrays and quantum wave filters. Each of these options has its own advantages and challenges, which are explored below. These architectures are even more mandatory as the sub-circuits scale down from the quantum electronic devices that are the focus of this work to true quantum computers (Section 2.1.3).

2.4.2 The Cellular Automaton Architecture

The interconnect challenge is perhaps even more serious than the quantum challenge to the advancement of electronics. Unless quantum-effect computers uses limited-interconnect architectures, it will not be possible to improve substantially, if at all, on the integration limits of conventional electronics. In the upper limit of integration density, only something like a nearest neighbor interconnection scheme is possible. An array of digital devices interacting with nearest-neighbors essentially describes the cellular automaton architecture [12]. It has been shown that, in theory, a cellular automaton array can be created to produce any desired digital function [12]. For example, Figure 2.6 shows a small inhomogeneous (differing interaction rules) 2-D cellular automaton array. 3-D arrays are also possible.

In a quantum-effect computer, because of the minute size of individual devices, it might seem that a quantum cellular automaton array (QCAA) must have periphery-only access - inputs must be supplied, and outputs monitored, only at the periphery of the cir-

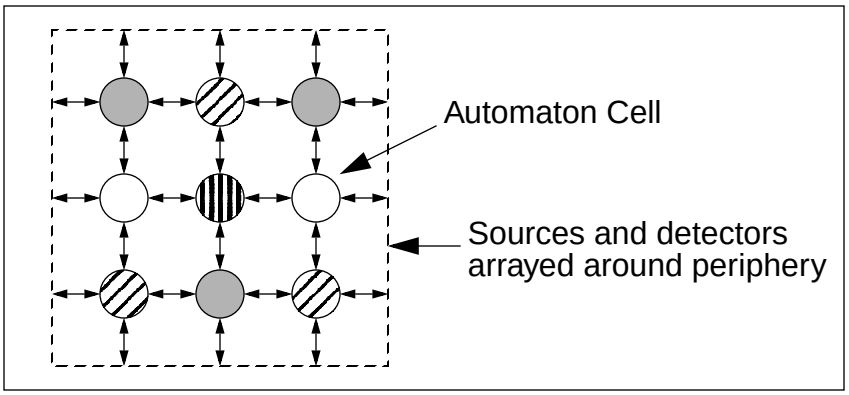


Figure 2.6: Inhomogeneous 2-D cellular automata array

Diagram is meant to indicate nearest-neighbor interaction of a 2-D array of cellular automata with differing interaction rules.

cuit. However, it may be possible to supply inputs optically to interior devices [22]. Capacitive coupling has been proposed for communication between nearest-neighbors within a QCAA [9], and tunneling might also be used. Both would eliminate all physical interconnects within the array. However, they are relatively weak (high attenuation) interactions. That is, signals (voltages for capacitive coupling, quantum waves for tunneling) would be seriously attenuated in the distance of very few cells. Thus, it is not clear how the effect of a peripheral input signal could successfully cascade through the array without external power contacts to the array interior. A single global contact above or below a 2-D QCAA offers some hope of accomplishing signal restoration, though a mechanism for doing so must be found. Alternatively, photon emission and absorption might allow a more lossless interaction between devices.

Note that the QCAA architecture inherently solves the cross-talk problem of conventional electronics to some extent. Cellular automaton array operation is *based* on nearest-neighbor interaction, using “cross-talk” to communicate between devices. In contrast, in conventional electronics, cross-talk is detrimental to proper circuit operation, and must be suppressed using adequate device isolation. Even in the QCAA circuit, the interaction must be limited, since devices must be isolated enough to maintain their own digital state, and much of the circuit function is accomplished through the design of the interaction rules between devices.

Finding some way to adequately isolate quantum-sized cells in a QCAA seems quite a challenge, especially since cell operation is based on waves, which are inherently difficult

to confine. The isolation between individual cells in the QCAA might be accomplished with heterojunctions perhaps not unlike grain boundaries in polycrystalline semiconductors. Finally, the cellular automaton architecture goes a long way towards solving the other challenges of conventional ULSI architectures as well. It can be stabilized against noise ("fail-soft") [37], and its inherent distributed processing, perhaps combined with the redundancy that is feasible with quantum-scale devices, can provide fault tolerance from single-cell hardware failures.

2.4.3 The Quantum Filter Architecture

An alternative to the quantum cellular automaton array for the quantum sub-circuits is the quantum wave filter (QWF). Externally, the QWF looks very similar to the QCAA, with periphery-only access. However, internally, the QWF would not have distinguishable, independent digital devices, but rather should be viewed as a single, extended, complex heterojunction structure designed to perform some desired transformation of the quantum wave. This quantum wave filter concept is at the same time more radical and more promising than the QCAA as a basis for a viable quantum-effect computer architecture.

To visualize how computing might be accomplished with a QWF, consider the quantum wavefunction of a carrier propagating through a quantum-scale system. Quantum effects that this carrier experiences are the result of its wavefunction interfering with (*i.e.*, scattering elastically off) the local energy bands and with itself as it propagates through the system. Each such interaction results in some predictable (via the Schrödinger equation) transformation of the carrier's wavefunction. By selectively combining many of these interactions, any desired total transformation of the wavefunction can (at least in principle) be generated. This kind of signal processing is analogous to optical computing [1], and so is not a purely theoretical concept.

Following this idea further, imagine implementing a binary adder as a QWF. There would be one input waveguide for each bit, and a wavefunction is sent down each waveguide that has a logic-1 input. Inside the QWF, the input waves interfere constructively and destructively, as in the quantum interference transistor, such that the quantum waves resulting on the output waveguides give the proper sum bit pattern for the existing input bit pattern. Note that interference (summation of two 180˚ out-of-phase waves) is essentially an XOR operation. NOT can be implemented simply as a 180˚ phase shift.

Also, the summation of in-phase waves is like an OR logic function. Finally, a $NOT^{1/2}$ (90˚ phase shift) can be created, so that two successive operations give NOT, although the usefulness of this logic function is not manifest. Other logic functions may be possible as well. In fact, the AND, OR, NAND, and NOR logic functions have been demonstrated in optical computing systems [39]. In any case, the QWF should be function-complete, so that it should be able to produce any desired logic function. Because the phase of the wave is very important to the proper function of the QWF, both the order and the locations of the interference junctions are significant. Note that the QWF vindicates quasi-equilibrium quantum devices (such as the quantum interference transistor) that were dismissed in favor of far-from-equilibrium quantum devices (such the RTD) based on the short-term approach of this work. In the long-term, quasi-equilibrium devices will undoubtedly be king of the quantum hill.

Based on the existence and initial success of optical computing, the QWF concept as expressed above appears to have a reasonable chance of success. However, two complications must be treated. First, in order for quantum waves to interfere, they must be phase-coherent. If independent carriers travel down each logic-1 input waveguide, the associated wavefunctions are not phase-coherent, so they would not produce any useful interference result. Since the quantum wave of a single carrier is phase-coherent with itself, one approach might be to just split that single carrier's quantum wave into the required number of "1" bits and send it into the inputs. This also won't work, because if only a single carrier goes in, only a single carrier can be detected at the output, regardless of the number of output bits that should be logic-1. Things get a little better by splitting the quantum wave of many carriers after they have traveled together for a while. Experiments in interfering such multi-carrier wavefunctions show stronger interference than expected assuming independent particles [32], but still not good enough to be used as the basis of a QWF. Thus, it must be admitted that some means of creating a phase-coherent quantum wave is essential. At first, this seems impossible, since the wavefunction, and particularly its phase information, has been thought to be not directly measurable, much less manipulated. However, experiments in electron holography [40] indicate that wavefunction phase can indeed be measured, which means that it may be possible to create a phase coherent multi-carrier quantum wave just as a coherent optical wave is created in a laser.

A second problem with the QWF concept is the large number of inputs and outputs

required. Even though the QWF may perform a very complex function, if it does so in the space of a single conventional electronic device, how can sufficient inputs and outputs be supplied for that function? Few useful and significant computations require only a few variable inputs. Fractal calculations are one example. A more typical case is a 10-bit adder, which requires fully 22 inputs (2 sets of 10 bits, and "power") and 11 outputs (10 bits and a carry). Clearly making over 30 contacts to a device the size of a conventional transistor is unrealistic, since making just three or four is increasingly difficult now. The solution seems to be to serialize the input and output, which should reduce the input count to perhaps 4 (2 inputs, 1 output, and power). Unfortunately, using serialized inputs and outputs might significantly delay the calculation, and requires the development and near-ubiquitous use of some form of quantum shift register.

The QWF architecture shares all of the advantages of the QCAA over conventional electronics (limited-interconnect architecture, high-function "devices", inherently parallel for speed, and distributed for fault-tolerance). It also avoids many of the remaining unknowns and challenges of the QCAA architecture. First, inter-device isolation is totally eliminated, which side-steps the very difficult problem of QCAA of providing some form of engineered of device isolation even at quantum scale integration. In fact, in the QWF, making any distinction between "devices" and "interconnects" (waveguides) is tenuous — both perform a significant operation on the propagating wavefunction.

The absence of isolation in the QWF also solves the related problem of the QCAA, that signal renormalization was needed due to signal attenuation between semi-isolated nearest-neighbor devices which need to communicate. The QWF simply admits that without gain, inter-device isolation is not desirable — the signal should pass unattenuated through the QWF and to the outputs. With the QWF, the input wave will either be transmitted to the output or reflected losslessly back to the input. In fact, ideally no power at all is used in the computation! How is it possible to perform a very complex computation while using zero energy, when a non-zero minimum energy has been derived for even a single switching event [41]? This is one of the amazing features of quantum phenomena: as long as only wave effects are used, no energy is expended. Another way to say this is that no energy is lost as long as no information is lost [41]. Pure quantum wave phenomena lose no information - given a wavefunction at any point in time, and its environment (the QWF, in this case), one can project exactly how the wavefunction will evolve into the

future (or how it looked in the past). Energy is only necessarily expended in supplying inputs and detecting the results of the computation.

2.4.4 General Comments

A complete discussion of quantum-effect computer architecture and operation is beyond the scope of this dissertation. In this overview, only two more issues are considered. First, recall from Section 2.4.1 the (as yet unsubstantiated) prediction that quantum-effect computers will have a hierarchical architecture, rather than being a single quantum circuit. The reason is inelastic scattering, also called the "decoherence problem" [17, 19, 42]. When a quanta scatters inelastically, it is essentially "detected" at that location. In quantum theory, the quantum wavefunction has "collapsed" (it is localized to a point), and the new quantum wave has no continuity or coherence with the past [43]. But coherence and continuity of the quantum wavefunction are essential for proper operation of wave phenomena. Therefore, predictable wave-based operation of quantum devices requires the bulk of the quantum circuit to be essentially free of inelastic scattering. This in turn essentially means that the quantum circuit must be smaller than the average distance between scattering events, which is called the elastic mean free path. Luckily, the elastic mean free path can be up to many tens of microns [25], though a few microns is more attainable. Of course, limiting the total integrated circuit size in a quantum circuit to a few microns is undesirable. A hierarchical architecture is the obvious solution, where the few-micron-sized quantum sub-circuits would be interconnected more-or-less conventionally in the larger quantum integrated circuit [12].

Second, consider the implications of the fact that wave phenomena are inherently *analog* in nature, but a replacement for *digital* electronics is being sought. Because wave effects are distributed, they usually can not, even in theory, produce a perfect digital result in any interference event. As a result, it may turn out that quantum-effect computers will only excel over conventional computers in "fuzzy" artificial intelligence applications. This application alone would be sufficient motivation to pursue quantum-effect computing. However, the analog nature of quantum wave phenomena should not automatically disqualify quantum-effect computers as the future of digital computing. Although wave processing is inherently analog, *detecting* a quanta is perhaps the only perfect binary event, giving either zero or one, and nothing else. Unfortunately, the exact *location* of detection

of a given quanta is probabilistic. However, a properly constructed quantum computer can, in theory, make the probability arbitrarily close to binary. Note that some "guessing" occurs in all digital systems. For example, voltage *ranges* are specified for acceptable signals in digital logic, maximum likelihood circuitry is now used in hard disk drive read circuitry, etc. In this sense, all computing systems are contrived to act as if they were binary.

2.5 Summary

This chapter summarized the field of quantum-effect computing: the idea of producing useful analog or digital signal processing using electronic devices whose operation is fundamentally based on quantum wave phenomena. Accomplishing digital computation with waves seems at least inefficient, if not impossible. However, initial research with optical computing and discrete quantum systems provides some clues to what may work. With the intent of applying this work in the short-term, the resonant tunneling diode was selected as the prototype quantum device for directing the development, testing the features, and benchmarking the performance of SQUADS. The RTD has many features to recommend it, including simplicity of structure and richness of physics. From the process of choosing this test device, several key features that SQUADS must handle have been specified: abrupt material changes, self-consistency, scattering, and high-speed transient operation.

The discussion in this chapter clearly shows that investigating discrete RTD-based devices is but a tiny step into the quantum realm. However, it is an undeniable and necessary first step. SQUADS was designed to illuminate this step and thereby to help direct future progress. Even with such tools, the approach toward the full realization of quantum electronics, much less true quantum computing, will continue to be difficult. In spite of the considerable effort that has already been focused on the development of quantum electronics as a viable successor to ULSI electronics, many significant barriers stand in the way of realizing this goal:

- conventional circuit architectures will not work,
- conventional computing models may not work,
- quantum physics is unfamiliar and unintuitive,
- quantum device simulators are rudimentary, and
- fabrication capability for quantum devices is also rudimentary.

Perhaps the biggest barrier to the development of quantum-effect computing is skepti-

cism. Based on the many unknowns indicated above, it is not difficult to understand this sentiment. A conceit of each generation is the belief that no significant discoveries remain to be made in physics and technology. This notion has never proved correct, and there is no indication that the pace of science and technology will even slow, much less stop, because of such beliefs. Admittedly, the debate is far from over concerning whether any technology will follow ULSI electronics. If there is a successor technology, the demands of the computing public will eventually force us to find it, and it *will* be based on quantum effects. Perhaps the best techniques available in the quest for quantum-effect computing, then, are the ability to suspend disbelief and to ignore conventional wisdom.

References

[1] T. E. Bell. "Optical computing: a field in flux." *IEEE Spectrum*, pages 34–57, Aug. 1986.

[2] E. Abraham, C. T. Seaton, and S. D. Smith. "The optical computer." *Scientific American*, 248(2):85–93, Feb. 1983.

[3] P. W. Smith and W. J. Tomlinson. "Bistable optical devices promise subpicosecond switching." *IEEE Spectrum*, pages 26–33, June 1981.

[4] B. C. Cole. "Optical ICs: the new alternative." *Electronics*, pages 39–44, 18 Nov. 1985.

[5] D. Bohm. *Quantum Theory*, chapter 3, 11, 12, 21. Prentice-Hall, Englewood Cliffs, NJ, 1963.

[6] T. K. Gaylord and K. F. Brennan. "Electron wave optics in semiconductors." *Journal of Applied Physics*, 65(2):814–820, 1989.

[7] M. C. Li. "Electron interference." *Advances in Electronics and Electron Physics*, 53:269–306, 1980.

[8] F. Capasso. "Band-gap engineering and interface engineering: From graded-gap structures to tunable band discontinuities." In F. Capasso and G. Margaritondo, editors, *Heterojunction Band Discontinuities: Physics and Device Applications*, chapter 10, pages 399–450. Elsevier Science Publishers, New York, 1987.

[9] R. Bate, G. Frazier, W. Frensley, and M. Reed. "An overview of nanoelectronics." *Texas Instruments Technical Journal*, pages 13–18, July-Aug. 1989.

[10] J. N. Randall, M. A. Reed, and G. A. Frazier. "Nanoelectronics: Fanciful physics

or real devices?" *Journal of Vacuum Science and Technology B*, 7(6):1398–1404, 1989.

[11] R. T. Bate. "Nanoelectronics." *Solid State Technology*, 32(11):101–108, 1989.

[12] R. T. Bate, G. A. Frazier, W. R. Frensley, J. W. Lee, and M. A. Reed. "Prospects for quantum integrated circuits." In *Quantum Well and Superlattice Physics*, pages 26–35. Proceedings of the SPIE, Vol. 792, 1987.

[13] F. A. Buot. "Mesoscopic physics and nanoelectronics: Nanoscience and nanotechnology." *Physics Reports*, 234(2-3):73–174, 1993.

[14] K. Hess and G. J. Iafrate. "Approaching the quantum limit." *IEEE Spectrum*, pages 44–49, July 1992.

[15] F. Capasso and s. Datta. "Quantum electron devices." *Physics Today*, 43(2):74–82, 1990.

[16] M. J. Kelly. "Low-dimensional devices: future prospects." *Semiconductor Science and Technology*, 5(12):1209–1214, 1990.

[17] C. H. Bennett. "Quantum information and computation." *Physics Today*, pages 24–30, Oct. 1995.

[18] S. Lloyd. "Quantum-mechanical computers." *Scientific American*, 273(4):140–145, 1995.

[19] D. P. DiVincenzo. "Quantum computation." *Science*, 270:255–261, Oct. 13 1995.

[20] D. Deutsch. "Quantum computation." *Physics World*, pages 57–61, June 1992.

[21] C. H. Bennett, G. Brassard, and A. K. Ekert. "Quantum cryptography." *Scientific American*, 267(4):50–57, Oct. 1992.

[22] S. Lloyd. "A potentially realizable quantum computer." *Science*, 261:1569–1571, Sep. 17 1993.

[23] J. Spector, H. L. Stormer, K. W. Baldwin, L. N. Pfeiffer, and K. W. West. "Electron focusing in two-dimensional systems by means of an electrostatic lens." *Applied Physics Letters*, 56(13):1290–1292, 1990.

[24] D. Bohm. *Quantum Theory*, page 237. Prentice-Hall, Englewood Cliffs, NJ, 1963.

[25] J. Spector, H. L. Stormer, K. W. Baldwin, L. N. Pfeiffer, and K. W. West. "Refractive switch for two-dimensional electrons." *Applied Physics Letters*, 56(24):2433–2435, 1990.

[26] T. K. Gaylord, E. N. Glytsis, and K. F. Brennan. "Semiconductor quantum wells as

electron wave slab waveguides." *Journal of Applied Physics*, 66(4):1842–1848, 1989.

[27] D. Bohm. *Quantum Theory*, page 240. Prentice-Hall, Englewood Cliffs, NJ, 1963.

[28] T. K. Gaylord, E. N. Glytsis, and K. F. Brennan. "Electron-wave quarter-wavelength quantum well impedance transformers between differing energy-gap semiconductors." *Journal of Applied Physics*, 67(5):2623–2630, 1990.

[29] D. D. Coon and E. Sorar. "Design of superlattice termination layers." *Applied Physics Letters*, 56(18):1790–1792, 1990.

[30] R. W. Keyes. "What makes a good computer device." *Science*, 230:138–144, Oct. 11 1985.

[31] S. Bandyopadhyay, M. R. Melloch, S. Datta, B. Das, J. A. Cooper Jr., and M. S. Lundstrom. "A novel quantum interference transistor (QUIT) with extremely low power-delay product and very high transconductance." *International Electron Devices Meeting (IEDM) Technical Digest*, pages 76–79, No. 4.1, 1986.

[32] S. Datta. "Quantum devices." *Superlattices and Microstructures*, 6(1):83–89, 1989.

[33] F. Capasso, K. Mohammed, and A. Y. Cho. "Resonant tunneling through double barriers, perpendicular quantum transport phenomena in superlattices, and their device applications." *Journal of Quantum Electronics*, 22(9):1853–1868, 1986.

[34] S. Luryi. "Coherent versus incoherent resonant tunneling and its implications for fast devices." *Superlattices and Microstructures*, 5(3):375–382, 1989.

[35] F. Capasso, S. Sen, F. Beltram, L. M. Lunardi, A. S. Vengurlekar, P. R. Smith, N. J. Shah, R. K. Malik, and A. Y. Cho. "Quantum functional devices: Resonant-tunneling transistors, circuits with reduced complexity, and multiple-valued logic." *IEEE Transactions on Electron Devices*, 36(10):2065–2082, 1989.

[36] E. R. Brown, J. R. Soderstrom, C. D. Parker, L. J. Mahoney, K. M. Molvar, and T. C. McGill. "Oscillations up to 712 GHz in InAs/AlSb resonant-tunneling diodes." *Applied Physics Letters*, 58(20):2291–2293, 1991.

[37] M. A. Reed. "Quantum semiconductor devices." In *Symposium on VLSI Technology*, pages 1–4, 1986.

[38] R. T. Bate. "The future of microstructure technology - the industry view." *Superlattices and Microstructures*, 2(1):9–11, 1986.

[39] S. D. Smith, A. C. Walker, F. A. P. Tooley, and B. S. Wherrett. “The demonstration of restoring digital optical logic.” *Nature*, 325:27–31, 1 Jan. 1987.

[40] S. Hasegawa, T. Kawasaki, J. Endo, A. Tonomura, Y. Honda, M. Futamoto, K. Yoshida, F. Kugiya, , and M. Koizumi. “Sensitivity-enhanced electron holography and its application to magnetic recording investigations.” *Journal of Applied Physics*, 65(5):2000–2004, 1989.

[41] R. Landauer. “Dissipation and noise immunity in computation and communication.” *Nature*, 335:779–784, 27 Oct. 1988.

[42] W. H. Zurek. “Decoherence and teh transition from quantum to classical.” *Physics Today*, pages 36–44, Oct. 1991.

[43] C. Cohen-Tannoudji, B. Diu, and F. Laloë. *Quantum Mechanics*, volume 1, page 220. John Wiley & Sons, New York, 1977.

Chapter 3

Quantum Device Simulation Approach

Many formulations of quantum mechanics have been proposed as mathematical foundations for numerical simulation of quantum systems. Each has its own strengths and weaknesses. The theoretical investigation of quantum electronics in the previous chapter, and especially the decisions made about the types of the quantum devices to be simulated, allows this chapter to describe the process by which the two formulations on which SQUADS is based were chosen. The choice hinges on which formulations of quantum mechanics offer the best combination of capabilities, accuracy, and computational efficiency in a numerical simulation. In addition, this chapter describes the plan of action and the guiding principles behind SQUADS' development process.

The first step in designing a plan of action for the development of SQUADS is to analyze the state of the field of quantum device simulation, which is done in Section 3.1. Section 3.2 then enumerates specifically the goals of this effort to develop a quantum device simulation capability. Section 3.3 examines conventional electronic device simulation methods to extract any insight and guidance it offers for the development of SQUADS. Next, the various formulations of quantum mechanics are compared in Section 3.4, concluding with the determination of which formulations are best suited to the task of simulating RTD-like quantum devices. Finally, Section 3.5 combines all of this analysis with the discussion of several additional issues to complete the plan of action and list the guiding principles for developing SQUADS.

3.1 State of Quantum Device Simulation

Quantum device simulation work to date has been rather disorganized. The main reason is that there is no software package that provides a base-line of functionality which researchers can use and enhance (for example, like PISCES [1, 2] for conventional device simulation). Luscombe and Frensley have advocated such a tool [3], and Frensley described initial work to realize it in a simulation program called BandProf [4]. However, BandProf has apparently not been mentioned in the literature in the five years since this first description, it was not widely available at the time (only one supported platform), and did not have all of the capabilities necessary for general quantum device simulation (*e.g.*, no transient capability was described). Work by other researchers have generally given no consideration at all to use of their simulation tools by others.

Due to the lack of a widely available and easily extensible quantum device simulation tool, each research team has had to implement the same base-line functionality themselves, before the enhancements of interest could be added and investigated. Of course, this new functionality is not available to anyone else, since each group's quantum device simulation tools have independently-evolved structures and interfaces. All of these facts have made advances in quantum device simulation much slower than necessary. This understanding had a defining impact on the goals and design principles of SQUADS, as described in the remainder of this chapter. As a result, SQUADS may become the first quantum device simulation tool which has the necessary characteristics to serve as a foundation for future quantum device simulation work, and thus to enable much more rapid advancements in the field.

3.2 Goals of Quantum Device Simulation

The overriding goal of this research is to develop a *general* quantum device simulator, which means that this tool should be able to efficiently model quantum device operation in any useful mode of operation. Based on the conclusions and decisions of the previous chapter, this goal can now be clarified. The quantum devices of interest in this work are externally similar to conventional electronic devices, so that the quantum devices can either enhance or replace conventional devices in existing architectures. Therefore, input and output signals of these quantum devices are voltages and currents. The resulting goal of this research is to develop the capability to simulate current flow through quantum

devices, either versus applied bias (current-voltage curve), time (switching response), or both (small-signal behavior).

Having stated this, the main goal of this *chapter* is to explain the choice of formulations of quantum mechanics that serve as the mathematical basis of SQUADS. An essential ingredient in understanding this choice is a statement of the goals of this work. Recall that the goal of quantum electronics for the purposes of this dissertation is to improve upon ULSI electronics by integrating quantum devices at quantum scale integration densities. The ideal progression of knowledge in this endeavor is to build from theory to simulation to experiment. The overview of the theory of quantum electronics in Chapter 2 has already been used to direct this simulation effort. In turn, the main purpose of quantum device simulation is to guide experiment. The one sentence goal of SQUADS, then, is to fill the knowledge gap between idealized quantum theory and highly expensive quantum experiment. The following list enumerates in more detail the ways in which SQUADS can accomplish this:

- SQUADS can do things which are not experimentally possible:
  - view the internal operation of a device,
  - view entities not easily measurable (e.g., a quantum wavefunction),
  - vary physical parameters/models independent of real materials and systems,
  - investigate systems not possible or feasible to produce experimentally, and
  - avoid experimental variation/uncertainty and measurement error.
- SQUADS can serve as an inexpensive substitute to experiment:
  - orders of magnitude faster and less expensive,
  - reduces the number of experimental iterations necessary,
  - mistakes are not costly - simply correct and re-run,
  - risk/benefit trade-off disappears - complete freedom to pursue new ideas, and
  - allows back-tracking or branching from an intermediate point at will.
- SQUADS can also serve as an adjunct to experiment:
  - replication of experiment to derive additional understanding,
  - verification of physical models and mechanisms, and
  - highlights non-ideal device operation.

Based on these points, several design goals for SQUADS can be stated. Most importantly for the choice of simulation methods, SQUADS should be a *general* device simula-

tor (i.e., not tied to a single device). Also, it should provide an internal view of the device state and operation, since, new ideas for quantum electronics will undoubtedly come from insight gained by "watching" the internal operation of quantum devices. This is information that experiment can't provide: experiment only gives aggregate terminal values (voltages and currents at contacts). The additional information is especially important when investigating new situations (such as the quantum realm) or designing new devices where the operation can't be extrapolated from previous results, or isn't understood based on the experimental results. In such cases, watching the internal operation is like turning the light on in a dark room or opening up a black box: reverse-engineering or guessing what's going on in the device is unnecessary - the device's operation physics becomes manifest.

The increasing importance of simulation with respect to experiment has been a steady trend in electronics research. Early electronics research was necessarily an experimental undertaking — computers of the time were not capable of *doing* simulations. As semiconductor devices have decreased in size and improved in performance, the cost of using experimental iteration for device development has increased dramatically. However, each generation of faster computers has made it possible and always more necessary to improve simulation tools and use them to a greater extent in research and development for the next generation of devices. Quantum electronics research has simply brought this trend much closer to completion. This change is due partly to the greatly increased power of computers to simulate physical systems, but more importantly in this case to the fact that numerical simulation can provide a detailed viewport into a world that is otherwise largely inaccessible: the quantum realm.

Despite the importance that simulation will play in quantum electronics research, the ultimate goal is to actually build useful quantum devices, circuits, and quantum electronic systems. Of course, reality will be different than simulation, so experimental results will direct simulator development, as discussed in Section 2.3.4. Only *after* a simulator is sufficiently developed does the influence arrow begin to reverse, so that simulations direct experiment. The importance of comparing to experimental results during simulator development can not be overstated, because experiment is the fire in which a simulator is tested. The process of simulator refinement and experimental verification is iterative. In fact, comparisons between simulations and experimental data in later chapters show examples of where SQUADS' development was partly directed by experimental results.

Another consideration that will direct the choice of mathematical basis for SQUADS is the desire to be able to link SQUADS with conventional electronic device simulators. Quantum effects are an increasing "nuisance" in shrinking conventional electronic devices, and will be unavoidable in the future. A quantum device simulator coupled with a conventional simulator can be used to investigate this, and thereby hopefully maintain reliable conventional device operation in spite of these "parasitic" quantum effects. These requirements are in addition to the one that SQUADS must be able to handle device characteristics that are considered important, as listed in Section 2.3.4.

3.3 Classical Electronic Device Simulation

Before considering possible quantum simulation methods, one final source of information can help direct the choice of quantum mechanics formulation as a basis for SQUADS: conventional electronic device simulation. Tools for conventional device simulation have been refined over many years so that they have exactly the qualities desired for SQUADS: accuracy and efficiency. Further, because a short-term focus has been chosen for this research, Chapter 2 concluded that the quantum devices SQUADS must handle are similar in many ways to conventional electronic devices. Thus, many of the characteristics and capabilities of classical simulation should be mirrored in SQUADS. The purpose of this section, then, is to analyze classical device simulation to further inform the choice of the best quantum device simulation method.

3.3.1 The Boltzmann Transport Equation

Conventional electronic device simulation is based on the Boltzmann transport equation (BTE) [5, 6]. The BTE specifies the evolution of the classical distribution function, $f_c(\boldsymbol{r}, \boldsymbol{p}, t)$, which is the density of carriers[1] at $(\boldsymbol{r}, \boldsymbol{p}, t)$, where $\boldsymbol{r}$ is the three-coordinate position, $\boldsymbol{p}$ is the three-coordinate momentum, and t is time. The fact that f_c is a phase-space function[2] is key to its usefulness as a basis for conventional device simulation. The BTE can be written

$$\frac{\partial f_c}{\partial t} + \left(\frac{\partial \boldsymbol{r}}{\partial t}\right)\frac{\partial f_c}{\partial \boldsymbol{r}} + \left(\frac{\partial \boldsymbol{p}}{\partial t}\right)\frac{\partial f_c}{\partial \boldsymbol{p}} = \left(\frac{\partial f_c}{\partial t}\right)_{\text{collision}}, \tag{3.1}$$

1. The number of carriers within the spatial volume $\delta\boldsymbol{r}$ and the momentum volume $\delta\boldsymbol{p}$.
2. A function of position and momentum (or velocity or wavevector).

which is often written as

$$\frac{\partial f_c}{\partial t} + \underbrace{\boldsymbol{v}\frac{\partial f_c}{\partial \boldsymbol{r}}}_{\text{diffusion}} + \underbrace{\boldsymbol{F}\frac{\partial f_c}{\partial \boldsymbol{p}}}_{\text{drift}} = \left(\frac{\partial f_c}{\partial t}\right)_{\text{collision}} . \tag{3.2}$$

In (3.2), $\boldsymbol{v}$ is the carrier velocity, and $\boldsymbol{F}$ is the force on the carriers.

The BTE would be very compute intensive if implemented in more than 1-D. Thus, simplified models of carrier transport which derive from the BTE are often more familiar. The hydrodynamic model [7] results after moderate simplifications of the BTE. Even more widely used is the “first order” drift-diffusion model [8]. For electrons, where $n = n(\boldsymbol{r}, t)$ is the electron concentration, this model is written as:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \bullet \boldsymbol{J}_n - U_n , \tag{3.3}$$

$$\boldsymbol{J}_n = q\mu_n n E_n + qD_n \nabla n . \tag{3.4}$$

Here, q is the electronic charge, $\boldsymbol{J}_n$ is the electron current density, U_n is the net electron recombination rate, μ_n is the electron mobility, and D_n is the electron diffusion constant. Similar equations hold for holes.

3.3.2 Strengths of the BTE

Now consider why the BTE and its simplifications are ideal for conventional device simulation, and thus what is desirable in a quantum device simulation formulation. First, simulation based on the BTE has all of the characteristics specified so far for SQUADS: it is a general formulation[3], it permits an internal view of device operation, and it can handle all of the device characteristics listed in Section 2.3.4, particularly, far-from-equilibrium, irreversible, and transient. Another critical feature of the BTE is that it lends itself to various levels of simplification, to allow either 1-D simulation with the BTE itself, or to make multi-dimensional simulation feasible using the hydrodynamic and drift-diffusion models. Further, by recasting the BTE in a path-integral (a.k.a., Monte-Carlo) form, even 3-D systems can be feasibly handled while still including the full complexity of the BTE [9].

The main reason the BTE has these features is that its state function f_c is a phase-space distribution function. As such, f_c contains all of the information of interest about the carriers, including position information (to calculate carrier densities) and velocity information

3. It is not limited to any single device or class of devices

(to calculate currents). For simulation of electronic computing devices, a phase-space distribution function is the most natural and efficient, yet complete, way to describe this information. Both the BTE and its classical distribution function are intuitive. They do not require opaque interpretation to understand. The results needed (e.g., carrier densities and currents) are easily and transparently calculated.

3.3.3 Implications for Quantum Device Simulation

All of these attributes of conventional simulation should be retained in choosing a quantum simulation method. Section 3.4 shows that some of the quantum transport formulations do not give the benefits of the BTE. Of course, the BTE itself can not be used for quantum system simulation because it is based on a classical, rather than quantum, formulation of physics. In particular, the BTE assumes that carriers obey the classical Newton's laws: they are point particles with a single momentum, they experience forces of a single value at a single location, and collisions are instantaneous [10]. These assumptions are what allow the use of a phase-space distribution function to describe the carrier distributions, the many benefits of which have already been described. Unfortunately, none of these assumptions and results are true at quantum dimensions, since particles on the quantum scale act like waves. The Heisenberg uncertainty principle declares that a particle's position and momentum can not be precisely known simultaneously. In fact, a particle usually does not have a single position or momentum value, but rather a *distribution* of such values. The result is that it is impossible to have a phase-space distribution function in a quantum mechanical representation of a system. This is very unfortunate, since *some* quantum formulation of carrier transport must be used, but not the phase-space state function that has made the BTE so ideal for classical electronic system simulation.

All of the quantum transport formulations considered below use *some* representation of the state of the system (e.g., the wavefunction $\Psi(\boldsymbol{r}, t)$ in the Schrödinger representation). Any of these representations can be used to determine the information needed for device simulation (carrier densities and velocities). The mathematical basis for SQUADS will be the formulation of quantum mechanics that gives this information most easily. This essentially means that the representation of the chosen formulation should be intuitive, like f_c, so that its interpretation is not opaque or convoluted. It is the purpose of the next section to determine which formulation of quantum mechanics has this quality through an

analysis of the various possible approaches to quantum device simulation.

3.4 Quantum Transport Formulations

This section finally answers the question of which of the many formulations of quantum mechanics (FQM) are best suited to quantum device simulation, and therefore will be used as a foundation for SQUADS. Due to the increasing interest in investigating quantum effects in electronic devices, several other researchers have also recently considered this question (see, *e.g.*, [11-13]). The set of FQMs which are considered appropriate for quantum device simulation is dynamic, and new formulations will undoubtedly be added over time, while others may even be dropped from the set as no longer competitive. The analysis in this section should therefore be considered a snap-shot of the current state of quantum device simulation. Future research will undoubtedly extend the capabilities and surmount the unique challenges of using some of the FQMs as a basis for numerical simulation. In fact, this research contributes to that dynamic.

3.4.1 Relationships Between Candidate Formulations

Figure 3.1 shows schematically the relationships between those FQMs which, at present, are most widely proposed, discussed, and/or used for quantum device simulation. These include the Schrödinger equation, transfer-matrix, density matrix, Green's functions, Wigner function, and path integral approaches. The respective state function[4] of each FQM is also shown. Based on the intention to simulate RTD-like devices, and the analysis of conventional device simulation in Section 3.3, several characteristics and features that SQUADS must have been specified. These include the use of an intuitive state function (for an internal view of device operation) and the ability to handle far-from-equilibrium, irreversible, transient, and open systems. Most of these requirements were already taken into account in generating Figure 3.1, because only those FQMs appropriate to quantum device simulation are shown.

FQMs not shown in Figure 3.1 include the force-force correlation function [14] and current-current correlation function [15], which are near-equilibrium, linear response analyses which do not use state functions; the Heisenberg matrix mechanics approach [16], which is mathematically equivalent to the Schrödinger equation, but is less intuitive; the

4. A function which describes the state (*e.g.*, position and velocity) of the carriers in the system.

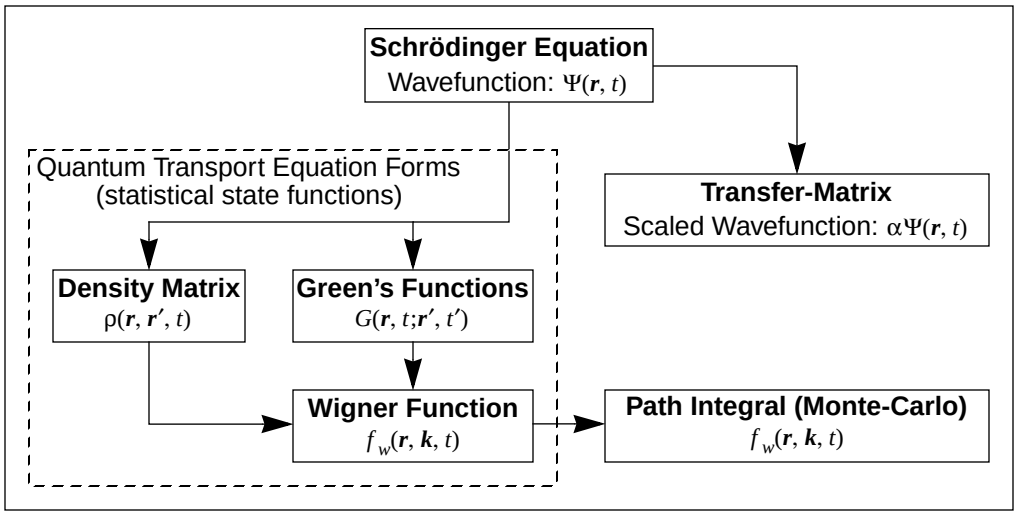


Figure 3.1: Family tree of relevant formulations of quantum mechanics

This flow chart shows the relationships between the main formulations of quantum mechanics that have been employed for electronic device simulation, as well as their respective state functions. Many other formulations of quantum mechanics are not shown.

related Langevin equation approach [17], which is too complex to solve except in the simplest cases; the scattering matrix approach [18, 19], which is similar to the transfer matrix approach included in Figure 3.1, but which includes scattering in a less satisfactory way than, for example, the Wigner function approach (also shown in Figure 3.1); and many others. The approaches to analyzing quantum systems, each of which has its own uses, are virtually innumerable.

3.4.2 Analysis of Formulations

This section compares the capabilities and characteristics of the six formulations of quantum mechanics shown in Figure 3.1, first with a summary of the comparison, and then with a more detailed discussion of the analysis underlying the summary.

3.4.2.1 Summary

To make a final choice of the best formulation for use in SQUADS, Table 3.1 grades each formulation in Figure 3.1 according to its ability to meet the requirements listed in the Section 3.4.1. The BTE is also listed in the table to show its excellent capabilities for simulating electronic devices dominated by classical mechanics, which capabilities will

ideally be mirrored in SQUADS. Because the list of FQMs has already been limited to those which meet most or all of the known requirements, the choice of the best approach for SQUADS will hinge on other more practical criteria. Thus, Table 3.1 also rates each formulation for its relative computational efficiency and the simplicity of interpretation (richness in intuitive information) of its state function. Based on all of the data, Table 3.1 prompts the conclusion that the Wigner function formulation is optimal for the simulation of RTD-type quantum electronic devices. For this purpose, it has all of the features and capabilities required, as well as an intuitive state function and acceptable computational complexity. An explanation and justification of the results of Table 3.1 for the six quantum mechanics formulations included in Table 3.1 is given in the following sections.

Table 3.1: Comparison of quantum system analysis approaches

Characteristic (5=good; 1 = poor)	Method of Quantum System Analysis						
	BTE	SE	TM	DM	GF	WF	PI
State Function-Based	yes	yes	yes	yes	yes	yes	yes
Far-From-Equilibrium	yes	yes	yes	yes	yes	yes	yes
Irreversibility (Scattering)	yes	no	no	yes	yes	yes	yes
Transient Simulation	yes	yes	no	yes	yes	yes	yes
Absorbing Boundaries	yes	no	yes	yes	yes	yes	yes
Computational Efficiency	3	4	4	3	2	3	2
Intuitive State Function	5	3	4	3	2	4	4
Suitability for SQUADS	1	3	3	3	2	4	3

3.4.2.2 The Schrödinger Equation

Most uses of the Schrödinger equation for quantum system simulation are based on a scaled single-particle wavefunction as the state function, since the exact many-particle wavefunction becomes unmanageably complex with more than just a few carriers. As indicated by Table 3.1, two features of quantum electronic devices have not been accurately treated with the Schrödinger equation approach: absorbing boundary conditions (*i.e.*, ohmic contacts) in a transient simulation, and inelastic scattering. Interestingly,

although the Schrödinger equation was first postulated fully 70 years ago [20], the increasing need for accurate numerical simulation of quantum devices has fueled recent advances in these areas [21-26]. Although the Schrödinger equation may have a brighter future for quantum device simulation, these limitations made it impossible to base a comprehensive quantum device simulator on this formulation of quantum mechanics.

3.4.2.3 The Transfer-Matrix Method

One way to surmount the absorbing boundary problem of the Schrödinger equation is to solve the equation in steady-state. The simplest result is the transfer-matrix method (TMM) [27-31], the most widely used method of quantum device simulation to date.[5] The TMM is based on the assumptions that particles enter and exit the system as continuous streams (beams) with amplitudes given by the fixed boundary conditions, that a particle beam entering at a given energy is perfectly phase-coherent, and that particle beams at different energies do not interact. The result is a state function for each particle beam which is simply a scaled, steady-state, single-particle wavefunction. The popularity of the TMM is due to its simplicity (in both theory and programming), and the relatively low computational requirements. However, because it is based directly on the Schrödinger equation, the TMM also can not handle irreversibility (inelastic scattering). Further, because *continuous* particle beams are assumed throughout the system, transient simulations are difficult or impossible to implement using the TMM.

3.4.2.4 The Quantum Transport Equation Approaches

The density matrix, Green function, and Wigner function formulations in Table 3.1 employ a statistical state function, rather than the exact many-quanta wavefunction (as with the Schrödinger equation) or one wavefunction per energy (as with the TMM). Quantum statistical mechanics does not attempt to retain all information about the evolutions and interactions of perhaps millions of distinct quanta, but rather deals with continuous *distributions* of particles and interactions, just as the classical distribution function does in the BTE. A statistical state function is thus a natural and efficient way to model particles in a many-body system [32]. A statistical approach should also be quite accurate — the myr-

5. This work differentiates between quantum *systems*, which are more commonly investigated via the Schrödinger equation, and quantum *devices*, which must have open, absorbing boundaries.

iad single-particle wavefunctions intermingle so completely that they can not be distinguished anyway. The statistical state function is usually formed by assuming total particle independence (the one-particle approximation). Several such state functions have been found to be useful, and the name of each formulation is given by the name of the particular state function employed. Each of these formulations also has its respective, but related, quantum transport equation (QTE), which specify how the state function evolves with time. As indicated in Table 3.1, each of the QTE approaches is able to handle all of the quantum device characteristics of interest, just as the BTE could for conventional electronic device simulation.

The main practical difference between the three QTE approaches is in their respective state functions, and this is where the Wigner function method (WFM) of quantum device simulation achieves its greatest superiority. The Wigner function [33], denoted f_w in this chapter, is a real-valued[6], phase-space[7] state function, just like the classical distribution function f_c in the BTE. Recall the argument in Section 3.3 that having a phase-space distribution function in quantum mechanics is impossible, because the Heisenberg uncertainty principle dictates that position and momentum information can not be known exactly simultaneously. In fact, the Wigner function is *not* a true phase-space *distribution* function, specifying the density of carriers at each position and velocity - if it were, f_w would be exactly equal to f_c! Rather the Wigner function is called a *quasi*-distribution function. Far away from quantum structures, f_w *is* equal to f_c. Where quantum effects are significant, the f_w must reflect these phenomena, and thus it will differ from f_c. However, all observables (e.g., carrier densities and current) are calculated from f_w just as from f_c. In fact, the Wigner function transport equation (WFTE), which in 3-D can be written[8]

$$\frac{\partial f_w}{\partial t} + \left(\frac{\hbar \boldsymbol{k}}{m^*}\right)\frac{\partial f_w}{\partial \boldsymbol{r}} + \frac{1}{\hbar}\iiint \frac{d\boldsymbol{k}'}{2\pi}\left(V(\boldsymbol{r}, \boldsymbol{k}-\boldsymbol{k}')\, f_w(\boldsymbol{r}, \boldsymbol{k}', t)\right) = \left(\frac{\partial f_w}{\partial t}\right)_{\text{collision}}, \tag{3.5}$$

is the quantum analog of the BTE (3.1), as the following interpretation of (3.5) shows:

$$\underbrace{\frac{\partial f_w}{\partial t}}_{} + \underbrace{\boldsymbol{v}_q \frac{\partial f_w}{\partial \boldsymbol{r}}}_{\text{diffusion}} + \underbrace{(\text{complicated term})}_{\text{drift}} = \left(\frac{\partial f_w}{\partial t}\right)_{\text{collision}}. \tag{3.6}$$

6. Real-valued state functions are rarity in quantum mechanics, and the Wigner function is unique in this respect among many-body formulations.
7. Phase-space function: a function of position and velocity.
8. The WFTE is derived in Chapter 5.

Because of the similarity of the WFM to conventional device simulation methods, and the intuitiveness of a phase-space state function, the WFM of quantum device simulation receives high marks in Table 3.1 in this regard.

In contrast to the Wigner function, both the density matrix [34] and the Green's functions [35, 36] are rather abstract state functions which correlate the state of the system at one point to that at another point. The interpretation of these functions is not intuitive. However, Figure 3.1 indicates that both the density matrix and Green's function formulations are related to the WFM, so their respective state functions can be, and often are, interpreted by conversion to the Wigner function (*e.g.*, after solution of their respective QTE) [13]. Thus, the Wigner function can be calculated from the density matrix as follows (in 3-D):

$$f_w(\boldsymbol{r}, \boldsymbol{k}, t) = \int_{-\infty}^{\infty}\int\int d\boldsymbol{r}'\rho(\boldsymbol{r} + \tfrac{1}{2}\boldsymbol{r}', \boldsymbol{r} - \tfrac{1}{2}\boldsymbol{r}', t)e^{-i\boldsymbol{k}\boldsymbol{r}'}. \tag{3.7}$$

Although the density matrix and Wigner function formulations are theoretically equivalent (they are related by a Fourier transform), the unfamiliar nature of the density matrix itself makes WFM much preferable. In contrast, the Green's function formalism is not directly equivalent to the WFM; it is more general. Not only are there four independent Green's functions [13] (and six Green's functions total), but each Green's function also contains more of the exact many-quanta wavefunction information than the Wigner function. The "double-time" correlation Green's function $G^<$ is most directly related to the Wigner function, but some information is integrated out, along with performing two (information-neutral) Fourier transforms and a change to center-of-mass coordinates, to arrive at f_w [37]. If $G^<(\boldsymbol{r}_1, t_1;\boldsymbol{r}_2, t_2)$ is the initial Green's function, then:

$$f_w(\boldsymbol{r}, \boldsymbol{k}, t) = \int\frac{d\omega}{2\pi}(-i)G^<(\boldsymbol{k}, \omega;\boldsymbol{r}, t)\,; \tag{3.8}$$

$$G^<(\boldsymbol{k}, \omega;\boldsymbol{r}, t) = \int\frac{d\omega}{2\pi}(-i)\iiint d\boldsymbol{r}'\, e^{-i\boldsymbol{k}\cdot\boldsymbol{r}'}\int dt\, e^{i\omega t'}G^<(\boldsymbol{r}', t';\boldsymbol{r}, t)\,; \tag{3.9}$$

$$G^<(\boldsymbol{r}', t';\boldsymbol{r}, t) = G^<(\boldsymbol{r}_1 - \boldsymbol{r}_2, t_1 - t_2;\tfrac{1}{2}(\boldsymbol{r}_1 + \boldsymbol{r}_2), \tfrac{1}{2}(t_1 + t_2))\,. \tag{3.10}$$

The Green's function formulation does have one important advantage over the WFM. Its powerful formalism allows for more general analytical derivations, presenting the possibility of more realistic simulation (i.e., with fewer simplifying assumptions) of quantum

systems than with the WFM. In fact, the Green's function formalism has produced the most general derivation and form of the WFTE yet [38]. However, solution of the exact Green's function QTE is currently intractable [39]. Simplifying assumptions and approximations have been used to arrive at tractable Green's function QTEs [40, 41], but for the requirements in Table 3.1, these are not currently preferable to the WFM. Again, as the cost of computing declines, the requirements of Table 3.1 may soon be met more agreeably by approximate versions of the Green's function QTE than by the WFM.

As a result of the abstract density matrix and Green's function state functions, it is also not obvious how to apply appropriate boundary conditions [40, 42]. The WFM, by contrast, can use the same boundary conditions as the BTE. This indicates another very important advantage of the WFM over the density matrix and Green's function approaches. A conventional electronic device simulator can be easily and naturally coupled to a WFM-based quantum device simulator, since the boundary conditions can be identical. This is not true of the density matrix and Green's function approaches.

3.4.2.5 Derivatives of the Wigner Function Method

Returning again to the WFM, note that, as with the BTE, it is currently only feasible to numerically solve the WFTE in 1-D. However, as shown in Chapter 2, the RTD is effectively a 1-D device, so this limitation does not disqualify the WFM as a basis for SQUADS. Recall that transport models derived from the BTE, specifically the hydrodynamic and Monte Carlo models, have allowed simulations of 2-D and 3-D systems, as discussed in Section 3.2. The fact that the WFTE is the quantum analog of the BTE has prompted the derivation of analogous transport models based on the WFTE. The Wigner function hydrodynamic (or moment) equation approach [43] offers a simplified formulation that makes 2-D simulations feasible. But since the (1-D) RTD was chosen as a prototype device, the more accurate WFM is preferred for SQUADS.

The path integral (or Monte-Carlo) approach [44] may allow multi-dimensional simulations as well, while including the full complexity of the WFTE. However, as indicated in Table 3.1, this simulation method is significantly more computationally demanding than the WFM, again making the latter preferable for this quantum device simulation effort. Also, difficulties arise in the tracing of the quantum trajectories of the path integral approach, especially where tunneling is involved. A quanta can disappear from one place

and appear in another without apparently having traversed the intervening space, calling into question the entire notion of quantum trajectories.

3.5 SQUADS Simulation Approach

This section explains the rationale for the remaining high-level choices faced in defining or directing SQUADS' development. It also combines the results of Section 3.4 into a well-defined plan for the design and use of SQUADS.

3.5.1 One-Dimensional versus Multi-Dimensional

Section 3.4 concluded that the optimal approach for simulating RTD-like (i.e., far-from-equilibrium, irreversible, dynamic, open) quantum devices, the Wigner function method, is only currently tractable for numerical solution in 1-D form. In accepting this limitation, the near-term approach has again been applied. First, recall that one of the main reasons the RTD was chosen as the prototype simulator test device was *because* it is a 1-D device, thus being relatively simple to fabricate and understand compared to multi-dimensional quantum devices. Further, the previous section argued that an equally capable multi-dimensional quantum simulator would require much more computational power than a one dimensional Wigner function approach, and is therefore currently infeasible to execute. Admittedly, multi-dimensional quantum simulation tools are necessary for more accurate quantum device simulation, and their realization is being pursued by other researchers. Available computing power will eventually make them as feasible as SQUADS is now. Much of the theory, computer code, and experience generated during the development of a 1-D simulator is applicable to the development of a related multi-dimensional simulation tool. In any event, the goal of this research is now stated formally: to develop a software tool for the accurate simulation of 1-D quantum devices (i.e., devices that are quantum scale in only one dimension).

3.5.2 Envelope Function versus Tight-Binding

Another decision which remains to be made is that of the energy band model that will be used in SQUADS. Two choices have become popular for quantum system modeling: the "tight-binding" model[9] [45] and the "envelope-function" model[10] [46]. Note that the Kronig-Penny model [47] is usually used mainly for analytical derivations. All three

energy band models are illustrated in 1-D in Figure 3.2. The envelope-function approach is most familiar, as standard valence and conduction band diagrams are based on this potential. Here the potential is assumed to be an average over a unit cell of the atomic lattice of the semiconductor. The carrier kinetics is treated almost the same as a free carrier, but with a modified mass called the effective mass, denoted m*. The envelope-function model is thus often called the effective mass model.

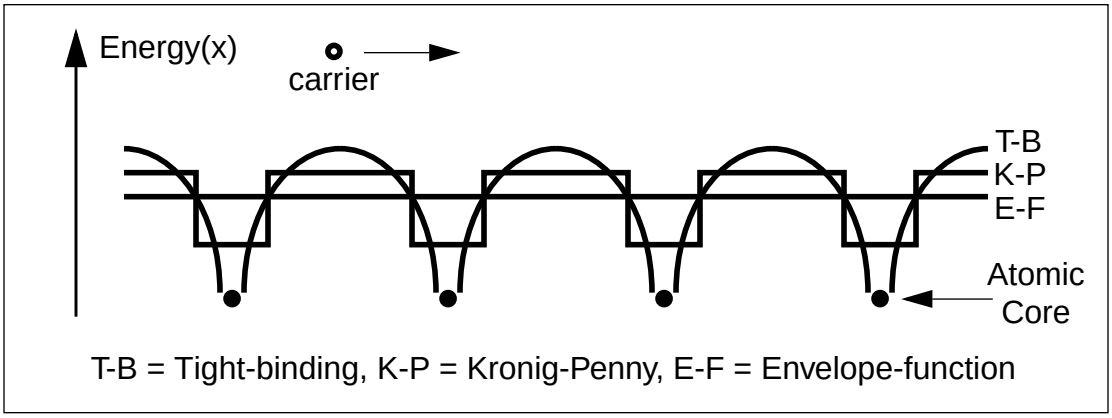


Figure 3.2: Energy band models for the conduction and valence bands

The tight-binding model most closely matches the physical atomic structure, but is also the most computationally demanding. the Kronig-Penny model is usually used mainly for analytical derivations. Due to its simplicity, this work uses the envelope-function model, which uses the average of the potential energy over a unit cell of the semiconductor lattice.

The tight-binding approach, as its name implies, takes the opposite extreme of a nearly-bound carrier. The potential is periodic, with deep energy wells at the atomic cores. Thus, electron concentration is highest near atomic cores, and is periodic. The tight-binding approach is theoretically more accurate, but it is significantly more computationally demanding. This presents another trade-off choice of accuracy versus computation time. The choice is not very difficult in this case, however. The tight-binding premise of nearly-bound carriers goes against the choice of far-from-equilibrium quantum devices, with carrier energies far above the energy band minimum. Further, the tight-binding approach can be very numerically difficult to implement, with successive multiplications of very small and very large numbers. Thus, the tight-binding approach was rejected in favor of the

9. Also known as the LCAO (linear combination of atomic orbitals) model
10. Also known as the “nearly-free” carrier model or the effective mass approximation.

envelope-function potential model. Like the choice of a 1-D simulator, this choice may also be subject to change as the cost of computing power continues to decline.

3.5.3 Two-Tiered Approach

During the development of SQUADS, it was decided that implementing an independent, less calculation-intensive quantum simulation approach in addition to the WFM would be very advantageous for several reasons listed below. The TMM was the obvious choice to fill this position: it has been the standard approach in quantum device simulation, it is relatively calculation-efficient, and it is an independent formulation of quantum mechanics from the WFM. Because the TMM is not able to handle irreversible and transient systems, it can not match the capabilities of the WFM in some cases. However, the TMM is very useful for the following roles in a two-tiered simulator scheme:

- Transfer-Matrix Method (TMM) roles in SQUADS:
  - efficient simulation of a wide range of structures to determine which merit more detailed study (by WFM simulation),
  - high resolution energy spectrum investigations,
  - a reality check on WFM results,
  - faster implementation and testing of simulator enhancements, and
  - 2nd and 3rd dimensions in multi-dimension simulations may be possible.
- Wigner Function Method (WFM) roles in SQUADS:
  - higher accuracy simulations to determine which devices merit experimental investigation,
  - simulations including inelastic scattering, and
  - transient simulations.

Note that without two independent simulation methods, no means would be available, except experiment, for checking the results of a simulation. In fact, Section3.1 argued that certain types of information available from simulations, such as the internal operation of the quantum device, can only be inferred from experiment. Even for those types of information that *can* be gleaned from experiment, the choice of implementing two independent simulation methods in SQUADS is also a response to the very expensive, uncertain, and time-consuming nature of experiment, as well as the sparsity and incompleteness of published results. The above list combines the results of Section 3.4 into a unified plan for the

design and use of SQUADS. Luscombe and Frensley [3] also argued that a "spectrum of modeling tools of varying degrees of sophistication is required to meet the needs of the various stages of quantum device development."

Given the intention to implement two independent formulations of quantum mechanics in SQUADS, it is advantageous to maintain as much common code between the TMM and WFM as possible. Functionality such as input file processing, output and plotting routines, current-voltage curve tracing, enforcing self-consistency, numerical methods, matrix manipulations, and others, would all be needed by virtually any simulation. By implementing this functionality in a modular fashion, it could be utilized by either simulation method where appropriate. Thus, maximizing code-sharing between the TMM and WFM was taken as another design goal for SQUADS.

3.5.4 Experimental Verification

The ultimate goal of this research was to create a numerical simulation tool to accurately predict the operation of any one-dimensional quantum device. The only way to determine if, or how well, that goal is met is to compare the predictions of SQUADS with actual experimental measurements. In fact, during SQUADS' development, knowledge of experimental results more often directed the course of this development than the reverse. Several examples of the impact of comparison with experiment on SQUADS development will be presented in later chapters. Two sources of experimental measurements are possible: in-house experiments and the published experimental results of other researchers. The advantages of in-house experimental work are that desired experimental structures and measurements can be specified, resulting in more certainty of what those structures and the associated measurements were. Further, a potentially wider range, or more targeted set, of measurements than are usually reported in the literature could be performed.

The choice between sources of experimental measurements was made with the understanding of the central goal of this project. With the focus of this work on the development of a quantum device simulator, fairly standard and well-understood quantum devices should be used for the development process. Thus, even with in-house experimental work, advancements in quantum device technology were not expected through this project. In fact, it seemed optimistic to expect to fabricate devices of equal quality to those of other researchers who had spent a great deal of time perfecting quantum device fabrication tech-

niques. Therefore, in spite of the significant advantages of doing (at least some) in-house experimental work, the interests of time dictated that using only published measurements of other researchers was optimal for achieving the central goal of the project.

3.5.5 Research Tool

One final point needs to be made explicit in defining the motivations, priorities, and principles which determined SQUADS development path. On one hand, SQUADS is not an end in itself, but is intended to be used eventually to guide experiment. On the other hand, because an accurate quantum device simulator is such an important tool, knowledge gained in its development and use is important in its own right. Thus, SQUADS is used not only to research quantum device *operation*, but also to research quantum device *simulation*. For this reason, many features and options that perhaps would not exist in a commercial simulation tool are maintained in SQAUDS simply for the knowledge they can give the researcher about various simulation issues, such as the memory use versus speed versus accuracy trade-off, the correct (or incorrect) functioning of a new feature, the importance of a physical effect (e.g., scattering or self-consistency), etc. Where ever a researcher might reasonably wonder about issues such as these, SQUADS was designed to allow the choice between various options so that the issue can be investigated as efficiently as possible. This is another example where SQUADS' modular structure and resulting flexibility and extensibility allows the substitution of one solution method for another, to a degree unmatched by any other quantum device simulator.

3.6 Summary

This chapter has described the goals and guiding principles used during the development of SQUADS. In short, SQUADS was developed as a general one-dimensional simulator for far-from-equilibrium, irreversible, open, transient quantum devices such as the resonant tunneling diode. SQUADS concentrates on providing access to information not available in experiment, such as the internal operation of the quantum device. SQUADS was also designed for flexibility and extensibility, to allow the investigation of quantum device *simulation*, in terms of alternative quantum mechanics formulations, numerical implementations, and quantum system characteristics.

Both the Wigner function and transfer-matrix methods of quantum device simulation

were chosen to accomplish these design goals. The TMM is an efficient method for fast initial simulations, for checking WFM simulation results, and for high-resolution energy spectrum investigations. The WFM allows a more complete description of real quantum systems, including scattering and transient operation, although at a higher computational cost. To make the implementation and upgrade of both simulations methods as efficient as possible, another design goal was to utilize code-sharing and modularity as much as possible. In the following two chapters, a detailed description is given of the numerical implementation of these two simulation methods, beginning with the transfer-matrix method in Chapter 4.

References

[1] M. R. Pinto, C. S. Rafferty, and R. W. Dutton. *PISCES-II - Poisson and Continuity Equation Solver.* Stanford University, 1984.

[2] M. R. Pinto, C. S. Rafferty, H. R. Yeager, and R. W. Dutton. *PISCES-IIB - Supplementary Report*. Stanford University, 1985.

[3] J. H. Luscombe and W. R. Frensley. “Models for nanoelectronic devices.” *Nanoelectronics*, 1:131–140, 1990.

[4] W. R. Frensley. “Development of an interactive design environment for heterostructure and quantum-well devices.” *IEEE Transactions on Electron Devices*, 38(12):2704–2705, 1991.

[5] M. R. Pinto. *Comprehensive semiconductor device simulation for silicon ULSI*. PhD thesis, Stanford University, Aug. 1990. p. 14.

[6] C. Kittlel and H. Kroemer. *Thermal Physics*, page 14. W. H. Freeman, New York, 2nd edition, 1980.

[7] C. L. Gardener. “Numerical simulation of a steady-state electron shock wave in a submicrometer semiconductor device.” *IEEE Transactions on Electron Devices*, 38(2):392–398, 1991.

[8] M. R. Pinto. *Comprehensive semiconductor device simulation for silicon ULSI*. PhD thesis, Stanford University, Aug. 1990. p. 23.

[9] M. R. Pinto. *Comprehensive semiconductor device simulation for silicon ULSI*. PhD thesis, Stanford University, Aug. 1990. p. 18.

[10] M. R. Pinto. *Comprehensive semiconductor device simulation for silicon ULSI.*

PhD thesis, Stanford University, Aug. 1990. p. 17.

[11] F. A. Buot. “Mesoscopic physics and nanoelectronics: Nanoscience and nanotechnology.” *Physics Reports*, 234(2-3):73–174, 1993.

[12] J. R. Barker. “Quantum transport theory for small-geometry structures.” In M. J. Kelley and C. Weisbuch, editors, *Springer Proceedings in Physics, Vol. 13: The Physics and Fabrication of Microstructures and Microdevices*, pages 210–230, Berlin, 1986. Springer-Verlag.

[13] D. K. Ferry. “Transport theory in ultra-submicron devices.” In H. Heinrich, G. Bauer, and F. Kuchar, editors, *Springer Series in Solid-State Sciences, Vol. 83: Physics and Technology of Submicron Structures*, pages 226–242, Berlin, 1988. Springer-Verlag.

[14] G. D. Mahan. *Many-Particle Physics*, chapter 7.3.A. Plenum Press, New York, 2nd edition, 1987.

[15] G. D. Mahan. *Many-Particle Physics*, chapter 3.8. Plenum Press, New York, 2nd edition, 1987.

[16] G. Baym. *Lectures on Quantum Mechanics*, chapter 5. Benjamin/Cummings, Menlo Park, CA, 1973.

[17] C. W. Gardiner. “Quantum noise and quantum langevin equations.” *IBM Journal of Research and Development*, 32(1):127–136, 1988.

[18] S. Datta, M. Cahay, and M. McLennan. “Scatter-matrix approach to quantum transport.” *Physical Review B*, 36(10):5655–5658, 1987.

[19] S. Bandyopadhyay and M. Cahay. “The generalized scattering matrix approach: an efficient technique for modeling quantum transport in relatively large and heavily doped structures,” 1991.

[20] E. Schrödinger. “Quantisierung als eigenwertproblem.” *Annalen der Physik*, 79:361–376, 1926.

[21] R. K. Mains and G. I. Haddad. “Improved boundary conditions for the time-dependent Schrödinger equation.” *Journal of Applied Physics*, 67(1):591–593, 1990.

[22] C. S. Lent and D. J. Kirkner. “The quantum transmitting boundary method.” *Journal of Applied Physics*, 67(10):6353–6359, 1990.

[23] L. F. Register, U. Ravaioli, and K. Hess. “Numerical simulation of mesoscopic systems with open boundaries using the multidimensional time-dependent

Schrödinger equation." *Journal of Applied Physics*, 69(10):7153–7158, 1991.

[24] J. R. Hellums and W. R. Frensley. "Non-Markovian open-system boundary conditions for the time-dependent Schrödinger equation." *Physical Review B*, 49(4):2904–2906, 1994.

[25] S. Bhobe, W. Porod, and S. Bandyopadhyay. "Modulation of impurity scattering rates by wavefunction engineering in quasi 2-D systems and its device applications." *Solid State Electronics*, 32(12):1083–1087, 1989.

[26] M. C. Yalabik. "Some ad-hoc methods for introducing dissipation to the Schrödinger equation." In S. P. Beaumont and C. M. S. Torres, editors, *Proceedings of a NATO Advanced Workshop on Science and Engineering of One- and Zero-Dimensional Semiconductors*, pages 83–89, Cadiz, Spain, Mar. 29-Apr. 1 1989. NATO Scientific Affairs Division.

[27] R. Tsu and L. Esaki. "Tunneling in a finite superlattice." *Applied Physics Letters*, 22(11):562–564, 1973.

[28] K. F. Brennan and C. J. Summers. "Theory of resonant tunneling in a variably spaced multiquantum well structure: An Airy function approach." *Journal of Applied Physics*, 61(2):614–623, 1987.

[29] L. A. Cury and N. Studart. "Resonant tunneling through Al(x)Ga(1-x)As-GaAs heterostructures." *Superlattices and Microstructures*, 4(2):245–250, 1988.

[30] J. R. Söderström, E. T. Yu, M. K. Jackson, Y. Rajakarunanayake, and T. C. McGill. "Two-band modeling of narrow band gap and interband tunneling devices." *Journal of Applied Physics*, 68(3):1372–1375, 1990.

[31] C. M. Tan, J. Xu, and S. Zukotynski. "Study of resonant tunneling structures: A hybrid incremental Airy function plane-wave approach." *Journal of Applied Physics*, 67(6):3011–3017, 1990.

[32] W. R. Frensley. "Wigner-function model of a resonant-tunneling semiconductor device." *Physical Review B*, 36(3):1570–1580, 1987.

[33] E. Wigner. "On the quantum corrections for thermodynamic equilibrium." *Physical Review*, 40:749–759, June 1 1932.

[34] W. R. Frensley. "Simulation of resonant-tunneling heterostructure devices." *Journal of Vacuum Science and Technology B*, 3(4):1261–1266, 1985.

[35] L. P. Kadanoff and G. Baym. *Quantum Statistical Mechanics*. Benjamin/Cum-

mings, Reading, MA, 1962.

[36] L. V. Keldysh. “(unknown).” *Soviet Physics JETP*, 20:1018, 1965.

[37] G. D. Mahan. *Many-Particle Physics*, pages 200–201. Plenum Press, New York, 2nd edition, 1987.

[38] F. A. Buot and K. L. Jensen. “Lattice Weyl-Wigner formulation of exact many-body quantum-transport theory and applications to novel solid-state quantum-based devices.” *Physical Review B*, 42(15):9429–9457, 1990.

[39] A. P. Jauho. “Nonequilibrium Green function techniques applied to hot electron quantum transport.” *Solid State Electronics*, 32(12):1265–1271, 1989.

[40] M. A. Alam, R. A. Morrisey, and A. N. Khondker. “Self-consistent analysis in the presence of phase-randomizing processes for double-barrier structures.” *Journal of Applied Physics*, 71(7):3077–3090, 1992.

[41] R. Lake and S. Datta. “High-bias quantum electron transport.” *Superlattices and Microstructures*, 11(1):83–87, 1992.

[42] W. R. Frensley. “Boundary conditions for open quantum systems driven far from equilibrium.” *Reviews of Modern Physics*, 62(3):745–791, 1990.

[43] J.-R. Zhou and D. K. Ferry. “Simulation of ultra-small GaAs MESFET using quantum moment equations.” *IEEE Transactions on Electron Devices*, 39(3):473–478, 1992.

[44] K. L. Jensen and F. A. Buot. “The methodology of simulating particle trajectories through tunneling structures using a Wigner distribution approach.” *IEEE Transactions on Electron Devices*, 38(10):2337–2347, 1991.

[45] C. Kittlel. *Introduction to Solid State Physics*, page 229. John Wiley & Sons, New York, 6th edition, 1986.

[46] S. Datta. *Quantum Phenomena*, volume VIII of *Modular Series on Solid State Devices*, page 9. Addison-Wesley, Reading, MA, 1986.

[47] C. Kittlel. *Introduction to Solid State Physics*, page 164. John Wiley & Sons, New York, 6th edition, 1986.

Chapter 4

The Transfer-Matrix Method

The intention of quantum device simulation — at least for quantum devices in this work, whose ambition is to replace conventional electronic devices — is to predict current flow through the device. The transfer-matrix method (TMM) accomplishes this by determining the transmission amplitude T of an incident wavefunction (the quantum manifestation of charge carriers) through the device, as depicted in Figure 4.1. The magnitude I of the incident wave is given by the number of incident carriers at the energy being considered. The TMM calculates T at the range of energies over which I is significant, and adds the results to arrive at total current flow through the device.

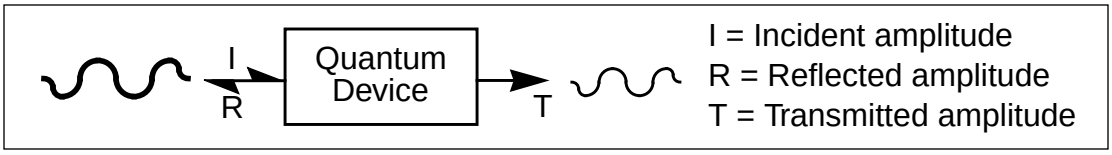


Figure 4.1: Transfer-matrix method overview

The transfer-matrix method determines current flow through a quantum device by calculating the transmission amplitude T, and thus the current, of many monoenergetic beams of carriers over the range of incident carrier energies.

This chapter describes the TMM in some detail, beginning with a review of the literature in terms of the accomplishments and state-of-the-art of this method in Section 4.1. Section 4.2 then presents an analytical derivation from the Schrödinger equation of the equations and expressions used in the TMM. A description of the basic implementation of

the TMM is then presented in Section 4.3. Several alternative implementations of the TMM are discussed in Section 4.4. Finally, Section 4.5 investigates these alternative TMM implementations, resulting in conclusions about how the TMM should be implemented for accuracy and efficiency.

In previous chapters, a relatively non-technical level of discourse has been maintained. As the theory and implementation of SQUADS are presented beginning in this chapter, the presentation will necessarily become more mathematical. However, it is undesirable, and in fact impossible, to include in this dissertation a comprehensive description of the implementation and internal workings of SQUADS. The reader interested in this level of discourse is referred to the *SQUADS Technical Reference* [1].

4.1 History and State of the Art

The TMM is currently the most widely used method of quantum electronic device simulation[1] for several reasons. It has been in existence for the longest (since 1962 [2]), its derivation is relatively simple, it is easy to understand (using the optical analogy), and it requires the least computational resources. The TMM isn't the *only* method of quantum device simulation because, as discussed in Section 3.3, it can not accurately treat scattering and transient operation. Because the TMM is widely used, its basic theory and implementation have been widely (if incompletely) described in the literature [2-6]. However, several alternative implementations have never been directly compared to the standard approach, and several complicating issues have apparently never been discussed. One purpose of this chapter is to properly discuss these issues.

Vassell *et al.* [3] enumerated eleven simplifying assumptions and approximations in the original description of the TMM [2], and other short-comings have been discussed elsewhere. The past 15 years have witnessed a continual attack on these approximations, and most have now been improved or removed. For example, position-dependent effective mass and more general structures were incorporated in [3] and most subsequent work. More accurate piece-wise linear (as opposed to piece-wise constant) potentials have been used by several researchers [7, 8]. Self-consistency has been included by simultaneous

1. It is necessary to differentiate between quantum *systems*, which are more commonly investigated via direct discretization of the time-dependent Schrödinger equation, and quantum *devices*, which must have open, ohmic boundaries. As discussed in Section 3.3, time-dependent Schrödinger equation has not adequately demonstrated the ability to treat this key feature of many-particle devices.

solution of the Poisson and Schrödinger equations [9-11]. The TMM has also been used with both tight-binding (see Section 3.5.2) [12, 13] and $\boldsymbol{k} \bullet \boldsymbol{p}$ approaches [14, 15], to represent the energy-velocity relationship more accurately. Interactions between energy bands (conduction, light hole, heavy hole, and/or split-off hole) have been incorporated in two-band [6, 12, 13], three-band [15, 16], and even four-band [14] models. The TMM has also been used to determine the discrete bound state energies of quantum wells [17]. Finally, a 1-D TMM calculation has been incorporated in a 2-D resonant tunneling transistor simulation [18].

The remainder of this chapter describes SQUADS' implementation of the transfer-matrix method of quantum device simulation. This implementation handles position-dependent effective mass, general device structures, piece-wise linear potentials, and self-consistency; and it includes a simplified multi-band capability. The last two features are described in Chapter 6. Many of the more subtle features of the TMM in SQUADS, which make it a robust and extensible quantum device simulation tool, will also be described.

4.2 Background

This section derives the basic analytic expressions and equations used in implementing the transfer-matrix method of quantum device simulation.

4.2.1 General Solutions of the Schrödinger Equation

The TMM is based on solving the time-independent, or steady-state, Schrödinger equation, as mentioned in Section 3.4.2.3. In 1-D,[2] the time-independent Schrödinger equation (TISE) in the effective mass approximation (see Section 3.5.2) is:

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x) + U(x)\psi(x) = E\psi(x). \tag{4.1}$$

In (4.1), the solution $\psi(x)$ is the quantum wavefunction of a charge carrier (*i.e.*, electron or hole) or (as in the TMM) a beam of carriers with effective mass m and energy E, $U(x)$ is potential energy (*i.e.*, the conduction or valence energy band minimum), and $\hbar$ is the reduced Planck's constant. Use of the effective mass (a.k.a. envelope function) approximation has important consequences in terms of the choice of position grid scheme, as dis-

2. Recall from Section 3.5.1 that this work took the practical step of limiting SQUADS to the simulation of one-dimensional quantum systems.

cussed shortly.

As described in Section 4.3, in the TMM, the TISE is solved multiple times for a beam of carriers at a sequence of closely-spaced energies.[3] Since the TISE is a second order differential equation, it has two independent solutions at a given energy E, which will be denoted $f(x)$ and $g(x)$. The general solution $\psi(x)$ is a linear combination of $f(x)$ and $g(x)$:

$$\psi(x) = a \cdot f(x) + b \cdot g(x). \tag{4.2}$$

The wavefunction $\psi(x)$ in any particular case is determined by the potential profile $U(x)$. For a constant or linear $U(x)$ (and few others), the TISE is analytically solvable. The TMM takes advantage of this fact. Of course, useful 1-D quantum devices don't have such simple potential profiles. Therefore, to solve the TISE analytically for real quantum devices, the simulation region is divided into a series of short regions, within which a constant or linear approximation to the potential is acceptable. Most implementations of the TMM, including the description in this section, use only constant potential regions. Even for a region of constant potential, there are three distinct solutions of the TISE, which are described below.

A region in which the carrier energy E is greater than the potential energy $U(x)$ is called a "classically allowed" region, since carriers are energetically allowed to exist in such regions according to classical (as well as quantum) physics. In a classically allowed region of *constant* potential (a CCA region), $E > U(x) = U$, and the wavefunction can be written as the sum of forward-travelling and backward-travelling plane waves:

$$\psi(x) = a \cdot e^{ikx} + b \cdot e^{-ikx}, \tag{4.3a}$$

where e is the base of the natural log and k is the "wavevector" of the quantum particle:

$$k = \frac{\sqrt{2m(E-U)}}{\tilde{\hbar}}. \tag{4.3b}$$

Wavevector k plays an important part in quantum system analysis, so its relationship to other, perhaps more familiar quantities will be given. The wavevector of a particle is a measure of its momentum p, velocity v, quantum wavelength λ, and kinetic energy K:

$$p = \tilde{\hbar}k \qquad v = \frac{p}{m} = \frac{\tilde{\hbar}k}{m} \qquad \lambda = \frac{2\pi}{k} \qquad K = \frac{p^2}{2m} = \frac{\tilde{\hbar}^2 k^2}{2m}, \tag{4.4}$$

where $\tilde{\hbar}$ is the reduced Planck constant and m is the effective mass of the particle.

A region in which the carrier energy E is *less* than the potential energy $U(x)$ is called

3. Technically, the TISE represents an eigenvalue equation at energy E.

a "classically forbidden" region, since carriers are energetically *not* allowed into such regions according to classical (although not quantum) physics. Quantum particles, such as electrons or holes, can travel into and through very narrow classically forbidden regions, a process called quantum tunneling. However, even quantum particles will inevitably be repelled from wide classically forbidden regions. In a classically forbidden region of *constant* potential (a CCF region), $E < U(x) = U$, and the wavefunction can be written as the sum of exponentially increasing and decreasing functions:

$$\psi(x) = a \cdot e^{\kappa x} + b \cdot e^{-\kappa x}, \tag{4.5a}$$

$$\kappa = \frac{\sqrt{2m(U-E)}}{\underset{\sim}{h}}. \tag{4.5b}$$

κ is the attenuation constant of the wavefunction as it penetrates into the CCF region.

Finally, a region in which the (fixed) carrier energy E is exactly equal to the potential energy $U(x)$ is called a constant "classically neutral" (CCN) region. For $E = U(x) = U$, the TISE (4.1) requires that the second derivative of the wavefunction be zero. Therefore, the wavefunction has the following general form in a CCN region:

$$\psi(x) = a \cdot x + b. \tag{4.6}$$

The treatment of CCN regions has apparently never been described in the literature. Admittedly, its occurrence is unlikely in TMM simulation, and most TMM implementations probably ignore this possibility altogether. However, a robust simulator (*i.e.*, one which will not crash or give erroneous results) must correctly handle this case.

4.2.2 Gridded Potential Profile

Figure 4.2 shows an approximate potential profile $U(x)$ of a resonant tunneling diode (the prototype quantum device used in this work). There is no analytic solution to the TISE for such a potential. As indicated above, a TMM simulation starts by dividing the quantum system into many small regions. Within each region, a constant approximation $U(x) = U$ is used for the potential. This *determines* the general solution of the TISE for that region. If a suitably fine position grid is chosen, and an appropriate choice is made for the potential function in each small region, the resulting piece-wise constant potential function and its resulting TISE solution will track the actual potential profile and its TISE solution arbitrarily closely. The remainder of this section describes the basic position grid scheme used in SQUADS and the choice of potential function for each grid region.

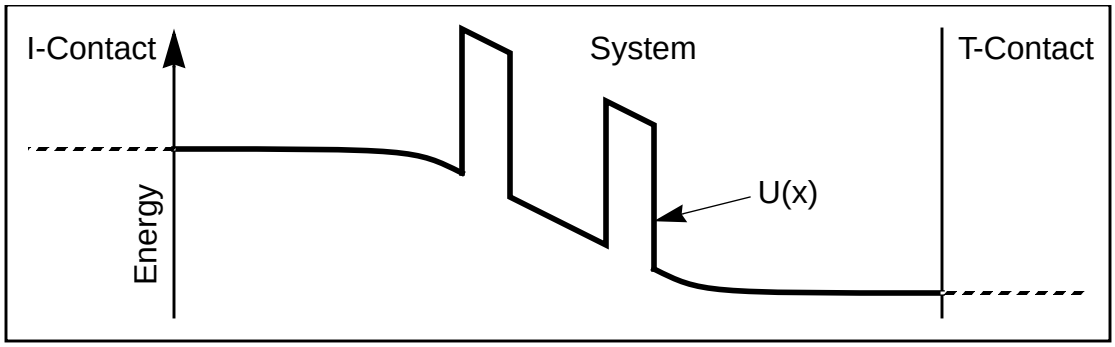


Figure 4.2: Typical quantum device potential energy profile

The potential energy profile U(x) shown for a resonant tunneling diode is relatively simple, but still far too complex to solve the TISE analytically.

Having argued that the solution of the TISE can be facilitated by dividing the system into many small regions, how should the position grid[4] be designed? To be consistent with SQUADS' use of the effective mass approximation, which averages the potential over a unit cell (i.e., the region "controlled" by a single atom), neither the wavefunction nor the potential should be resolved to any finer degree than the atomic spacing. In fact, the position grid should be designed to have exactly one grid point per atomic spacing.

Thus, as shown in Figure 4.3, SQUADS uses a uniform position grid with node points (denoted x_n) separated by a distance Δx, starting at $x_0 = 0$, and ending at $x_N = L$, where L is the width of the system being simulated. N is the number of grid regions between the contacts (one less than the number of nodes), so that $\Delta x = L/N$. As discussed in the previous paragraph, Δx should be equal to the lattice spacing of the material. Of course, different materials typically have slightly different lattice constants, so it is not possible with a uniform grid to exactly mirror the lattice of real quantum devices, which are always multi-material systems. The discrepancy should be fairly small, however.

SQUADS uses the grid node scheme depicted in Figure 4.3, as well as device structure information and the applied bias, to calculate the potential U_n at each grid node x_n. The potential values U_n at the internal device nodes may be approximated using a suitable algorithm, such as a simple linear model for the potential profile, or calculated self-consistently (*i.e.*, consistent with the carrier density), as discussed in Chapter 6. To solve the TISE in each region, the TMM must translate the U_n values at the grid nodes into $U_n(x)$ functions for each region between the nodes. The logical choice for $U_n(x)$ is to use the

4. The set of points at which physical quantities, such as carrier density, will be calculated.

average of the potential at two bounding nodes n and n-1 :

$$U_n(x) = \frac{U_n + U_{n-1}}{2} . \tag{4.7}$$

This interpolation scheme results in a stepped, rather than smooth, potential profile, but with atomic grid spacing, the error should not be large. More accurate potential interpolation schemes are discussed in Sections 4.4.1 and 4.4.3.

Note from (4.3b) and (4.5b) that the effective mass m must also be known in the grid region to determine the TISE general solution in that region. SQUADS supplies to the TMM the value of the effective mass at the grid nodes, m_n. Therefore, SQUADS uses the average of the effective masses at the two bounding nodes for that of the enclosed region:

$$m_n(x) = \frac{m_n + m_{n-1}}{2} . \tag{4.8}$$

For boundary conditions on the electrostatic potential $U(x)$, as indicated in Figure 4.4, SQUADS defines the Fermi energy at the left contact (called the incident contact, for now) as the reference energy. The potential at the right contact (called the transmitted contact) is set by the applied bias V_a. Thus, contact potentials are:

$$U_0 = -E_{\mathrm{FI}}, \tag{4.9a}$$

$$U_N = -qV_a - E_{\mathrm{FT}}, \tag{4.9b}$$

where q is the electron charge, E_{FI} is the Fermi energy (relative to the energy band) at the incident contact, and E_{FT} is that at the transmitted contact.

The electrostatic potential boundary conditions depicted in Figure 4.4 differ from the standard boundary conditions in quantum device simulation [3, 8, 10, 19-21]. The standard approach is to take the energy band minimum (rather than the Fermi level) at the I-

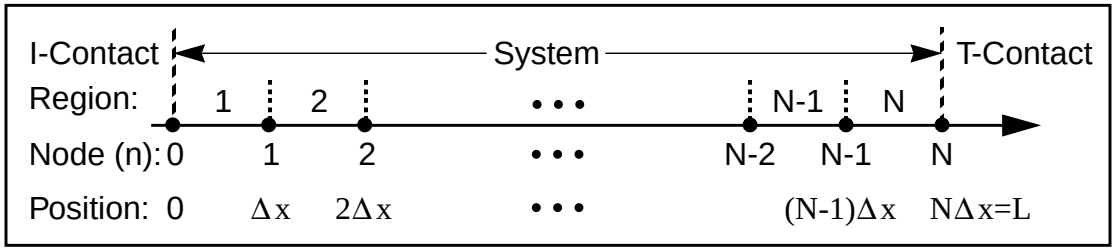


Figure 4.3: SQUADS position grid scheme

SQUADS uses a uniform position grid, $x_n = n\Delta x$, at which points device parameters (e.g., band offset, doping) are supplied and simulation results (e.g., carrier density, wavefunction) are calculated.

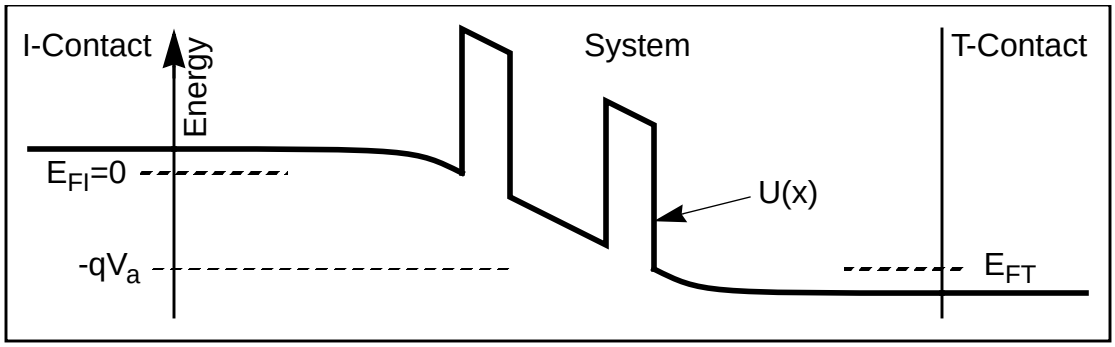


Figure 4.4: Typical potential with applied bias and boundary conditions

SQUADS uses the I-contact Fermi level as the energy reference ($E_{FI} = 0$), rather than the I-contact electrostatic potential [U(0) = 0].

contact as the reference ($U_0 = 0$). SQUADS uses the boundary conditions shown in Figure 4.4 (also used in [6, 11]) for several reasons:

- If multiple energy bands are included, each of which may have a different band minimum at the contacts, there is no reason to prefer one band minimum over another as reference. In contrast, there is only one Fermi level at each contact.
- At equilibrium (zero applied bias), it is reasonable to expect the reference on one side of the device to equal the analogous point on the other side. Even using a single energy band as reference, with any material difference (band offset, effective mass, or doping), the band minima at the two contacts will not be equal. However, the Fermi levels at the two contacts are always equal at equilibrium.
- By extension, bias V_a is applied between the two Fermi levels, not the band minima at the contacts. When the contact materials are different, the T-contact band minimum is $-qV_a + \delta U$ (where δU is the relative band offset at the T-contact) below the I-contact, but the T-contact Fermi level is exactly qV_a below E_{FI}.

Although the Fermi level is in a sense a more “fundamental” entity of a device than the band minimum (based on the above points), SQUADS recognizes that using the I-contact band minimum as reference in the cases of a single band is more user-friendly. Therefore, while all calculations in SQUADS are performed using the I-contact Fermi level as reference, when only one energy band is included in the simulation, SQUADS shifts potential profile plots such that the I-contact minimum is at 0 energy.

4.2.3 Wavefunction Matching Conditions

Given the potential approximation $U_n(x)$ for each region using (4.7), the general solution of the TISE (*i.e.*, the wavefunction) in each region is known, based on the results of Section 4.2.1. But the TISE is not yet completely solved. In particular, the *coefficients* in the general solutions for each region [see (4.2)] are as yet unknown. In fact, there are 2 unknown coefficients for each region! However, the wavefunctions in adjacent regions can be related using "matching conditions" at the interface between the two regions. These matching conditions arise from an analysis of the Schrödinger equation, which shows that every physically viable wavefunction will be continuous, and its probability current density[5] must be constant (in steady-state), everywhere in the system.

The continuity requirements must be satisfied even across interfaces between regions with different TISE general solutions. In other words, the wavefunctions of adjacent regions must be matched, as well as their probability current densities, at the interface between the regions. For example, at interface n, the matching conditions are:[6]

$$a_n f_n(x_n) + b_n g_n(x_n) = a_{n+1} f_{n+1}(x_n) + b_{n+1} g_{n+1}(x_n), \tag{4.10a}$$

$$\frac{a_n}{m_n}\frac{\partial}{\partial x} f_n(x_n) + \frac{b_n}{m_n}\frac{\partial}{\partial x} g_n(x_n) = \frac{a_{n+1}}{m_{n+1}}\frac{\partial}{\partial x} f_{n+1}(x_n) + \frac{b_{n+1}}{m_{n+1}}\frac{\partial}{\partial x} f_{n+1}(x_n). \tag{4.10b}$$

Since the wavefunction coefficients are the unknowns (the TISE general solutions f and g are known, given the region potential), (4.10a) and (4.10b) are solved for a_{n+1} and b_{n+1} in terms of a_n and b_b:

$$a_{n+1} = i_{11,n} a_n + i_{12,n} b_n, \tag{4.11a}$$

$$b_{n+1} = i_{21,n} a_n + i_{22,n} b_n, \tag{4.11b}$$

where the i values are constants. In matrix form, these equations are written

$$\begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} = \begin{bmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{bmatrix}_n \begin{bmatrix} a_n \\ b_n \end{bmatrix} \equiv [i_n] \begin{bmatrix} a_n \\ b_n \end{bmatrix}. \tag{4.12}$$

Note that the 2x2 matrix $[i_n]$ essentially transfers the wavefunction coefficient relationship across interface n, which is the origin of the term "transfer-matrix". Detailed expressions for the terms in transfer matrix $[i_n]$ are derived in [1].

---

5. Probability current density is analogous to "normal" charge current density.
6. These wavefunction matching conditions, although universally used in the TMM, were recently disputed by Harrison and Kozlov [22].

In SQUADS, $[i_n]$ is actually calculated as the product of two 2x2 matrix factors,

$$[i_n] = [i_{n+}][i_{n-}] \quad , \tag{4.13}$$

where $[i_{n-}]$ is associated with region n just before interface n and $[i_{n+}]$ is associated with region $n+1$ just after it. With three types of regions (CCA, CCF, and CCN), there are *nine* different functional forms for $[i_n]$, since this transfer matrix includes factors relating to the TISE solutions in both regions bounding node n. In general, to implement r region types in a TMM simulator, there are r^2 forms of $[i_n]$ to code.[7] However, there are only r forms each of $[i_{n-}]$ and $[i_{n+}]$. In using this separation of $[i_n]$, modification and addition of region types are greatly simplified, making SQUADS easily extensible.

Because of the central role of (4.12) in the TMM, a graphical depiction is shown in Figure 4.5. Using (4.12), the coefficients can be mathematically related between regions $n-1$ and n, and between regions n and $n+1$. Then the relationship between each pair of regions can be combined to relate the coefficients in region $n-1$ to those in region $n+1$. By extension, the wavefunction coefficients in any region can be related mathematically to those in any other region using the appropriate transfer matrices.

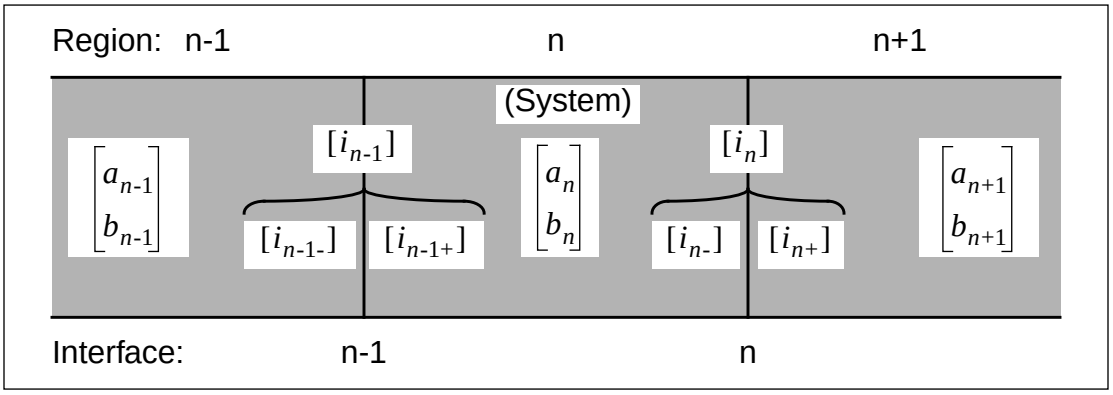


Figure 4.5: Relating wavefunction coefficients across an interaface

The wavefunction coefficients in adjacent regions, say n and n+1, are related by the 2x2 transfer matrix $[i_n]$, which is composed of two factors, $[i_{n-}]$ and $[i_{n+}]$.

4.2.4 System Transmission Matrix

The I-contact and T-contact can be thought of as regions of the device just like any of the internal grid regions. The main result of Section 4.2.3 was that matching conditions

7. In Section 4.4.3, a fourth region type will be added, giving 16 possible forms of $[i_n]$.

allow the wavefunction coefficients in any region to be related mathematically to those in any other region. Thus, consider constructing a relationship between the coefficients in the I-contact and those in the T-contact by combining the relationships across each grid interface in turn. This net relationship across the entire device, called the system transmission matrix (STM), has a special place in the TMM, for reasons that will be apparent in Section 4.3.1. This section describes the simplest method of calculating the STM.[8]

4.2.4.1 Basic STM Calculation

As indicated above, the STM is a composite of the matching conditions (4.12) relating the wavefunction coefficients across each interface in order, from one end of the device to the other. To form the STM, first use (4.12) to relate the coefficients across the $n = 0$ interface (between the I-contact and first device region) and the $n = 1$ interface:

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix} = \begin{bmatrix} i_0 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} = \begin{bmatrix} i_0 \end{bmatrix} \begin{bmatrix} I \\ R \end{bmatrix}, \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} = \begin{bmatrix} i_1 \end{bmatrix} \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}. \tag{4.14}$$

Now substitute the expression for the $n = 1$ region coefficients from the first equation into the second, to relate the wavefunction coefficients in the I-contact to those in the second device region:

$$\begin{bmatrix} a_2 \\ b_2 \end{bmatrix} = \begin{bmatrix} i_1 \end{bmatrix} \left(\begin{bmatrix} i_0 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} \right) = \begin{bmatrix} i_1 \end{bmatrix} \begin{bmatrix} i_0 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}. \tag{4.15}$$

Note that the parentheses can be removed in (4.15) using the associative property of matrix multiplication[9] [23]. However, matrix location (order) in the product must be maintained, since matrix multiplication is not commutative[10] [23]. Finally, note that the product of two 2x2 matrices is again a 2x2 matrix.

Following the approach indicated in (4.15), the coefficient relationship is expanded to include additional interfaces. By bridging the relationship between the I- and T-contact regions (i.e., across all interfaces from $n = 0$ to $n = N$), the STM is determined. Thus:

$$\begin{bmatrix} a_{N+1} \\ b_{N+1} \end{bmatrix} = \begin{bmatrix} i_N \end{bmatrix} \begin{bmatrix} i_{N-1} \end{bmatrix} \cdots \begin{bmatrix} i_1 \end{bmatrix} \begin{bmatrix} i_0 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} = \left(\prod_{n=0}^{N} \begin{bmatrix} i_n \end{bmatrix} \right) \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} \equiv \begin{bmatrix} \mathrm{STM} \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}. \tag{4.16}$$

8. Other approaches to calculating the STM, which are often more efficient although their derivations are more complex, are described in Section 4.4.
9. If A, B, and C are matrices such that the product ABC can be performed, then (AB)C = A(BC).
10. If A and B are matrices, in general, AB ≠ BA.

The matrix product in parentheses is the STM:

$$\left[\text{STM}\right] = \left[i_N\right]\left[i_{N\text{-}1}\right]\cdots\left[i_1\right]\left[i_0\right] = \prod_{n=0}^{N}\left[i_n\right]. \tag{4.17}$$

The transfer matrices $[i_n]$ are constants, depending only on the general solutions of the TISE in the associated regions of the device [1]. Thus, the STM is a constant 2x2 matrix, where each of the 4 elements is (in general) a complex number.

Every significant computation in a numerical simulator is also an opportunity, even an obligation, to find efficiencies that can be exploited. The computation of the STM is certainly such a computation, and SQUADS uses several “tricks” to make this computation more efficient. The most obvious is discussed in the remainder of this section. In many simulated devices, there are flat-band regions (where no energy band bending occurs) of significant extent adjacent to one or both contacts (*e.g.*, see Figure 4.4). Since the potential function in the flat-band region is the same as that in the adjacent contact, the TISE solutions (including coefficients) is also the same as that in the contacts.[11] Thus, the STM calculation can be confined to the active region over which band bending occurs, with exactly the same result. This can often reduce the length of the STM calculation by 50%. Suppose the flat-band regions are before interface $n1$ on the I-contact side, and after interface $n2$ on the T-contact side. SQUADS simply treats this as a general case of the STM derivation above, where $n1 = 0$ and $n2 = N$. The STM equation (4.17) now is

$$\left[\text{STM}\right] = \left[i_{n2}\right]\left[i_{n2\text{-}1}\right]\cdots\left[i_{n1+1}\right]\left[i_{n1}\right] = \prod_{n=n1}^{n2}\left[i_n\right]. \tag{4.18}$$

4.2.4.2 STM Calculation Complications

There are two important complications to the basic STM calculation that must be handled to ensure a robust and accurate TMM simulator. Both complications involve tunneling through extended classically forbidden (CF) regions between the two (classically allowed) contacts. In a real system, extended CF regions would just result in essentially total reflection of the quantum wave $(T = 0)$. In a numerical simulator, these cases require the consecutive multiplication of exponentially large and small numbers. This can result in numeric overflow (crashing or invalidating the calculation) or rounding error

11. Note that effective mass must also be constant to take advantage of a flat-band region, since the TISE solution depends on effective mass. SQUADS does check for effective mass variations.

(translating into an exponentially larger error in the STM result). To prevent this, if an extended CF region is encountered during the STM calculation, SQUADS mimics the real system result by terminating the STM calculation and taking $T = 0$ (total reflection) for that wave. Of course, this gambit should only be used when the transmission truly is negligible. Most waves will tunnel through some amount of CF region (such as the tunnel barriers in an RTD), but they can still contribute significantly to current flow. The task, therefore, is to establish a measure to judge the effective width of a CF region.

From Section 4.2.1, the wavefunction in a CCF region is

$$\psi(x) = a \cdot e^{\kappa x} + b \cdot e^{-\kappa x}. \tag{4.19}$$

Therefore, the proper measure of the "tunnel width" of a CCF grid region is the exponential "decay phase" of that region:

$$\Delta\phi_n \equiv \kappa_n \Delta x. \tag{4.20}$$

The total decay phase of a series of CCF grid regions is

$$\phi_d = \sum_{x_d} \kappa_n \Delta x, \tag{4.21}$$

where x_d is the position where the series of CCF regions starts. If ϕ_d exceeds a specified maximum decay phase $\phi_{\max}$ before the series of CCF regions ends, then the STM calculation is terminated and the transmission amplitude T for that energy is taken to be 0. SQUADS uses $\phi_{\max} = 20$, which gives a wavefunction decay factor of $e^{-20} \approx 2 \times 10^{-9}$. Nine orders of magnitude of attenuation ensures that these wavefunctions can not contribute significantly to current flow.

The second potential problem in the basic STM calculation involves a more direct source of numerical overflow. In a CCF region, the STM calculation requires the calculation of numbers of the form $e^{\kappa x}$. Depending on how large x and κ might be, a numerical overflow may occur in forming this number. Example calculations show that this situation is unlikely, but quite possible, depending on the maximum double precision number on the machine being used. A robust simulator must therefore protect against it. Actually, the total reflection conditions $(R = 1, T = 0)$ discussed above occur well before numerical overflow in most cases. In the few remaining cases, the decay rate of the wavefunction must be extremely high, so the best solution is simply to implement the total reflection gambit a little early. In SQUADS, the exponent limit was set at 10 below the maximum (typically about 700), leaving a few orders of magnitude breathing room to do computa-

tions with large numbers which are just below the limit. Rather then being “hard-wired”, this exponent limit is calculated at run time based on the maximum double precision number of the machine on which SQUADS is being executed.

4.3 Quantum Device Simulation Using the TMM

The first paragraph of this chapter described the TMM in very general terms. In Section 4.2, the foundation for a mathematical description of the TMM has been laid. This section completes the picture. Several important simulation tasks in the TMM investigation of quantum device operation are described in this section. The most basic task is the calculation of the current-voltage curve, which is described in Section 4.3.1. Other tasks described include the calculation of the wavefunction (Section 4.3.2), energy spectrum of carriers in the device (Section 4.3.3), and carrier density profile (Section 4.3.4).

4.3.1 Current-Voltage Curve Simulation

Perhaps the most basic goal of a TMM simulation is to determine the I-V characteristic of a quantum electronic device. Current density is independent of position in a steady-state system such as that modeled by the *time-independent* Schrödinger equation. Thus, current density can be calculated from the wavefunction at any point in the device. However, both coefficients of the wavefunction must be known, not just the general solution of the TISE, at some location in order to determine current. This section shows how the TMM determines both wavefunction coefficients at a single point in the device, and with these how current is calculated.

4.3.1.1 Determining the Transmission Amplitude

Figure 4.6 shows the abstraction of a typical RTD simulated by the TMM. Consider a quantum wave of kinetic energy $E - U_0$, which is incident on the system from the I-contact, is partially reflected back into the I-contact, and partially transmitted through the system into the T-contact. The TMM calculation assumes (for the moment) that there is no incident wave from T-contact. The obvious point at which to determine the wavefunction coefficients is one of the device contacts. To accomplish this, the general solution of the TISE in the contacts must be specified. SQUADS assumes ideal ohmic contacts (no potential drop outside the device),[12] so $U(x)$ in each contact is a constant. In particular, from the

boundary conditions (4.9a) and (4.9b), $U(x < 0) = U_o$ and $U(x > L) = U_N$.

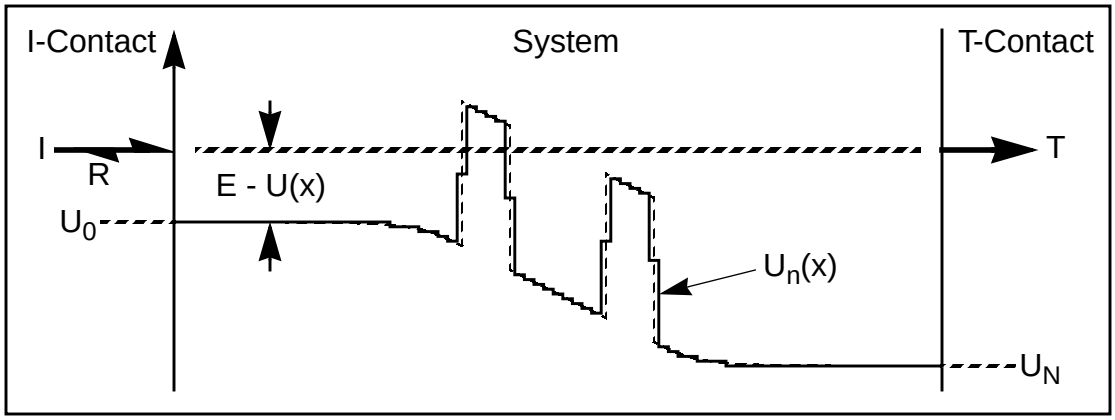


Figure 4.6: TMM quantum system abstraction

Each TMM calculation determines the transmission amplitude T and reflection amplitude R of an incident beam of energy E and amplitude I.

On the I-contact side (region 0), incident wavefunctions must have positive kinetic energy $(E > U_o)$. In other words, the I-contact is a CCA region, and the solution of the TISE in this region is given by (4.3a):

$$\text{I-Contact:} \quad \psi_0(x \le 0) = a_0 \cdot e^{ik_0 x} + b_0 \cdot e^{-ik_0 x}. \tag{4.22}$$

On the T-contact side, there are two possibilities. If the T-contact is a classically forbidden region (*i.e.*, $E < U_N$, usually due to a negative applied bias), the wavefunction will be totally reflected back to the I-contact, and there will be no transmission $(T = 0)$. Consequently, waves with $E < U_N$ do not contribute to current flow, and need not be considered in a current calculation. The only non-trivial case is where $E > U_N$, where the solution of the TISE in the T-contact is also given by (4.3a):

$$\text{T-contact:} \quad \psi_{N+1}(x \ge L) = a_{N+1} \cdot e^{ik_{N+1} x} + b_{N+1} \cdot e^{-ik_{N+1} x}. \tag{4.23}$$

As derived in Section 4.2.4, the relationship between the I-contact and T-contact wavefunction coefficients is written as

$$\begin{bmatrix} a_{N+1} \\ b_{N+1} \end{bmatrix} = \left[\text{STM}\right] \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}, \tag{4.24}$$

where T_{11} through T_{22} are complex numbers resulting from the multiplication of all of the

12. For accurate comparison of simulation and experiment, the system width L should be large enough to naturally accommodate all band bending between the contacts.

$[i_n]$ factors of the STM, as described in Section 4.2.4. In the I-contact, the coefficient of the FTW (*i.e.*, the incident amplitude) is denoted $a_0 = I$, and that of the BTW (the reflected amplitude) is $b_0 = R$. In the T-contact, the coefficient of the FTW (the transmitted amplitude) is denoted $a_{N+1} = T$. These changes give

$$\begin{bmatrix} T \\ b_{N+1} \end{bmatrix} = \begin{bmatrix} \text{STM} \end{bmatrix} \begin{bmatrix} I \\ R \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} I \\ R \end{bmatrix}. \tag{4.25}$$

(4.25) is still two equations in four unknowns, so it is still impossible to solve for anything - the entire exercise of generating the STM seems to gain nothing. Actually, because the I- and T-contacts are at the boundaries of the system, two boundary conditions on the TISE can be supplied (since it is a second-order differential equation) as the constraints necessary to make (4.25) solvable. Concerning boundary conditions, the TMM first assumes, as stated previously, that there is no incident wave (BTW) from the T-contact side $(b_{N+1} = 0)$. Second, a normalized incident wave $(I = 1)$ is used.[13] Thus, the normalized solutions in the contacts are:

$$\psi_0(x \le 0) = e^{ik_0 x} + R \cdot e^{-ik_0 x}, \tag{4.26a}$$

$$\psi_{N+1}(x \ge L) = T \cdot e^{ik_{N+1} x}. \tag{4.26b}$$

and (4.25) finally becomes two equations in two unknowns:

$$\begin{bmatrix} T \\ 0 \end{bmatrix} = \begin{bmatrix} \text{STM} \end{bmatrix} \begin{bmatrix} 1 \\ R \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} 1 \\ R \end{bmatrix} \tag{4.27}$$

$$\Rightarrow \quad R = -\frac{T_{21}}{T_{22}} \tag{4.28}$$

$$\Rightarrow \quad T = T_{11} + T_{12} R. \tag{4.29}$$

Note that, since T_{11} through T_{22} are complex numbers, the reflection amplitude R and transmission amplitude T are also complex, having both magnitude and phase. Thus, given the STM, the wavefunction *including coefficients* can be determined in either of the contacts. Since the wavefunction in the T-contact is simpler, it is invariably used in the current calculation. From another standpoint, it is more intuitive to calculate current flow from the transmission amplitude T, due to their conceptual similarity.

---

13. Both of these maneuvers are corrected later in current density calculation, which properly scales the result given the actual incident carrier distribution at both contacts.

A quantity related to the transmission amplitude T is the transmission coefficient:[14]

$$\Theta = (k_T/k_I)[T^*(E)T(E)] \ , \tag{4.30}$$

where T^* is the complex conjugate of T. The transmission spectrum $\Theta(E)$ (transmission coefficient versus incident wave energy) contains significant information in its own right, independent of the current calculation. For example, sharp peaks in $\Theta(E)$ indicate resonant energies in the device. Example transmission spectra are shown in Section 4.5.1.

4.3.1.2 Calculating Current

Upon determining the transmission spectrum $\Theta(E)$, current flow can be calculated using a modified Tsu-Esaki formula [24],

$$J(V_a) = \frac{qm\beta}{2\pi^2\tilde{h}^3}\int_0^\infty \Theta(E) ln\left(\frac{1 + exp[(E_{FI} - E)/\beta]}{1 + exp[(E_{FT} - E - qV_a)/\beta]}\right)dE \ , \tag{4.31}$$

where V_a is the applied bias across the device; $\beta = k_B\theta$, with θ being the temperature and k_B being the Boltzmann constant; and the E_F's are the I- and T-contact Fermi energies (w.r.t. the energy band).

(4.31) corrects for the two boundary conditions (given below (4.25)) used in calculating the transmission amplitude. To do this, it uses the fact that the transmission coefficient is the same for carriers incident from the left and right at energy E. Thus, it not only properly scales the current from the normalized transmission amplitude, but it also accounts for carriers incident from both contacts. More accurate (and more complicated) expressions for the TMM current have been derived [25, 26] and challenged [27]. These expressions are not currently implemented in SQUADS. However, SQUADS does use a slight generalization of (4.31) and (4.30), allowing the effective mass of the two contact materials to be different. This requires the ln term to be split up, giving:

$$\begin{aligned} J(V_a) &= \frac{qm_I\beta}{2\pi^2\tilde{h}^3}\int_0^\infty \Theta(E) ln\{1 + exp[(E_{FI} - E)/\beta]\}dE \\ &\quad - \frac{qm_T\beta}{2\pi^2\tilde{h}^3}\int_0^\infty \Theta(E) ln\{1 + exp[(E_{FT} - E - qV_a)/\beta]\}dE \ , \end{aligned} \tag{4.32}$$

where

14. Technically, the ratio of incident current to transmitted current.

$$\Theta(E) = \left(\frac{k_T}{k_I}\right)\left(\frac{m_I}{m_T}\right)[T^*(E)T(E)] \quad . \tag{4.33}$$

Another simple generalization is to allow the contacts to be at different temperatures, making $\beta \rightarrow \beta_I, \beta_T$, as appropriate. SQUADS does not currently implement this feature.

4.3.1.3 I-V Curve Simulation Overview

The complete procedure for calculating current-voltage curve of a quantum device using the TMM is as follows:

1) At a given incident wave energy E and applied bias V_a, compute the system transmission matrix $[\mathrm{STM}]$ using (4.17).
2) Calculate the transmission *amplitude* T for that incident wave using (4.29).
3) Determine the transmission *coefficient* of the system Θ from (4.33).
4) Repeat steps (1) - (3) to determine $\Theta(E)$ over the range of energies at which there are significant incident carriers from either contact.
5) Use (4.32) to calculate the current density at that applied bias.
6) Finally, repeat steps (1) - (5) over a desired range of applied biases, yielding the current-voltage curve.

Step (4) requires elaboration. In quantum devices like the RTD, transmission resonances (sharp peaks in the transmission spectrum) can be very narrow. Further, most of the current flow may be due to carriers at these resonant energies, since off-resonance transmission can be exponentially small. Thus, the energy spacing in the $\Theta(E)$ calculation must be very small to adequately resolve these resonances and accurately calculate current flow. Of course, doing more computation than necessary is almost as bad as doing too little. To this end, the range of energies should be restricted to only those which could possibly carry a significant amount of current. The resulting energy range is depicted in Figure 4.7. Of course, there is no point in calculating T for carriers incident at energies below the band minimum at the T-contact ($E < U_N$)—these will be totally reflected back to the I-contact [$T = 0$]. Also, there is no point computing T at very high energies where the number of incident carriers is negligible. In SQUADS, $E_{\max}$ is set at 15 $k_B\theta$ above the higher Fermi level (or 15 $k_B\theta$ above the higher band minimum if it is above the higher Fermi level). Thus, for purposes of calculating current flow via the TMM, SQUADS typi-

cally uses 1,000 to 10,000 energy points distributed evenly from the higher band minimum up to $E_{\max}$. Several TMM-simulated I-V curves are shown in Section 4.5.

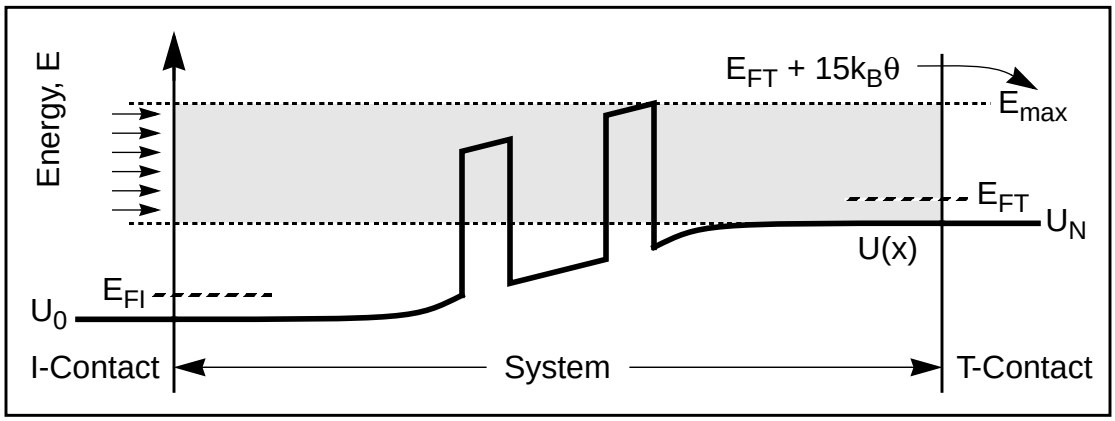


Figure 4.7: Energy range for T(E) calculation in I-V curve simulation

The shaded area shows the incident energy range over which the transmission spectrum $\Theta(E)$ is determined for the current calculation. Energies where total reflection would occur [$E < \max(U_0, U_N)$] or where there are neligible incident carriers [$E > \max(E_{\mathrm{FI}}, E_{\mathrm{FT}}, U_0, U_N) + 15k_B\theta$] are ignored. [$\theta$ is temperature.]

4.3.2 Calculating the Wavefunction

In the transfer-matrix method, the quantum wavefunction is the fundamental container of carrier information. In the TMM, to go beyond the terminal, I-V characteristics of a quantum device, an internal view of device operation requires calculation of the wavefunction. A single wavefunction shows how carriers at particular energy are behaving. Wavefunctions at a range of energies can be used to determine the energy spectrum (Section 4.3.3) and carrier density profile (Section 4.3.4). This section describes the calculation of the wavefunction for a continuous beam of carriers incident on the quantum device at energy E. This is exactly the entity for which the transmission coefficient Θ was calculated for the I-V curve, as described in the previous section. Although it is necessary to calculate wavefunctions incident from either contact, for a given wavefunction calculation, one contact (right or left) is the I-contact while the other is the T-contact. Thus, this contact naming scheme can (and will) be retained.

In Section 4.3.1, when $\Theta(E)$ was calculated, it was only necessary to determine the wavefunction (for each energy) at a single point: the T-contact. Now the task is to deter-

mine the wavefunction at all points, or rather, all grid points in the system being simulated. As shown in Section 4.3.2.1, the results of the $\Theta(E)$ calculation can often (but not always) be used to quickly determine the wavefunction at the device's internal grid nodes. Sections 4.3.2.2 and 4.3.2.3 then consider two complications that arise in the wavefunction calculation, and explain how they are addressed in SQUADS.

4.3.2.1 Basic Wavefunction Calculation

The task at hand is to calculate the wavefunction (numerical values - not functions and coefficients) at the grid nodes of the simulated system. From Section 4.2.1 the wavefunction in region n at node n can be written as

$$\psi(x_n) = a_n f_n(x_n) + b_n g_n(x_n). \tag{4.34}$$

Note that the wavefunction matching conditions (4.10a) and (4.10b) used in the STM calculation use the function values $f_n(x_n)$ and $g_n(x_n)$. It is a simple matter to store these values for later use during the wavefunction calculation.

What *isn't* known during the STM calculation are the wavefunction coefficients a_n and b_n. But as discussed in Section 4.2.3, given both wavefunction coefficients in any region, the wavefunction matching conditions can be used to determine the wavefunction coefficients in any other region. Thus, the STM calculation is a prerequisite to the calculation of the wavefunction: it gives the wavefunction coefficients in both contacts, and it gives all of the wavefunction coefficient relationships. The I-contact wavefunction coefficients are used as the "seed" to calculate the coefficients at all internal grid nodes using the matching conditions.

Of course, the matching conditions, like the TISE solution values, used during the STM calculation must also be stored for later use in the wavefunction calculation. This storage is minimal (typically 50 KB) if the wavefunction calculation is completed immediately after each STM calculation, and the storage is then reused for the wavefunction calculation at the next energy. The STM calculation is therefore completed as follows:

$$\left[p_0\right] \equiv \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \tag{4.35a}$$

$$\left[p_{n+1}\right] \equiv \left[i_n\right]\left[p_n\right] \qquad (n = 1, 2, \ldots, N) \tag{4.35b}$$

$$\Rightarrow \left[\mathrm{STM}\right] = \left[p_{N+1}\right]. \tag{4.35c}$$

Here, $[p_n]$ relates the region 0 coefficients across interface n to the region n+1 coefficients. After the STM calculation, both region 0 coefficients are known ($a_0 = 1$ and $b_0 = R$). Thus,

$$\begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} = [p_n] \begin{bmatrix} 1 \\ R \end{bmatrix}, \tag{4.36}$$

and the wavefunction coefficients of all other regions are quickly calculated.

This wavefunction calculation algorithm assumes that a product matrix $[p_n]$ is available from the STM calculation for every interface $n = 0, 1, \ldots, N$. However, Section 4.2.4 showed that, in many cases, the STM calculation can be significantly shortened by taking advantage of flat-band regions near the contacts. One result is that product matrices are not calculated and stored for grid nodes in the flat-band regions. However, rather than retreating to a full (and more expensive) STM calculation, SQUADS again takes advantage of the flat-band regions to achieve a more efficient computation. At grid nodes in the flat-band regions, the wavefunction is simply an extension of that in the contacts:

$$\text{Flat-band I-contact wavefunction: } \psi(x) = e^{ik_I x} + R \cdot e^{-ik_I x} \qquad (x < x_{n1}), \tag{4.37}$$

$$\text{Flat-band T-contact wavefunction: } \psi(x) = T \cdot e^{ik_T x} \qquad (x > x_{n2}). \tag{4.38}$$

The STM calculation produced values for R and T, so these functions can be easily evaluated at the relevant grid points x_n.

Recalling from Section 4.2.4 all of the effort required to calculate the STM, it should be clear from the above description that calculation of the wavefunction requires relatively little additional effort. Unfortunately, several complications, described below, make the wavefunction calculation more difficult in some cases than this simple picture portrays.

4.3.2.2 Classically Forbidden T-Contact

The main complication in the calculation of the wavefunction stems from the fact that the transmission coefficient calculation is only done at incident energies that are above both contact energy band minima. Thus, in Figure 4.7, the transmission coefficient was not calculated for $E < U(L)$, since the result would be zero (total reflection). In fact, the meaning of a system transmission matrix, not to mention its calculation, is dubious for carriers incident at these energies where the T-contact is classically forbidden. However, it is often necessary (*e.g.*, in the calculation of the carrier density) to determine the wavefunction

even when the incident beam is eventually totally reflected. This section discusses the calculation of the wavefunction in this case of a classically forbidden (CF) T-contact. Figure 4.8 shows an example of such a wavefunction.

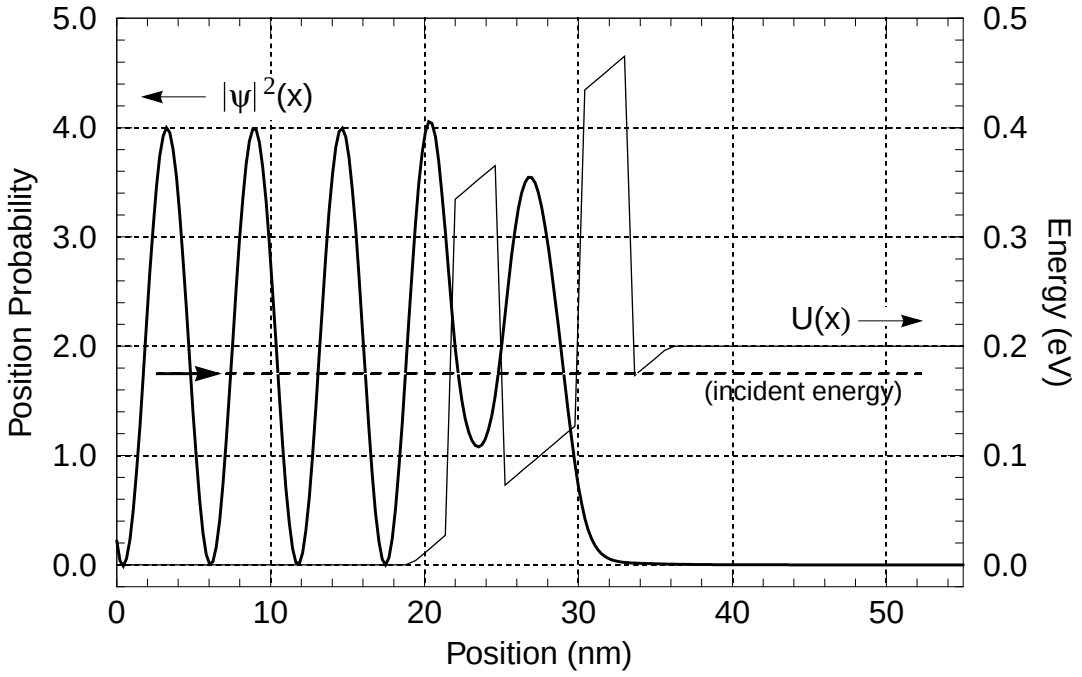


Figure 4.8: Wavefunction incident at energy below T-contact minimum

The position probability is shown for a wavefunction incident on a reverse-biased RTD at an energy just below U_N. In this case, T=0 and |R|=1.

To calculate the wavefunction when the T-contact is classically forbidden [$U(L) > E$], SQUADS changes the standard procedure only slightly. The STM calculation proceeds and the product matrices $[p_n]$ are formed just as with a classically allowed T-contact:

$$[p_n] = [i_n][p_{n\text{-}1}] \ , \tag{4.39}$$

$$\begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} = [p_n] \begin{bmatrix} 1 \\ R \end{bmatrix} . \tag{4.40}$$

In order to determine R from this equation (to seed the full wavefunction calculation), either a_{n+1} or b_{n+1} must be specified in some region so that there are only two unknowns in (4.40). Essentially, a new second boundary condition is needed. From Section 4.2.1, if the T-contact is a CF region, the TISE solution here is:

$$\psi_{N+1}(x) = a_{N+1} \cdot e^{\kappa_{N+1} x} + b_{N+1} \cdot e^{-\kappa_{N+1} x} . \tag{4.41}$$

Since the T-contact region extends (in the simulation abstraction) to infinity, a physically permissible wavefunction in a CF T-contact region can not have an exponentially growing component, or the wavefunction would grow without bound. The coefficient a_{N+1} must therefore be zero. This will serve as the second boundary condition.

SQUADS could calculate the product matrices up to the T-contact and then set a_{N+1} to 0 in (4.40), which would indeed give R. As usual, however, this is a special case where the length of the T-contact flat-band region is zero. In general, if the T-contact has a flat-band region after node $n2$, the wavefunction coefficients are constant in this region. Thus, the coefficient of the exponentially growing piece can be set to zero as the STM calculation crosses into the CF T-contact flat-band region at node $n2$. Further, product matrix $[p_{n2}]$ is the STM, since it relates the wavefunction coefficients in the I-contact to those in the T-contact flat-band region, and thus in the T-contact itself. Then (4.40) becomes

$$\begin{bmatrix} 0 \\ b_{n2+1} \end{bmatrix} = \left(\prod_{n\,=\,n1}^{n2} \begin{bmatrix} i_n \end{bmatrix} \right) \begin{bmatrix} 1 \\ R \end{bmatrix} = [p_{n2}] \begin{bmatrix} 1 \\ R \end{bmatrix} = [\mathrm{STM}] \begin{bmatrix} 1 \\ R \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} 1 \\ R \end{bmatrix}. \tag{4.42}$$

R and b_{n+1} are found as:

$$R = -\frac{T_{11}}{T_{12}}, \tag{4.43a}$$

$$b_{n2+1} = T_{21} + T_{22}R = T_{21} - \frac{T_{11}T_{22}}{T_{12}} = \frac{T_{12}T_{21} - T_{11}T_{22}}{T_{12}}, \tag{4.43b}$$

which are different than the expressions for R and T for a classically-allowed T-contact. Given R, SQUADS uses the same approach as in Section 4.3.2.1 to calculate the wavefunction for nodes $0 \le n < n2$. For nodes $n2$ through N, the TISE solution for the T-contact CF region is simply evaluated at the grid nodes x_n:

$$\psi(x_n) = b_{n2+1} \cdot e^{-\kappa_T x_n}. \tag{4.44}$$

4.3.2.3 Quantum Turning Points

There is yet a further complication in the wavefunction calculation for certain CF T-contacts, or, in fact, where the device includes *any* extended CF region for some incident energies, even if the T-contact is classically allowed. In the latter case, recall from Section 4.2.4.2 that the STM calculation was abandoned and the transmission amplitude set to 0 when an extended CF region was encountered. Again, it is necessary to compute the wave-

function even in such cases, so the computation can not simply be abandoned. The trick will be to introduce what will be called a quantum turning point (QTP) as the second boundary condition. This artifice will terminate the STM calculation cleanly, enable the calculation of the reflection amplitude, and thereby the calculation of the wavefunction. This section describes SQUADS implementation of QTPs and the calculation of the wavefunction in these cases.

A quantum turning point is herein defined as the point at which an incident wave has traversed enough CF space that its decay phase exceeds ϕ_{max}, as defined in Section 4.2.4.2. Recall from the STM calculation that at this point, the wavefunction amplitude has decayed by about 9 orders of magnitude from its incident value, and has thus effectively been totally reflected back to the I-contact. To minimize both unnecessary computation and numerical error of multiplying more exponentials together, SQUADS mathematically inserts (for the calculation of this one wavefunction) a totally reflecting barrier to reflect the tiny remainder of the wavefunction, thus terminating the wavefunction at this quantum turning point. This introduces negligible error, since the wavefunction is already highly attenuated.

To calculate the wavefunction in the case of a QTP, the procedure is almost identical to that for CF T-contact: at node n, the STM calculation gives:

$$\begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} = [p_n] \begin{bmatrix} 1 \\ R \end{bmatrix}. \tag{4.45}$$

When the STM calculation finds that the decay phase has exceeded ϕ_{max} during grid interval $n = n_{qtp}$, a quantum turning point is inserted at the next interface (node n_{qtp}). A totally reflecting interface is created by assuming that region n_{qtp+1} (just after the QTP) has an infinitely high potential, $U_{qtp+1}(x) \approx \infty$. This has the unfortunate effect of making some of the matrix elements in $[i_{qtp}]$ infinite. Thus, implementing the QTP boundary condition requires a little finesse.

To implement a second boundary condition, SQUADS uses the fact that the wavefunction must be continuous, even at an infinite discontinuity in the potential energy $U(x)$, such as at a QTP. In particular, since the wavefunction is 0 just after the QTP, it must go to zero just before it. Thus, in region n_{qtp} at the QTP, the general solution of the TISE is:

$$\psi_{qtp}(x_{qtp}) = a_{qtp} \cdot f_{qtp}(x_{qtp}) + b_{qtp} \cdot g_{qtp}(x_{qtp}) = 0, \tag{4.46}$$

$$\Rightarrow \qquad a_{\text{qtp}} \cdot f_{\text{qtp}}(x_{\text{qtp}}) = -b_{\text{qtp}} \cdot g_{\text{qtp}}(x_{\text{qtp}}) \equiv X, \tag{4.47}$$

where X is just a place-holder name. (4.47) will be used as the second boundary condition in the case of a QTP. To incorporate it, first write the product matrix $[p_{\text{qtp-1}}]$, as the relationship between the I-contact coefficients to those in region n_{qtp} (just before the QTP):

$$\begin{bmatrix} a_{\text{qtp}} \\ b_{\text{qtp}} \end{bmatrix} = [p_{\text{qtp-1}}] \begin{bmatrix} 1 \\ R \end{bmatrix}. \tag{4.48}$$

Next, multiply both sides of (4.48) by the following 2x2 matrix:

$$\begin{bmatrix} f_{\text{qtp}}(x_{\text{qtp}}) & 0 \\ 0 & g_{\text{qtp}}(x_{\text{qtp}}) \end{bmatrix} \equiv [i_{\text{qtp}}], \tag{4.49}$$

$$\begin{bmatrix} a_{\text{qtp}} \cdot f_{\text{qtp}}(x_{\text{qtp}}) \\ b_{\text{qtp}} \cdot g_{\text{qtp}}(x_{\text{qtp}}) \end{bmatrix} = [i_{\text{qtp}}][p_{\text{qtp-1}}] \begin{bmatrix} 1 \\ R \end{bmatrix} \equiv [\text{STM}] \begin{bmatrix} 1 \\ R \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} 1 \\ R \end{bmatrix}. \tag{4.50}$$

Now use (4.47), and solve for R:

$$\begin{bmatrix} X \\ -X \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} 1 \\ R \end{bmatrix}, \tag{4.51}$$

$$\Rightarrow \qquad R = -\left(\frac{T_{11} + T_{21}}{T_{12} + T_{22}}\right). \tag{4.52}$$

Having calculated R, the same approach as in Section 4.3.2.1 is used to calculate the wavefunction at node 0 through $n_{\text{qtp-1}}$. For nodes n_{qtp} through N, the wavefunction is zero, since the QTP is totally reflecting. Note that in the rare cases where a QTP is inserted due to impending numerical overflow, as described at the end of Section 4.2.4.2, the QTP is treated no differently for purposes of wavefunction calculation that a "normal" QTP. Finally, note that for all cases of wavefunction calculation (partial transmission, CF T-contact, or QTP), the STM calculation always produced some reflection amplitude R. In contrast, only in the first case was there a transmission amplitude T. Therefore, for consistency, the wavefunction calculation in SQUADS always uses the I-contact coefficients as the "seed" to initiate the wavefunction calculation.

4.3.3 Calculating the Energy Spectrum

A useful derivative of the wavefunction calculation is the calculation of the energy spectrum $P(E)$ of carriers in various regions of the device. The energy spectrum is simply

the number of carriers (probability density) versus energy at a given location. SQUADS allows specific quantum devices to be defined, as well as regions for that device. For example, the basic resonant tunneling diode has three device regions: the quantum well and the two contacts (separated by the tunnel barriers). During an energy spectrum calculation, SQUADS records the maximum amplitude of each wavefunction calculated in each region. This capability can be used, for example, to determine the energy and energy-width of resonant states in the quantum well of the RTD. The wavefunctions with the highest amplitude are at the resonant energy. Resonant states are critical to the operation of quantum devices, and being able to locate and map these states in detail is an important (albeit rare) feature of quantum device simulators.

Normally, the transmission spectrum $\Theta(E)$ gives essentially the same information as the energy spectrum $P(E)$, since $\Theta(E)$ has peaks exactly at the resonant energies. However, as should be clear from the wavefunction calculation discussion in Section 4.3.2, the transmission spectrum is not available for incident energy ranges where a CF T-contact or QTP occurs. However, there may still be resonant states in the device for wavefunctions incident at these energies. Chapter 8 describes such a case, and the energy spectrum feature of SQUADS was crucial in clarifying some particularly interesting RTD behavior in this case.

4.3.4 Calculating the Carrier Density Profile

A more common use of the wavefunction calculation in simulations is for the determination of the carrier density profile $c(x)$ (density of carriers versus position) in the device. $c(x)$ is useful in determining how an electronic device is operating, and it is an essential ingredient in implementing self-consistency [agreement between $c(x)$ and $U(x)$] for more accurate simulations (see Chapter 6). This section describes the calculation of the carrier density profile using the transfer-matrix method of quantum device simulation.

The basic strategy of the carrier density calculation is to add up the densities due to individual wavefunctions over the energy range where the number of incident carriers is significant. The carrier density calculation is thus similar to the current calculation, but there are some important differences. Figure 4.9 helps to illustrate the similarities and differences. With the current calculation, waves are incident at only one contact, since the transmission coefficient at a given energy is the same in either direction (see Section

4.3.1.2). Further, any energy at which the T-contact is classically forbidden (dark shading in Figure 4.9) is not considered, and any calculation during which a QTP is encountered (medium shading in Figure 4.9) is simply abandoned. In contrast, in the carrier density calculation, all energies at which there are a significant number of incident carriers must be considered. This includes those cases where the opposite contact is classically forbidden or where a QTP is encountered. The result is that the wavefunctions of carriers incident from each contact must be considered separately. The distribution and range of incident energies are identical with the carrier density and current calculations, however.

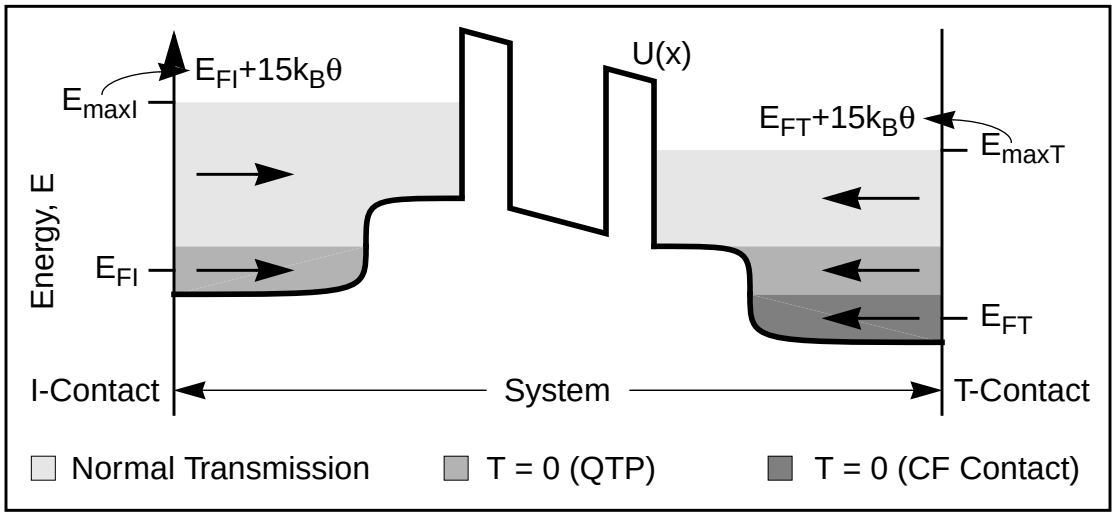


Figure 4.9: Classes of carriers during carrier density calculation

Even carriers which have no transmission ($T = 0$, due to a QTP or CF contact) must be included in the carrier density calculation. Carriers incident from each contact must be considered separately, since the two contacts include different energy ranges.

Considering the carrier density calculation, then, given wavefunction $\psi(x)$, which is the normalized position probability amplitude of a beam of carriers at energy E, the carrier density due to this wavefunction is simply $|\psi(x)|^2$. The total carrier concentration requires a summation (integration) over all wavefunctions from each contact, each multiplied by the respective number of carriers incident at that energy. In SQUADS, the integration is actually carried out over wavevector k (4.4), rather than E, as in the current calculation, which enables an integration only over incoming waves from each contact. Once again, SQUADS modifies the standard formula for carrier density [10, 21] by allowing different effective masses at the two contacts, giving:

$$c(x) = \frac{2m_I\beta}{h^2}\int_0^\infty \left|\psi_{IT}(x)\right|^2 ln\{1 + exp[(E_{FI} - E_I)/\beta]\}dk$$
$$+ \frac{2m_T\beta}{h^2}\int_{-\infty}^0 \left|\psi_{TI}(x)\right|^2 ln\{1 + exp[(E_{FT} - E_T - qV_a)/\beta]\}dk \quad . \qquad (4.53)$$

4.3.5 Calculating the Wigner Function

One final feature of the TMM simulator in SQUADS is the ability to view the results as a Wigner function $f_w(x, k)$. As discussed in Chapter 3, the Wigner function is a very intuitive and efficient way to view the state and operation of a quantum device. Like the carrier density $c(x)$, the Wigner function contains a composite view of all wavefunctions computed. However, the Wigner function also shows the number of carriers at each wavevector (which is proportional to velocity) as well as position, so it contains much more information about carrier behavior than the carrier density profile. To maintain a reasonable complexity in this discourse, only the basics of the Wigner function calculation from individual wavefunctions are presented here. The interested reader is referred to [1] for details of the derivations and numerical implementation of this calculation in SQUADS.

The Wigner function $f_w(x, k)$ is calculated from the wavefunctions via an intermediate entity: the density matrix $\rho(x_1, x_2)$. The density matrix calculation is very similar to the carrier density calculation[15], being essentially a summation of all wavefunctions of significant amplitude in the system. However, instead of adding probabilities, the density matrix calculation uses a *correlation* of the wavefunction amplitude at one point with that at another point. That is:

$$\rho(x_1, x_2) = \frac{2m_I\beta}{h^2}\int_0^\infty [\psi_{IT}(x_1)\psi_{IT}^*(x_2)] ln[f(E_I)]dk$$
$$+ \frac{2m_T\beta}{h^2}\int_{-\infty}^0 [\psi_{TI}(x_1)\psi_{TI}^*(x_2)] ln[f(E_T)]dk \ , \qquad (4.54)$$

where $f(E_I) \equiv 1 + \exp[(E_{FI} - E_I)/\beta]$ and $f(E_T) \equiv 1 + \exp[(E_{FT} - E_T)/\beta]$. Note that $\rho(x_1, x_1) = c(x_1)$. The Wigner function is calculated from the density matrix via a series of Fourier transforms at successive fixed positions x:

15. In fact, SQUADS uses the same subroutine to calculate the carrier density and the density matrix, with only a few lines of code specific to each calculation.

$$f_w(x, k) = \int_{-\infty}^{\infty} e^{-iky} \rho(x + y/2, x - y/2) dy. \quad (4.55)$$

It is not difficult to show [1] that the Wigner function $f_w(x, k)$ is a real (as opposed to complex) function. Figure 4.10 shows the Wigner function for the RTD in Figure 4.8, but at a positive bias of 0.4 V. The Wigner function will be described in more detail in Chapter 5, but for now note the beam (small ridge) of carriers which have tunneled through the RTD and are exiting the RTD to the right.

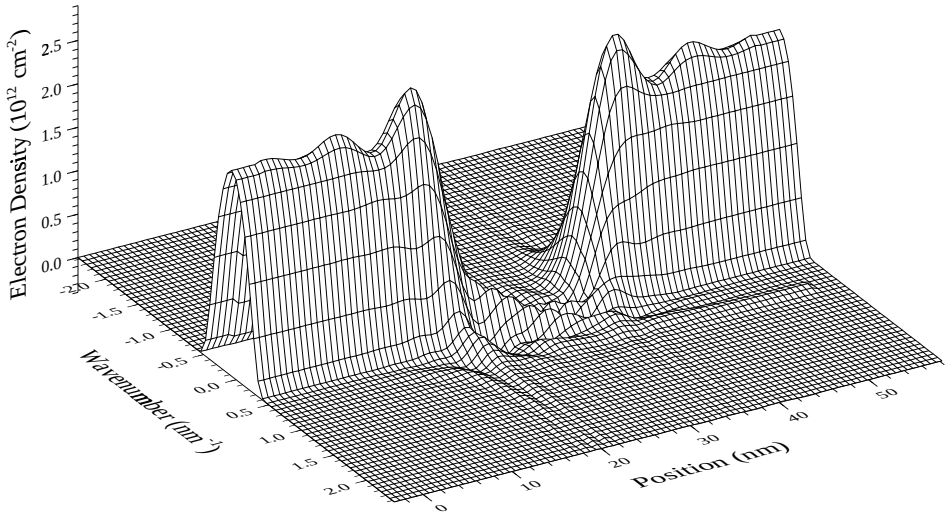


Figure 4.10: Wigner function calculated from TMM wavefunctions

The TMM-calculated Wigner function $f_w(x, k)$ is shown for an RTD at 0.4 V bias. The Wigner function shows the number of carriers versus position and velocity (actually wavevector) in the device. The small beam of carriers travelling at high velocity into the T-contact have tunneled through the quantum well state.

4.4 Alternative Implementations

This section treats several alternative implementations of the transfer-matrix method of quantum device simulation. As shown, these alternatives are often more accurate or efficient than the standard implementation described in Sections 4.2 and 4.3.

4.4.1 Node-Centered Regions

One simple but significant modification of the TMM is to use node-centered grid regions, where grid interfaces are half-way between nodes rather than at the nodes, as in the (perhaps universally used) node-bounded gridding scheme described in Section 4.2.2. The main benefit of node-centered gridding is depicted in Figure 4.11, which shows a simple potential profile $U(x)$, the points U_n extracted from $U(x)$, and the piece-wise-constant, node-bounded and node-centered approximations to the potential as determined from the U_n. Node-bounded gridding uses $U_n(x) = \frac{1}{2}(U_{n-1} + U_n)$, which results in poor fidelity to the actual potential profile near heterojunctions, as shown. In contrast, node-centered gridding uses $U_n(x) = U_n$: the potential approximation of the region is just the potential at the node in the center of that region. Node-centered gridding tends to give much better agreement between the actual and piece-wise-constant potentials, even though the grid points x_n and potential values U_n are exactly the same as for node-bounded gridding. The fidelity of node-centered gridding is especially good when device layer widths are some multiple of the atomic spacing in the real materials, and the position grid nodes are spaced one lattice constant apart, as discussed in Section 4.2.2. This puts material interfaces half-way between nodes, which coincides with the node-centered region interfaces.

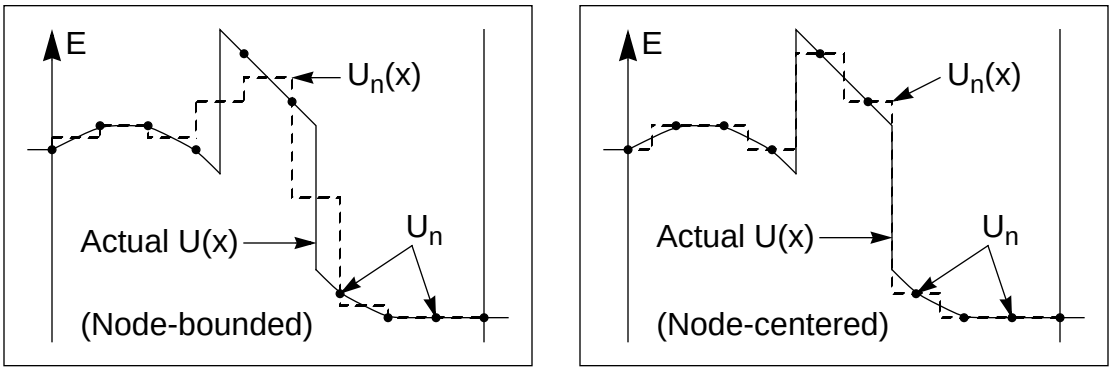


Figure 4.11: Node-bounded/node-centered gridding and TMM potentials

TMM potential approximations $U_n(x)$ for node-bounded and node-centered gridding schemes are shown. Node-centered gridding tends to give significantly better approximation of the actual potential $U(x)$ near heterojunctions.

There is also no confusion with node-centered gridding about what to use for the effective mass in each region. As with the U_n values, effective mass values m_n are supplied to the TMM simulator at the grid nodes. For node-centered gridding, m_n in region n

is just that at the associated node. For node-bounded gridding, the average of the effective masses at the two bounding nodes had to be used, as discussed in Section 4.2.2.

Since node-centered gridding is clearly superior to the standard node-bounded gridding scheme, the obvious question is why node-centered gridding isn't the standard. The fact that the grid nodes and grid interfaces are not coincident with node-centered gridding results in a more complicated TMM implementation and computation. The main complications of node-centered gridding include the following:

- The transfer-matrix terms $[i_n]$ are slightly more complicated.
- Wavefunction values used for the STM calculation (at internode points) can not be reused during the wavefunction calculation (at node points).
- More care is required in checking for QTPs and numeric overflow.
- The normalization algorithm (see Section 4.4.2.2) is more complicated.
- Linear potential interpolation (see Section 4.4.3) is more complicated.

These difficulties require more careful coding, and are slower for some calculations, but the use of node-centered gridding has significant benefits for the simulation. Therefore, SQUADS does implement this gridding scheme. The details and complications of node-centered gridding [1] will not be covered in detail here. However, the last two issues listed above will be mentioned again in Sections 4.4.2.2 and 4.4.3. Sections 4.5.1 and 4.5.2 compare the accuracy and computational efficiency, respectively, of the node-bounded and node-centered gridding schemes.

4.4.2 Alternate STM Calculation Algorithms

The STM calculation scheme described in Section 4.2.4 is the most straight-forward way to develop a relationship between the I-contact and T-contact wavefunction coefficients, and thus to solve for the current or the rest of the wavefunction. However this scheme [2, 7, 17, 19, 5], hereafter called the interface algorithm, is not the only reasonable STM calculation method. This section describes two others, the region algorithm and the normalization algorithm, which are often more efficient and more numerically robust (have less chance of numerical overflow). A comparison of the computational efficiency of these algorithms is given in Section 4.5.2. Both the normalization algorithm [4] and the region algorithm [3, 8] have been used by other groups. Most publications which use the transfer-matrix method do not describe the STM calculation algorithm used.

4.4.2.1 Region Algorithm

The region algorithm follows the standard interface algorithm derivation in most details, but it is designed for fast computation of the STM, and thus current. In the interface approach, the STM was calculated as (4.17):

$$[\mathrm{STM}] = [i_N][i_{N-1}]\cdots[i_1][i_0] = \prod_{n=0}^{N}[i_n]. \tag{4.56}$$

From Section 4.2.3, each 2x2 transfer matrix $[i_n]$ is the product of two 2x2 matrix factors, one associated with region n just before interface n and the other associated with region n+1 just after it, as indicated in Figure 4.12. The region formulation just regroups these matrix factors as follows:

$$\begin{aligned}
[\mathrm{STM}] &= [i_N][i_{N-1}]\cdots[i_1][i_0] \\
&= [i_{N+}][i_{N-}]\cdot[i_{(N-1)+}][i_{(N-1)-}]\cdot\ldots\cdot[i_{1+}][i_{1-}]\cdot[i_{0+}][i_{0-}] \\
&= [i_{N+}]\cdot[i_{N-}][i_{(N-1)+}]\cdot\ldots\cdot[i_{2-}][i_{1+}]\cdot[i_{1-}][i_{0+}]\cdot[i_{0-}] \\
&= [i_{N+}][r_N][r_{N-1}]\cdots[r_2][r_1][i_{0-}] \\
&= [i_{N+}]\left(\prod_{n=1}^{N}[r_n]\right)[i_{0-}],
\end{aligned} \tag{4.57}$$

where the matrix factors $[r_n]$ are associated with region n, as shown in Figure 4.12.

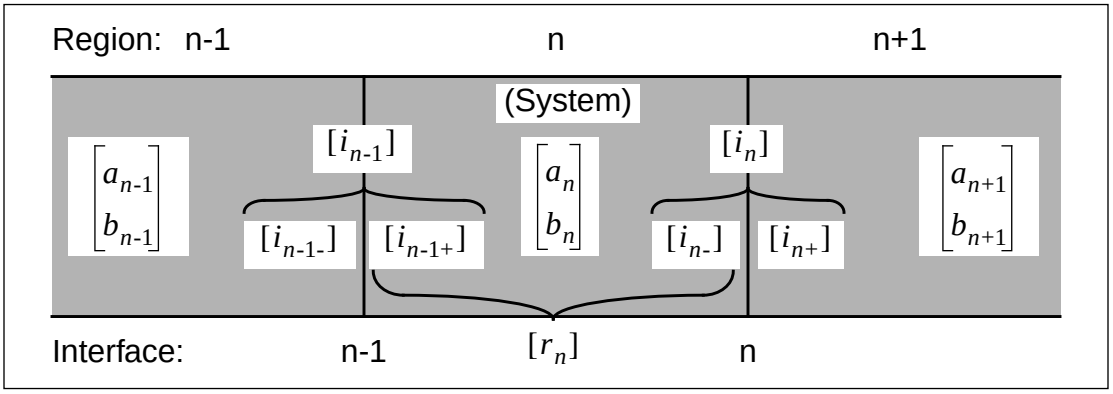


Figure 4.12: STM factor matrix for the region algorithm

The region STM calculation algorithm groups the matrix factors so that the two matrices associated with region n are used to compute transfer matrix $[r_n]$.

Further details of the region algorithm [1] will not be reproduced here, although the basic motivations and issues involved in its use will be described briefly. (4.57) would seem to entail the same amount of calculation as (4.17) for the interface algorithm. However, since there are only three region types (CCA, CCF, and CCN), there are only three forms for $[r_n]$, while there are nine forms of $[i_n]$ (one for each region pair). Further, the $[r_n]$ can be simplified much further than the $[i_n]$, since all factors of $[r_n]$ involve solutions of the TISE in a single region. Finally, numerical overflow, which is a potential problem in a CCF region with the interface algorithm (see Section 4.3.2), is improved with the region formulation. Whereas any $[i_n]$ involving a CCF region contains terms of the form $exp(\kappa_n x_n)$ (see [1]), terms in $[r_n]$ are of the form $exp(\kappa_n \Delta x)$, where Δx is the (small) position grid spacing. Flat-band regions and quantum turning points are handled the same in the region algorithm as in the interface algorithm (see Section 4.3).

Because $[r_n]$ is, in general, simpler than $[i_n]$, the region algorithm enables quicker calculation of the STM, and thus current, than the interface algorithm. However, the region formulation is not well suited to quick calculation of the wavefunction. Each region matrix $[r_n]$ transfers the wavefunction coefficient relationship *across* region n and halfway through the next interface, rather than into region $n+1$, Thus, the product matrices produced during the region STM calculation do not give the wavefunction coefficients in the regions, although these can be computed with additional work.

4.4.2.2 Normalization Algorithm

The third STM calculation algorithm implemented in SQUADS is called the normalization algorithm. Its advantages are similar to those of the region algorithm, but it can be used for efficient calculation of both the STM and the wavefunction. Describing the normalization algorithm requires rewriting the transfer-matrix equation (4.12)

$$\begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} = \begin{bmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{bmatrix}_n \begin{bmatrix} a_n \\ b_n \end{bmatrix} \equiv \begin{bmatrix} i_n \end{bmatrix} \begin{bmatrix} a_n \\ b_n \end{bmatrix} \tag{4.58}$$

as

$$\begin{bmatrix} a_{n+1} f_{n+1}(x_n) \\ b_{n+1} g_{n+1}(x_n) \end{bmatrix} = \begin{bmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{bmatrix}_n \begin{bmatrix} a_n f_n(x_n) \\ b_n g_n(x_n) \end{bmatrix} \equiv \begin{bmatrix} i_n \end{bmatrix} \begin{bmatrix} a_n f_n(x_n) \\ b_n g_n(x_n) \end{bmatrix}. \tag{4.59}$$

In words, the transfer-matrix equations in the normalization algorithm relate not the wave-

function *coefficients* (a and b) across the interface, but each coefficient *times* its respective TISE solution (af and bg). Although (4.59) doesn't look simpler than (4.58), the i_{11} - i_{22} terms *are* much simpler to calculate for each of the solutions of the TISE discussed in Section 4.2.1.

As was the case with the region algorithm, the news is not all good for the normalization algorithm. This formulation relates af and bg, not just a and b, across the device, and f and g vary across grid regions. To create a continuous chain relationship between the solution in the I-contact and that in the T-contact, it is necessary to include region matrices to incorporate this change. Luckily, the region matrices are quite simple for each of the region types considered in Section 4.2.1. Also, the elements in the region matrix for a CCF region are of the form $exp(\kappa_n \Delta x)$, as in the region algorithm, so the numerical overflow danger is minimal with the normalization algorithm.

Once again, the complications and details of implementing the normalization algorithm [1] will not be given here, although a few of the issues will be mentioned. Flat-band contact regions are handled in the normalization STM calculation with a single region matrix for a entire region. Quantum turning points are handled as in the other STM algorithms. Calculating the wavefunction is actually slightly easier with the normalization formulation, since the two pieces of the wavefunction (af and bg) are given by the process, rather than just the coefficients. However, calculating the wavefunction with node-centered regions and the normalization algorithm requires the multiplication of two half-region matrices to get the value of the wavefunction at the center of the region.

4.4.3 Piece-Wise Linear Interpolation

As mentioned in Section 4.2.1, the TISE (4.1) is analytically solvable for only a few potential functions $U(x)$. Thus far in this chapter, only constant potential regions have been considered, but linear and parabolic potentials also yield an analytically solvable TISE. It is not difficult to imagine that using such potential regions could produce a better piece-wise approximation of the actual potential. Many groups [7, 8, 19, 28-30] have implemented the TMM for linear potential regions, claiming a significant improvement in the accuracy of the TMM.[16] To test these claims and possibly achieve better accuracy, the ability to handle linear potential regions was implemented in SQUADS. This section over-

16. Apparently parabolic potentials have not been used yet in the TMM.

views this implementation. Later sections (4.5.1 and 4.5.2) will consider the accuracy and computational efficiency of the piece-wise-linear potential approximation scheme.

The solution of the time-independent Schrödinger equation (4.1) for a linear potential $U(x) = cx + d$ is a linear combination of the Airy functions, Ai and Bi [31]:

$$\psi(x) = a \cdot \mathrm{Ai}(z) + b \cdot \mathrm{Bi}(z)\,, \tag{4.60a}$$

$$z \equiv \gamma x + \beta \qquad \gamma = \left(\frac{2mc}{\tilde{h}^2}\right)^{1/3} \qquad \beta = \frac{\gamma(d-E)}{c}\,. \tag{4.60b}$$

Since the TMM is supplied the potential values U_n at the grid nodes, the obvious scheme for creating linear potential grid regions is simply to connect the points:

$$U_n(x) = \left(\frac{U_n - U_{n\text{-}1}}{x_n - x_{n\text{-}1}}\right)(x - x_{n\text{-}1}) + U_{n\text{-}1} = c_n x + d_n \quad . \tag{4.61}$$

This adopts the node-bounded gridding scheme, and is shown on the left in Figure 4.13. Also shown in Figure 4.13 is the node-centered, piece-wise-linear scheme, which clearly has less error than all other potential approximations schemes yet considered. However, implementing node-centered gridding with linear regions is problematic, requiring the estimation of the potential derivative at the grid points. The location of all abrupt band offsets must be known and accounted for in this estimation, or the approximation will be very poor near heterojunctions. For all other gridding and potential approximation schemes, only the points U_n had to be known. Due to the difficulty of handling general device structures, the node-centered, piece-wise-linear combination is not implemented in SQUADS.

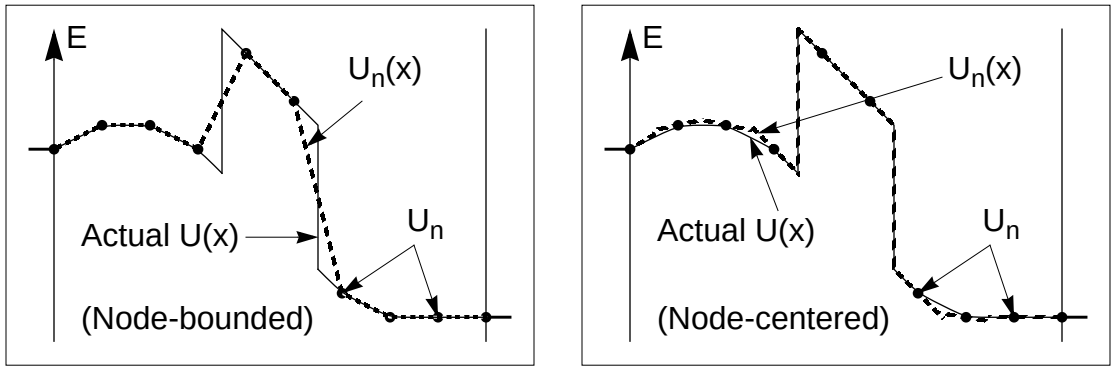


Figure 4.13: Node-bounded/node-centered regions with linear potentials

TMM potential approximations $U_n(x)$ for the node-bounded and node-centered gridding schemes with linear potential regions are shown. As with constant potential regions, node-centered gridding tends to give a significantly better approximation of the actual potential U(x) near heterojunctions, but its implementation is problematic.

Given the TISE solution in each region, the calculation of the STM, current, wavefunctions, carrier density, and Wigner function proceed just as before. Again, details are given in [1], and will not be repeated here. A few significant details will be mentioned. First, note that the addition of a fourth region type increases the number of possible interface transfer matrices to 16 in both the interface and normalization STM algorithms. This makes the division of the interface transfer matrix into two parts (each of which has only four possible forms) quite essential (see Section 4.2.3). Adding a fifth region type (e.g., parabolic) in the future would further necessitate this approach. A second issue is that the transfer-matrices do not simplify with linear regions as they did with constant regions when using the region and normalization formulations. This will be apparent in the comparison of algorithm efficiencies in Section 4.5.2. Quantum turning points are also more difficult to locate, since determining the decay phase is more complicated. Finally, numerical overflow is a possibility with linear regions, regardless of the STM algorithm (because simplification is not possible), even when a QTP is not indicated. In particular, the Bairy function and its derivative grow exponentially with z. This occurs, for example, when c becomes very small (a nearly constant potential). This numerical overflow is avoided in SQUADS by using a constant potential region when z exceeds a (platform-dependent) calculated limit, as recommended in [8].

4.5 Simulation Results

This section serves a dual purpose. First, it presents the results of three investigations of the relative merits of various alternative implementations for transfer-matrix method simulation. Second, this section shows the basic TMM simulation capabilities of SQUADS. The three investigations include a comparison of linear versus constant potential region simulation results (Section 4.5.1), a comparison of the computational efficiencies of the various STM calculation algorithms (Section 4.5.2), and an investigation of the significance of using a position-dependent effective mass (Section 4.5.3). All of these simulations use the resonant tunneling diode (RTD) shown in Figure 4.14 (and previously in Figures 4.8 and 4.10). Simulation parameters used are given below the figure.

4.5.1 Accuracy of Linear versus Constant Potential Regions

As mentioned in Section 4.4.3, many groups using the TMM for quantum device sim-

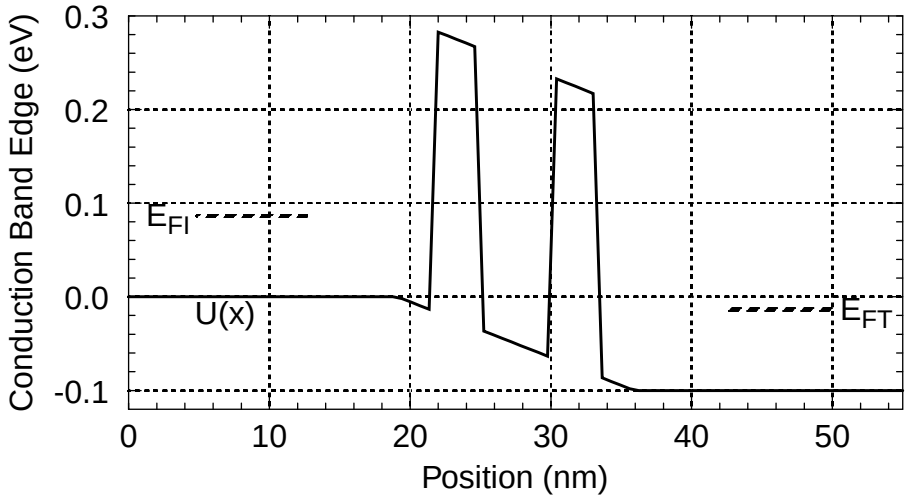


Figure 4.14: Conduction band profile of RTD used in TMM investigations

The test RTD is shown at a bias of 0.1 V, and the potential is assumed to drop linearly across the central "active" region. The RTD is composed of a 5 nm GaAs quantum well between 3 nm $Al_{0.3}Ga_{0.7}As$ tunnel barriers and 3 nm GaAs spacer layers. The GaAs contact layers are 19 nm each, giving a total simulation width of L = 55 nm. Except in Section 4.5.3, electron effective mass is assumed constant at $0.0667m_0$, and permittivity is also assumed constant at $12.9\varepsilon_0$. Also, these simulations use 86 position points, 10,000 energy points, and a temperature of 77K. Doping on both sides is $N_d=2e^{18}/cm^3$.

ulation, claim that using a piece-wise-linear (PWL) potential instead of a piece-wise-constant (PWC) potential significantly improves TMM simulation accuracy. This section presents the results of several TMM simulations to determine, at least for the RTD and simulation parameters in Figure 4.14, whether these claims are valid.

The first set of simulation results are shown in Figure 4.15, which compares the transmission spectrum $\Theta(E)$ (transmission coefficient versus energy) for the three potential approximation schemes discussed in this chapter: a PWC potential (both node-bounded (NB) and node-centered (NC) gridding) and a PWL potential. The main result is that differences in the transmission spectrum are quite small throughout the energy range considered. In the critical first transmission peak (through which most of the current flows), the shape and size of the peak are indistinguishable for the three schemes, and the location of the peak varies by only 1-2 meV. The second transmission valley does show some difference, but there are virtually no incident carriers at this high of energy. Another conclusion is that the PWC/NC scheme is (as expected) preferable than the PWC/NB approach since the PWC/NC spectrum is closer to the (presumably) more accurate PWL result. Finally,

the relative locations of the transmission peaks are not difficult to understand. The confining "strength" of the barriers is highest with the PWC/NC scheme and lowest in the PWC/NB case. A more confining quantum well shortens the wavelength and thus raises the energy of the resonant state. Of course, transmission is greatest at the resonant energy.

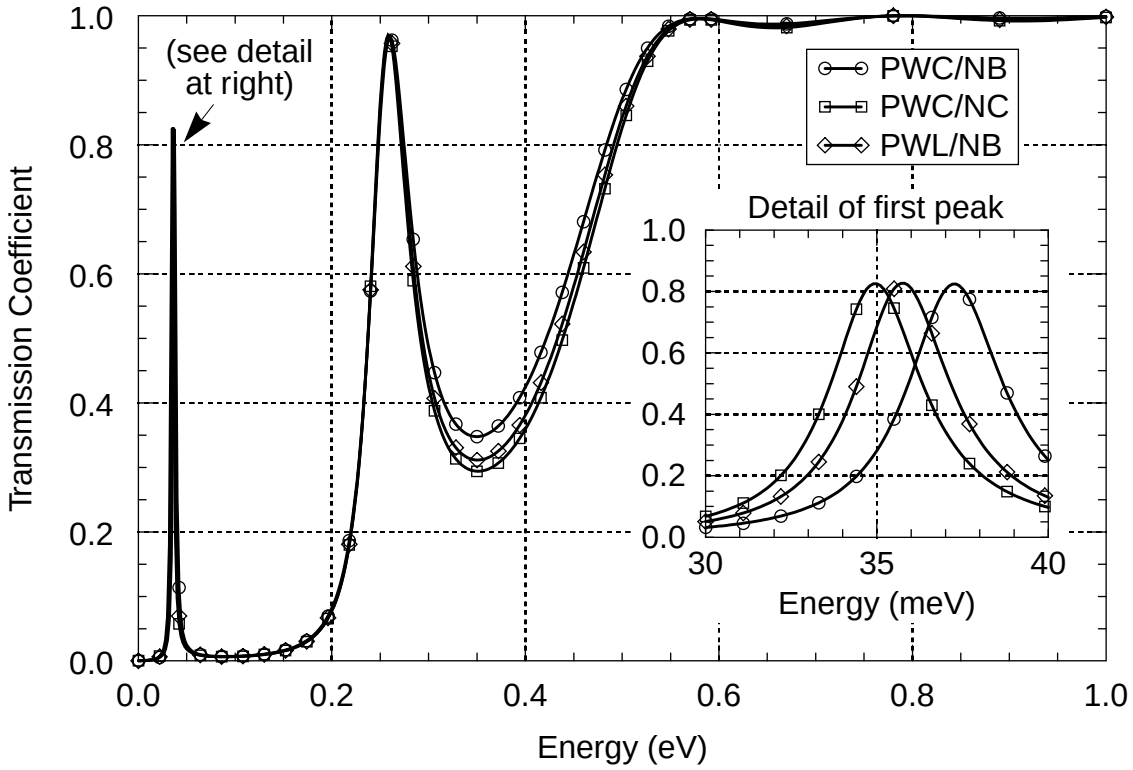


Figure 4.15: Transmission spectra for the three potential approximations

The transmission spectra (transmission coefficient versus incident energy) are given for TMM simulations of the test RTD at 0.1 V for a piece-wise constant, node-bounded (PWC/NB) potential; a piece-wise constant, node-centered (PWC/NC) potential; and a piece-wise linear, node-bounded (PWL/NB) potential. The inset details the first transmission resonance (peak). The PWC/NC spectrum is closer than the PWC/NB result to the (presumably most accurate) PWL/NB curve. However, all diferences are relatively small. For example, the peaks in the first transmission resonance are separated by only about 2 meV.

A more definitive indicator of electronic device operation is the current-voltage (I-V) curve. Therefore, Figure 4.16 shows I-V curves for the RTD in Figure 4.14 using the three potential approximation schemes. Once again, the differences between the PWC and PWL simulations are relatively small. The inevitable conclusion from these simulations is that using a PWL potential does not change the simulation results significantly for this (very

typical) device, and therefore is not significantly more accurate. As the next section shows, using a PWL potential also comes at a high computational price. SQUADS therefore uses the PWC/NC scheme as default. It is worth noting, however, that more sophisticated, gridding algorithms could use a single large linear region for extended, relatively linear potentials. In fact, the RTD in Figure 4.14 could be exactly represented by 5 linear potential regions between two constant potential regions, which would result in an extremely fast simulation. This avenue is quite worthy of further investigation.

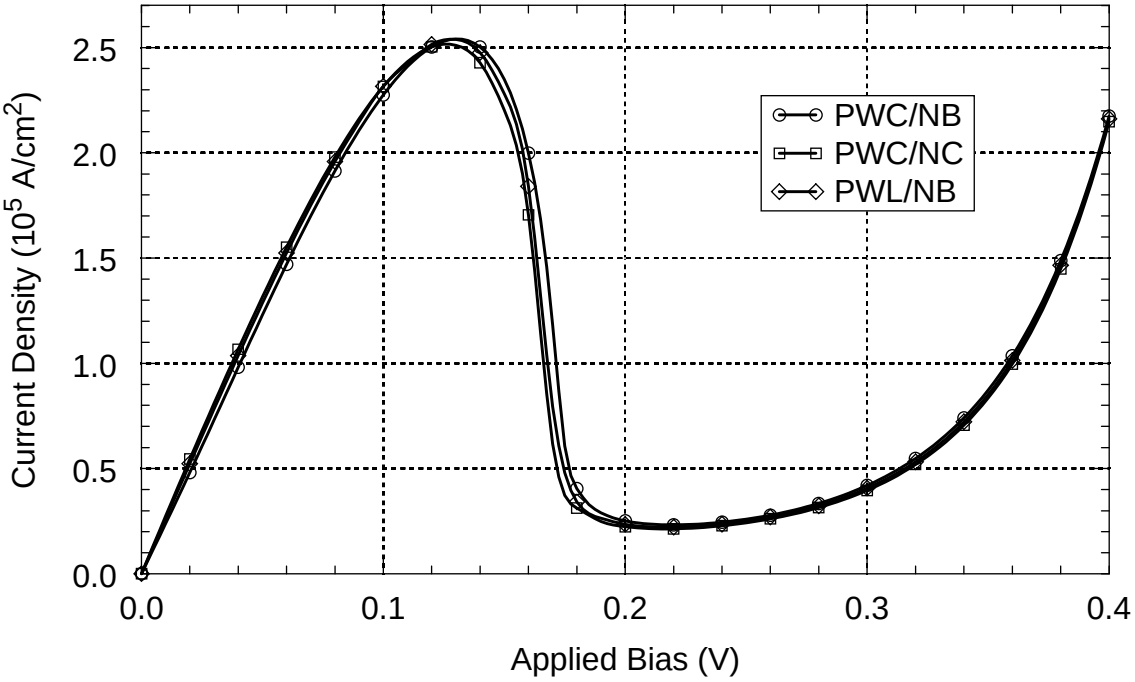


Figure 4.16: I-V curves for the three potential approximations

Current-voltage curves for the test RTD are shown for TMM simulations using a piece-wise constant, node-bounded (PWC/NB) potential; a piece-wise constant, node-centered (PWC/NC) potential; and a piece-wise linear, node-bounded (PWL/NB) potential. The differences are small, meaning that the PWL scheme did not yield significantly more accurate results in this case.

4.5.2 Efficiency of STM Calculation Algorithms

This chapter described the three system transmission matrix (STM) calculation algorithms implemented in SQUADS and used by other researchers: the interface, region, and normalization algorithms. This section compares the relative computational efficiencies of these alternative STM calculation algorithms (which give identical results for the STM), both for current density and carrier density calculations. Note that determining the carrier

density requires calculating the wavefunction at all points, and is therefore a superset of the current density calculation (which only requires the wavefunction to be known at a single point).[17] Also, note that each of the STM algorithms can be used with any of the three potential approximation schemes investigated in the previous section. All nine resulting combinations will be compared.

Table 4.1 summarizes the results of these simulations, giving the computation times (in seconds) for all nine combinations and for both the current density and carrier density calculations at a single bias point (0.1 V) for the RTD in Figure 4.14. The first conclusion from these results is that the most efficient STM calculation algorithm depends on the goal of the simulation. If only current density is needed (e.g., for the I-V curve), then the region algorithm is optimal for all three potential approximation schemes. For carrier density calculations, the normalization algorithm is optimal, except for the PWL potential case. However, the commonly used interface algorithm is not far behind the most efficient algorithm in any of the calculations, and is therefore the best single choice. As a result, SQUADS uses the interface algorithm unless instructed otherwise. Finally, note that the PWL potential requires 2.5 - 5 times as much CPU time as the equivalent PWC simulation. Therefore, as mentioned in the previous section, using a PWL potential is not worth the effort for general quantum device simulation, contrary to claims in the literature. Undoubtedly, further innovations in these (and possibly other) STM calculation algorithms could significantly modify these conclusions about the most efficient algorithm.

4.5.3 Constant versus Variable Effective Mass

Sections 4.5.1 and 4.5.2 investigated two aspects of the numerical implementation of the transfer-matrix method of quantum device simulation. This section presents a more device-oriented example of the use of TMM simulation. In particular, this section investigates the importance of using a position-dependent effective mass for accurate quantum device simulation. All previous simulations in this chapter, and many quantum device simulations in the literature, assume a constant effective mass. Of course, in real systems, effective mass varies with material. For example, the test RTD in Figure 4.14 should use an effective mass of about $0.092m_o$ in the tunnel barriers, not the $0.0667m_o$ that has been

17. SQUADS does not implement them this way, however. The current density is implemented as an integral over energy, while the carrier density is implemented as an integral over wavevector.

Table 4.1: TMM computation times for STM calculation algorithms

Computation times (in seconds) are shown for TMM simulations of the test RTD at 0.1 V for the three potential approximation schemes implemented in SQUADS [piece-wise constant, node-bounded (PWC/NB); piece-wise constant, node-centered (PWC/NC); and piece-wise linear, node-bounded (PWL/NB)], and for all three STM calculation algorithms (interface, region, and normalization). The region algorithm is the most efficient for the current density calculation, while the normalization algorithm is optimal for the carrier density calculation. However, the interface algorithm is all-around performer. Simulations using the PWL potential were 3 times slower on average than the equivalent PWC simulations, while giving little additional accuracy. CPU times are averaged over 3 runs on a DECstation 5000-200.

Calculation	Potential Approximation	STM Calculation Algorithm		
		Interface	Region	Normalized
Current Density	PWC/NB	24.9	22.3	26.5
	PWC/NC	24.4	21.6	30.2
	PWL/NB	92.0	88.1	146.4
Carrier Density	PWC/NB	74.4	87.9	67.7
	PWC/NC	76.7	90.2	76.2
	PWL/NB	198.2	257.0	289.2

assumed throughout the RTD. Using the correct barrier effective mass will make the barriers more opaque, which will decrease current. Figure 4.17 compares the TMM simulated I-V curve for a position-dependent effective mass to that simulated previously with a fixed effective mass. Clearly, the use of a position-dependent effective mass is crucial for accurately modeling quantum devices. This conclusion also has important consequences for the implementation of a Wigner function method simulator, as discussed in Chapter 5.

4.6 Summary

This chapter has described the transfer-matrix method of quantum device simulation and its implementation in SQUADS. Although the TMM is the simplest and least computationally demanding means of simulating quantum devices, this chapter shows (even without presenting most of the details) that these are relative figures of merit. Specific

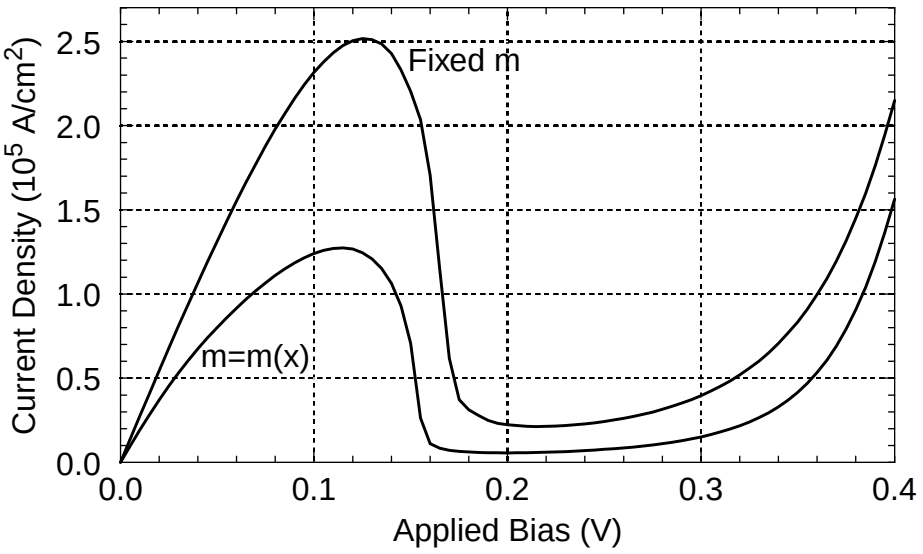


Figure 4.17: RTD I-V curves for constant and position-dependent mass

Current-voltage curves for the test RTD are shown for TMM simulations using either a fixed or a material-dependent (and thus position-dependent) effective mass. Since the I-V curves differ significantly, use of the material-dependent effective mass is important for accurate quantum device simulations.

results included conclusions about the most efficient algorithms for computing the system transmission matrix, the starting point for all TMM analysis. Another result contradicted claims in the literature: using a piece-wise-linear (instead of a piece-wise-constant) potential was found to make little difference in the simulation result, and required about a factor of 3 more computation time. Therefore, the piece-wise-linear modification was not worth the additional computation. Finally, it was shown that using a position-dependent effective mass, instead of assuming a fixed effective mass as is done quite often in the literature, can produce very inaccurate simulation results, and therefore should be avoided.

Perhaps more important than these particular conclusions is the evidence these investigations provide that SQUADS provides a foundation for the study of quantum device simulators (as well as quantum device operation) that is broad (many alternative implementations to work with), strong (well-tested), efficient, and extensible. SQUADS handling of several of the complications of TMM simulation were described, including the incorporation of contact flat-bands for efficiency, the implementation of CCN regions and quantum turning points for robustness and accuracy, and other protective measures taken against numerical overflow and round-off error.

The TMM is a very useful method of quantum device simulation, allowing the tracing

of current-voltage curves and the calculation of carrier density profiles. However, one very important variable has been conspicuously absent in this entire chapter: time. The TMM is based on the time-*independent* Schrödinger equation. As a result, the TMM can not speak to any transient effect in, or operation of, quantum devices. On this basis alone, and in spite of its capabilities and efficiency, the TMM can not serve as the (sole) basis of a general quantum device simulator. In contrast, the Wigner function method of quantum device simulation, whose implementation in SQUADS is described in the next chapter, does have the necessary capabilities. Nevertheless, the TMM in SQUADS serves several important functions in quantum device analysis, as outlined in Section 3.5.3, including:

- efficient simulation of a wide range of structures to determine which merit more detailed study (by Wigner function method simulation),
- high resolution energy spectrum investigations,
- a reality check on Wigner function method results, and
- faster implementation and testing of simulator enhancements.

References

[1] B. A. Biegel. *SQUADS Technical Reference*. (Unpublished), Stanford University, 1996.

[2] R. Tsu and L. Esaki. “Tunneling in a finite superlattice.” *Applied Physics Letters*, 22(11):562–564, 1973.

[3] M. O. Vassell, J. Lee, and H. F. Lockwood. “Multibarrier tunneling in GaAlAs/GaAs heterostructures.” *Journal of Applied Physics*, 54(9):5206–5213, 1983.

[4] C. Schwartz. “An efficient algorithm for calculating bound- and resonant-energy spectra.” In *Proceedings of the SPIE, Vol. 792*, pages 257–262, Bay Point, FL, 20-22 Mar. 1987. The International Society for Optical Engineering.

[5] H. Mizuta, T. Tanoue, and S. Takahashi. “A new triple-well resonant tunneling diode with controllable double-negative resistance.” *IEEE Transactions on Electron Devices*, 35(11):1951–1956, 1988.

[6] J. R. Söderström, E. T. Yu, M. K. Jackson, Y. Rajakarunanayake, and T. C. McGill. “Two-band modeling of narrow band gap and interband tunneling devices.” *Journal of Applied Physics*, 68(3):1372–1375, 1990.

[7] K. F. Brennan and C. J. Summers. “Theory of resonant tunneling in a variably spaced multiquantum well structure: An Airy function approach.” *Journal of Applied Physics*, 61(2):614–623, 1987.

[8] C. M. Tan, J. Xu, and S. Zukotynski. “Study of resonant tunneling structures: A hybrid incremental Airy function plane-wave approach.” *Journal of Applied Physics*, 67(6):3011–3017, 1990.

[9] H. Ohnishi, T. Inata, S. Muto, N. Yokoyama, and A. Shibatomi. “Self-consistent analysis of resonant tunneling current.” *Applied Physics Letters*, 49(19):1248–1250, 1986.

[10] M. Cahay, M. McLennan, S. Datta, and M. S. Lundstrom. “Importance of space-charge effects in resonant tunneling devices.” *Applied Physics Letters*, 50(10):612–614, 1987.

[11] B. Zimmermann, E. Marclay, M. Ilegems, and P. Gueret. “Self-consistent calculations of tunneling currents in n+ ga.” *Journal of Applied Physics*, pages 36–44, Oct. 1991.

[12] K. V. Rousseau and K. L. Wang. “Gamma- and X-state influences on resonant tunneling current in single- and double-barrier GaAs/AlAs structures.” *Applied Physics Letters*, 54(14):1341–1343, 1989.

[13] J. C. Chiang and Y.-C. Chang. “Resonant tunneling of electrons in si/ge strained-layer double-barrier tunneling structures.” *Applied Physics Letters*, 61(12):1405–1408, 1992.

[14] D. Mui, M. Patil, J. Chen, S. Agarwala, N. S. Kumar, and H. Morkoc. “Modelling of the i-v characteristic of single and double barrier tunneling diodes using a k-p band model.” *Solid State Electronics*, 32(11):1025–1031, 1989.

[15] K. Fobelets, R. Vounckx, and G. Borghs. “Matrix formalism for the triple-band effective-mass equation.” *Semiconductor Science and Technology*, 8:1815–1821, 1993.

[16] Y. Fu, Q. Chen, and M. Willander. “Resonant tunneling of holes in Si/GeSi.” *Journal of Applied Physics*, 70(12):7468–7473, 1991.

[17] R. M. Kolbas and J. N. Holonyak. “Man-made quantum wells: A new perspective on the finite square-well problem.” *American Journal of Physics*, 52(5):431–437, 1984.

[18] M. Tomizawa, H. Taniyama, and A. Yoshii. “Two-dimensional simulation for resonant tunneling transistor.” *IEEE Transactions on Electron Devices*, 41(6):883–887, 1994.

[19] L. A. Cury and N. Studart. “Resonant tunneling through Al(x)Ga(1-x)As-GaAs heterostructures.” *Superlattices and Microstructures*, 4(2):245–250, 1988.

[20] R. K. Mains and G. I. Haddad. “Time-dependent modeling of resonant-tunneling diodes from direct solution of the Schrödinger equation.” *Journal of Applied Physics*, 64(7):3564–3569, 1988.

[21] K. K. Gullapalli, A. J. Tsao, and D. P. Neikirk. “Multiple self-consistent solutions at zero bias and multiple conduction curves in quantum tunneling diodes incorporating N-N+N- spacer layers.” *Applied Physics Letters*, 62(23):2971–2973, 1993.

[22] W. A. Harrison and A. Kozlov. “Matching conditons in effective-mass theory.” In *Proceedings of the International Conference on the Physics of Semiconductors*, Beijing, China, August 10-14 1992.

[23] H. Anton. *Elementary Linear Algebra*, chapter 1.5, pages 30–31. John Wiley & Sons, New York, 4th edition, 1984.

[24] E. T. Yu and T. C. McGill. “III-V/II-VI double-barrier resonant tunneling structures.” *Applied Physics Letters*, 53(1):60–62, 1988.

[25] D. D. Coon and H. C. Liu. “Tunneling currents and two-body effects in quantum well and superlattice structures.” *Applied Physics Letters*, 47(2):172–174, 1985.

[26] K. M. S. V. Bandara and D. D. Coon. “Derivation and correction of the Tsu-Esaki tunneling current formula.” *Journal of Applied Physics*, 66(2):693–696, 1989.

[27] A. M. Kriman, N. C. Kluksdahl, and D. K. Ferry. “Scattering states and distribution functions for microstructures.” *Physical Review B*, 36(11):5953–5959, 1987.

[28] D. C. Hutchings. “Transfer matrix approach to the analysis of an arbitrary quantum well structure in an electric field.” *Applied Physics Letters*, 55(11):1082–1084, 1989.

[29] M. R. Young, N. G. Demas, C. A. Ventrice, and D. P. Kanousis. “Analytic calculation of electron transmission probability for planar-doped potential barrier devices.” *Journal of Applied Physics*, 71(1):498–502, 1992.

[30] C. S. Y. Leung, D. J. Skellern, and B. C. Sanders. “Numerical calculation of the resonant tunneling current for biased quantum well-barrier structure: A trapezoidal

model." In *International Semiconductor Device Research Symposium*, pages 681–684, 1993.

[31] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*, page 446. Dover, New York, Dec. 1972. 10th printing.

Chapter 5

The Wigner Function Method

The Wigner function method (WFM) of quantum device simulation was introduced in Section 3.4.2.4. The WFM is based on solving the Wigner function transport equation (WFTE), which describes charge carrier action in a quantum system in the same way that the Boltzmann transport equation [1] does for classical systems. In particular, the WFTE describes the evolution of the Wigner function $f_w(x, k)$, which contains both density and velocity information of carriers in a quantum system. The WFM solves for $f_w(x, k)$ at a discrete set of points in the simulated system. Given $f_w(x, k)$, it is a simple matter to calculate aggregate quantum device operation measures such as current and carrier density.

This chapter details the numerical implementation of the WFM in SQUADS. As in Chapter 4, the level of mathematical complexity is mitigated by placing more detailed derivations in the *SQUADS Technical Reference* [2]. The outline of this chapter is as follows. Section 5.1 gives a brief review of the history of WFM simulation. Section 5.2, describes in some detail the analytical equations involved in the Wigner function formulation of quantum mechanics, including the Wigner function transport equation. The discretization of the WFTE for numerical solution is detailed in Section 5.3. Section 5.4 discusses the advanced memory utilization schemes used in SQUADS to reduce the relatively large memory footprint of WFM simulations. Finally, Section 5.5 presents the results of several simulations produced with the WFM capability in SQUADS.

5.1 History and State of the Art

The Wigner function formulation of quantum mechanics was derived over 60 years ago [3], although its use in quantum system simulation spans only the last 15 years. However, because of the vast knowledge of numerical simulation for other purposes, and with the rapid advance of computational capabilities, the functionality and accuracy of WFM simulators have improved greatly over this relatively short period. This section provides a brief review of the history and state of the art of WFM simulation. Jensen and Ganguly [4] also give a brief review of WFM simulation research.

The first useful numerical implementation of the WFM was accomplished by Kluksdahl et al. [5, 6], who simulated a Gaussian wave packet in a quantum structure, including a simple scattering model. Frensley [7-9] reported the first successful simulation of RTDs, using two significant improvements to the WFM: upwind spatial differencing and boundary conditions (to simulate ohmic, dissipative boundaries) and backward Euler time differencing (to avoid divergent transient simulations). Both groups later added self-consistency to their WFM implementations [10, 11]. Jensen et al. further advanced the WFM by using a second-order time and spatial difference schemes [12, 13]. They then used these features in quantum particle-trajectory studies [14, 15], self-consistent, transient RTD simulations [16, 17], and field emission simulations [4, 18].

During just the past five years, Tsuchiya et al. [19] implemented a position-dependent effective mass capability. Gullapalli et al. corrected this model [20], and also investigated improved spatial difference schemes [21]. Miller and Neikirk demonstrated a multi-band formulation of the WFM [22]. Wu and Wu implemented the WFM including an in-plane magnetic field [23]. Zhou et al. investigated the use of quantum moment equations derived from the Wigner function formulation [24] (similar to derivations based on the Boltzmann transport equation for classical systems). Finally, Mains and Haddad [25] recently proposed a significantly different (and purportedly more accurate) numerical implementation of the WFM, although this approach has yet to be demonstrated.

The remainder of this chapter describes SQUADS' implementation of the Wigner function method of quantum device simulation. This implements virtually all of the schemes used by other researchers, but in a single numerical tool. One of the conclusions of this chapter is the determination of the relative strengths and weaknesses of various numerical implementations of the WFM.

5.2 Analytical Description

This section develops the theory and concepts necessary to describe the Wigner function method. It largely follows the approach used in developing the background for the transfer matrix method in Section 4.2. In several cases, the reader is referred to that section, rather than repeating its details here. Some material *is* repeated in order to introduce notation appropriate to the WFM.

5.2.1 The Wigner Function Transport Equation

In Chapter 3, the Wigner function was denoted $f_w(\boldsymbol{r}, \boldsymbol{k}, t)$ to differentiate it from the classical distribution function $f_c(\boldsymbol{r}, \boldsymbol{p}, t)$, since both can be written with the same independent variables. There should be no confusion in this and future chapters, so the Wigner function will hereafter be written as f instead of f_w to simplify notation. With this notation, the fairly general[1] 3-D form of the Wigner function transport equation (WFTE) in Equation (3.5) becomes:

$$\frac{\partial f}{\partial t} + \left(\frac{\hbar \boldsymbol{k}}{m^*}\right)\frac{\partial f}{\partial \boldsymbol{r}} + \frac{1}{\hbar}\iiint \frac{d\boldsymbol{k}'}{2\pi}\left(V(\boldsymbol{r}, \boldsymbol{k}-\boldsymbol{k}')\, f(\boldsymbol{r}, \boldsymbol{k}', t)\right) = \left(\frac{\partial f}{\partial t}\right)_{\text{collision}} . \tag{5.1}$$

To assess the memory requirements of solving the discrete WFTE, assume 100 points are required in each dimension (position and wavenumber) to adequately resolve physical processes.[2] With typical position grid spacing of 0.5 nm, this would allow the simulation of a 50 nm quantum region. In the numerical solution of the WFTE, there is one unknown and equation for each node point. A 3-D simulation would then have 100^6 = 1 trillion equations and unknowns. Even *storing* the Wigner function $f(\boldsymbol{r}, \boldsymbol{k})$ in this case would require 8 TB, and the equations would require at least 100 times as much storage. Thus, numerically solving the 3-D WFTE is beyond present computing technology. In 2-D, storage requirements for the numerical equations would still be in the 100 GB range. It is clear why in Chapter 3 this work was limited to quantum simulation in 1-D. Even in 1-D, the WFTE is relatively formidable to solve numerically for quantum systems of interest, as later sections of this chapter show. It is the goal of SQUADS to implement those features that are both necessary for accuracy and feasible for numerical solution on a scientific

1. This equation does make two important simplifying assumptions: that effective mass is position-independent, and that particles do not interact directly.
2. 100 points may be overkill for some dimensions, but it will be inadequate for others.

workstation. This goal demonstrates the classic accuracy-versus-efficiency trade-off. It is the purpose of this section to introduce the form of the WFTE used in SQUADS after describing the other choices made in implementing the WFM in SQUADS based on a realistic evaluation of this trade-off.

The derivation of the form of the WFTE used in SQUADS is rather involved, so the details [2] will not be repeated here. The resulting equation is:

$$\frac{\partial}{\partial t} f(x, k, t) + \frac{\hbar k}{m}\frac{\partial}{\partial x} f(x, k, t) + \frac{1}{\hbar}\int \frac{dk'}{2\pi} V(x, k - k') f(x, k', t) - \frac{\partial}{\partial t} f(x, k, t)\bigg|_c = 0, \qquad (5.2)$$

where $f(x, k, t)$ is the 1-D Wigner function (particles/cm^2) at position x, wavenumber k, and time t; $\hbar = h/2\pi$ is the reduced Planck constant. Also, $V(x, k)$, called the non-local potential, is calculated from the real potential $U(x)$ via a Fourier transform [2], which in this case simplifies to:

$$V(x, k) = 2\int_0^\infty dy \sin(ky)\left[U(x + \tfrac{1}{2}y) - U(x - \tfrac{1}{2}y)\right] . \qquad (5.3)$$

Finally, the scattering term $(\partial f/\partial t)_c$ almost universally used in WFM simulations (where scattering is included at all) is the relaxation-time approximation:

$$\frac{\partial}{\partial t} f(x, k, t)\bigg|_c = \frac{1}{\tau}\left[f(x, k, t) - \frac{f^{eq}(x, k, t)}{c^{eq}(x, t)} c(x, t)\right], \qquad (5.4)$$

where τ is the relaxation time, c is carrier density, and "eq" indicates equilibrium.

Although the derivation of (5.2) will not be detailed here, it is necessary to mention the approximations made in this derivation. These approximations are described in more detail in [2].

- The WFTE in (5.2) was derived from the effective mass form of the Schrödinger equation - the same Schrödinger equation used in the transfer-matrix method.
- A single energy band minimum and associated effective mass were assumed. However, SQUADS correctly treats multiple, non-interacting energy bands.
- Carriers were assumed to interact only as a distribution. As a result, $f(x, k, t)$ is the scaled one-particle Wigner function for a mixed state (*i.e.*, a carrier in a superposition of energy states).
- The effective mass was assumed to be position-independent.
- The scattering rate was assumed to be governed by a single relaxation time τ throughout the device. Scattering therefore is independent of the initial and final energy (*i.e.*, wavevector).

To ease the introduction to the WFM, several additional simplifications are made in the description of the WFM in this chapter. First, self-consistency is not enforced in this chapter, although it is treated in detail in Chapter 6. Also, the boundary conditions are taken to be fixed (time-independent), and given by the equilibrium Fermi-Dirac distribution. Finally, although the scattering term is included in the derivations in this chapter, simulations in this chapter assume zero scattering. Some examples of the importance of this effect are given in Chapters 6 and 8.

5.2.2 Gridding and the Potential Profile

As with the time-independent Schrödinger equation (4.1) used in the transfer matrix method, the WFTE (5.2) used in the Wigner function method also can not be solved analytically for the kinds of potential profiles $U(x)$ occurring in even the simplest quantum devices, such as the resonant tunneling diode. Therefore, a numerical solution of the WFTE must be attempted at a finite number of position points. SQUADS uses the same algorithms in the WFM as in the TMM (see Section 4.2.2) to select the position grid points x_i, calculate electrostatic potential U_i at these points, and determine the electrostatic potential boundary conditions. These algorithms are summarized below.

SQUADS uses a uniform position grid with node points $x_i = i\Delta x$, where $i \in \{0, 1, \ldots, N_x\}$, as shown in Figure 5.1. The total simulation width is $L = N_x\Delta x$. As discussed in Section 4.2.2, Δx should be equal (as near as possible) to the lattice spacing of the material. Note by comparison of Figures 4.3 and 5.1 that the most commonly-used names for the two contacts are different in the TMM and WFM, and that grid regions (in contrast to the grid nodes) no longer have a significant role in the WFM.

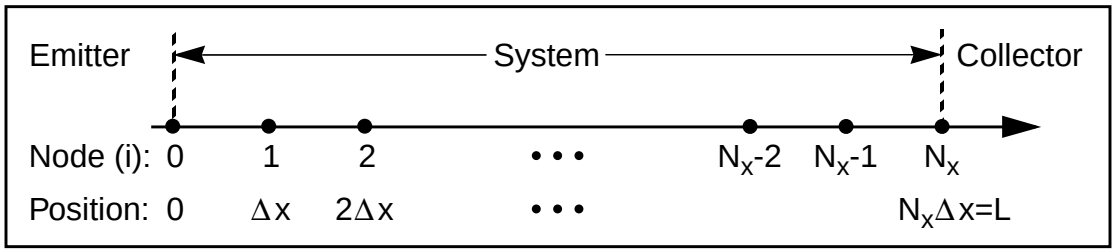


Figure 5.1: SQUADS position grid scheme

SQUADS uses a uniform position grid, $x_i = i\Delta x$, at which points device parameters (e.g., band offset, doping) are supplied and simulation results (e.g., carrier density, current density) are calculated.

The electrostatic potential U_i at grid points x_i must be supplied to the WFM simulator to perform the numerical simulation. As discussed in Section 4.2.2, SQUADS defines the Fermi energy at emitter contact as the reference, so that the potential at the collector contact is set by the applied bias V_a. Thus, contact potentials are:

$$U_0 = -E_{\mathrm{FE}}, \tag{5.5a}$$

$$U_{Nx} = -qV_a - E_{\mathrm{FC}}, \tag{5.5b}$$

where q is the electron charge, E_{FE} is the Fermi energy (relative to the energy band) at the emitter contact, and E_{FC} is that at the collector contact. The potential values at the internal grid nodes are supplied to the WFM simulator by SQUADS using a suitable algorithm. The examples in this chapter use a linear profile algorithm, leaving the discussion of self-consistency (*i.e.*, consistency between the potential profile and the carrier density profile) to be presented in detail in Chapter 6.

5.2.3 Boundary Conditions

Another issue to be settled before performing a WFM simulation is the determination of boundary conditions (BCs). The WFTE (5.2) contains a first-order spacial derivative, meaning that the solution (*i.e.*, the WF) must be specified at one position point (for all k) to make the WFTE solvable. However, there are *two* system boundary points to choose from. The issue of how to specify the BCs in a physically-based manner, while not over-constraining the system, has been discussed at length by several researchers [10, 26, 27]. As stated in Section 3.3, the ability to use conventional boundary conditions with the WFM is one of its strongest advantages. The almost universally agreed upon choice in WFM simulations is to implement ideal ohmic contacts, with the carrier reservoir on each side of the system being in equilibrium with the local potential. Only the distribution of carriers *entering* the system at each electrode is specified, while the distribution of carriers exiting the system at each electrode is determined by the simulation, and is irreversibly absorbed into the contact reservoirs. As indicated in Section 5.2.1, the BCs for the WFM will be taken as simply half of an equilibrium Fermi-Dirac distribution[3] at each electrode, integrated over transverse momenta to make them appropriate for 1-D. These BCs were first applied to the WFTE by Frensley [9]:

3. SQUADS also implements *drifted* Fermi-Dirac BCs, which automatically maintain current continuity at the contacts. No examples of these BCs are contained in this manuscript, however.

$$f(x = 0, k > 0) = \alpha \ln\{1 + \exp[-\beta(K(k) - E_{\mathrm{FE}})]\} \ , \tag{5.6a}$$

$$f(x = L, k < 0) = \alpha \ln\{1 + \exp[-\beta(K(k) - E_{\mathrm{FC}})]\} \ , \tag{5.6b}$$

where E_{Fe} and E_{Fc} are the Fermi energies (a.k.a. Fermi levels) at the contacts (see Figure 5.2), K is kinetic energy, and:

$$\alpha \equiv \frac{m^*}{\pi \tilde{h}^2 \beta} \qquad \beta \equiv \frac{1}{k_B T} \qquad K(k) = \frac{\tilde{h}^2 k^2}{2m^*}. \tag{5.7}$$

To determine the Fermi level at each contact (or at any other point, assuming equilibrium), SQUADS takes the Joyce-Dixon approximation [28] as an initial guess, and uses a Newton iteration to determine the exact value. At equilibrium, the Fermi level is correct when the carrier concentration equals the doping density.[4]

5.2.4 Carrier and Current Density

Having specified the potential profile in the device given the applied bias, and with the necessary boundary conditions, it is now possible to solve the WFTE for the Wigner function $f(x, k, t)$. The details of solving the WFTE are covered in Section 5.3. Assuming the Wigner function has been computed, this section shows how to calculate from it the other information needed about quantum device operation, namely current and carrier densities.

Section 3.3 stated that all observables (quantum-speak for physical quantities) can be calculated from the Wigner function just as they are from the classical distribution function. The carrier density and current density are two examples. Since the classical distribu-

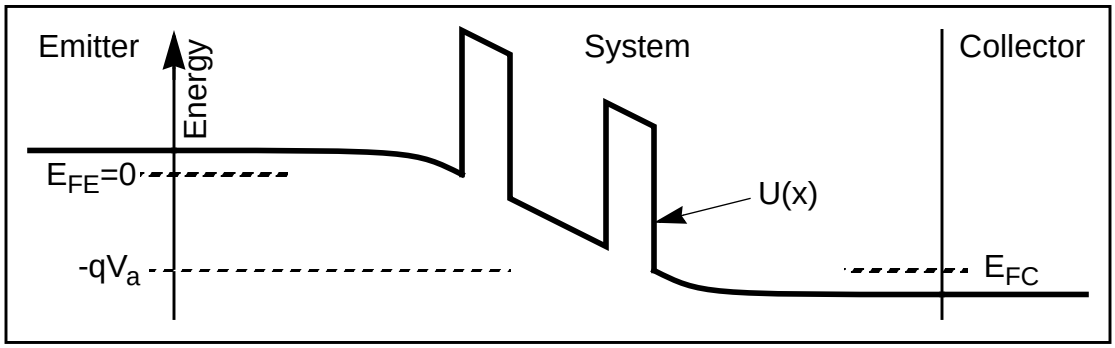


Figure 5.2: Typical potential with applied bias and boundary conditions
SQUADS uses the emitter Fermi level as the energy reference ($E_{\mathrm{FE}} = 0$), rather than the emitter electrostatic potential [$U(x = 0) = 0$].

4. SQUADS assumes full dopant ionization and no local electric fields in this calculation.

tion function $f_c(x, k)$ gives the density of carriers versus position and wavevector, integrating over wavevector gives the carrier density $c(x)$ versus position alone [29]. Section 3.1.2 of the *Squads Technical Reference* [2] gives a more rigorous derivation of calculating the carrier density from the Wigner function, with the result:

$$c(x, t) = \frac{1}{2\pi}\int dk\, f(x, k, t). \tag{5.8}$$

The analytic expression for calculating the current density $J(x, t)$ from the Wigner function is likewise given in [2], in Section 3.1.3. Two approaches are used therein, the less rigorous of which calculates current density as (charge) x (density) x (velocity), integrating over velocity (wavenumber). The more rigorous derivation requires the steady-state carrier density and current density to satisfy the continuity (or conservation of charge) equation. In both cases, the current density is found to be

$$J(x, t) = \frac{q\hbar}{2\pi m^*}\int dk\, kf(x, k, t)\,. \tag{5.9}$$

Note that in steady-state, current density is independent of position, since charge is not accumulating or depleting anywhere in the device. This fact will be used in deriving numerical expressions for current density.

This completes the analytical description of WFM simulation. In summary, a WFM simulation involves three steps: determine the potential profile at the given applied bias, solve the WFTE (5.2) for the Wigner function, and calculate desired quantities such as current density from the Wigner function. The following section discusses how these steps are implemented in SQUADS.

5.3 Numerical Implementation

As with the time-independent Schrödinger equation used in the transfer matrix method, even the simplified WFTE in (5.2) can only be solved analytically for a few very simple cases, such as a single electron in a constant potential $U(x, t) = U_o$. For any useful quantum device, the WFTE must be solved numerically, producing a numerical approximation (rather than a functional expression) for the Wigner function. Thus, the three steps described above of a WFM simulation must be converted into numerical expressions and algorithms suitable for execution on a digital computer.

At least two general approaches have been applied to solving the WFTE for useful

quantum systems: the method of moments and numerical solution. With the moment-method [24, 30], the Wigner function at each position is assumed to be a perturbed version of the equilibrium Wigner function, and the WFTE is simplified based on this assumption. This approach may be useful for multi-dimensional device simulation. For essentially 1-D quantum systems such as the RTD, the more general and accurate approach of numerical solution is feasible. By this approach, the solution of the WFTE (which is the Wigner function) is sought at a discrete set of points. The procedure for calculation of the potential at the grid nodes was discussed in Section 5.2.2. The remaining two steps are described in this section, as implemented in SQUADS. In particular, this section describes the procedure of calculating the Wigner function and other device operation information on a discrete domain.

5.3.1 Discretization of the Independent Variables

In numerical simulation of physical systems, the first step in solving the relevant equation(s) is discretization of the problem domain (*i.e.*, all independent variables). With the (inherently steady-state) transfer-matrix method, there were only two independent variables in the problem domain: position x and energy E. The state function[5] in the TMM is the quantum wavefunction, $\psi(x, E)$.[6] Recall that typically thousands of wavefunctions had to be simulated at closely-spaced energies to calculate current. With the Wigner function method of quantum device simulation, the Wigner function $f(x, k, t)$ is the single, *aggregate* state function, encapsulating position and velocity information of *all* of the carriers in the system. However, it contains three independent variables: position x, wavevector k, and time t. Each of these variables must be converted into a set of discrete points at which all functions [such as the Wigner function $f(x, k, t)$] will be computed. Frensley [9] has done an excellent job of describing how the discretization of these independent variables is chosen for the WFM. This section is an abbreviated presentation of the decisions involved and the resulting discretizations.

Sections 4.2.2 and 5.2.2 have already explained SQUADS' use of a uniform, 1-D position grid, with node points $x_i = i\Delta x$, where $i \in \{0, 1, \ldots, N_x\}$ and $L = N_x\Delta x$. In defining the wavevector grid below, another motivation for using a uniform position grid arises

5. The state function contains the state (*e.g.*, location, velocity) of charge carriers in the system.
6. The energy-dependence of ψ was usually implicit in Chapter 4.

in the WFM.

In order to specify the wavevector grid (k-grid), consider first the calculation of the non-local potential (NLP) in (5.3), which is rewritten here:

$$V(x, k) = 2\int_0^\infty dy \sin(ky) [U(x + \tfrac{1}{2}y) - U(x - \tfrac{1}{2}y)] \ . \tag{5.10}$$

When the problem domain is discretized, the Fourier transform above becomes a discrete Fourier transform (DFT). Properties of the DFT make the specification of the k-grid and y-grid strongly interdependent. First, the DFT ideally takes a discrete function defined on a uniform grid as input and produces the same as output: both the y-grid and k-grid should be uniform.[7] To minimize computation and maximize reuse of $U(x)$ values, a uniform x-grid is used, with $\Delta y = 2\Delta x$.[8] In fact, it was argued for other reasons that $U(x)$ should be defined only on a uniform x-grid, so this works out perfectly. The number of points N_y in the y-grid is as yet unspecified. Thus, the y-grid as defined so far is:

$$y_{i'} = 2i'\Delta x \qquad i' \in \{0, 1, \ldots, N_y\} \qquad N_y = ?. \tag{5.11}$$

The *range* of the k-grid is also determined by the properties of the DFT [31]:

$$(k_{\max} - k_{\min}) = 2\pi/\Delta y = \pi/\Delta x \ . \tag{5.12}$$

One is free to choose $k_{\max}$ or $k_{\min}$ as appropriate to the problem, since a DFT function is periodic with period $(k_{\max} - k_{\min})$. In this case, carriers flowing in both directions must be modeled, and recalling that wavevector is proportional to velocity, SQUADS uses a k-range centered around 0:

$$k_{\max} = -k_{\min} = \tfrac{1}{2}(k_{\max} - k_{\min}) = \pi/2\Delta x \ . \tag{5.13}$$

Note that an unsuitably large value of Δx may result in a small k-range that misses some high-energy carriers. To prevent this, N_x must be large enough to produce a small Δx and thereby a large enough $k_{\max}$ to capture all of the significant carrier action. It turns out that with Δx equal to the lattice spacing as discussed previously, $k_{\max}$ is almost certainly sufficiently large for an accurate simulation.

The number of points N_k in the k-grid has not yet been chosen. The only constraint here is that a discrete function and its DFT will have the same number of points.[9] Since N_y

---

7. Although DFTs can be computed on non-uniformly-spaced data, the simplicity and accuracy of the DFT calculation are greatly improved if the data is uniformly spaced.

8. The relationship $\Delta y = 2\Delta x$ is universally used in WFM simulation, but its necessity is not manifest. However, for implementing self-consistency via the Newton method (see Chapter 6), this relationship is even more difficult to avoid.

is as yet unspecified, N_k is selected as desired, and N_y will be determined by this choice. The result is:

$$\Delta k = \frac{(k_{\max} - k_{\min})}{N_k} = \frac{\pi}{N_k \Delta x} . \tag{5.14}$$

For the actual values k_j, SQUADS follows the analysis of Frensley [9]. He observed that solving the discretized WFTE is more complicated if $k = 0$ is taken as one of the k_j. Thus, the k-grid is designed to straddle 0, meaning that N_k is even, and:

$$k_j = \left\{ \frac{\pi}{N_k \Delta x} [j - \tfrac{1}{2}(N_k\text{-}1)] \right\} \qquad j \in \{0, 1, \ldots, N_k\text{-}1\} \qquad N_k \text{ is even} . \tag{5.15}$$

The phase-space (*i.e.*, position-velocity) grid scheme used for WFM simulations in SQUADS is summarized graphically in Figure 5.3 [9]. In addition to the position-wavevector grid, Figure 5.3 also shows the "incident-particle" boundary conditions used for the WFM, as discussed in Section 5.2.3.

The final independent variable in the WFTE is time t, which is only used in transient simulations. To perform a transient simulation, the solution at $t = 0$ is first determined by solving the WFTE in steady-state mode (*i.e.*, with the transient term $\partial f / \partial t$ set to 0). The solution is then is advanced in small time steps Δ_t, with the transient term included, until the completion criterion for the simulation is reached (either steady-state or N_t time steps). SQUADS uses a fixed Δ_t, but the determination of an appropriate value for Δ_t depends strongly on the form of the discrete transient term, which issue is discussed in Section 5.3.3.5. The discretized time-domain used in SQUADS, as far as currently known, is:

$$t_n = n\Delta t \qquad n \in \{0, 1, \ldots, N_t\} \qquad \Delta t = ?. \tag{5.16}$$

5.3.2 The Discrete WFTE Matrix Equation

This section formalizes the notation used in writing the discrete WFTE. Stating the discrete WFTE in a standard form is a significant step towards its solution, since general solution approaches can then be applied. For simplicity, the steady-state case is considered first.[10] The steady-state WFTE is written:

---

9. Actually, they have the same number of *degrees of freedom*. For example, if a function is real and its DFT is imaginary, the DFT function will have only $N/2$ points, although each point has real and imaginary parts. The DFT and IDFT are information-conserving operations.

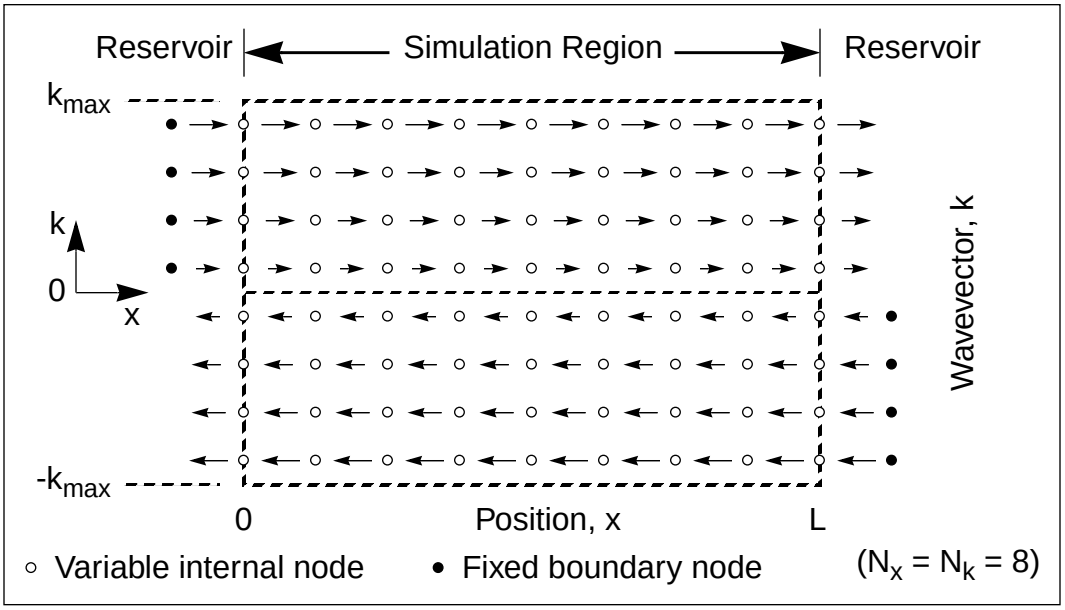


Figure 5.3: WFM phase-space grid scheme

The phase space (*i.e.*, position-wavevector) grid used to discretize and solve the WFTE is shown for the simple case of $(N_x = N_k = 8)$. There is one equation for each internal, unknown (x, k) pair (open circles). The incoming boundary conditions are shown as filled circles. The number of position points, N_x, wavevector points (N_k), and the position grid spacing (Δx) are all independently specified. The k-grid includes both positive and negative velocities, straddles zero due to numerical concerns, and $k_{\max} = \pi/2\Delta x$.

$$\frac{\hbar k}{m}\frac{\partial}{\partial x} f(x, k) + \frac{1}{\hbar}\int \frac{dk'}{2\pi} V(x, k - k') f(x, k') + \frac{1}{\tau}\left[\frac{f^{\text{eq}}(x, k)}{c^{\text{eq}}(x)} c(x) - f(x, k)\right] = 0. \qquad (5.17)$$

In the discrete domain, the steady-state Wigner function becomes:

$$(x, k) \Rightarrow f(x_i, k_j) \equiv f_{i,}. \qquad (5.18)$$

Solving the steady-state discrete WFTE means computing the value of the Wigner function (a real number) at each node point (x_i, k_j) in the domain (see Figure 5.3). Thus, there are $(N_x + 1)N_k$ unknowns to calculate. Of course, there must be one equation for each unknown in order to find a unique solution. For example, a typical quantum device simulation might have $N_x = N_k = 100$, resulting in 10,100 equations and unknowns! The process of converting (5.17) into one independent equation corresponding to each point in the domain is the essence of discretization.

10. Each time step in a transient simulation is virtually identical to a steady-state simulation.

Note that direct numerical solution of multiple equations and unknowns generally requires that the equations be linear in the unknowns.[11,12] Describing the WFM thus requires the introduction of the concepts of solving large sets of linear equations, called a linear system, as well as the associated matrix notation. This introduction is accomplished using a simpler set of N equations in the unknowns x_i, where $i \in \{1, 2, \ldots, N\}$. A complete set[13] of linear equations in these unknowns can be written

$$a_{i,1}x_1 + a_{i,2}x_2 + \ldots + a_{i,N}x_N = b_i \qquad i \in \{1, 2, \ldots, N\}, \tag{5.19}$$

or

$$\sum_{i'}^{N} a_{i,i'}x_{i'} = b_i \qquad i \in \{1, 2, \ldots, N\}, \tag{5.20}$$

where the $a_{i,i'}$ are constants. In matrix notation, (5.20) is:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,N} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,N} \\ \ldots & \ldots & \ldots & \ldots \\ a_{N,1} & a_{N,2} & \ldots & a_{n,N} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_N \end{bmatrix}, \tag{5.21}$$

or:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{5.22}$$

(5.22) formally introduces notation used throughout this thesis: bold, upper-case letters represent matrices (2-D arrays), and bold lower-case letters are vectors (1-D arrays). Given a set of equations in the form of (5.21), any of several linear system solving algorithms (such as Gaussian elimination) can be applied to the task of solving for the $\{x_i\}$. One or more of these algorithms will be used to solve the discrete WFTE.

Returning to the discretization of the WFTE, then, note that the Wigner function represents a 2-D array of unknowns: $f_{i,j}$, with $i \in \{0, 1, \ldots, N_x\}$ and $j \in \{0, 1, \ldots, N_k\text{-}1\}$. In the form of (5.20), a complete set[14] of linear equations in these unknowns is

$$\sum_{i'=0}^{N_x} \left(\sum_{j'=1}^{N_k} (a_{i,j;i',j'}) f_{i',j'} \right) = b_{i,j} \qquad \begin{matrix} i \in \{0, 1, \ldots, N_x\} \\ j \in \{1, 2, \ldots, N_k\} \end{matrix}, \tag{5.23}$$

11. Linear equations can not involve the unknowns raised to any power other than zero or unity.
12. The indirect (iterative) solution of a non-linear WFTE is treated in Chapter 6.
13. A *complete* set of equations simply means that there is one equation for each unknown.
14. In this case, a *complete* set of equations means that there must be one equation for each (i, j) pair (i.e., each phase-space node).

or in matrix notation:

$$\boldsymbol{A}\,\boldsymbol{f} \;=\; \boldsymbol{b}. \tag{5.24}$$

In (5.24), $\boldsymbol{A}$ is a square matrix of $a_{i,\,j;i',\,j'}$ coefficients with $N_{xk} \equiv (N_x + 1)N_k$ rows and columns, $\boldsymbol{f}$ is a vector of N_{xk} unknown Wigner function values denoted $f_{i,\,j}$, and $\boldsymbol{b}$ is a vector with N_{xk} elements denoted $b_{i,\,j}$ containing any constants in each equation (e.g., boundary conditions). Note from (5.23) that although there are two indices, i and j, to scan through, there is no fundamental difference between this case and the simpler one of the 1-D x_i unknowns in (5.20). One can simply think of (i, j) as a single index to step through, solve for the unknowns $f_{i,\,j}$ as 1-D vector, and consider them as a 2-D array afterwards. In Section 5.4.1, it is shown to be advantageous to order the Wigner function unknown values by successively setting x_i, scanning through the k_j, and then moving to x_{i+1}. Based on this, Figure 5.4 depicts the layout of the WFTE matrix equation (5.24) for the very simple case of $N_x = 2$, $N_k = 4$.

The formal task of discretization of (5.17) is to determine the coefficients of the square matrix $\boldsymbol{A}$ in Figure 5.4. Each term in (5.17) may contribute to each coefficient $a_{i,\,j;i',\,j'}$

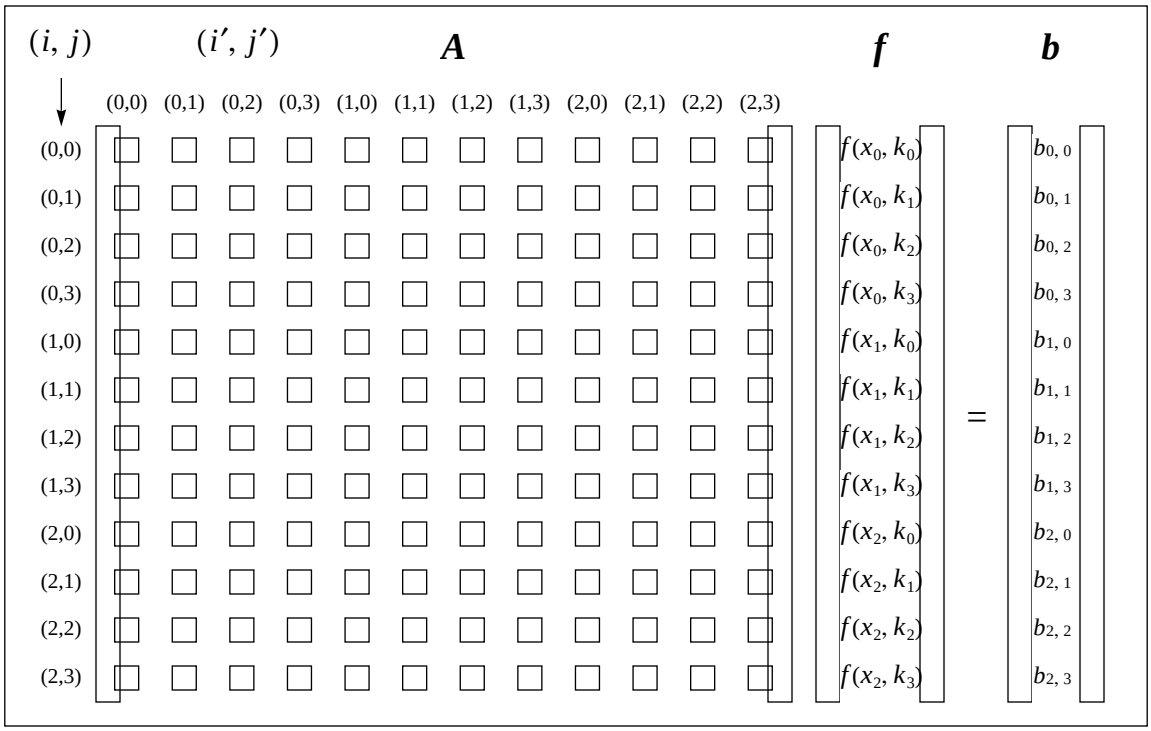


Figure 5.4: Discrete WFTE matrix equation

Although the Wigner function has two independent variables, its values can be "unfolded" into a 1-D vector $\boldsymbol{f}_{i,\,j}$ for the purposes of solving the discrete WFTE.

(i.e., in equation (i, j), and multiplying unknown $f_{i', j'}$). In addition to $\boldsymbol{A}$, the constant vector $\boldsymbol{b}$ in Figure 5.4 must be specified to solve for $\boldsymbol{f}$. Comparing (5.17) and (5.24), it appears that $\boldsymbol{b}$ will be zero, since there are only terms that multiply the unknowns. However, in general there will be constants in some of the terms on the RHS of (5.17) which must be moved to the LHS of the equation (since they don't multiply any unknown) and thus become the constant vector $\boldsymbol{b}$. By molding the WFTE into the form of Figure 5.4 and (5.23), it has become a set of N_{xk} equations which are linear in the unknowns $f(x_i, k_j) \equiv f_{i,j}$. As mentioned previously, these equations can be solved numerically using linear system solving algorithms. Having formalized this plan, the motivation for, and goal of, the discretization of the WFTE into a set of linear equations should now be clear. The actual discretization is taken up in the following section.

5.3.3 Discretization of the WFTE

This section finally tackles the discretization of the WFTE (5.17). Since the discretization of each term has its own complications and alternative discretization schemes, each term is treated in summary form in separate sections below. Most of the details of the discretization process are given in the *SQUADS Technical Reference* [2]. To further minimize the complexity of this presentation, some short-hand notation is first introduced in Section 5.3.3.1. The diffusion, drift, and scattering terms are then treated in Sections 5.3.3.2, 5.3.3.3, and 5.3.3.4 respectively. The transient term, whose discretization is presented in Section 5.3.3.5, adds perhaps the most complexity and notation. Significant observations about the WFTE discretization are given in Section 5.3.3.6.

5.3.3.1 Short-Hand Notation

To simplify the notation in the remainder of this (and later) chapters, a few additional symbols will be introduced. The following defines the transient operator $\boldsymbol{T}$, diffusion operator $\boldsymbol{K}$, drift operator $\boldsymbol{P}$ and scattering operator $\boldsymbol{C}$:

$$\boldsymbol{T}[\, f(x, k, t)] \equiv \frac{\partial}{\partial t} f(x, k, t), \tag{5.25}$$

$$\boldsymbol{K}[\, f(x, k)] \equiv \frac{\hbar k}{m^*} \frac{\partial}{\partial x} f(x, k), \tag{5.26}$$

$$\boldsymbol{P}[\, f(x, k)] \equiv \frac{1}{\hbar} \int \frac{dk'}{2\pi} V(x, k - k')\, f(x, k'), \tag{5.27}$$

$$\boldsymbol{S}[\,f(x,k)] \equiv \frac{1}{\tau}\left[\frac{f^{\mathrm{eq}}(x,k)}{c^{\mathrm{eq}}(x)}c(x) - f(x,k)\right]. \tag{5.28}$$

With these, the transient WFTE is

$$(\boldsymbol{T} + \boldsymbol{K} + \boldsymbol{P} + \boldsymbol{S})[\,f(x,k,t)] \;=\; 0\,, \tag{5.29}$$

and the steady-state WFTE (5.17) can be written

$$(\boldsymbol{K} + \boldsymbol{P} + \boldsymbol{S})[\,f(x,k)] \;=\; 0\,. \tag{5.30}$$

5.3.3.2 Diffusion Term

The discretization of the diffusion term is tackled first. From (5.26):

$$\boldsymbol{K}[\,f(x,k)] \equiv \frac{\hbar k}{m^*}\frac{\partial}{\partial x}f(x,k) \Rightarrow \boldsymbol{K}[\,f_{i,\,j}]. \tag{5.31}$$

Several discrete forms of the diffusion term are possible, depending on the difference scheme used for the spatial derivative of the Wigner function. However, in order to couple the incoming boundary conditions into the solution, but not the outgoing boundary values, an upwind difference scheme (UDS) must be used, as argued and implemented first by Frensley [9]. Upwind differencing means using a backward difference for $k < 0$ and a forward difference for $k > 0$. Frensley used a first-order UDS (UDS1), while Jensen and Buot used a second-order UDS (UDS2). SQUADS implements both UDS1 and UDS2, as well as a third-order UDS (UDS3), facilitating a direct comparison of the accuracy and computational cost of each.

The derivation of the discrete expressions for the three UDS discretizations are given in [2]. The results are:

$$\boldsymbol{K}_1[\,f_{i,\,j}] \;=\; \frac{\hbar k_j}{m}\frac{1}{\Delta x}\begin{cases} f_{i+1,\,j} - f_{i,\,j} & (j \le \frac{1}{2}N_k) \\ -f_{i-1,\,j} + f_{i,\,j} & (j > \frac{1}{2}N_k) \end{cases}, \tag{5.32}$$

$$\boldsymbol{K}_2[\,f_{i,\,j}] \;=\; \frac{\hbar k_j}{m}\frac{1}{2\Delta x}\begin{cases} -f_{i+2,\,j} + 4f_{i+1,\,j} - 3f_{i,\,j} & (j \le \frac{1}{2}N_k) \\ f_{i-2,\,j} - 4j_{i-1,\,j} + 3f_{i,\,f} & (j > \frac{1}{2}N_k) \end{cases}, \tag{5.33}$$

$$\boldsymbol{K}_3[\,f_{i,\,j}] \;=\; \frac{\hbar k_j}{m}\frac{1}{6\Delta x}\begin{cases} 2f_{i+3,\,j} - 9f_{i+2,\,j} + 18f_{i+1,\,j} - 11f_{i,\,j} & (j \le \frac{1}{2}N_k) \\ -2f_{i-3,\,j} + 9f_{i-2,\,j} - 18f_{i-1,\,j} + 11f_{i,\,j} & (j > \frac{1}{2}N_k) \end{cases}, \tag{5.34}$$

where, from (5.15):

$$k_j \;=\; \left\{\frac{\pi}{N_k\Delta x}[j - \tfrac{1}{2}(N_k + 1)]\right\}. \tag{5.35}$$

In spite of the strong case mandating an upwind difference scheme, it is possible to use a non-UDS at interior nodes as long as the boundary conditions are coupled in correctly with a UDS. At least two groups have taken this approach, because the UDS is not the most accurate approximation to the derivative at a point. One group [10, 32] used a 2nd-order central difference scheme (CDS2), but changed to UDS1 at the outgoing boundary nodes. Another group [21] used a hybrid UDS2/CDS2 scheme (denoted HDS22 herein), but used UDS2 at the outgoing boundary. To enable the investigation and comparison of these difference schemes as well UDS, SQUADS implements CDS2, CDS4, and CDS6, and also allows any hybrid combination of a UDS and a CDS:

$$\mathrm{HDS}ij \;=\; \frac{1}{\alpha+\beta}(\beta\mathrm{UDS}i + \alpha\mathrm{CDS}j) \;. \tag{5.36}$$

For details on these discretizations, see [2]. The error introduced by changing the discretization scheme near the boundaries has apparently not been investigated by other researchers yet, so this will be a subject for investigation in Section 5.5.

It is worth recalling at this point that the WFTE (5.2) was derived with the simplification of a position-independent effective mass. The WFTE for a spatially-varying effective mass has a much more complicated diffusion term [2, 19, 20]. As a result, this form of the WFTE is not currently implemented in SQUADS. Also recall that TMM simulations in Chapter 4 (see Figure 4.17) demonstrated the importance of properly including a position-dependent effective mass. As an attempt to recover some of the physics of a position-dependent effective mass, Frensley [9] used a simple model which simply moved the effective mass inside the discrete derivative. For UDS1, (5.32) becomes:

$$\boldsymbol{K}_1[f_{i,j}] \;=\; (\underset{\sim}{\hbar}k_j)\frac{1}{\Delta x}\begin{cases} \dfrac{f_{i+1,j}}{m_{i+1}} - \dfrac{f_{i,j}}{m_i} & (j \le \frac{1}{2}N_k) \\ -\dfrac{f_{i-1,j}}{m_{i-1}} + \dfrac{f_{i,j}}{m_i} & (j > \frac{1}{2}N_k) \end{cases} . \tag{5.37}$$

To test its accuracy, this simple position-dependent effective mass model was also implemented in SQUADS (for all difference schemes), and will be investigated in Section 5.5.

5.3.3.3 Drift Term

The discretization of the drift term in the WFTE is summarized in this section. From (5.27), the drift term is:

$$\boldsymbol{P}[\, f(x,k)] \equiv \frac{1}{\tilde{h}}\int \frac{dk'}{2\pi}\, V(x, k-k')\, f(x,k') \Rightarrow \boldsymbol{P}[\, f_{i,j}]\,. \tag{5.38}$$

The derivation of the discrete drift term is rather complicated, and is again detailed in [2], with the following result:

$$\mathbf{P}[\, f_{i,j}] \;=\; \frac{1}{\tilde{h}} \sum_{j'=1}^{N_k} V_{i,\,j-j'}\; f_{i,j'}, \tag{5.39}$$

where the discrete non-local potential is:

$$V_{i,j''} \equiv \frac{2}{N_k} \sum_{i'=1}^{N_k/2} \sin\!\left[\frac{2\pi}{N_k} i'j''\right] [U_{i+i'} - U_{i-i'}]\ . \tag{5.40}$$

The derivation in [2] also shows that $N_y = \frac{1}{2}N_k$, which relationship was unknown when the discretization of the independent variables was discussed in Section 5.3.1.

As with the diffusion term, alternatives to the standard drift term have been suggested. Jensen [33] proposed a Simpson integration rule (triangular smoothing) for the discrete integration, with the intent of making abrupt changes in the potential profile (*i.e.*, hetero-junctions) have a somewhat muted effect on the high-energy tails of the Wigner function. Gullapalli et al. [21] instead proposed a rectangular-smoothed integration, with the result that the real potential $U(x)$ has a decreasing effect on the NLP $V(x_0)$ as $(x - x_0)$ increases. In contrast, in the standard NLP calculation (5.41), the effect $U(x)$ remains in full force to a distance $(x - x_0) = N_k \Delta x/2$, beyond which the effect drops immediately to 0. The standard calculation assumes no scattering, and the cut-off distance is somewhat arbitrary. Frensley [27] discusses the rationales for employing alternative NLP s from a theoretical perspective. To determine concretely the effects of these approaches, all three drift term discretization schemes are implemented in SQUADS, and the effects of using each are analyzed in Section 5.5.

5.3.3.4 Scattering Term

Compared to that of the other WFTE terms, the discretization of the scattering term derived in [2] is relatively simple, mostly because the investigation of alternative implementations of scattering in the WFTE have not yet been investigated with SQUADS (or apparently by any other researchers). From Section 5.2.1, then, the analytical scattering term used in SQUADS is:

$$\boldsymbol{S}[f(x,k,t)] \equiv -\frac{1}{\tau}\left[f(x,k,t) - \frac{f^{\mathrm{eq}}(x,k,t)}{c^{\mathrm{eq}}(x,t)} c(x,t) \right]. \tag{5.41}$$

The expression for calculating carrier density from the Wigner function is given in Section 5.3.4:

$$c_i = \frac{\Delta k}{2\pi} \sum_{j'=1}^{N_k} f_{i,j'}, \tag{5.42}$$

so that (5.41) becomes

$$\boldsymbol{S}[f_{i,j}] = \frac{1}{\tau}\left[\frac{f^{\mathrm{eq}}_{i,j}}{c^{\mathrm{eq}}_i} c_i - f_{i,j} \right]. \tag{5.43}$$

5.3.3.5 Transient Term

Finally, this section presents the discretization of the transient term. This task is described in detail in [2]. Unlike the discretization of the other WFTE terms, the transient term can not be discretized in isolation. Instead, transient discretization results in a modified WFTE, although the discretization effort for the other terms can still be used. For comparison, the unmodified WFTE from (5.29) can be written:

$$0 = -\frac{\partial}{\partial t} f(t_n) + \boldsymbol{L}[f(t_n)], \tag{5.44}$$

where $f(t_n)$ is the Wigner function at time t_n, and one final operator has been defined:

$$\boldsymbol{L} \equiv \boldsymbol{K} + \boldsymbol{P} + \boldsymbol{S}. \tag{5.45}$$

Suppose the solution $f(t_n)$ at time t_n is known, and that at time $t_{n+1} = t_n + \Delta t$ is to be determined. Five reasonable forms of the transient operator are implemented in SQUADS, again to allow direct comparison for the purpose of determining the optimal approach. In the earliest implementations of the WFM [5], both first and second-order forward (or explicit) Euler transient terms were tested. These give the following WFTEs:

- First-order forward Euler (first-order Taylor series expansion):

$$\left.\frac{f}{t}\right|_{t=t_n} \approx \frac{f(t_n + \Delta t) - f(t_n)}{\Delta t} = \boldsymbol{L} f(t_n) \tag{5.46}$$

$$\Rightarrow \qquad f(t_n + \Delta t) = [1 + \Delta t \boldsymbol{L}] f(t_n). \tag{5.47}$$

- Second-order forward Euler (second-order Taylor series expansion):

$$(t_n + \Delta t) \approx f(t_n) + \Delta t \frac{\partial f}{\partial t}\bigg|_{t = t_n} + \frac{1}{2}\Delta t^2 \frac{\partial^2 f}{\partial t^2}\bigg|_{t = t} , \tag{5.48}$$

$$\Rightarrow \qquad f(t_n + \Delta t) = \left[1 + \Delta t \boldsymbol{L} + \frac{\Delta t^2}{2}\boldsymbol{L}^2\right] f(t_n). \tag{5.49}$$

Implementation of the first two terms on the RHS of (5.49) are obvious, given the derivations for $\boldsymbol{K}$, $\boldsymbol{P}$, and $\boldsymbol{S}$ above. However. the implementation of $\boldsymbol{L}^2$ is rather complicated, so its full detail [2] will not be repeated here. Note that the two forward Euler transient forms require an initial Wigner function, produced using a steady-state solution of the WFTE or a transfer matrix calculation of the Wigner function, as discussed in Section 4.3.5. However, forward Euler approaches are very computationally efficient, as they do not require the solution of a matrix equation to determine the Wigner function $f(t_n + \Delta t)$ at the next time step, but simply the multiplication of a matrix and vector. The stability (tendency to diverge) of forward Euler approaches is always a concern.

Frensley later argued [9] that an implicit (or backward) Euler transient term should be used, since only such has a bounded error which does not grow to infinity over long simulation times, in contrast to the forward Euler approaches. Frensley used a first-order backward Euler, and SQUADS also implements a second-order backward Euler, which appear very similar to the forward Euler expressions above:

- First-order backward Euler (first-order Taylor series expansion):

$$\frac{f}{t}\bigg|_{t = t_n} \approx \frac{f(t_n + \Delta t) - f(t_n)}{\Delta t} = \boldsymbol{L} f(t_n + \Delta t, \tag{5.50}$$

$$\Rightarrow \qquad [1 - \Delta t \boldsymbol{L}] f(t_n + \Delta t) = f(t_n). \tag{5.51}$$

- Second-order backward Euler (second-order Taylor series expansion):

$$(t_n) \approx f(t_n + \Delta t) - \Delta t \frac{\partial f}{\partial t}\bigg|_{t = t_n + \Delta t} + \frac{1}{2}\Delta t^2 \frac{\partial^2 f}{\partial t^2}\bigg|_{t = t_n + \Delta} , \tag{5.52}$$

$$\Rightarrow \qquad \left[1 - \Delta t \boldsymbol{L} + \frac{\Delta t^2}{2}\boldsymbol{L}^2\right] f(t_n + \Delta t) = f(t_n). \tag{5.53}$$

Note that although the forward and backward Euler equations appear quite similar, the former only requires a matrix-vector multiplication (computation proportional to $N_x N_k^2$) per time step, while the latter requires the full solution of a set of linear equations (computation proportional to $N_x N_k^3$) per time step [9]. Solving the steady-state WFTE always

requires the full solution of a linear system of equations.

Finally, SQUADS also implements the second-order Cayley (*a.k.a.* Crank-Nicholson) form of the transient term, which was first proposed by Jensen and Buot [13]. In this case, the WFTE is:

$$\left.\frac{\partial f}{\partial t}\right|_{t=t_n} \approx \frac{f(t_n+\Delta t)-f(t_n)}{\Delta t} \approx \frac{1}{2}\boldsymbol{L}_n[\,f(t_n+\Delta t)+f(t_n), \tag{5.54}$$

$$0 = \left[1-\frac{\Delta t}{2}\left(\frac{\boldsymbol{L}}{i\hbar}\right)\right][\,\boldsymbol{f}(t_{n+1})] - 4f(t_n)\ , \tag{5.55}$$

where $\boldsymbol{f}(t_{n+1}) \equiv f(t_{n+1}) + f(t_n)$. With the Cayley form of the transient term, the matrix equation is actually solved for $\boldsymbol{f}(t_{n+1})$. After doing so, the previous solution is subtracted out to get the new Wigner function, $f(t_{n+1})$.

Frensley [9] discussed in some detail the considerations involved in the proper selection of the time step Δt. Kluksdahl et al. [10] argued that stability of simulations using a forward Euler scheme requires that $\Delta t \le \Delta x / v_{\max}$, where $v_{\max}$ is the highest velocity of any carriers in the simulation, which often requires time steps smaller than 0.1 fs. In contrast, the backward Euler approach is inherently stable (error is bounded), but a relatively small time step (typically 1 fs) will keep the error small. The net result is that the time step for a forward Euler simulation must often be at least 10 times smaller than that of a backward Euler simulation. However, except in specific cases, the backward Euler simulation requires a factor of N_k (typically 50-200) more computation. Thus, the forward Euler approach may be computationally preferable in most cases, contrary to the arguments of Frensley [9]. The computational efficiency, as well as the stability and accuracy, of the five transient simulation approaches are investigated in simulations presented in Section 5.5.

5.3.3.6 The Discrete WFTE

This section combines the results of the previous sections to describe the full discrete WFTE. In the discrete domain, the WFTE will be solved numerically as a matrix equation:

$$\boldsymbol{A}\boldsymbol{f} = \boldsymbol{b}. \tag{5.56}$$

This section discusses which entries in the coefficient array $\boldsymbol{A}$ and constant vector $\boldsymbol{b}$ are non-zero. For illustration, this section uses $N_x = 5$, $N_k = 6$, and UDS2 for the diffusion term. The steady-state case is considered first, followed by a descriptions of the minor modifications to the matrix equation for solving the transient WFTE. The discrete WFTE

for steady-state is:

$$0 = \boldsymbol{L}[f_{i,j}] = (\boldsymbol{K}+\boldsymbol{P}+\boldsymbol{S})[f_{i,j}]$$

$$= \frac{\tilde{\hbar}k_j}{2m\Delta x}\begin{cases} -f_{i+2,j} + 4f_{i+1,j} - 3f_{i,j} & \left(j \le \tfrac{1}{2}N_k\right) \\ f_{i-2,j} - 4j_{i-1,j} + 3f_{i,f} & \left(j > \tfrac{1}{2}N_k\right) \end{cases}$$

$$+ \frac{1}{\tilde{\hbar}}\sum_{j'=1}^{N_k} V_{i,j-j'}\, f_{i,j'} + \frac{1}{\tau}\left[\frac{f_{i,j}^{\mathrm{eq}}}{c_i^{\mathrm{eq}}}\left(\frac{\Delta k}{2\pi}\sum_{j'=1}^{N_k} f_{i,j'}\right) - f_{i,j}\right], \tag{5.57}$$

where $0 \le i \le N_x$, $0 \le j \le N_k$-1, and the non-local potential V is given in (5.40). The non-zero coefficient structure for the discrete WFTE, to be explained below, is shown in Figure 5.5.

Understanding the matrix structure in Figure 5.5 is non-trivial. The drift term $\boldsymbol{P}$ supplies non-zero coefficients to every column where $i = i'$, resulting in solid $N_k \times N_k$ blocks of non-zero coefficients along the main diagonal. The scattering term $\boldsymbol{S}$ contributes to these same coefficients. The diffusion term $\boldsymbol{K}$ only has non-zero coefficients for $j = j'$, and only for $i' \approx i$. Thus, the diffusion term produces coefficients along the main diagonal (the dashed line in Figure 5.5) and on one or more out-lying diagonals, depending on the difference scheme used. Finally, note that for the first two and last two blocks of equations, some of these diagonals of diffusion coefficients "fall off the edge" of the coefficient matrix. This simply means that they are coefficients for values of the Wigner function outside the simulation region, which values have been specified through boundary conditions. In other words, these Wigner function values are known, so the coefficient multiplied by the WF value moves into the RHS constant vector $\boldsymbol{b}$. For the steady-state case, these few boundary condition cases produce the only non-zero elements of $\boldsymbol{b}$.

Now consider the discrete *transient* WFTE. For example, the first-order backwards Euler WFTE is:

$$[1 - \Delta t\boldsymbol{L}][f_{i,j,n+1}] = [f_{i,j,n}]\,. \tag{5.58}$$

The non-zero coefficient structure of this equation is identical to that of the steady-state WFTE. In the transient equation above, the steady-state coefficients (and the boundary conditions) are multiplied by $-\Delta t$, and 1 is added to each coefficient on the main diagonal of $\boldsymbol{A}$. Also $f_{i,j,n}$ from the previous solution is added to the corresponding element of the constant vector $\boldsymbol{b}$. The Cayley transient equation is nearly the same, but instead of solving directly for the updated Wigner function f_{n+1}, the unknown vector holds $f_{n+1} \equiv f_{n+1} + f_n$.

The second-order Euler schemes add additional non-zero terms, so that blocks with outlying half-diagonals in Figure 5.5 become nearly full $N_k \times N_k$ blocks, as described in [2].

5.3.4 Discrete Carrier and Current Densities

Once the discrete Wigner function has been calculated, other carrier-related information can be computed from it, such as the carrier density and current density. This section presents discrete expressions for these quantities. Deriving the discrete carrier density expression is straight-forward. The analytical expression for the carrier density was given in (5.8):

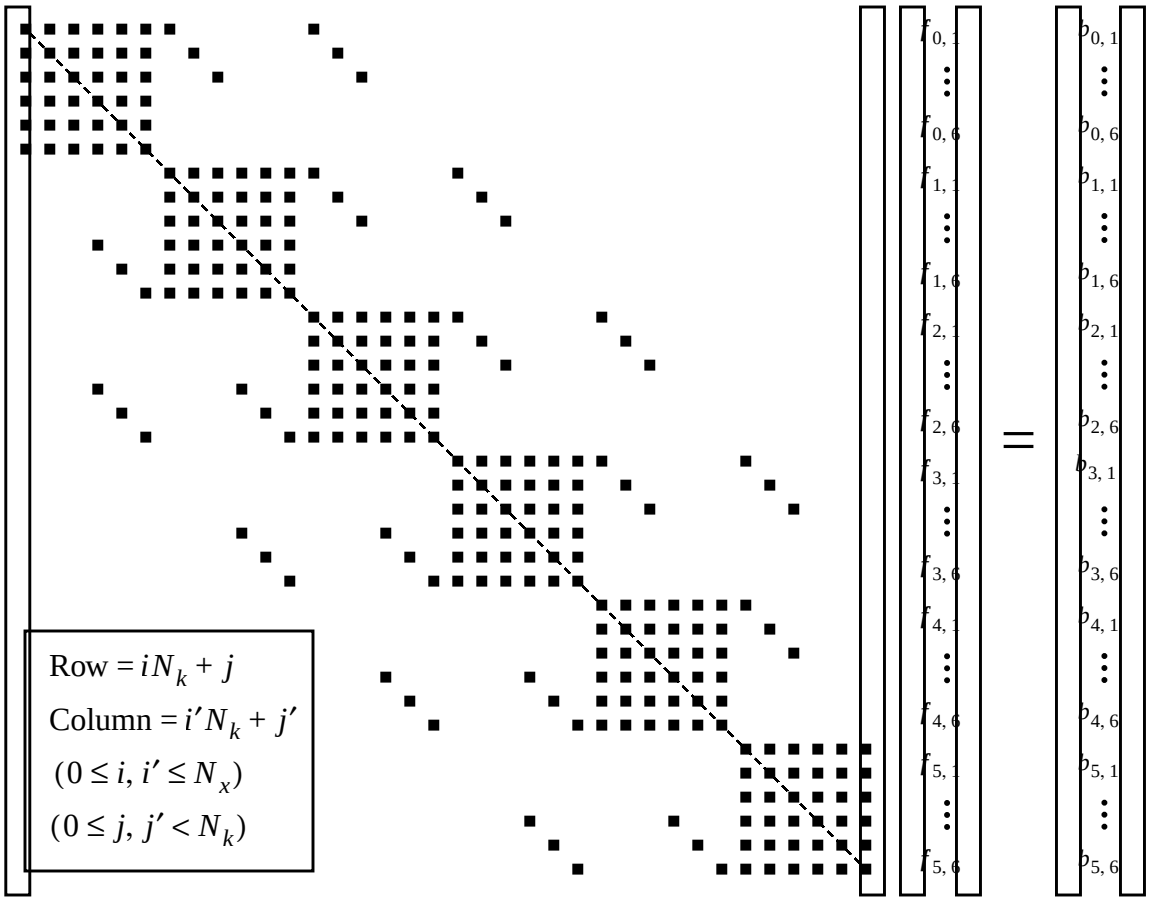


Figure 5.5: Discrete WFTE coefficient matrix structure

Like the discretizations of many differential equations, the discrete WFTE has a banded structure, which enables the employment of optimized matrix storage and solution techniques. The drift and scattering terms in the WFTE produce the coefficient blocks along the diagonal, the diffusion term produces terms along the main diagonal (the dashed line) and one or more out-lying diagonals. The transient term also adds to the main diagonal coefficients.

$$c(x, t) = \frac{1}{2\pi}\int dk\, f(x, k, t) \Rightarrow c(x_i, t_n) = c_{i, n}. \tag{5.59}$$

In the discrete domain, $dk \Rightarrow \Delta_k$, and the integral becomes a summation:

$$c_{i, n} = \frac{\Delta_k}{2\pi}\sum_{j=1}^{N_k} f_{i, j, n}. \tag{5.60}$$

The derivation of the discrete current density expression is more complicated, so the complete details are only given in [2]. Recall the analytical expression for current density in (5.9):

$$J(x, t) = \frac{q\underline{h}}{2\pi m}\int dk\, kf(x, k, t) \Rightarrow J(x_i, t_n) = J_{i, n}\,. \tag{5.61}$$

SQUADS follows the approach of Frensley [9], who noted that current density, being a vector, is most appropriately and accurately defined at the centerpoint *between* position grid nodes. Frensley also pointed out that the discrete expression for current density depends on the form of the diffusion operator $\boldsymbol{K}$. Since the diffusion operator options in SQUADS are almost innumerable, only a few of the discrete current density expressions will be given.

$$I^{\mathrm{UDS1}}_{i+1/2} = \frac{-q\underline{h}\Delta k}{2\pi m}\left(\sum_{j \le N_k/2} k_j f_{i+1, j} + \sum_{j > N_k/2} k_j f_{i, j}\right). \tag{5.62}$$

$$I^{\mathrm{UDS2}}_{i+1/2} = \frac{-q\underline{h}\Delta k}{2\pi m}\frac{1}{2}\left(\sum_{j \le N_k/2} k_j(-f_{i+2, j} + 3f_{i+1, j}) + \sum_{j > N_k/2} k_j(3f_{i, j} - j_{i-1, j})\right). \tag{5.63}$$

$$J^{\mathrm{UDS2/CDS2}}_{i+1/2} = \frac{-q\underline{h}\Delta k}{4\pi(\alpha+\beta)m}\sum_{j=1}^{N_k} k_j \begin{cases} \alpha f_{i, j} + (\alpha+3\beta) f_{i+1, j} - \beta f_{i+2, j} & (j \le \tfrac{1}{2}N_k) \\ \alpha f_{i+1, j} + (\alpha+3\beta) f_{i, j} - \beta f_{i-1, j} & (j > \tfrac{1}{2}N_k) \end{cases}. \tag{5.64}$$

Because the sign of current flow is fairly arbitrary, SQUADS assumes the sign of the current flow is the same as that of the bias applied at the collector $(x = L)$. Although this goes against convention, there should be no confusion about device operation, since SQUADS only models one carrier type (electrons or holes) in a given device.

5.4 Efficient Solution of the Discrete WFTE

Solution of the discrete WFTE in Figure 5.5 requires the use of a set of simple mathematical operations to transform the coefficient matrix into the identity matrix (all ones along the main diagonal, and zeros everywhere else). The RHS vector $\boldsymbol{b}$ undergoes the

same operations. At the conclusion, the Wigner function values can be read from RHS vector. For example, the first equation will then be:

$$(1.0) f_{0,0} = b_{0,0}. \tag{5.65}$$

This section describes the procedures used to efficiently solve the discrete WFTE.

In spite of the approximations and simplifications made in the derivation of the discrete WFTE, as discussed in Section 5.2.1, the numerical solution of this equation is still quite computationally demanding, both in terms of memory requirements and CPU usage. However, SQUADS' implementation of the WFTE solution uses techniques that allow both of these demands to be greatly mitigated, as discussed in this section. An understanding of the basic concepts and mechanics of solving systems of linear equations [34, 35] is assumed. Throughout this section, a typical "test case" simulation with $N_x = N_k = 100$ and UDS2 for the diffusion term will be used to evaluate the memory and CPU requirements of solving the discrete WFTE. For illustration purposes, the $N_x = 5$, $N_k = 6$, UDS2 example of Figure 5.4 will again be used.

5.4.1 Memory Management Schemes

A significant concern during the development of the WFM in SQUADS was the relatively high amount of computer memory required for the solution of the discrete WFTE. Consider, for example, the memory requirements of solving the test case discrete WFTE matrix equation used in this section. Since there are $(N_x + 1)N_k \equiv N_{xk}$ unknown values of the Wigner function to solve for, there are N_{xk} rows and columns of coefficients in the $\boldsymbol{A}$ matrix. Solving this system of equations accurately requires double-precision coefficients, each of which occupy 8 bytes of storage. Thus, for the test case simulation, solving the discrete WFTE appears to require $8N_{xk}^2 \approx 8(100^4) = 800$ MB of memory! Clearly, the memory requirements seem to make finding the Wigner function (and thus simulating a quantum system) infeasible by the WFM. However, this section describes the matrix storage and solution schemes used in SQUADS to reduce memory requirements to only a small fraction of 800 MB, while still retaining full accuracy in the solution.

The first job in minimizing the memory usage of a matrix equation is to determine the structure (a.k.a. sparsity) of the coefficient matrix. In other words, which coefficients of the matrix are initially non-zero, and also which will become non-zero during the solution of the system of equations? The fewer non-zero coefficients a system of equations has, the

smaller the memory requirements and solution time of the system. For the illustration example, the location of non-zero and fill-in[15] coefficients in the matrix equation is as shown in Figure 5.6. Note that most of the coefficients in $\boldsymbol{A}$ are initially 0, and remain so during solution of the matrix equation. In fact, this example gives a fill factor[16] of only about 33%. For the larger (but typical sized) test case, the fill factor is only about 2%.

Obviously, there is no point in storing coefficients that are always zero. To avoid this, sparse matrix storage schemes are devised to store as few null coefficients as possible. When there is a structure or pattern (as opposed to randomness) to the non-zero coefficient locations, this structure can usually be exploited to produce not only highly efficient stor-

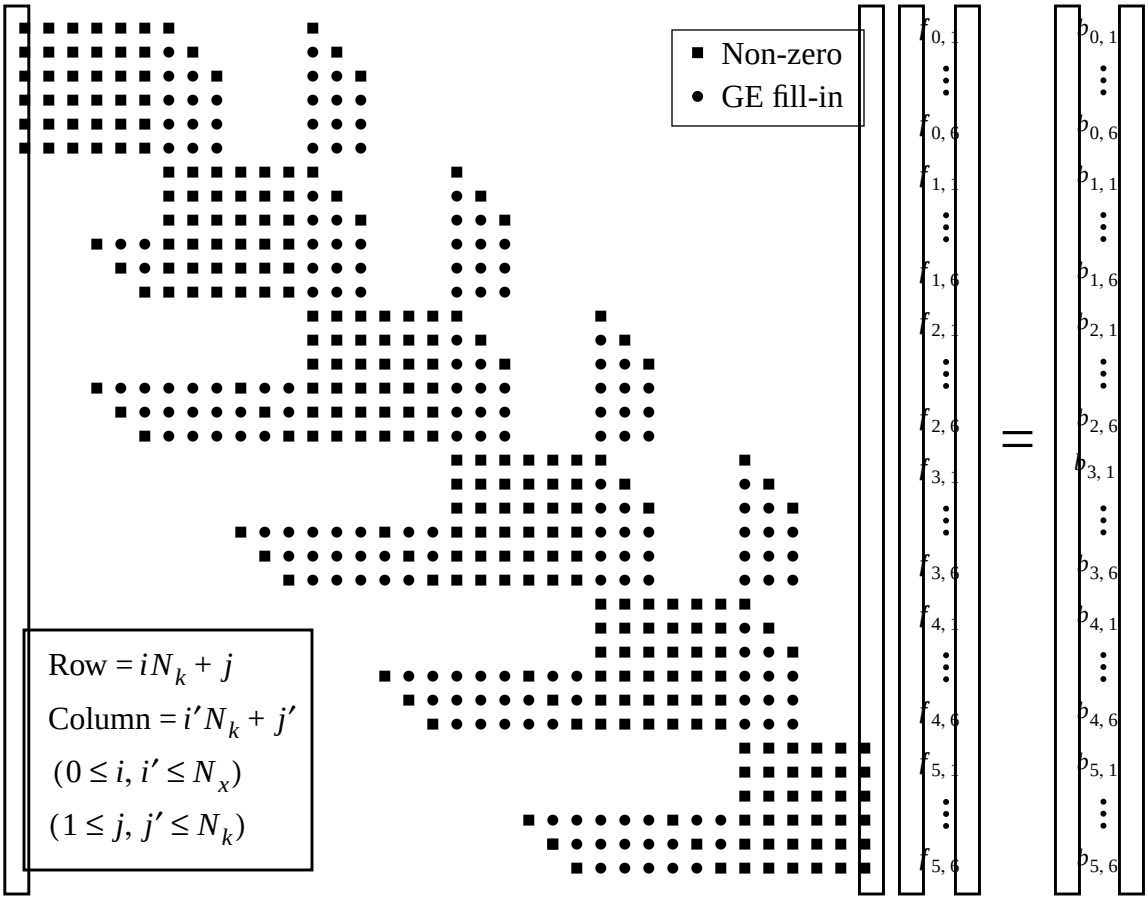


Figure 5.6: Discrete WFTE matrix equation coefficient/fill-in structure

Fill-in (filled circles) for a standard a standard Gaussian elimination solution of the discrete WFTE is shown. The bandwidth does not increase, and due to the structure of the initial non-zero coefficinets, fill-in is not too severe.

15. Coefficients that become non-zero during a Gaussian elimination solution of the system.
16. Ratio of non-zero plus fill-in coefficients to total matrix size.

age schemes, but also very efficient solution schemes. The most common scheme for improving the fill factor of "banded" matrices[17] such as that in Figure 5.5 is "diagonal" storage, where each diagonal (a.k.a. band) that contains non-zero coefficients is stored in a successive column of the storage array. The discrete WFTE coefficient matrix has a bandwidth of $4N_k + 1$, although the bandwidth depends strongly on the discrete form used for the diffusion term $\boldsymbol{K}$. Using the diagonal storage scheme, the coefficients for a typical (*i.e.*, not near the top or bottom of the coefficient matrix) block of N_k equations would be as shown in Figure 5.7. Note that the fill factor has increased to roughly 60%, and is essentially independent of N_x and N_k (but not $\boldsymbol{K}$). The full diagonal coefficient matrix for the WFTE has $4N_k + 1$ columns and N_{xk} rows. Thus, total memory usage for the test case drops to 32 MB - still a large amount, but certainly acceptable for a scientific workstation.

A side issue can now be considered and put to rest. In Section 5.3.1, it was stated without proof that it is advantageous to order the unknowns in the matrix equation by setting x_i, scanning through the k_j, and then moving to x_{i+1}, as shown by the unknown vector in Figures 5.5 and 5.6. If the opposite ordering of unknowns had been used (*i.e.*, i instead of j in the inner loop), the non-zero coefficients would be spread across the entire matrix, making the diagonal storage scheme useless. Even though the same number of coefficients are initially non-zero, more coefficients would fill in during the solution of the matrix equation, resulting in a coefficient matrix fill-factor for the test case of $5/16 \approx 31\%$, versus just over 2% previously. The CPU cost of solution is even more adversely affected.

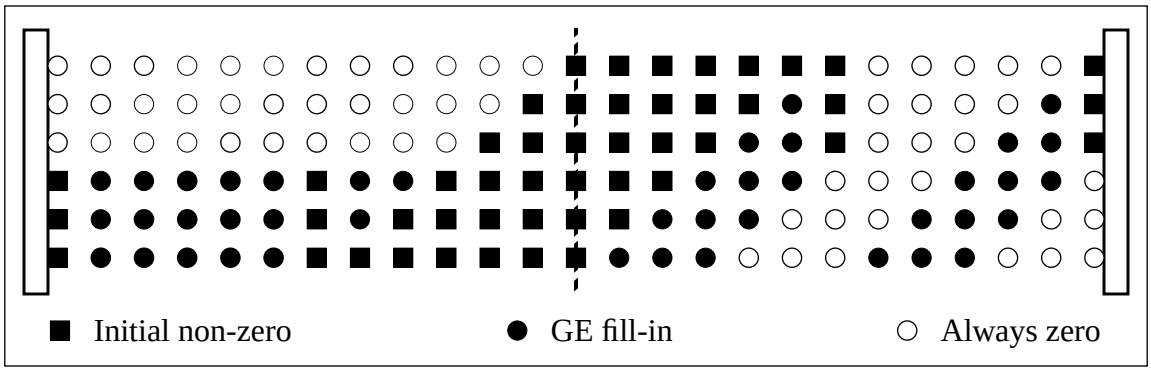


Figure 5.7: WFTE coefficient matrix structure with diagonal storage

A typical block of N_k equations is shown. There are $N_x + 1$ such blocks in the discrete WFTE matrix equation. The main diagonal is shown as a dashed line.

17. Banded simply means that there are 1 or more diagonals of non-zero coefficients, usually surrounding the main diagonal, outside of which all coefficients are zero.

These observations adequately justify the unknown ordering used in SQUADS (and by all other WFM researchers).

SQUADS uses diagonal storage for some types of WFM simulation, as will be discussed in Section 5.4.2, but most WFM simulations use a more efficient storage scheme. Note from Figure 5.5 that, because of the diagonal structure of the coefficient matrix, only a relatively small "window" of rows interact at a given time during solution by having initial non-zero coefficients in a given column. Thus, it is not necessary to calculate and store all of the coefficients before starting to solve the system. Consider Figure 5.8, which shows the top portion of the coefficient matrix from Figure 5.5 after the Gauss-Jordan elimination of all coefficients in first N_k columns. Following this elimination step, the remaining non-zero blocks of coefficients in the first N_k rows are stored for use during the back-substitution phase of matrix solution, the remaining coefficients in the WGE matrix are shifted up and left by N_k, and N_k new rows of coefficients are calculated and inserted into the newly-vacated bottom rows. The process repeats until all N_{xk} equations have been similarly eliminated, at which point, back-substitution follows on the stored coefficients. This windowed Gaussian elimination (WGE) scheme reduces memory requirements to just 8 MB for the test case (plus a small amount for the WGE matrix). This is a factor of 100 improvement over storage requirements for the full coefficient matrix!

This section has considered only the discrete WFTE matrix structure for a UDS2 diffusion term. It should be clear from Section 5.3.3 that SQUADS already implements many discretization schemes, and others could be added in the future. In general, each discretization scheme can result in a different non-zero coefficient structure, and therefore require a different storage and solution scheme. Nevertheless, SQUADS' goal is to produce nearly optimal performance and storage requirements for each and every discretization scheme it implements. In order to accomplish this, SQUADS was written in a flexible and general (as opposed to hard-wired) manner. In particular, before the matrix solution is initiated, SQUADS takes the initial non-zero structure of a block of N_k rows of coefficients and actually performs a simulation to determine how the coefficient matrix will fill-in during the WGE procedure. Based on this simulation, optimal storage and solution algorithms are chosen for that WFTE solution.

5.4.2 Computation Time and Accuracy

This section takes up the issue of how to quickly and accurately solve the discrete WFTE matrix equation. From a user standpoint, once memory usage is acceptable for the available hardware, minimizing computation time and getting an accurate result are always more important than reducing memory usage. As it happens, almost every effort described in the previous section to reduce memory usage produced a concomitant reduction in computation time. CPU time decreased because smaller amounts of memory needed to be appropriated, there was less swapping of memory to and from disk, and the smaller working blocks of data could be handled more efficiently while in memory. Perhaps the most important step in this regard was the adoption of the WGE scheme, with the result that virtually all of the calculation occurred in a relatively small data array that could often fit entirely within the (very fast) level-2 cache memory on many workstations. But in addition to memory usage reduction, there are other issues to consider and avenues to explore in the effort to reduce computation time. These are discussed below as they relate to the solution of the discrete WFTE in SQUADS.

One way to improve the speed of solving a matrix equation is to use pre-packaged

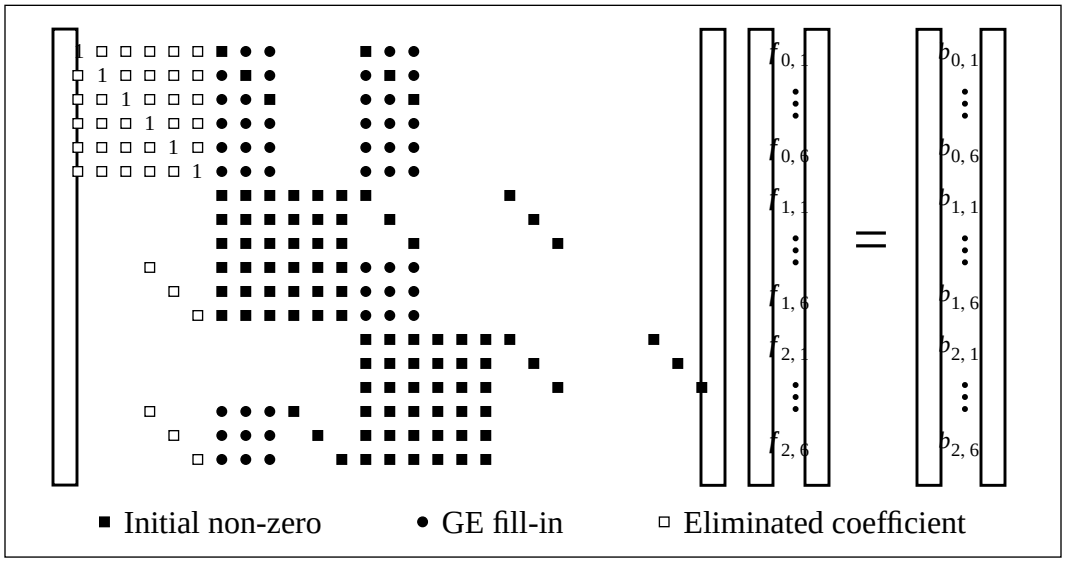


Figure 5.8: Windowed Gauss-Jordan elimination of discrete WFTE

With this procedure, only a small window of the discrete WFTE matrix equation is considered at a time. The coefficients in the first N_k columns are fully eliminated, the remaining coefficients in the first N_k *rows* are stored for the back-solve step, the remaining coefficients in the window are shifted up and to the left by N_k positions, and N_k new rows of coefficients are added to the bottom the the window.

matrix solving routines. Typically, a great deal of time, effort, and knowledge goes into the development of these routines, with the result that they achieve excellent speed and accuracy for a wide array of coefficient matrix structures. For this reason, one should usually employ the highest level pre-packaged routines that are appropriate to the task. At the highest level are complete matrix solving packages. However, there are some significant disadvantages to the use of such routines: enhancing/modifying them is often time-consuming or illegal; they usually can not achieve the performance of specialized code which takes advantage of the structure of the coefficient matrix; and they require standard data storage schemes, such as diagonal or full-block storage. The result for our case is that these routines would require 4-5 times the storage, and probably twice the CPU time, as SQUADS' optimized scheme. In order to retain the WGE optimized scheme, the highest level pre-packaged routines which are appropriate are the BLAS (basic linear algebra subprograms) routines. Typically, these routines ship with, and optimized for, each workstation. During compilation, SQUADS incorporates BLAS if it is available, and uses equivalent replacement routines otherwise. SQUADS simulations which use BLAS are typically 20% faster than those which use the replacement functions. In fact, when SQUADS was converted from in-line code to the BLAS replacement functions, a similar speed gain was realized, mainly because more aggressive optimization is possible with smaller, self-contained code blocks.

Although the scheme for solving the discrete WFTE described in Section 5.4.1 resulted in a much smaller and faster solution of this matrix equation, some disadvantages of this scheme and the resulting specialized solution algorithm should be mentioned. First, more effort may be required in maintaining and upgrading the code as compared to a more general matrix equation solver. Also, the optimized storage and solution algorithms must be modified if the sparsity structure changes. Actually, this task is well automated by the fill-in simulation discussed in Section 5.4.1. As a result, all that must be done is to supply the correct non-zero coefficient structure to the WGE routine, and optimal storage and solution algorithms will be used.

One important technique which is invariably used to help assure an accurate solution of a matrix equation is pivoting (exchanging two rows) [34, 35]. This technique reduces numerical error by re-arranging equations at each GE step such that multipliers are always smaller than unity. This requires that the coefficient on the diagonal at each GE step is

larger than any below it in that column.[18] For example, solving the two equation system

$$\begin{aligned} 001x_1 + 1000x_2 &= 2000.001 \\ 000x_1 + 0.001x_2 &= 1000.002 \end{aligned} \quad \Rightarrow \quad \begin{bmatrix} 0.001 & 1000 \\ 1000 & 0.001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2000.001 \\ 1000.002 \end{bmatrix} \tag{5.66}$$

is more likely to produce an accurate result if the two equations are swapped before performing GE. In partial differential equations (PDEs) that represent physical systems, the initial coefficient matrix for the discrete form is usually "well-behaved"[19], and the discrete WFTE is no exception. However, the GE process removes any such assurances, and could even result in a 0 on the diagonal, at which point the matrix equation is unsolvable without pivoting.

The problem with pivoting is that, in general, it expands the bandwidth of the coefficient matrix, resulting in higher storage requirements and CPU solution time. Consider the discrete WFTE coefficient structure in Figure 5.5. If diagonal storage is used for this array, pivoting would, in general, expand the bandwidth of the matrix. However, if pivoting is restricted to the remaining rows in each block of N_k, the bandwidth does not expand very much. In fact, with block storage and the WGE algorithm developed in the previous section, pivoting only within the first N_k rows of the WGE matrix does not increase storage or computation at all (see Figure 5.8). Test simulations indicated that pivoting by this scheme was only undertaken about 5% of the time, which indicates that the discrete WFTE is quite well-behaved. However, the probabilistic inevitability of near-zero diagonal terms makes the use of pivoting essentially mandatory to assure numerical accuracy. SQUADS therefore implements the limited pivoting scheme described above. For particularly ill-behaved simulations, SQUADS allows the specification of more aggressive pivoting, although the computation time increases dramatically.

Another standard matrix equation solution technique which has not yet been mentioned is lower-upper decomposition (LUD) [34], which is an alternative to Gaussian elimination. This approach is useful when a fixed coefficient matrix $\boldsymbol{A}$ can be manipulated once to then quickly solve a series of systems $\boldsymbol{Af} = \boldsymbol{b}$ with different RHS vectors $\boldsymbol{b}$. The question is, are there any situations where $\boldsymbol{A}$ is fixed but $\boldsymbol{b}$ changes from one WFTE solution to the next? It turns out that there are, but only under rather restrictive circumstances.

---

18. This is actually partial pivoting. Complete pivoting swaps rows and columns to put the largest remaining coefficient below and to the right of the current pivot coefficient into the pivot position.
19. That is, it has relatively large coefficients in the pivot position on the main diagonal.

Some restrictions are easily met: that N_x, N_k, and the discretization schemes of all WFTE terms must not change between successive solutions. This would normally be the case anyway, since changes in any of these would require a lot of overhead effort. A more significant restriction is that the relaxation time τ must be fixed (if scattering is included). This is assumed in SQUADS anyway. The one "problem" restriction is that the energy bands (*i.e.*, the potential profile) must remain fixed between successive solutions. This last restriction is met by only transient simulations where self-consistency (see Chapter 6) is not enforced, and where the applied bias remains fixed for several time steps. The two types of WFM simulations which fall in this category are Gaussian wave packet simulations (which currently assume fixed energy bands) and switching simulations (where the bias is switched at $t = 0$ and the system is allowed to evolve with this fixed applied bias). For both cases, the transient term of the WFTE is the source of the varying RHS vector, since $\boldsymbol{b}$ holds the previous WFTE solution (see Section 5.3.3.5).

The test case simulation will be used to compare the CPU and storage requirements of the LUD and WGE solution approaches. For a UDS2 simulation, $5\frac{1}{2}N_xN_k^2$ FLOPS (floating-point operations) are required for each LUD solve (after the first), compared to $1\frac{1}{2}N_xN_k^3$ FLOPS for the WGE solution (and first LUD solve). For the test case simulation, the WGE approach is therefore about 27 times slower than LUD! The main disadvantage of the LUD approach is that it requires storage of all non-zero and fill-in coefficients in $\boldsymbol{A}$. For LUD simulations, SQUADS uses a block-diagonal storage scheme, requiring 32 MB for the test case simulation, a factor of 4 larger than that used by the WGE approach. The speed advantage of the LUD approach is sufficient justification for its extra storage requirements, but only for simulations which meet its restrictions. Thus, SQUADS implements both the LUD and WGE solution schemes, and automatically uses the appropriate one for the simulation type. One final note is that, because the coefficient matrix for transient simulations is inherently diagonally-dominant,[20] the overhead of checking for pivoting makes little sense, and is therefore not implemented when the LUD approach is used. This allows higher order BLAS routines to be used in the LUD approach than can be used with the WGE approach, improving the relative computational efficiency of the LUD even further.

---

20. That is, the diagonal term is much larger than the others on each line, in the case of the transient WFTE, because all other terms are multiplied by Δ_t.

5.4.3 Other Solution Schemes

Two other researchers have developed alternate algorithms for minimizing the storage required to solve the discrete WFTE, and therefore might serve as alternatives to the optimized scheme described in the previous sections. Jensen and Ganguly's approach [4] requires the same amount of memory as SQUADS' optimized scheme. However, it is somewhat more complicated to describe and implement, and it would not mesh well with the LUD solution (i.e., it has less code overlap). However, it does allow for the use of higher level matrix solution codes, which usually implies easier maintenance and upgradeability. Invariably, "higher level" also implies not taking full advantage of the matrix structure, resulting in higher CPU time to solution. Therefore, Jensen's approach was not implemented in SQUADS.

Another approach to reducing memory usage was used by Jansen et al. [36]. This method uses an iterative conjugate-gradient (CG) algorithm to find the steady-state solution of the discrete WFTE. Although this approach requires the same order of computation time as the non-LUD approach in SQUADS, it only requires $48N_xN_k$ bytes of storage for DSO = 1, compared to $4N_xN_k^2$ for the algorithm used in SQUADS. Since the CG algorithm can not be used to directly find the transient response, it has not been implemented in SQUADS. However, SQUADS has already begun to use specialized code for various simulation modes (*e.g.*, the LUD approach for non-self-consistent, transient simulations), and the CG approach would be an excellent candidate for further specialization of SQUADS for steady-state (including self-consistent) simulations.

5.5 Simulation Results

This section demonstrates many of the basic capabilities of the Wigner function method of quantum device simulation in SQUADS, as described in preceding sections of this chapter. This demonstration is accomplished through the investigation of several key implementation details of the WFM. First, Section 5.5.1 describes the simulation of the evolution of a Gaussian wave packet (like a charge packet) in bulk semiconductor. This scenario also has an analytic solution, enabling the comparison in Sections 5.5.2 and 5.5.3 of the merits of alternative diffusion and transient term discretization schemes described in this chapter. The optimal discretization approach is then applied in the remaining simulations of this chapter. Section 5.5.4 then describes the quantum device (a resonant tunneling

diode) and simulation parameters used in further WFM simulation investigations. Several steady-state (Section 5.5.5) and transient (Section 5.5.6) WFM simulations for this RTD produce other conclusions about accurate WFM simulation. Finally, Section 5.5.7 compares WFM and TMM simulations of this RTD with experimental measurements, and attempts to explain the absence in the literature of direct comparisons between quantum device simulations and experimental measurements.

5.5.1 Gaussian Wave Packet Simulations

A seemingly infinite selection of discretization options for the WFTE have been implemented in SQUADS and discussed in previous sections of this chapter. This section investigates, as efficiently as possible, which discretization approach is "optimal". In other words, which discretization approach offers the best combination of accuracy, efficiency, and robustness?[21] Published results of such comparative investigations are rare, in contrast to the claims [9, 12, 19-21, 25-27] of the relative superiority of one discretization approach over another. Of the three measures of merit, efficiency is relatively less important, since an incorrect simulation result (due to poor accuracy or robustness) is useless, no matter how quickly it was computed. Further, the capabilities of computational hardware continue to increase rapidly, and computation tasks that are unacceptably expensive today will likely become feasible in the near future.

The best way to judge relative accuracy of numerical calculations is to simulate a system that also has an analytic solution. Any difference between the numeric and analytic results is due to numerical error in the simulation. One of the few non-trivial cases where an analytic solution to the WFTE exists is the propagation of a Gaussian wave packet (GWP) in bulk semiconductor[22] [10, 32, 37]. Although this scenario seems unrelated to Wigner function simulation of quantum devices such as the resonant tunneling diode, the only significant differences in terms of the computation are the initial condition and the boundary conditions. The computation of the Wigner function proceeds identically in either case.

Note that with flat potential GWP simulations, no comparison of various potential term discretizations can be accomplished. However, there are few position-dependent

21. "Accuracy" indicates nearness to the correct result, while robustness indicates the reliability that a vastly inaccurate and physically incorrect result will not be produced.

22. A flat potential $U(x) = 0$ is assumed throughout the simulation region.

potentials for which the WFTE is analytically solvable. Further, after scattering elastically (as opposed to dissipative scattering) off a position-dependent potential, the resulting GWP is invariably too "noisy" (lots of fine structure) to assign any significance to numeric/analytic discrepancies. Also, since scattering is not included in the analytic calculation, it must be turned off in the numerical simulation. GWP simulations *are* able to compare the theoretical accuracies of the diffusion and transient terms, however. These terms have many alternatives, and are otherwise very difficult to accurately compare.

From [38] the wavefunction for a GWP is:

$$\psi(x, t) = \left[\frac{1}{2\pi a^2(1 + i\beta t)^2}\right]^{1/4} e^{i(k_0 x - \omega_0 t)} \exp\left[-\frac{(x - v_0 t)^2}{4a^2(1 + i\beta t)}\right] \quad , \tag{5.67}$$

where v_0 is the average velocity, a is the minimum position spread, and

$$\beta \equiv \frac{\hbar}{2ma^2}, \; v_0 \equiv \frac{\hbar k_0}{m} = \frac{2\omega_0}{k_0} \; . \tag{5.68}$$

For generality, SQUADS also allows the center of the GWP to be at a location other than $x = 0$ at $t = 0$ by replacing x with $x - x_0$ in (5.67).

For a GWP, the wavefunction already represents a mixed state (*i.e.*, it has components at many energies), so the wavefunction and density matrix are identical. The Wigner function for a GWP is calculated from the density matrix via a Fourier transform, as described in [2], with the final result:

$$\begin{aligned} f_W(x, k, t) &= 2\exp\left[-\frac{(x - x_0 - v_0 t)^2}{2a^2(1 + \beta^2 t^2)}\right] \\ &= \times \exp\left\{-2a^2(1 + \beta^2 t^2)\left[(k - k_0) - \frac{\beta t(x - x_0 - v_0 t)}{2a^2(1 + \beta^2 t^2)}\right]^2\right\} \quad , \end{aligned} \tag{5.69}$$

where x_0 is the position-center of the GWP at $t = 0$. Note that the initial condition ($t = 0$) for a GWP simulation is:

$$f_W(x, k, 0) = 2\exp\left[-\frac{(x - x_0)^2}{2a^2}\right]\exp[-2a^2(k - k_0)^2] \quad . \tag{5.70}$$

Starting from this initial condition, the GWP will spread and decrease in amplitude with time.

The GWP simulations in the Sections 5.5.2 and 5.5.3 use $a = 5\Delta x$ in (5.69), where Δx is the position grid spacing. GaAs is the assumed material, so it is appropriate to take

$\Delta x = 0.565$ nm (to make the grid spacing equal to the physical atomic spacing, for reasons discussed in Section 4.2.2). The simulations also use $Nx = Nk = 100$. The basic approach in this investigation is to compare simulated and analytic results for various WFTE discretization schemes after 20 fs of GWP evolution. As an example, Figure 5.9 shows the GWP at $t = 0$ and $t = 20$ fs (twenty 1 fs steps) from a typical simulation.

5.5.2 Diffusion Term Discretization Comparison

The diffusion term discretization schemes implemented in SQUADS were described in Section 5.3.3.2. To compare the accuracy and computational efficiency of these discretization approaches, Cayley discretization will be used for the transient term, with a 1 fs time step for all simulations. Table 5.1 summarizes the results of 45 simulations comparing every diffusion term discretization implemented in SQUADS[23] for three cases:

1) GWP centered at $x_0 = L/2$ with zero group velocity. This simulation

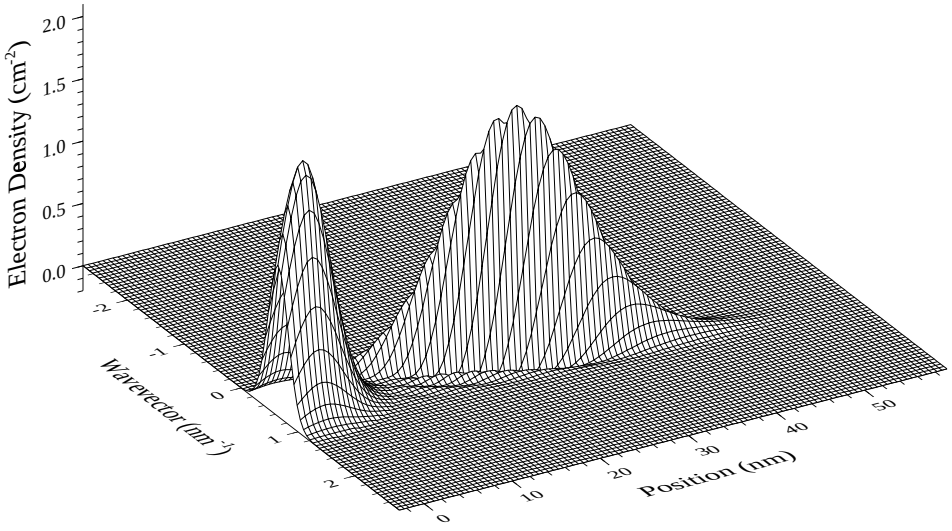


Figure 5.9: Gaussian wave packet simulation typical result

The initial ($t = 0$) and final ($t = 20$ fs) GWPs are combined in this plot. The initial GWP was centered at $x_0 = 0$ and traveling in the $+x$ direction with average wavevector $k_0 = k_{max}/2$. After 20 fs the GWP is approaching the opposite end of the simulation region, and the faster and slower components (*i.e.*, higher and lower wavevector, respectively) have spread the GWP out in the x-dimension. The ripples and negative regions in the final GWP are the result of numerical error.

23. Fifteen diffusion term discretization schemes are implemented in SQUADS. In these simulations, the hybrid difference schemes are formed as HDS = UDS + 2 CDS ($\alpha = 2, \beta = 1$ in (5.36)).

Table 5.1: WFTE diffusion term discretization scheme accuracy, efficiency
The maximum percent error is shown (determined from the analytic result) after a 20 fs WFM simulation of a Gaussian wave packet using various discretization schemes for the diffusion term of the WFTE. For these discretization schemes, UDS = upwind difference scheme, CDS = central difference scheme, and HDS = hybrid difference scheme (combination of a UDS and a CDS). Trailing numbers indicate the order of the difference scheme. The simulations are 1) a stationary GWP in the center of the simulation region, 2) a moving GWP within the simulation domain, and 3) a moving GWP interacting with both boundaries. Motion of the GWP introduces significant error into the simulation, while interaction with the boundaries does not. HDS22 is the optimal diffusion term difference scheme in terms of accuracy and computational efficiency.

Diffusion Term Discretization	$x_0 = L/2$ $k_0 = 0$	$x_0 = L/4$ $k_0 = k_{max}/4$	$x_0 = 0$ $k_0 = k_{max}/2$	Relative Compute Time	Memory Usage (MB)
UDS1	10.7%	40.0%	47.3%	1.0	25
UDS2	6.18%	23.2%	29.3%	2.4	41
UDS3	4.34%	19.1%	37.9%	4.7	57
CDS2	2.90%	16.7%	38.4%	1.5	33
CDS4	3.81%	15.7%	36.5%	3.4	49
CDS6	3.83%	15.6%	36.5%	6.1	65
HDS12	4.58%	21.8%	31.7%	1.0	25
HDS14	5.02%	22.0%	30.8%	2.4	41
HDS16	5.03%	22.0%	30.8%	4.7	57
HDS22	3.86%	15.6%	35.9%	2.4	41
HDS24	4.58%	17.1%	34.1%	2.4	41
HDS26	4.61%	17.1%	34.1%	4.7	57
HDS32	3.35%	17.0%	38.4%	4.7	57
HDS34	3.98%	16.7%	36.9%	4.7	57
HDS36	4.00%	16.7%	36.9%	4.7	57

allows a determination of the optimal diffusion term discretization absent the complications of interaction with the boundary conditions.

2) GWP initially centered at $x_0 = L/4$, with a group velocity equivalent

to $k_0 = 0.25k_{\max}$. This allows a determination of the effect of high-speed carrier motion on WFM simulation efficiency.

3) GWP initially centered at $x_0 = 0$, with a group velocity equivalent to $k_0 = 0.5k_{\max}$. This shows the significance of the interaction with both the incoming and outgoing boundaries.

From the data in Table 5.1, these GWP simulations indicate that on average, HDS22 is the optimal diffusion term discretization approach (low error and computational cost), followed by CDS2 and HDS12. The simulations also show that error improves with higher order UDS, but not with higher order CDS. Also note that error increases significantly when the average velocity of the GWP is non-zero. In UDS1, this error manifests itself as excessive spreading and amplitude decay of the GWP, while in other discretization approaches, a large part of the error is due to oscillations (including regions of negative values) that form in the GWP. The relatively large error of UDS1 makes it unsuitable for accurate quantum device simulation, making a second-order scheme a minimum.

Recall from Section 5.3.3.2 that both the CDS and HDS discretizations require changing the difference scheme at outgoing boundaries. Frensely [27] argued that this could introduce significant additional error in WFM simulations. However, these simulations indicate that error does not significantly increase due to changing the discretization approach at the boundary. Finally, note that error for each of these simulations is relatively high. Clearly, more effort should be applied in the future to understanding the sources of this numerical error in the standard WFM implementation (used in SQUADS and all other WFM code to date).

5.5.3 Transient Approach Comparison

In addition to the many diffusion term implementations, SQUADS also implements several approaches for accomplishing transient simulations. These transient term discretizations include first-order and second-order forward Euler (FE1 and FE2), first-order and second-order backward Euler (BE1 and BE2), and Cayley (a.k.a. Crank-Nicholson). This section uses Gaussian wave packet simulations similar to those in the previous section, but this time the goal will be to determine the relative accuracies of the transient discretization approaches. To this end, HDS22 discretization will be used for the diffusion term in all simulations, and the transient term discretization will be varied. Note from Section 5.3.3.5

that FE2 and BE2 include a second-order position derivative. For the associated difference scheme, SQUADS implements UDS1, UDS2, CDS2, CDS4, and the corresponding HDS schemes [2]. A series of simulations similar to those in the previous section indicated that UDS1 was the optimal difference scheme (low computational cost and error) for the second derivative term. Therefore, all FE2 and BE2 simulations in this section use UDS1 for the second derivative.

The following GWP simulations were used in this investigation of the accuracy of the transient term discretization schemes:

1) GWP in the center of the simulation region with zero average velocity ($x_0 = L/2, k_0 = 0$),
2) GWP entering the simulation region with a moderate average velocity ($x_0 = 0, k_0 = k_{\max}/4$), and
3) Repeats of the above simulations with the time increment reduced by a factor of ten ($\Delta t = 0.1\,\text{fs}$) and the number of time steps increased by a factor of ten ($NT = 200$).

The numerical results (error compared to the exact analytical result) for each of the five transient discretizations are shown in Table 5.2. The conclusion from these results is very clear: the Cayley discretization, first used by Jensen and Buot [12], is optimal over a wide range of simulation conditions. It strength is due to its unitary nature [26], which attribute tends to maintain the total number of carriers in the GWP better than the other transient approaches. Frensley [27] suspected that the Cayley discretization was more accurate than the more widely-used first-order backward Euler, but these simulations provide the first hard proof. They also show the superiority of Cayley over FE1 and FE2, which were used in some early WFM simulations [10, 32] due to a relatively low computational cost.

Note that the use of a smaller time step dramatically reduced the error of all simulations. These results suggest that 0.1 fs should be considered a reasonable time step in WFM simulations. Due to the high computational cost and the lack of concrete evidence prior to these simulations, all transient WFM simulations except those in [10] have used a time step of 1 fs or larger. Computational cost concerns also resulted in the use of a 1 fs time step in this work, although the rapidly increasing power of scientific workstations should make the use of a 0.1 fs time step quite feasible in the near future.

Another conclusion from these and similar GWP simulations is that FE1, FE2, and

BE2 are not robust, having a tendency to diverge (error grows exponentially) after some length of time. Using smaller time steps will delay divergence, but this reduces the computational advantage of the forward Euler approaches. It is not apparent before running a simulation what time step will be required, although guidelines were discussed in Section 5.3.3.5. The only way to make certain a transient simulation is not diverging is to run additional time steps. Finally, implementation of the second-order transient schemes is much more complex than Cayley. For all of these reasons, Cayley discretization is recommended for the transient term in WFM simulation. In fact, Cayley is used in the remaining simulations of this section and in all later chapters of this work.

5.5.4 WFM Simulations of RTDs with SQUADS

Further investigation of the WFM implementation in SQUADS requires the simulation of actual quantum devices, rather than the bulk semiconductor regions used in the Gaussian wave packet simulations above. This section discusses the selection of a quantum device to be used for the simulations presented in the remainder of Section 5.5, as well as

Table 5.2: WFTE transient term discretization scheme accuracy
The maximum percent error is shown (determined from the analytic result) after a 20 fs WFM simulation of a Gaussian wave packet using various discretization schemes for the transient term of the WFTE. The discretization schemes include first- and second-order forward Euler (FE1 and FE2), first- and second-order backward Euler (BE1 and BE2), and Cayley. The simulations are 1) a stationary GWP in the center of the simulation region, 2) a moving GWP starting at the $x = 0$ boundary. Both 1.0 fs and 0.1 fs time steps are simulated. Note that FE1 and FE2 are unreliable (may diverge). Cayley is the optimal transient term difference scheme in terms of accuracy and computational efficiency. A time step of 0.1 fs seems necessary for accuracy, but it entails a huge computational cost.

Transient Scheme	$\Delta t = 1.0$ fs, $NT = 20$		$\Delta t = 0.1$ fs, $NT = 200$	
	$x_0 = L/2$ $k_0 = 0$	$x_0 = 0$ $k_0 = k_{max}/4$	$x_0 = L/2$ $k_0 = 0$	$x_0 = 0$ $k_0 = k_{max}/4$
FE1	14.2%	5.16e10	0.790%	31.5%
BE1	7.32%	49.3%	1.073%	15.4%
Cayley	3.86%	15.8%	0.516%	3.20%
FE2	75.3%	8.11e23	0.684%	7.09%
BE2	6.48%	47.1%	0.693%	8.03%

the choice of simulation parameters used in these simulations. In Section 2.3.4, the RTD was selected as the default quantum simulator test device. One important reason for this choice was the fact that RTDs have a wealth of experimental measurements to which simulations can be compared for accuracy. Several factors must be considered in choosing a particular experimentally measured RTD for comparison to a WFM simulation. Of course, the RTD must have both a structure and a material system that are accurately known. Also, the distance between contacts of the device (denoted L in this work) should be small as possible, so that WFM simulations of the structure are computationally feasible. Finally, the RTD must have adequately reported measurement details (*e.g.*, ambient temperature, contact parasitics, and circuit model of measurement apparatus) and results.

Based on these considerations, the experimental RTD described in [39] was chosen for this investigation of WFM quantum device simulation, having met most of the requirements listed above. The layer structure and energy band offsets of this GaAs/$Al_{0.3}Ga_{0.7}As$ RTD are depicted in Figure 5.10. The RTD lateral area is given as 1 - 5 μm diameter, with a typical diameter of 3 μm. The experimental measurements (and thus all simulations) were carried out at 100 K. Unless stated otherwise, scattering is included in the simulations below, using a relaxation time constant for GaAs at 100 K of 441 fs.[24] The 30% aluminum content of the barriers produces a conduction band offset of approximately 0.23 eV [40]. Except where stated otherwise, this chapter assumes that effective mass is *not* position-dependent. Since most of the device is GaAs, the GaAs bulk effective mass (in the Γ band) of 0.067 m_o was used. Finally, the relative permittivity was taken as 13.1 in GaAs and 10.06 in AlAs, with a linear variation with respect to aluminum fraction [41].

Finally, the WFM simulation parameters will be listed. The optimal diffusion and transient term discretizations were found to be HDS22 and Cayley, respectively. As discussed in Section 5.3.1, the position grid spacing should be equal to the lattice spacing of the material, giving $\Delta x = 0.565$ nm for GaAs. Some length of the heavily doped contact regions must be included in the simulation domain in order that the assumed classical boundary conditions (see Section 5.2.3) are far away from the strong quantum effects near the tunnel barriers and quantum well. These simulations used 26.3 nm contact regions giving $N_x = 120$. All simulations used $N_k = 100$. The choice of transient simulation param-

24. The relaxation time is estimated using logarithmic interpolation between two values for GaAs given in [15] at 77 K and 300K.

eters will be discussed in Section 5.5.6.

In order to compare the WFM simulation results to experimental measurements, it is necessary to use an energy band profile $U(x)$ that closely matches the experimental case. The proper way to achieve this is by enforcing self-consistency, but this added complication is not discussed until Chapter 6. Instead, the simulations in this chapter (like the TMM simulations of Chapter 5) assume a linear potential profile across the active region of the device, with flat energy bands in the contacts, as indicated in Figure 5.10. To approximately mirror the experimental energy band profile, the active region in these simulations is defined to include a 3 nm accumulation region and a 12 nm depletion region.

5.5.5 Steady-State Simulations

Up to this point, this chapter has taken a conceptual approach to the discussion of quantum device simulation using on the Wigner function method. That approach moder-

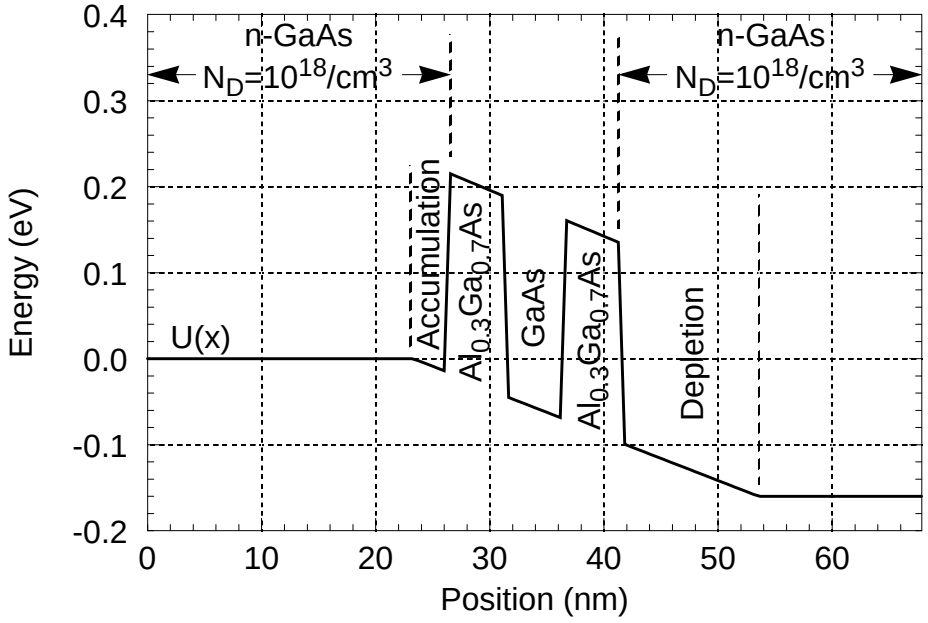


Figure 5.10: Conduction band profile of RTD used in WFM simulations

The conduction band of the RTD used in all remaining simulations of this chapter is shown at a bias of 0.16 V. To approximate the experimental conduction band profile, a 3 nm accumulation region and a 12 nm depletion region are specified. The potential is dropped linearly across the "active" region. The RTD is composed of a 5 nm GaAs quantum well between 5 nm AlGaAs tunnel barriers and 26.3 nm GaAs contact layers, giving a total simulation width of L = 67.6 nm. Contact layer doping is $N_d = 10^{18}/\text{cm}^3$; the other layers are undoped.

ates beginning in this section, as simulations of a real quantum device, the RTD described in the previous section, are presented and analyzed. This more practical approach begins with a plot in Figure 5.11 of the Wigner function for this RTD at 1 V bias. This relatively high bias was chosen for this plot so that the beam of carriers tunneling through the double barrier structure and propagating into the left contact was large (due to a high current flow) and more distinct from the equilibrium carriers (due to a high energy separation). Figure 5.11 can be compared to the TMM-computed Wigner function in Figure 4.10, and shows that the WFM captures the basic quantum physics of RTDs.

Another essential quantum device simulation result is the carrier density profile, which in WFM simulation is calculated by integrating the Wigner function over the wavevector domain, as described in Section 5.3.4. As a typical example Figure 5.12 shows the carrier density for the simulated RTD at a bias of 0.16 V, which is near resonance (peak current) for this RTD. Note the quantum exclusion of carriers not only from the barrier regions, but also from the regions just outside the barriers, even on the accumulation region side (in this case, on the left). By contrast, a classical simulation would predict a high carrier density all the way up to the emitter barrier. Also note that there is a substantial density of carriers in the quantum well at this bias, since the discrete quantum well state is still above the emitter minimum. Thus, carriers are able to tunnel from the emitter into the quantum

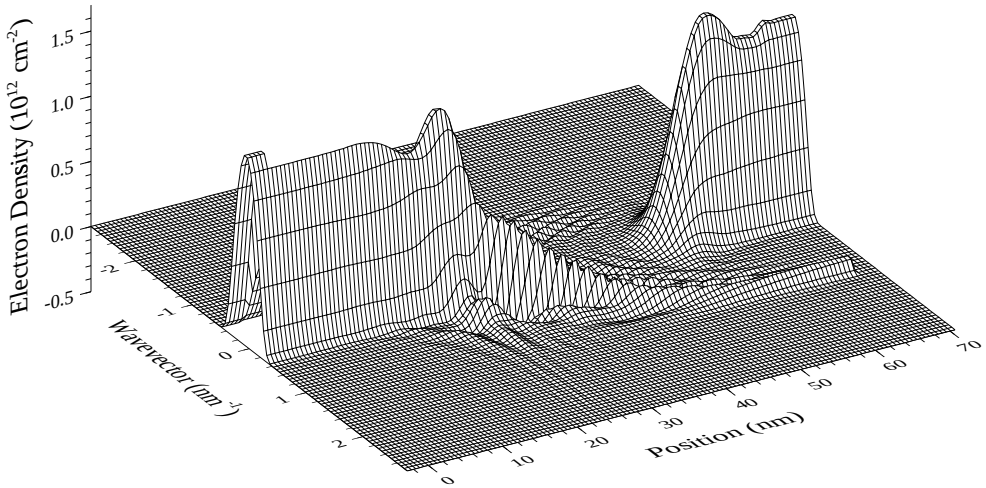


Figure 5.11: RTD Wigner function at high bias

The simulated Wigner function $f_w(x, k)$ is shown for an RTD at 1 V bias. The Wigner function shows the number of carriers versus position and wavevector (proportional to velocity) in the device. The beam of carriers travelling at high velocity into the right contact have tunneled through the RTD.

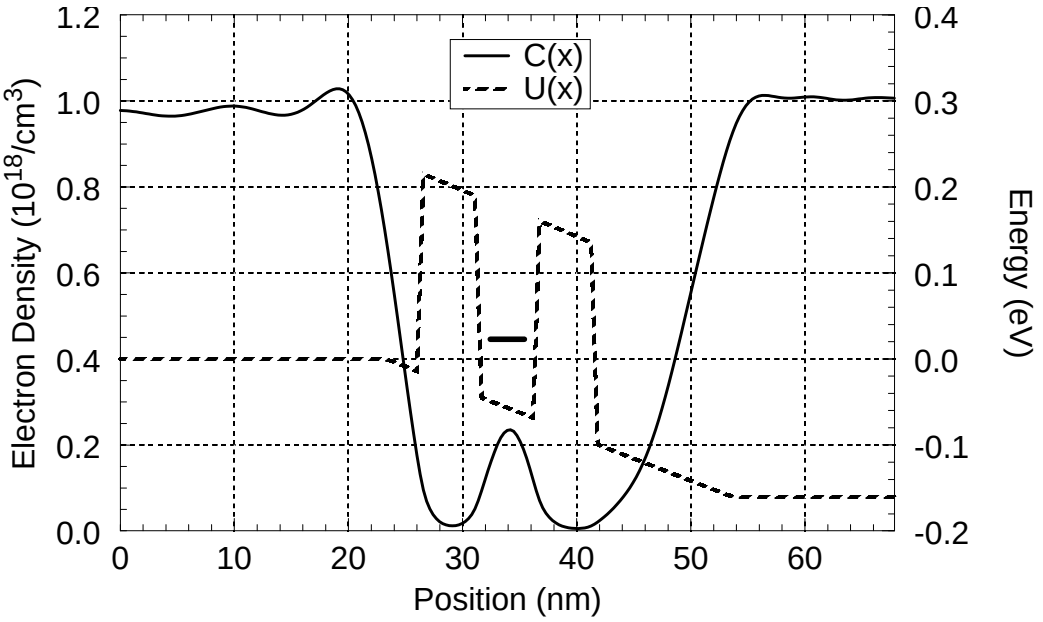


Figure 5.12: RTD carrier density profile near resonance (peak current)

The conduction band profile is also shown (dashed line). Quantum exclusion of carriers is evident in the barriers and in the accumulation region (in this case, just outside the left barrier). However, the substantial density of carriers in the quantum well at this bias indicates that the discrete quantum well state is still above the emitter minimum, and being filled by carriers from the emiter.

well state and then to the collector.

Having demonstrated in the above discussion that the WFM reproduces the basic quantum (and classical) physics of RTDs, the remainder of this section summarizes the results of first-time investigations of three aspects of WFM simulation. The first seeks to ascertain the significance of inelastic scattering (*i.e.,* energy dissipation) in quantum device operation and simulation. To this end, Figure 5.13 shows simulated RTD I-V curves both with and without scattering. The conclusion is that scattering is very important before resonance (peak current). In this region of operation, the Fabry-Perot resonance effect [42] enhances tunneling current if carriers maintain phase coherence through the entire tunneling process. The ability to include scattering is one of the main advantages of the WFM over the TMM, the latter still being the mainstay of quantum device simulation. In spite of the importance of scattering in accurate quantum device simulation, even well below room temperature (these simulations use a 100 K device temperature), the development of more accurate scattering models than the relaxation time approximation have not

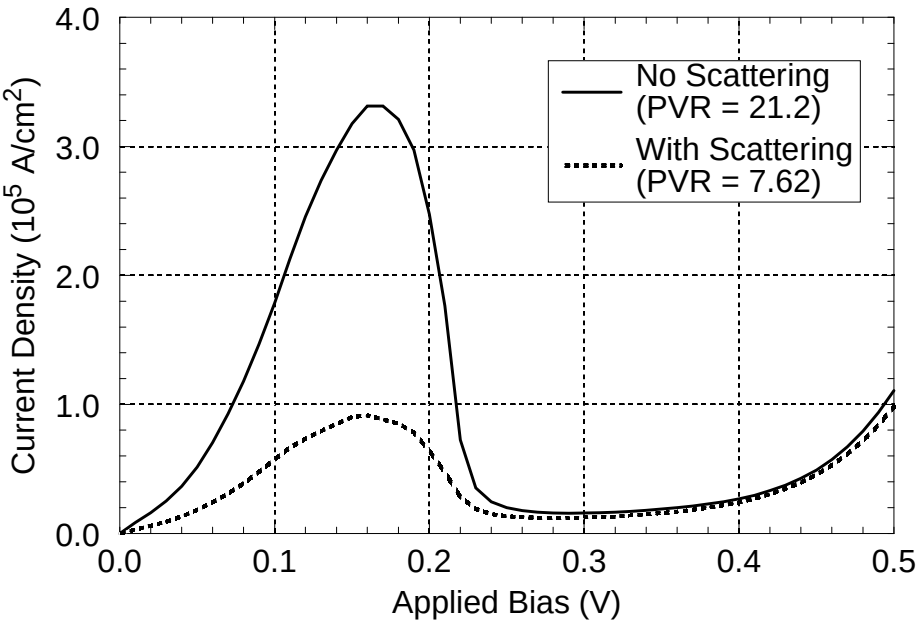


Figure 5.13: Simulated RTD I-V curves with and without scattering

The I-V curves show that the inclusion of scattering in simulations significantly reduces current when the RTD is operating in the Fabry-Perot regime (up to and including the current peak). After resonance, scattering has little effect on current, since coherence is irrelevant. PVR is the peak to valley current ratio, an important figure of merit for RTDs.

been attempted in the WFM. Based on the results in Figure 5.13, this oversight should certainly be addressed in the future.

The next WFM investigation concerns the inclusion of a position-dependent effective mass (PDEM). For reasons that will shortly be apparent, the simulations in this chapter, and most WFM simulations by other researchers, assume a position-*independent* effective mass. As discussed in Section 5.3.3.2, the correct derivation and implementation of a PDEM is quite involved [19, 20], and is therefore not implemented in SQUADS. However, the simple PDEM model used by Frensley [9] was implemented in SQUADS to determine its efficacy in reproducing the effects of a PDEM. The RTD I-V curve with this PDEM model is shown in Figure 5.14, clearly demonstrating that the Frensley PDEM is unacceptable for accurate quantum device simulation. Thus, in order to incorporate a PDEM, the correct (and very complicated) implementation must be used. One advantage of the TMM over the WFM is that correctly implementing a PDEM is much easier. In fact, one conclusion from the TMM simulations in Chapter 4 (see Figure 4.17) was that including a

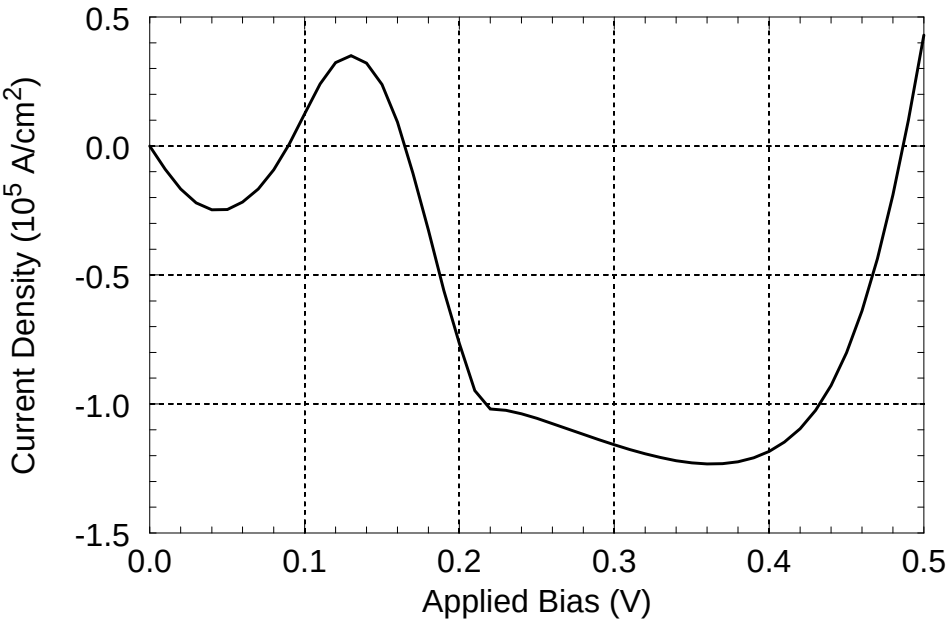


Figure 5.14: RTD I-V curve with simple variable effective mass model

This simulation is identical to the I-V curve in Figure 5.13 including scattering, except that the Frensley position-dependent effective mass model has been turned on here. The regions of negative current (power production) clearly demonstrate that this model is not acceptable for accurate quantum device simulation. The simulation does assume a more conventional appearance at biases above 0.5 V.

PDEM is necessary for accurate quantum device simulation.

Various options for implementing the diffusion and transient terms of the WFTE have been discussed and compared, and only a single approach is common for implementing the scattering term. The remaining term of the WFTE yet to be investigated is the drift term. Therefore, as a final investigation of the steady-state WFM, recall from Section 5.3.3.3 that three different algorithms have been used to calculate the non-local potential (NLP) in the drift term. Only the standard model is mathematically correct,[25] but the modifications were proposed to address other concerns, such as smoothing out abrupt transitions in the potential profile. Figure 5.15 gives simulated RTD I-V curves for each of the NLP models. Note that scattering was not included in this simulation, so that differences in the simulation results would not be obscured. Even so, the differences between the three I-V curves are not large. Thus, the NLP modifications should not significantly degrade accuracy, and thus may be used without undue concern. Whether they accomplish their

25. See the caveat in Section 5.5.7.

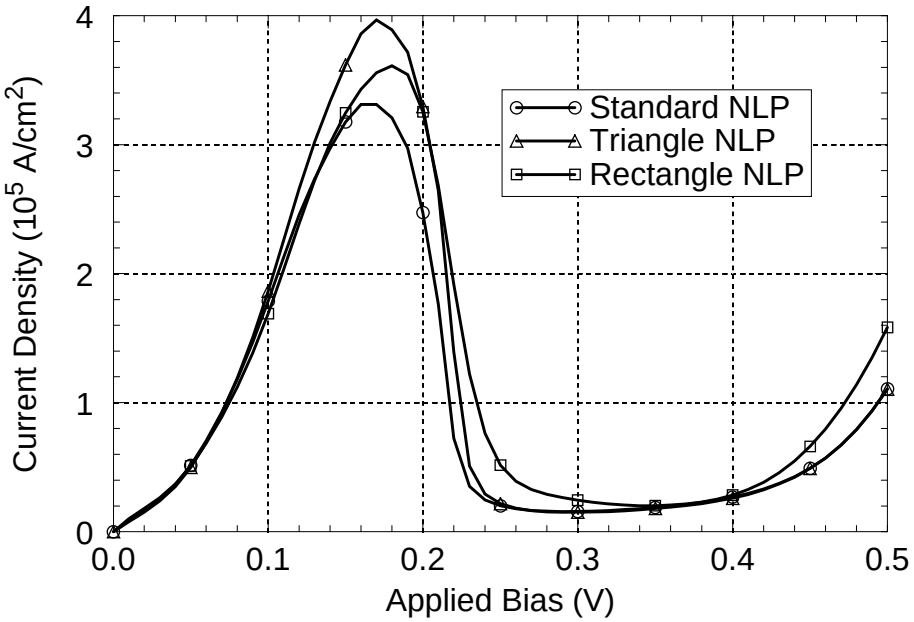


Figure 5.15: RTD I-V curves for different drift term implementations

Compared to the effects of alternative diffusion, transient, and scattering terms, the alternative drift terms [based on modified non-local potential (NLP) calculations] produce relatively little quantitative difference in the simulatied I-V curve. Thus, the alternative (rectangle- and triangle-smoothed) NLP forms can be used without undue concern about introducing a large error in the simulation result.

ancillary purposes is well worth further investigation, but this question will not be addressed here.

5.5.6 Transient Simulations

In addition its ability to include dissipation, another important advantage of the WFM over the TMM is its transient simulation capability, which will be discussed and demonstrated in this section. Since numerous transient simulations are detailed in Chapters 6-8, only a cursory look at transient WFM simulation will be given here. In particular, the traditional transient RTD simulations of switching from peak to valley and vice-versa will be described. For the RTD being investigated in this chapter, the peak and valley applied biases are 0.16 V and 0.28 V, respectively. Figure 5.16 shows the position-averaged current after instantaneously switching the RTD between these two biases. Note that steady-state is essentially reached in about 600 fs when switching from peak to valley, while the opposite switching event takes about 800 fs. The current pulse and high-frequency oscilla-

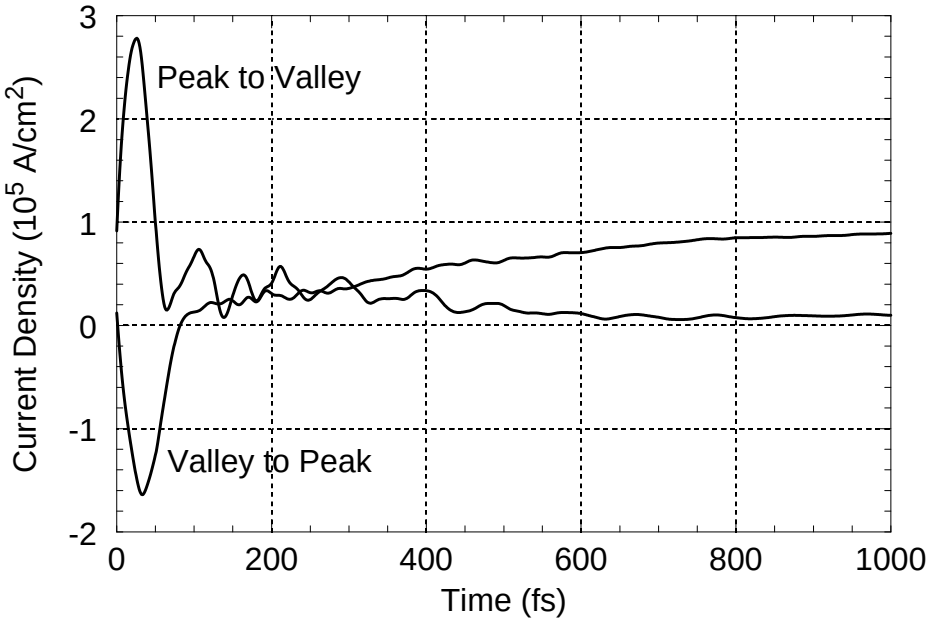


Figure 5.16: Transient current after switching RTD between peak and valley

The simulated RTD is switched instantaneously at $t = 0$ from 0.16 V to 0.28 V applied bias (peak to valley operation) and vice versa. The RTD takes only 600 to 800 fs to essentially reach steady-state, demonstrating that RTDs are inherently fast devices. The origin of the current pulse after switching is discussed in Chapter 7, and the cause of the high-frequency oscillations is discussed in Chapter 8.

tions will be discussed in detail in Chapters 7 and 8, respectively. The simple message, however, is that RTDs and similar quantum devices can operate at very high speed, as was claimed in Section 2.3.4.

5.5.7 Comparison to Experiment

It has been stated several times in this work that the *raison d'etre* of electronic device simulators is to predict how electronic devices (whether existing or proposed) will operate. In order to determine how well a simulation tool meets this goal, its predictions must be compared to experimental measurements of *real electronic devices*. And yet, in a literature review of WFM simulation papers, only one paper [10] out of roughly 50 provided both experimental and WFM simulation results (I-V curves) for the same device. It happens that this is the same device described in Section 5.5.4 and used in the RTD simulations above. In that paper, the simulation results appear to agree quite well with experimental measurements. The obvious question is: why are comparisons between WFM simulations and actual device measurements not published? The answer will be equally obvious when

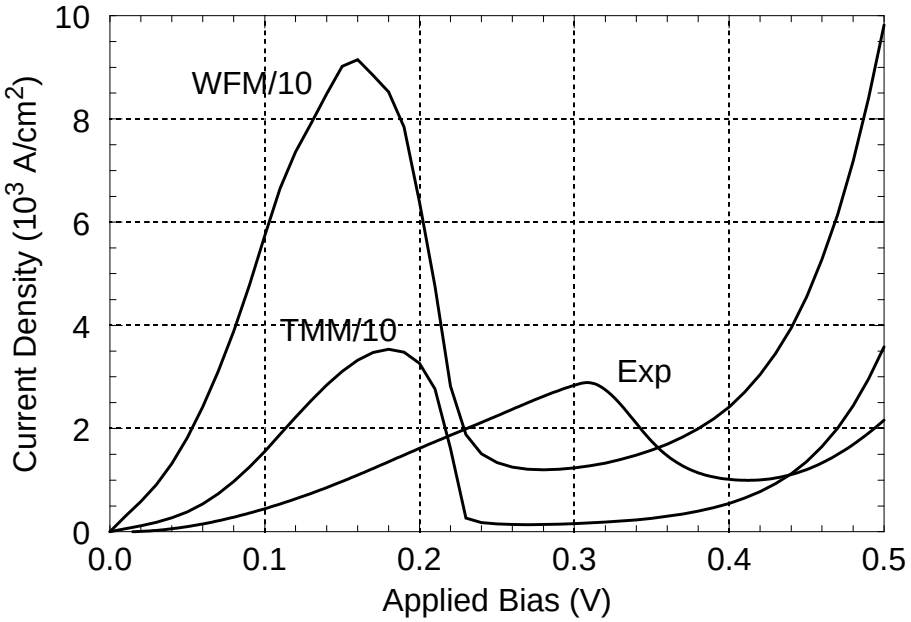


Figure 5.17: Comparison of experimental and simulated RTD I-V curves

Wigner function method (WFM) and transfer matrix method (TMM) simulations are shown (each scaled down by a factor of 10) along with the experimental curve. The simulated peak currents are more than an order of magnitude larger than the experimental value. Some of the discrepancy can be attributed to inaccuracies in the simulation methods (especially the WFM), and some is undoubtedly due to non-idealities in the experimental device and measurement.

the experimental and simulation results for the RTD used in this chapter are juxtaposed.

Figure 5.17 gives the experimental I-V curve from [39] as well as I-V curves from both WFM and TMM simulations. Note that the simulated I-V curves have been scaled down by a factor of ten. Thus, the simulated peak currents are more than a factor of 10 larger than the experimental value.[26] Two other values of interest, the peak and valley voltages, are 100 mV (about 33%) larger in the experimental measurements. Thus, although the simulations have correctly reproduced the basic physics of the RTD, the quantitative accuracy of the simulations leaves much to be desired. It is for this reason that comparisons between WFM simulations and experimental results are not published: the comparison would call into question the usefulness of the simulator.

---

26. Note that the lateral area of the experimental RTD was not directly specified in [39]. The range of sizes was given as 1-5 μm diameter mesas, with a typical diameter of 3 μm. Figure 5.17 calculates current density from total current assuming the typical diameter. Also, when a position-dependent effective mass is included in the TMM, the simulated peak current is a factor of 3 smaller.

In order to improve the agreement between quantum device simulations and experiment, both experimentalists and simulation tool developers will have to cooperate. On the experimental side, the structure of the device must be accurately determined and reported. For example, in RTDs, even single atomic layer error or variation in the tunnel barrier widths produces a huge current change, since tunneling current varies exponentially with barrier width. Similarly, an inexactly known quantum well width will change the peak and valley voltages. Other important intrinsic device characteristics are the doping profile and the lateral area. For example, since a *range* of sizes was given for the RTD simulated above, the current density in the experimental device was unknown by a factor of 25. Also, the measurement details must be accurately described. For example, device temperature may differ from the ambient, there may be significant parasitics (inductance, capacitance, and resistance) in the measurement circuit,[27] and there may be parasitics associated with the device itself (such as surface leakage current on the sides of a mesa, or a surface potential on the sides of a mesa, which would reduce the effective lateral area of the device).

On the simulation side, quantum device simulations must be able to account for external parasitics. Since quantum devices are inherently very small, and thus very sensitive to external conditions and non-idealities, the effects listed in the previous paragraph must be accounted for in, and studied with, simulation. This can be accomplished by deriving an equivalent circuit model for the quantum device, as is done in Chapter 8, and using this model in a larger circuit simulation including suspected or known parasitics. In this way the significance of the parasitics can be understood and their effects mitigated.

Since the TMM is a well-behaved and widely accepted simulation approach, much of the discrepancy between simulation and experiment is likely to be due either to non-idealities in the experimental system, or functionality (such as scattering) that can not be implemented in a TMM simulation. The WFM, on the other hand, is still at a formative stage of development, and its discrepancy with the TMM (as well as experiment) are likely due largely to inaccuracies in the WFM. Many advancements can be made in the WFM which should produce better agreement with experiment, as detailed in Section 9.3. Some of these include a fundamentally different and more accurate implementation of the WFTE [25, 27], simulations with a much higher number of wavevector points, interband interac-

27. The fact that the experimental peak and valley voltages were at higher biases than in simulations, even when self-consistency (see Chapter 6) was enforced, indicates a series resistance between the measurement probes and metal contacts, or between the contacts and the device.

tions, and 2-D simulation capability. In spite of the discrepancies between the WFM and TMM, the two quantum device simulation approaches still produce qualitatively similar results. A comparison of the Wigner functions produced by each simulation method for an RTD at high bias (Figures 4.10 and 5.11) fully verifies this.

In summary, although there is presently a significant quantitative discrepancy between quantum device simulations and experiment, it is still important to make direct comparisons between the two. Through this exercise, the accuracy of quantum device simulators will improve more quickly, enhancing its usefulness in quantum electronics research and development.

5.6 Summary

This chapter has described the Wigner function method of quantum device simulation and its implementation in SQUADS. As discussed in Section 3.5.3, the capabilities and potential of the Wigner function method compliment those of the TMM in SQUADS. Whereas the TMM is well-suited to efficient, wide-ranging simulations, the WFM is required for transient quantum device simulations, and for any simulations where scattering is to be included. To the user, the functionality of a simulator is largely determined by the range of output information it can produce. By this measure, the WFM as implemented in SQUADS can investigate the operation of a quantum device at a single bias, with plots of the Wigner function, carrier density profile, and potential profile; over a range of biases, with I-V and Q-V plots (charge in each device region versus bias); and under transient operation, including time-logs of terminal and average currents, 3-D plots of current density and charge density versus position and time, and Q-T plots (charge in each device region versus time). Most of the more advanced features of the WFM in SQUADS will be used in the investigations in Chapters 6-8.

As discussed in Section 3.5.5 and amply demonstrated in this Chapter, SQUADS was designed to allow the investigation of quantum device *simulators*, as well as quantum *devices*. As a result, a wide range of alternative implementations of the discrete Wigner function transport equation, which is the basis of the WFM, are available in SQUADS. The implementation and comparison of these alternatives enabled the determination in this chapter of their relative accuracies and efficiencies. From Gaussian wave packet simulations, optimal WFTE diffusion and transient term implementations were determined. Fur-

ther simulations of resonant tunneling diodes showed the importance of including scattering in the simulation. Finally, the three proposed drift term implementations were shown to produce little difference in the simulation result, so that alternate forms can be used as desired without a large increase in simulation error.

Another important contribution of this chapter derives from the rather unimpressive comparison between simulation (both WFM and TMM) and experimental I-V curves for an RTD. To date, the under-reported dark secret of quantum device simulation is that it is not quantitatively accurate. Publications show quantum device simulation results which look qualitatively reasonable, but which invariably make no comparison to measured data for the same device. The reason is that the quantitative match is currently very poor. However, unless and until comparisons between simulation and experiment are made, there will be no way to judge the accuracy of the simulator, and no public discussion and investigation of how the accuracy might be improved. As a result, the improvement of quantum device simulation would be much slower than necessary. By disclosing the current limitations of quantum device simulation beyond the few groups who pursue this endeavor directly, this work should help to accelerate the process of quantum device simulation development.

Finally, in spite of the quantitative disagreement between the WFM and TMM, both simulation approaches have demonstrated their ability to provide qualitative insight into quantum device operation. With this understanding, Chapters 6-8 describe three in-depth investigations, based mainly on the WFM capabilities of SQUADS, of quantum device simulation and the operation of RTDs.

References

[1] C. Kittlel and H. Kroemer. *Thermal Physics*, page 14. W. H. Freeman, New York, 2nd edition, 1980.

[2] B. A. Biegel. *SQUADS Technical Reference*. (Unpublished), Stanford University, 1996.

[3] E. Wigner. “On the quantum corrections for thermodynamic equilibrium.” *Physical Review*, 40:749–759, June 1 1932.

[4] K. L. Jensen and A. K. Ganguly. “Numerical simulation of field emission and tunneling: A comparison of the wigner function and transmission coefficient

approaches." *Journal of Applied Physics*, 73(9):4409–4427, 1993.

[5] U. Ravaioli, M. A. Osman, W. Potz, N. Kluksdahl, and D. K. Ferry. "Investigation of ballistic transport through resonant-tunneling quantum wells using Wigner function approach." *Physica B*, 134:36–40, 1985.

[6] N. Kluksdahl, W. Potz, U. Ravaioli, and D. K. Ferry. "Wigner function study of a double quantum barrier resonant tunneling diode." *Superlattices and Microstructures*, 3(1):41–45, 1987.

[7] W. R. Frensley. "Transient response of a tunneling device obtained from the wigner function." *Physical Review Letters*, 57(2):2853–2856, 1986.

[8] W. R. Frensley. "Quantum transport simulation of the resonant-tunneling diode." In *International Electron Devices Meeting (IEDM) Technical Digest*, pages 571–574, 1986.

[9] W. R. Frensley. "Wigner-function model of a resonant-tunneling semiconductor device." *Physical Review B*, 36(3):1570–1580, 1987.

[10] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer. "Self-consistent study of the resonant-tunneling diode." *Physical Review B*, 39(11):7720–7735, 1989.

[11] W. R. Frensley. "Effect of inelastic processes on the self-consistent potential in the resonant-tunneling diode." *Solid State Electronics*, 32(12):1235–1239, 1989.

[12] K. L. Jensen and F. A. Buot. "Numerical simulation of transient response and resonant-tunneling characteristics of double-barrier semiconductor structures as a function of experimental parameters." *Journal of Applied Physics*, 65(12):5248–5250, 1989.

[13] K. L. Jensen and F. A. Buot. "The methodology of simulating particle trajectories through tunneling structures using a Wigner distribution approach." *IEEE Transactions on Electron Devices*, 38(10):2337–2347, 1991.

[14] K. L. Jensen and A. K. Ganguly. "Quantum transport simulations of electron field emission." *Applied Physics Letters*, 55(7):669–671, 1989.

[15] K. L. Jensen and F. A. Buot. "The effects of scattering on current-voltage characteristics, transient response, and particle trajectories in the numerical simulation of resonant tunneling diodes." *Journal of Applied Physics*, 67(12):7602–7607, 1990.

[16] K. L. Jensen and F. A. Buot. "Numerical simulation of intrinsic bistability and

high-frequency current oscillations in resonant tunneling structures." *Physical Review Letters*, 66(8):1078–1081, 1991.

[17] F. A. Buot and K. L. Jensen. "Intrinsic high-frequency oscillations and equivalent circuit model in the negative differential resistance region of resonant tunneling devices." *COMPEL*, 10(4):241–253, 1991.

[18] K. L. Jensen and F. A. Buot. "The numerical simulation of particle trajectories in quantum transport and the effects of scattering and self-consistency on the performance of quantum well devices." In *International Electron Devices Meeting (IEDM) Technical Digest*, pages 771–775, 1990.

[19] H. Tsuchiya, M. Ogawa, and T. Miyoshi. "Simulation of quantum transport in quantum devices with spatially varying effective mass." *IEEE Transactions on Electron Devices*, 38(6):1246–1252, 1991.

[20] K. K. Gullapalli and D. P. Neikirk. "Incorporating spatially sarying effective-mass in the Wigner-Poisson model for AlAs/GaAs resonant-tunneling diodes." In *Proceedings of the 3rd International Workshop on Computational Electronics*, pages 171–174, 1994.

[21] K. K. Gullapalli, D. R. Miller, and D. P. Neikirk. "Simulation of quantum transport in memory-switching double-barrier quantum-well diodes." *Physical Review B*, 49(4):2622–2628, 1994.

[22] D. R. Miller and D. P. Neikirk. "Simulation of intervalley mixing in double-barrier diodes using the lattice Wigner function." *Applied Physics Letters*, 58(24):2803–2805, 1991.

[23] G. Y. Wu and K.-P. Wu. "Electron transport in a resonant-tunneling diode under the effect of a transverse magnetic field: A quantum theory in the Wigner formalism." *Journal of Applied Physics*, 71(3):1259–1264, 1992.

[24] J.-R. Zhou and D. K. Ferry. "Simulation of ultra-small GaAs MESFET using quantum moment equations." *IEEE Transactions on Electron Devices*, 39(3):473–478, 1992.

[25] R. K. Mains and G. I. Haddad. "An accurate re-formulation of the wigner function method for quantum transport modeling." *Journal of Computational Physics*, 112:149–161, 1994.

[26] F. A. Buot and K. L. Jensen. "Lattice Weyl-Wigner formulation of exact many-

body quantum-transport theory and applications to novel solid-state quantum-based devices." *Physical Review B*, 42(15):9429–9457, 1990.

[27] W. R. Frensley. "Boundary conditions for open quantum systems driven far from equilibrium." *Reviews of Modern Physics*, 62(3):745–791, 1990.

[28] W. B. Joyce and R. W. Dixon. "Analytic approximations for the fermi energy of an ideal fermi gas." *Applied Physics Letters*, 31(5):354–356, 1977.

[29] S. Selberherr. *Analysis and Simulation of Semiconductor Devices*, page 12. Springer-Verlag, New York, 1984.

[30] A. M. Kriman, J.-R. Zhou, N. C. Kluksdahl, H. H. Choi, and D. K. Ferry. "Effective potential for moment-method simulation of quantum devices." *Solid State Electronics*, 32(12):1603–1607, 1989.

[31] C. D. McGillem and G. R. Cooper. *Continuous and Discrete Signal and System Analysis*, page 169. Holt, Reinhart, and Winston, New York, 2nd edition, 1984.

[32] N. C. Kluksdahl, A. M. Kriman, and D. K. Ferry. "The role of visualization in the simulation of quantum electronic transport in semiconductors." *Computer*, pages 60–66, Aug. 1989.

[33] Private conversation with K. L. Jensen.

[34] G. H. Golub and C. F. Van Loan. *Matrix Computations*, chapter 3. The John Hopkins University Press, Baltimore, 2nd edition, 1989.

[35] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*, chapter 2. Cambridge University Press, Cambridge, MA, 2nd edition, 1992.

[36] R. E. Jansen, B. Farid, and M. J. Kelly. "The steady-state self-consistent solution to the nonlinear Wigner-function equation; a new approach." *Physica B*, 175:49–53, 1991.

[37] S. Collins, D. Lowe, and J. R. Barker. "The quantum mechanical tunneling time problem revisited." *Journal of Physics C*, 20:6213–6232, 1987.

[38] M. A. Morrison. *Understanding Quantum Physics: A User's Manual*. Prentice-Hall, Englewood Cliffs, NJ, 1990.

[39] T. C. L. G. Sollner, P. E. Tannenwald, D. D. Peck, and W. D. Goodhue. "Quantum well oscillators." *Applied Physics Letters*, 45(12):1319–1321, 1984.

[40] T. C. L. G. Sollner, W. D. Goodhue, P. E. Tannenwald, and D. D. Peck. "Resonant

tunneling through quantum wells at frequencies up to 2.5 thz." *Applied Physics Letters*, 43(6):588–590, 1983.

[41] S. Adachi. "Gaas, alas, and algaas: Material parameters for use in research and device applications." *Journal of Applied Physics*, 58(3):R1–R29, 1985.

[42] F. Capasso, K. Mohammed, and A. Y. Cho. "Resonant tunneling through double barriers, perpendicular quantum transport phenomena in superlattices, and their device applications." *Journal of Quantum Electronics*, 22(9):1853–1868, 1986.

Chapter 6

Quantum Self-Consistency

Self-consistency in electronic device simulation means ensuring that the carrier density profile in the simulated device is consistent with its potential profile, as dictated by Poisson's equation. In previous simulations in this dissertation, the potential profile was approximated by some simple algorithm, and no attempt was made after solving the Wigner function transport equation or Schrödinger equation to assure that the resulting carrier profile was consistent with the assumed potential. However, to accurately model real quantum devices, enforcing self-consistency is essential. This chapter details the implementation of self-consistency in SQUADS, and demonstrates its significant effect on device operation predictions. For two reasons, most of this chapter is dedicated to self-consistency in Wigner function method (WFM) simulation. First, scattering must also be included for accurate self-consistent simulations, and only the WFM can include scattering. Second, SQUADS implements four alternative implementations of self-consistency for the WFM, while only one of these approaches is possible with the TMM.

The organization of this chapter is as follows. Section 6.1 presents the background information necessary for an understanding of the implementation of quantum self-consistency. The next three sections (6.2-6.4) present the analytical formulation and numerical implementation of each of the four WFM self-consistency approaches. Section 6.5 then simulates the self-consistent I-V curve of an RTD as a test case to compare the efficiency (computational cost), accuracy (ability to correctly reproduce device physics), and robustness (reliability) of these iteration methods. Finally, Section 6.6 describes briefly the

implementation of self-consistency in TMM simulation, and gives a few associated simulation results.

6.1 Background

As described in Chapter 5, the Wigner function method of quantum device simulation models a quantum system by computing the evolution of the Wigner function $f(x, k, t)$ according to the Wigner function transport equation (WFTE):

$$\frac{\partial}{\partial t} f(x,k,t) + \frac{\hbar k}{m}\frac{\partial}{\partial x} f(x,k,t) + \frac{1}{\hbar}\int \frac{dk'}{2\pi} V(x, k-k') f(x,k',t) - \frac{\partial}{\partial t} f(x,k,t)\Big|_c = 0 \qquad (6.1)$$

To enforce self-consistency in the WFM [1-5], the Poisson equation (PE) relating the potential profile to the carrier density profile must be satisfied simultaneously with the WFTE. In 1-D, the PE can be written:

$$\frac{d}{dx}\left[\varepsilon(x)\frac{d}{dx}u(x)\right] = q\rho(x) = q^2[C(x) - c(x)] , \qquad (6.2)$$

where ε is permittivity, u is the (Hartree, or mean-field) potential, q is the electronic charge, c is the free electron density, and C is the fixed charge density (*e.g.*, ionized dopants). The conduction band minimum is calculated from the potential:

$$U(x) = u(x) + \delta U(x), \qquad (6.3)$$

where δU is the (fixed) heterostructure band offset and $U(x)$ is the potential used in calculating the non-local potential $V(x, k)$ in the WFTE. To complete the WFTE - PE interdependence, the carrier density is calculated from the Wigner function using:

$$c(x) = \frac{1}{2\pi}\int dk\, f(x,k). \qquad (6.4)$$

Conceptually, a self-consistent simulations proceeds as follows: the PE uses the carrier density profile $c(x)$ to determine the energy band profile $U(x)$ of the device, and the WFTE uses $U(x)$ to determine (among other things) $c(x)$. The relationship between carrier density and energy bands is non-linear. Finding the simultaneous solution of the WFTE and PE therefore requires iteration. Section 5.4 described the high computational cost of solving the WFTE, so an n iteration solution of the WFTE will be *n times* as expensive. The result is that the computational efficiency of the WFM self-consistency iteration method is critically important. To investigate and compare efficiency and other strengths and weaknesses, SQUADS implements four basic self-consistency iteration methods for

the WFM: steady-state Gummel, transient Gummel, steady-state Newton, and transient Newton.

Due to the difficulty of implementing and maintaining multiple self-consistency iteration approaches in a numerical simulator, most researchers using the WFM rely on a single implementation, usually the steady-state or transient Gummel approach, in their quantum device research. Test case simulations in Section 6.5 illustrate the dangers of this practice, and show how to take advantage of the complementary strengths of both steady-state and transient iteration methods where appropriate. SQUADS' modular structure makes it ideally suited to the implementation and comparison of alternative simulation approaches, such as the comparison of self-consistency iteration methods in this chapter. Only by presenting the theory, numerical implementation, and simulation examples for all of these simulation alternatives in a cohesive framework is this comparison possible.

In selecting specific simulation examples for this comparison, obviously only transient iteration methods are suitable for time-dependent investigations, such as switching, small-signal, or large-signal simulations. However, for the very basic electronic device simulation task of tracing the current-voltage (I-V) curve, steady-state methods are also suitable. Therefore, the accurate generation of the I-V curve for the "prototypical" quantum device, the resonant tunneling diode (RTD) [6-8], served as the test case for evaluating the four self-consistency iteration methods. In fact, this device and simulation task have been the most common in the Wigner function simulation literature. Figure 6.1 shows a "typical" measured RTD I-V curve [7]. Some features of note in this I-V curve are a negative differential resistance region and a bistable region. The "plateau" shape in the negative differential resistance region is actually the time-average of a very fast oscillating current. The ability of the various self-consistency iteration methods to efficiently and reliably reproduce these features will be the basis for their comparison.

6.2 Discretization of the Poisson Equation

To solve the WFTE - PE system, the simulation domain must be discretized, and, accordingly, these two equations. Details of WFTE discretization are described in Chapter 5. This section describes most of the PE discretization, leaving only those details which differ between the self-consistency iteration methods for the following sections. The PE has been discretized in two ways by researchers investigating the WFTE - PE system: the

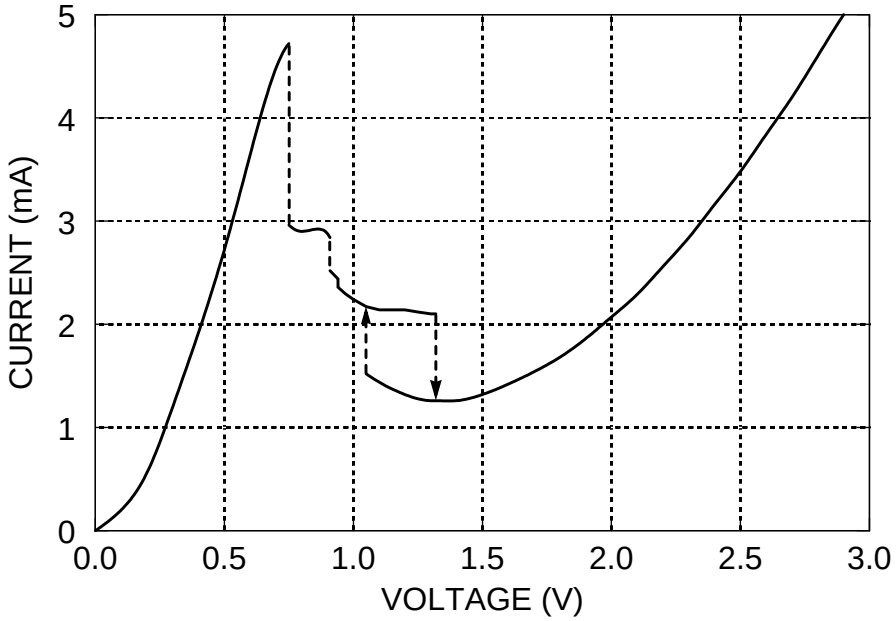


Figure 6.1: Experimental RTD I-V curve

Experimental RTD I-V curve [7] showing the characteristic negative differential resistance region and plateau structure between 0.8 V and 1.3 V. The plateau current is actually the time-average of a high-frequency oscillating current. [Permission to reprint data given by T.C.L.G Sollner.]

direct form [9] and the differential (a.k.a. Newton) form [3]. The appropriate form of PE depends on the self-consistency iteration method, as discussed in Sections 6.3 and 6.4. Both forms are described here.

Recall from Chapter 5 that the position dimension of the WFM simulation domain is discretized as:

$$x_i = i\Delta_x, \qquad i \in \{0, 1, \ldots, N_x\}, \qquad (L = N_x\Delta_x), \tag{6.5}$$

where L is the width of the simulation region. With this discretization, the direct Poisson equation can be written, for a position-dependent permittivity:[1]

$$(1 + a)u_{i+1} - 2u_i + (1 - a)u_{i-1} = (q^2\Delta_x^2/\varepsilon_i)[C_i - c_i] \quad , \tag{6.6a}$$

$$a \equiv (\varepsilon_{i+1} - \varepsilon_{i-1})/(4\varepsilon_i) \quad . \tag{6.6b}$$

Details of the derivation of (6.6a) are given in [10].

The Newton form of the PE is more complicated but more flexible. Newton equations

1. For the position-independent permittivity assumed in the simulations of Section 6.5, $\alpha = 0$, and the discrete PE (in both direct and differential forms) is even simpler.

are inherently iterative, seeking to find the solution to a non-linear system by successively better approximations. To derive the Newton PE, first define the "Poisson function" $P(u)$, which is based on the PE and must evaluate to 0 when the self-consistent potential and carrier density are supplied as input. From (6.2):

$$P^{(n)}(u) \equiv \frac{d}{dx}\left[\varepsilon(x)\frac{d}{dx}u^{(n)}(x)\right] - q^2[C(x) - c^{(n)}(x)] \quad . \tag{6.7}$$

In (6.7), n is the iteration index (which for transient simulations is also the time step). A Newton iteration is a 2-step process. First, the Newton PE system of equations is solved for δu, the *change* in the potential:

$$\left[\frac{\partial P^{(n)}(u)}{\partial u^{(n)}}\right][\delta u^{(n+1)}(x)] = -[P^{(n)}(u)] \quad . \tag{6.8}$$

Then the potential is updated:

$$u^{(n+1)}(x) = u^{(n)}(x) + \delta u^{(n+1)}(x) \quad . \tag{6.9}$$

If the Newton iteration converges to the self-consistent solution, $P^{(n)}(u)$ converges to 0, and therefore so will the updates, δu. The converged self-consistent potential will be denoted $u^*(x)$.

In discrete form, the Newton PE becomes:

$$\left[\frac{\partial P_i^{(n)}}{\partial u_{i'}^{(n)}}\right][\delta u_i^{(n+1)}] = -[P_i^{(n)}(u)] \quad , \tag{6.10a}$$

$$u_i^{(n+1)} = u_i^{(n)} + \delta u_i^{(n+1)} \quad , \tag{6.10b}$$

where:

$$P_i^{(n)}(u) = (1+a)u_{i+1}^{(n)} - 2u_i^{(n)} + (1-a)u_{i-1}^{(n)} - (q^2\Delta_x^2/\varepsilon_i)[C_i - c_i^{(n)}] \quad , \tag{6.11}$$

$$\frac{\partial P_i^{(n)}}{\partial u_{i'}^{(n)}} = (1+a)\delta_{i+1,\,i'}^{(n)} - 2\delta_{i,\,i'}^{(n)} + (1-a)\delta_{i-1,\,i'}^{(n)} + \left(\frac{q^2\Delta_x^2}{\varepsilon_i}\right)\left[\frac{\partial c_i^{(n)}}{\partial u_{i'}^{(n)}}\right] \quad , \tag{6.12}$$

and $\delta_{i,\,i'}$ is the Kronecker delta function.[2] Note that $\partial c/\partial u$ is left unspecified for now since its value depends on which self-consistency iteration method is used. It is not difficult to show [10] that the direct PE, (6.6a), is a special case of the Newton PE, where $\partial c/\partial u = 0$. SQUADS implements the direct PE using this special case of the differential

2. The Kronecker delta function is unity when the two subscripts are equal, and zero otherwise. In (6.12), for example, $\delta_{i+1,\,i'}$ is unity where column i' equals row $i+1$, and zero otherwise.

PE. To make the admittedly abstract Newton PE (6.10a) a little more concrete, Figure 6.2 shows the discrete, direct PE in differential (*i.e.*, Newton) matrix form for $N_x = 10$.

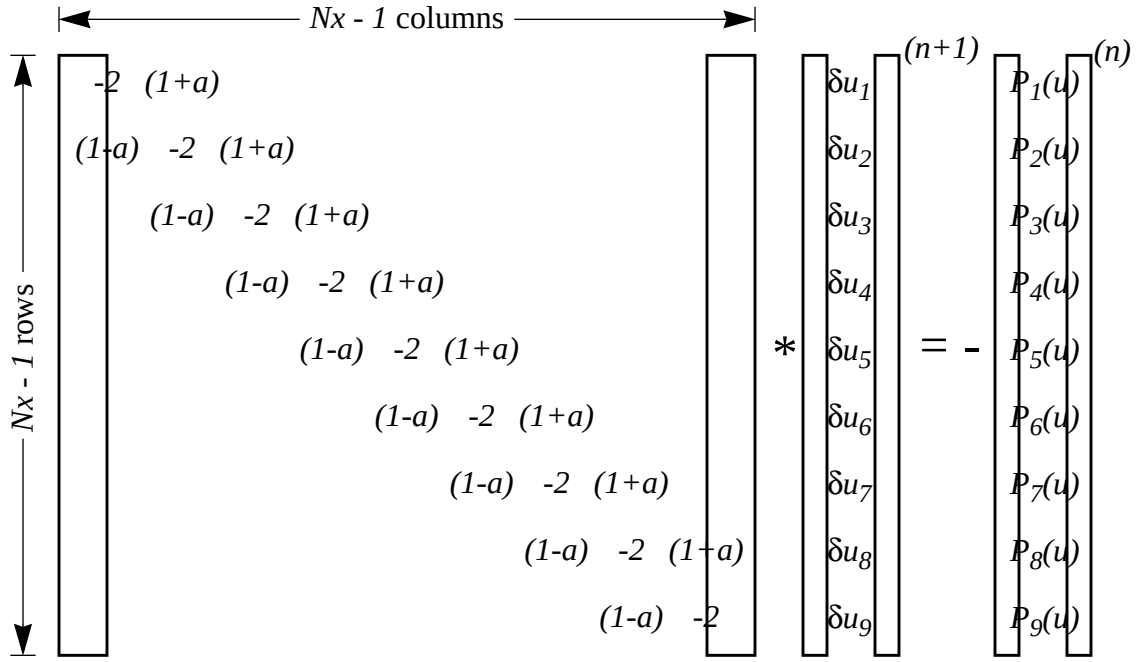


Figure 6.2: Discrete, direct Poisson equation in matrix form

The example above assumes $Nx = 10$. Because the $\partial c/\partial u$ feedback term is zero in the direct PE, the coefficient matrix is constant (iteration independent), so the iteration superscript is omitted for this term. Also note that since u_0 and u_{Nx} are fixed boundary conditions, $\delta u_0 = \delta u_{Nx} = 0$, and equations corresponding to these points do not appear in the discrete PE.

As discussed above, enforcing self-consistency (*i.e.*, finding the simultaneous solution of the WFTE and PE) is an iterative process. Solving the PE with the carrier density profile $c^{(n)}(x)$ as input yields an updated potential profile $u^{(n+1)}(x)$. Solving the WFTE with this updated potential produces a new carrier profile $c^{(n+1)}(x)$. However, it is often possible to predict $c^{(n+1)}(x)$ approximately even without solving the full (and very expensive) WFTE. An inexpensively computed prediction could be used to make $u^{(n+1)}(x)$ closer to the self-consistent solution $u^*(x)$, and thus allow $u^*(x)$ to be found with fewer iterations (and solutions of the WFTE). In fact, this inexpensive prediction is exactly the purpose of the $\partial c/\partial u$ term. That is, the $\partial c/\partial u$ term (which should be based on a distillation of the WFTE) provides some corrective feedback to achieve faster convergence to the self-consistent operating point. By taking $\partial c/\partial u = 0$, the direct PE does not attempt to use this

prediction, while the Newton PE does, with $\partial c / \partial u \neq 0$.

An unresolved issue is what should be used for the initial potential profile u^0 in the first solution of the Newton PE and the WFTE at each bias point in an I-V curve simulation. For steady-state I-V curve simulations, SQUADS uses linear extrapolation from $u^*(x)$ at the previous two bias points.[3] Transient I-V curve tracing is one continuous simulation, so the final potential profile $u^*(x)$ at one bias point is used to compute u^0 at the next. In particular, when the bias is incremented in a transient simulation, the potential profile is incremented linearly across the entire device (see Section 7.1).[4]

The combination of the WFTE and PE, when discretized for numerical solution, constitutes a non-linear system of equations. The self-consistency iteration methods offer a means of solving this non-linear system (which can't be solved directly) by iteratively solving a set of linear equations (which *can* be solved directly). The following two sections detail the remainder of the numerical implementation of four self-consistency iteration methods for the WFTE - PE system. Because the mathematics of the steady-state and transient approaches of each method (Gummel or Newton) are similar, the two Gummel approaches are described together in Section 6.3, and the two Newton approaches in Section 6.4. However, tracing the self-consistent operating points along the I-V curve, which task has been chosen for this iteration method comparison, is very different for the transient and steady-state approaches. The steady-state approaches try to locate the self-consistent operating point in as few iterations as possible, while the transient approaches seek to follow the actual time-dependent operation of the device until it evolves to steady-state. Therefore, when running simulations, the converse pairing is more appropriate, so in Section 6.5 the two steady-state methods are considered together followed by the two transient methods.

6.3 Gummel (Plug-in) Approach

The Gummel (*a.k.a.* plug-in) approach [11] to solving the WFTE - PE system is almost universally used to add self-consistency to the WFTE. This is due to the simplicity of the Gummel approach, since the two equations are solved independently [12], and the

3. At the first bias point, linear band bending is used, and at the second, a linear potential is added to $u^*(x)$ at the first bias point.

4. Again, linear band bending is used to initialize the potential profile at the first bias point.

PE is numerically much simpler to implement and solve than the WFTE. For the steady-state Gummel method [2, 3], the steady-state WFTE and the PE are iteratively and alternately solved, plugging-in one equation's solution as input for the other. When the Wigner function and potential stop changing (within specified convergence criteria), the self-consistent operating point has been reached. For the transient Gummel method [1, 13] the only mathematical difference is that the transient WFTE is used, so that each iteration is a time step. That is, one alternately time-steps the WFTE and updates the potential using the PE until steady-state operation is reached (again, within specified convergence criteria). The transient Gummel iteration is initiated by solving the WFTE once in steady-state mode.

Now consider whether the direct or Newton form of the PE (*i.e.*, zero or non-zero $\partial c/\partial u$ term in (6.12)) should be used for the steady-state and transient Gummel iteration methods. Test simulations showed that a steady-state Gummel iteration often diverges (consecutive Wigner function and $u(x)$ solutions oscillate wildly) unless some corrective feedback is supplied through a non-zero $\partial c/\partial u$. Thus, the Newton PE must be used for the steady-state Gummel method. In general, there is no exact, closed-form expression for $\partial c/\partial u$ for a quantum system. This is why the WFTE is solved - it accounts for quantum effects such as tunneling and reflection, along with non-equilibrium carrier transport, to relate the energy bands to carrier concentration. So, an approximate form for $\partial c/\partial u$ is sought that is easy to compute but still produces self-consistency convergence. To this end, SQUADS uses the classical, equilibrium expression for $\partial c/\partial u$. Any justification for the choice of $\partial c/\partial u$ must be based on the transport equation. In this case, the boundary conditions on the WFTE supply carriers to the device according to the classical relationship, even though quantum processes and non-equilibrium transport will distort this relationship as the distance from the contacts increases. Also, scattering (usually included in the WFTE in self-consistent simulation) tends to produce the classical result.

The standard approach (cf. [3]) in deriving $\partial c/\partial u$ is to assume equilibrium, classical Maxwell-Boltzmann statistics:

$$c(u) = N_c \exp[(u - u_0)/(k_B T)] \ , \tag{6.13}$$

$$\frac{\partial c}{\partial u} = \frac{c(u)}{k_B T} . \tag{6.14}$$

Under this assumption, the carrier density c_i at a given position x_i depends only on the

potential u_i at the same point. Thus, the discrete form of (6.14) is:

$$\frac{\partial c_i}{\partial u_{i'}} = \frac{c_i}{k_B T}\delta_{i,\,i'}. \tag{6.15}$$

In other words, the $\partial c/\partial u$ term in (6.12) only modifies the main diagonal elements in the coefficient matrix of Figure 6.2. Using (6.15) as the feedback term in the Newton PE results in relatively slow but reliable convergence to the self-consistent operating point.

Note that the boundary conditions in (5.6a) and (5.6b) for the WFTE are based on Fermi-Dirac statistics, not Maxwell-Boltzmann statistics. Test simulations showed that using Fermi-Dirac statistics to derive $\partial c/\partial u$ can significantly accelerate the convergence speed of the steady-state Gummel method. SQUADS uses the Joyce-Dixon approximation [14] to relate c and u according to Fermi-Dirac statistics. To determine $\partial c/\partial u$, we write $u(c)$, derive $\partial u/\partial c$, and invert. Thus:

$$r \equiv c/N_c, \tag{6.16}$$

$$u - u_0 = k_B T\left[\ln(r) + \sum_{m=1} a_m r^m\right], \tag{6.17}$$

$$\frac{\partial u}{\partial c} = \frac{\partial u}{\partial r}\frac{\partial r}{\partial c} = k_B T\left[\frac{1}{r} + \sum_{m=1} \frac{a_m}{m} r^{m\text{-}1}\right]\frac{1}{N_c}, \tag{6.18}$$

$$\frac{\partial c}{\partial u} = \frac{N_c}{k_B T}\left[\frac{1}{r} + \sum_{m=1} \frac{a_m}{m} r^{m\text{-}1}\right]^{-1}. \tag{6.19}$$

Although the $\partial c/\partial u$ term in (6.19) is more complicated than that in (6.15), it still only modifies the main diagonal elements in the Newton PE coefficient matrix of Figure 6.2.

Test simulations also showed that using Joyce-Dixon terms above $m = 3$ does not improve convergence speed. In fact, in cases where $r_i \gg 1$ for one or more position nodes x_i, including higher order terms may render the steady-state Gummel method non-convergent. Therefore, SQUADS uses a third-order Joyce-Dixon approximation by default. If the iterates P_i^n are not converging towards 0, the iteration drops back to Maxwell-Boltzmann statistics (zeroth-order Joyce-Dixon approximation) until progress towards convergence is maintained for several iterations. The algorithm by which the Joyce-Dixon order is dynamically chosen to accelerate convergence of the steady-state Gummel method in SQUADS is now rather complicated, being based more on experience than theory. Only the standard (*i.e.*, Maxwell-Boltzmann) form of $\partial c/\partial u$ has been used in previous steady-state Gummel iterations of the WFTE - PE system. Section 6.5.5 shows that the acceler-

ated convergence algorithm described above greatly decreases the computational cost of the steady-state Gummel iteration method.

In contrast to the steady-state Gummel method, the transient Gummel method seeks to follow the exact evolution of the device. Since there is no closed form for $\partial c / \partial u$ in a general quantum system, and because the approximations typically used (such as those used with the steady-state Gummel method) are only heuristically correct, using them in the transient Gummel method is more likely to create physics than to model it. To avoid this, the direct PE ($\partial c / \partial u = 0$) must be used. For the transient Gummel method, then, each iteration starts with the exact potential profile for the carrier density at the current time point, the system is evolved one time step with the transient WFTE, and then the potential is adjusted for the new (but only slightly different) carrier density. The results of particular transient and steady-state Gummel simulations are presented in Section 6.5.

6.4 Full Newton Approach

With the Gummel approach to solving the WFTE - PE system, two independent (*i.e.*, uncoupled) sets of linear equations are alternately solved, one derived from the WFTE and resulting in an updated Wigner function, and the second derived from PE and producing an updated potential. With the full Newton formulation [15], a *combined* (*i.e.*, coupled) WFTE - PE linear system is solved to produce simultaneous updates of both the Wigner function and potential. The advantage of the full Newton approach is that changes in one solution directly affect the outcome of the other, so the corrective feedback that had to be approximated in the steady-state Gummel method is inherent in the Newton formulation. This tends to produce much faster convergence with a steady-state Newton method than with the steady-state Gummel method. Like the transient Gummel method, the transient Newton method seeks to follow the exact evolution of the quantum system, so it evolves to the steady-state operating point only as quickly as a real device would. However, the transient Newton method should be more accurate than the transient Gummel method, though by how much is not yet clear.

Use of the Newton formulation for quantum self-consistency [16] requires the definition of a WFTE function $W(F)$, analogous to the PE function defined in (6.7). For this purpose, simply use (5.29):

$$W(F) \equiv (\boldsymbol{T} + \boldsymbol{K} + \boldsymbol{P} + \boldsymbol{S})[F] + (4/\Delta_t) f_{i,j,n} . \tag{6.20}$$

The Newton formulation for the WFTE - PE system solves the following system:

$$\begin{bmatrix} \dfrac{\partial W}{\partial F} & \dfrac{\partial W}{\partial u} \\ \dfrac{\partial P}{\partial F} & \dfrac{\partial P}{\partial u} \end{bmatrix}^{(n)} \begin{bmatrix} \delta F \\ \delta u \end{bmatrix}^{(n+1)} = -\begin{bmatrix} W(F) \\ P(u) \end{bmatrix}^{(n)}, \tag{6.21}$$

where the left-most matrix is the Jacobian and P is the Poisson function defined in (6.7). After each solution of (6.21), the unknowns are updated as:

$$F^{(n+1)} = F^{(n)} + \beta(\delta F)^{(n+1)}, \tag{6.22a}$$

$$u^{(n+1)} = u^{(n)} + \alpha(\delta u)^{(n+1)}. \tag{6.22b}$$

As with the Gummel iteration methods, convergence towards the steady-state, self-consistent operating point with the Newton iteration methods is determined by monitoring the progress of the Poisson function $P^{(n)}(u)$ and update $\delta u^{(n)}$ iterates towards 0.

The update scaling factors α and β in (6.22a) and (6.22b) are used only for the steady-state Newton method. Because the transient Newton method attempts to exactly follow the transient operation of the device, one must not modify the updates that are computed. Even for the steady-state Newton method, these update factors are ideally unity, though they can be reduced to some fraction when the iterates are not converging. Frensley [4] used $\alpha = 0.5$ and $\beta = 0.1$. However, for the simulations reported herein, in the few cases when the steady-state Newton method could not locate the self-consistent operating point, reducing α and β did not help, and in fact usually made convergence less likely. Thus, the simulations in this work always used $\alpha = \beta = 1$. Instead, where convergence was not occurring with the steady-state Newton method, SQUADS uses the steady-state Gummel method until the iteration begins converging again. Finally, since the Newton update in (6.22a) requires a Wigner function to update *from*, both steady-state and transient Newton simulations begin with a single steady-state Gummel solution of the WFTE. Initialization of the potential profile was discussed in Section 6.2.

The full Newton equation (6.21) must be discretized for numerical solution. In discrete form (6.21) is:

$$\begin{bmatrix} \dfrac{\partial W_{i,j}}{\partial F_{i',j'}} & \dfrac{\partial W_{i,j}}{\partial u_{i'}} \\ \dfrac{\partial P_i}{\partial F_{i',j'}} & \dfrac{\partial P_i}{\partial u_{i'}} \end{bmatrix}^{(n)} \begin{bmatrix} \delta F_{i,j} \\ \delta u_i \end{bmatrix}^{(n+1)} = -\begin{bmatrix} W_{i,j}(F) \\ P_i(u) \end{bmatrix}^{(n)} . \tag{6.23}$$

Expressions for $W_{i,j}$ and $P_{i,j}$ were given in Section 6.2. The Jacobian blocks have yet to be determined. Actually, the Jacobian block for $\partial W/\partial F$ is identical to the coefficient matrix used for the WFTE solution of a Gummel iteration, although the unknowns in the Newton formulation are $\delta F_{i,j,n}$ instead of $F_{i,j,n}$. The only difference between the Gummel WFTE coefficient matrix and the Newton $\partial W/\partial F$ Jacobian block is that terms which become boundary conditions with the Gummel formulation are zero with the Newton formulation, since δF^{bc} is zero. Thus, these terms do not appear in the right-hand-side vector as in the Gummel methods.

The $\partial P/\partial u$ Jacobian block is also slightly different than the PE coefficients used in the Gummel formulation. In particular, with the Newton formulation, there is no need to approximate the effect of the change in potential on the carrier concentration through $\partial c/\partial u$. This relationship is taken care of exactly through the off-diagonal Jacobian blocks. The $\partial P/\partial u$ block is therefore the same as that used for the direct PE:

$$\frac{\partial P_i}{\partial u_{i'}} = (1+a)\delta_{i+1,i'} - 2\delta_{i,i'} + (1-a)\delta_{i-1,i'} . \tag{6.24}$$

The more interesting Jacobian blocks in this case are the off-diagonal ones, if only because expressions for these have (to our knowledge) never been published, although Frensley has used the steady-state Newton method to solve the WFTE - PE system [4]. The $\partial W/\partial u$ block is somewhat complicated, due to the convoluted way in which the u_i enter into the computation of the non-local potential, as shown in (5.40). Note from the relationship between the band edge U and the potential energy u in (6.3) that

$$\frac{\partial W_{i,j}}{\partial u_{i'}} = \frac{\partial W_{i,j}}{\partial U_{i'}}. \tag{6.25}$$

After some effort [10], the $\partial W/\partial u$ Jacobian block is:

$$\frac{\partial W_{i,j}}{\partial u_{i'}} = \frac{-4\pi}{N_k h} \sum_{j''=1}^{N_k} f_{i,j''} \sin\left[\frac{2(i-i')(j-j'')\pi}{N_k}\right], \tag{6.26a}$$

$$(1 \le |i'-i| \le N_k/2) . \tag{6.26b}$$

The Jacobian block for $\partial P / \partial f$ is much simpler. Recalling from (6.4) how carrier concentration c is calculated, and using the definition of the discrete Poisson function in (6.11):

$$\frac{\partial P_i}{\partial f_{i',\,j'}} = \frac{q^2 \Delta_x}{2\varepsilon N_k} \delta_{i,\,i'}. \tag{6.27}$$

Combining all of these results, Figure 6.3 gives an example of the structure and size of the discrete full Newton equation for $N_x = 7$, $N_k = 6$. Since the $\partial W / \partial f$ Jacobian block is identical in the Gummel and Newton formulations, and because this block is by far the largest in the Jacobian matrix, one might expect that solving the WFTE - PE system by the two approaches should require roughly the same storage and CPU time. This is not at all the case, especially in SQUADS, where the storage and solution of the discrete WFTE (and thus the $\partial W / \partial f$ Jacobian block) have been highly optimized. The result is that the Newton formulation requires typically twice the storage and five times as much CPU time per loop as the Gummel formulation. Performance data are presented in the next section for all self-consistency iteration methods along with simulation results.

6.5 Results and Discussion

6.5.1 Simulated Device and Parameters

Simulations in this chapter (and Chapters 7 and 8) use the RTD device structure and simulation parameters of Jensen and Buot [9] as a test case. The simulated RTD, depicted in Figure 6.4 at equilibrium, is composed of a 5 nm undoped GaAs quantum well between 3 nm undoped $Al_{0.3}Ga_{0.7}As$ tunnel barriers and 3 nm undoped GaAs spacer layers. The GaAs contact layers are 19 nm each, giving a total device width of $L = 55$ nm. The electron effective mass is assumed constant at $0.0667 m_0$, and the permittivity is also taken as constant at $12.9\varepsilon_0$. Finally, these simulations assumed $N_x = 86$, $N_k = 72$, $\Delta_t = 1$ fs, and $\tau = 525$ fs [17] at $T = 77$ K.

6.5.2 Convergence Criteria

The choice of convergence criteria for the WFTE - PE iteration presents a dilemma: too loose of criteria and the predicted self-consistent operating point is not trustworthy; too tight and the number of iterations required for convergence may rise dramatically. This

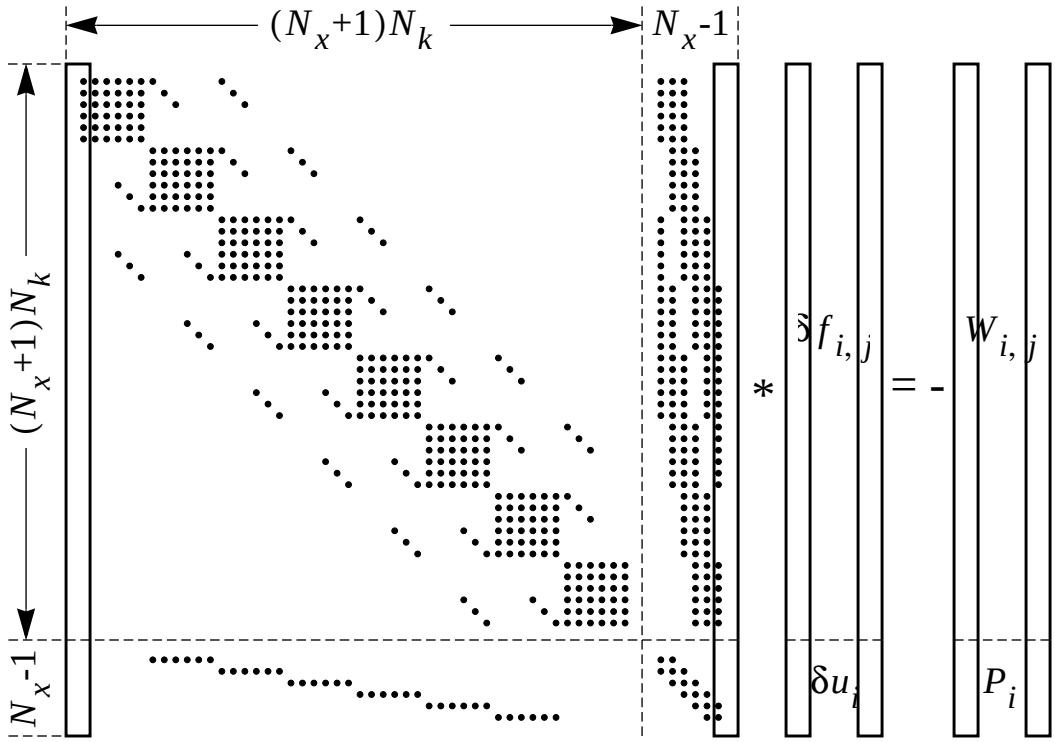


Figure 6.3: Full-Newton WFTE-PE matrix equation

The Jacobian matrix block sizes and non-zero coefficient structure are shown for $N_x = 7$, $N_k = 6$.

work errs on the side of too much computation rather than too little: convergence criteria were relatively strict.

For steady-state simulations the proper convergence criterion is simply to verify that the (direct) Poisson equation is satisfied to a high degree. These simulations required that:

$$P_i(u) < 10^{-8}\,\text{eV} \qquad (0 < i < N_x)\,. \tag{6.28}$$

This convergence criteria, although necessary, is not sufficient in all cases. To assure that consecutive solutions are not oscillatory, and for steady-state Newton simulations where $P(u)$ is always very small (if update constant α is unity), it is also necessary to require that the potential update at any point be very small:

$$\delta u_i < 10^{-6}\,\text{eV} \qquad (0 < i < N_x)\,. \tag{6.29}$$

These relatively strict convergence criteria are feasible for the steady-state iteration methods because convergence tends to be very fast. Some researchers [3] have used criteria like (6.29) as their only indication of self-consistency, but this is not sufficient. It is possible,

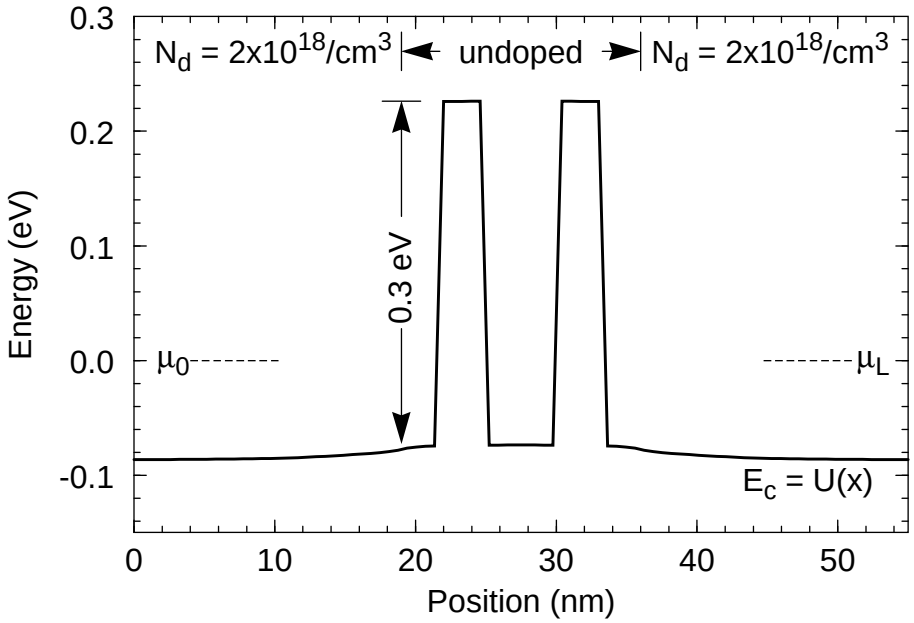


Figure 6.4: GaAs RTD used in self-consistency simulations

Shown are the equilibrium self-consistent conduction band, Fermi levels, and doping. The 0.3 eV $Al_{0.3}Ga_{0.7}As$ tunnel barriers are 3 nm thick, and the GaAs quantum well width is 5 nm. The center 17 nm of the device (including 3 nm outside each tunnel barrier) are undoped.

especially with an approximate iteration method such as the steady-state Gummel approach, for the potential updates to be small without actually having reached the self-consistent solution.

The convergence criteria in (6.28) and (6.29) are also enforced for the transient iteration simulations in this work, but they are inadequate to guarantee that the steady-state, self-consistent operating point has been reached. The δu criterion is not especially revealing in a transient simulation because of the approximate proportionality of δu to the time step, Δt. [A small time step gives little time for carriers to move, resulting in a correspondingly small change in the potential.] Also, because transient simulations tend to oscillate around the steady-state operating point as they relax towards it, satisfying the $P(u)$ criterion does not guarantee that a simulation has reached steady-state. A more definitive convergence criterion for transient simulations, also used by Jensen and Buot [5], is based on the fact that the discrete current density for the WFTE is defined such that it is position-independent at steady-state, as discussed in [10]. Thus, a WFTE transient simulation can be said to have reached steady-state when the variation in current density δJ over

the width of the device drops below some relatively small value. In this work, current densities were on the order of 10^5 A/cm^2, so the final transient simulation convergence criterion is:

$$\delta J \equiv (J_{\max} - J_{\min}) < 1000 \text{ A/cm}^2 \ . \tag{6.30}$$

This criterion is less strict than one might prefer, but tightening it results in excessively long simulation times. When (6.30) is satisfied in a transient simulation, a steady-state simulation using the final potential profile usually differs from the actual steady-state current density by less than 10 A/cm^2. Therefore, when it is necessary to verify controversial transient simulation results in this work, transient simulations will be run in which all three convergence criteria are several orders of magnitude tighter.

6.5.3 Steady-State Iteration Method Simulations

One purpose of this chapter is to examine when the (physically-based) transient iteration methods are required to accurately reproduce the operation of an RTD, and when the computationally more efficient steady-state iteration methods may be used. The test device for this work was selected because of the very interesting I-V curve simulated by Jensen and Buot [5], who used the transient Gummel method to implement self-consistency. Their simulations produced an I-V curve similar in shape to the experimental curve in Figure 6.1 (although for a different RTD). In fact, they even observed persistent current oscillations for all biases in the plateau region of the I-V curve, concluding that [5] “intrinsic oscillations have a dominant influence on the plateaulike structure and hysteresis in the *I-V* characteristics.” Subsequent work by Buot and Rajagopal [18, 19] described the physics behind this behavior.

Based on the results obtained by Jensen and Buot, it was not clear that the steady-state Gummel and Newton iteration method simulations would converge in the plateau region, since persistent oscillations indicate that no stable, self-consistent operating point exists. Although unstable equilibrium points should exist in this region, an otherwise convergent steady-state iteration method could be rendered non-convergent. In fact, both the accelerated Gummel and the Newton simulations were unable to converge at some challenging points in the plateau. However, by automatically using the standard Gummel iteration method in these cases, the steady-state iteration methods did find self-consistent operating points over the entire simulated bias range. The resulting I-V curve (Figure 6.5) was very

similar to that of Jensen and Buot (also shown), and identical for the two steady-state iteration methods. The hysteresis loop in the I-V curve required the simulation of both the up-trace (0.0 V to 0.4 V) and the down-trace (0.4 V to 0.0 V).

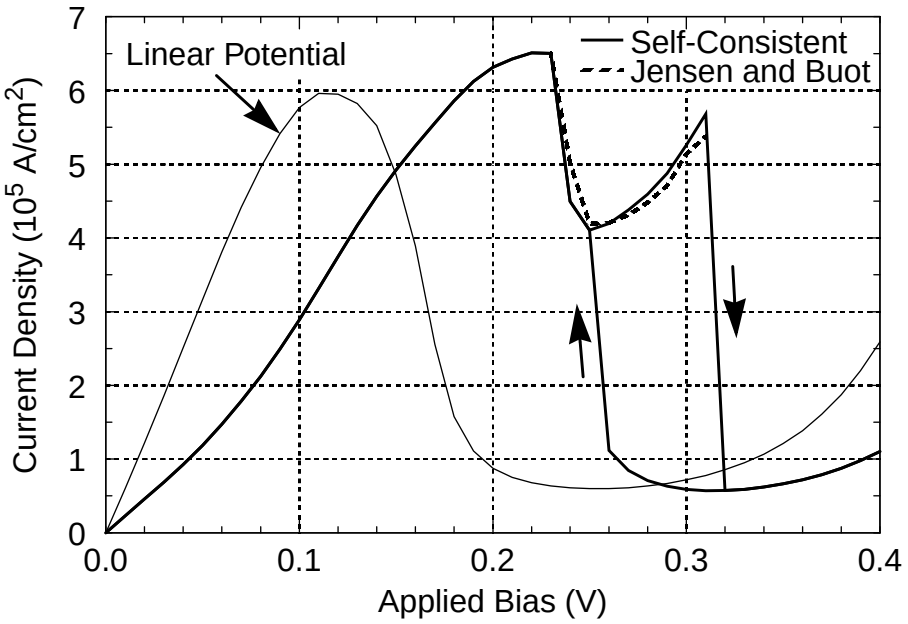


Figure 6.5: Steady-state simulated RTD I-V curve

Both Gummel and Newton steady-state self-consistency iteration methods are shown. Jensen and Buot's up-trace [5] (where different) and a non-self-consistent (linear potential) I-V curve are shown for comparison. [Permission to reprint data given by K. L. Jensen.]

It seems contradictory that the steady-state iteration methods found steady-state operating points in the plateau (0.24 V to 0.31 V on the up-trace and 0.25 V to 0.24 V on the down-trace), while the transient simulation of Jensen and Buot did not. One possible explanation is that these oscillations, although persistent, are not perpetual. Jensen and Buot's conclusion that oscillations are required for the plateau to occur seem to rule this out. If the oscillations are perpetual, the simultaneous WFTE and PE solutions found by the steady-state iteration methods must be unstable equilibrium operating points. Thus, given any impulse or even numerical noise, a system prepared according to the steady-state solution will begin to oscillate in a transient simulation. Determining whether one or both of these explanations are correct can only be accomplished with transient iteration simulations, which are described in the following section.

Before moving on to transient simulations, one conclusion can already be drawn based

on the RTD I-V curves in Figure 6.5. Also shown in Figure 6.5 is a non-self-consistent simulated I-V curve for the RTD of Figure 6.4. This simulation assumed a linear potential drop across the undoped (central) region of the RTD, and therefore did not require solution of the PE, and only a single WFTE solution per bias point. Comparing these simulated I-V curves with the experimental one in Figure 6.1 (for a different RTD structure), it is clear that although the linear potential simulation was able to predict a negative differential resistance region, that is about the limit of its usefulness. On the other hand, the similarity between the simulated self-consistent I-V curve the experimental curve clearly shows that enforcing self-consistency is necessary to reproduce some of the salient physics of real RTDs. The open question at this point is whether the computationally expensive transient iteration methods can add any further detail.

6.5.4 Transient Iteration Method Simulations

To compare self-consistency iteration methods, and now to investigate the nature of the plateau operating points, the transient Gummel iteration method was used to simulate the I-V curve of the RTD in Figure 6.4 over the same bias range as for the steady-state simulations. A maximum of 4,000 iterations (4 ps) per bias point was allowed. If the transient simulation did not converge in this time (*e.g.*, due to sustained oscillations), the simulation moved to the next bias point anyway. Surprisingly, although the current oscillations observed by Jensen and Buot did occur in the plateau region, the simulation converged for all bias points except the first three in the plateau (0.24 V, 0.25 V, and 0.26 V). Further, the resulting I-V curve (except for those three points) was indistinguishable from the steady-state curve, as one would expect (assuming the convergence criteria are strict enough).

The oscillations in the plateau region were progressively more persistent at lower biases. Whereas only 1,300 iterations were required to reach convergence at 0.31 V, fully 3,800 iterations were required at 0.27 V. The 0.26 V bias point was apparently on course to convergence at 4,000 iterations. Indeed, further evolution resulted in full convergence after a total of 7,008 iterations. To demonstrate these oscillations, Figure 6.6 shows the complete plot of collector current versus time at 0.26 V on the up-trace. Both the oscillation amplitude and the convergence criteria decreased very regularly over the course of the simulation, with a decay constant of 0.2/ps. For example, for the oscillation amplitude:

$$A(t) \approx 0.8\times10^{5} e^{-(0.2t/1\text{ps})} \text{A/cm}^2 \,. \tag{6.31}$$

Although the ultimate fates of the remaining points, 0.24 V and 0.25 V on both curve traces, were inconclusive after 4,000 iterations, extrapolation from the results and trends for the other plateau points suggested that their oscillations would simply be even more persistent, but not perpetual.

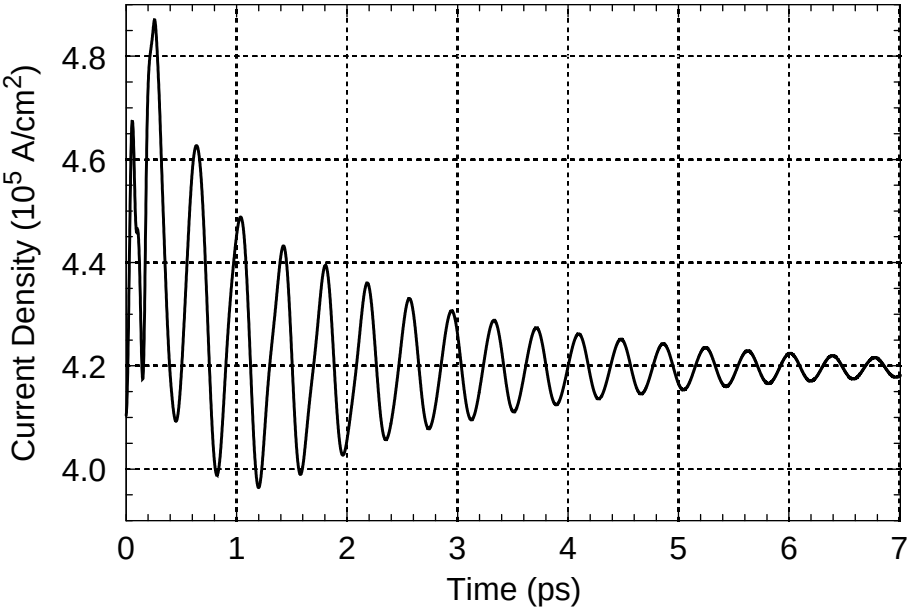


Figure 6.6: Damped oscillatory current in quasi-stable plateau region

Shown is the simulated transient collector current as RTD evolves to steady-state after switching from 0.25 V to 0.26 V, showing that the RTD is stable at this bias.

The expectation that the transient RTD simulation would eventually reach steady-state for 0.24 V and 0.25 V turned out to be incorrect. Further evolution (in either curve-trace direction) led to oscillations of constant amplitude by about 8,000 iterations at both biases. For example, Figure 6.7 shows the transient current at 0.24 V on the up-trace. These simulations were allowed to run for several thousand more iterations to make certain that the oscillations were not slowly decreasing, as had been expected. Data on the final oscillations at these two points (independent of the trace direction) are given in Table 6.1.

One additional test was employed to assure that the 0.24 V and 0.25 V bias points were unstable. As suggested in the previous section, transient Gummel simulations were run starting from the fully-converged steady-state Gummel solution at 0.24 V and 0.25 V. Now the expectation was that the simulations would diverge (i.e., oscillation amplitude

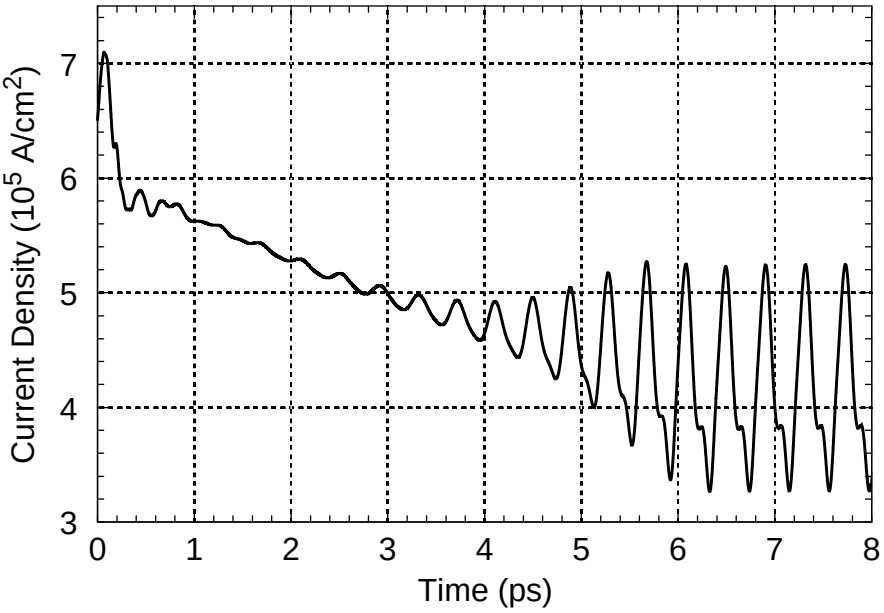


Figure 6.7: Unstable operation in NDR region of plateau

Simulated transient collector current after switching from 0.23 V to 0.24 V, showing sustained oscillations.

Table 6.1: Oscillation statistics for unstable operation

Collector current final oscillation data (after 10 ps) at applied biases of 0.24 V and 0.25 V. Current density from the steady-state simulations is appended for comparison.

Oscillation Parameter	0.24 V	0.25 V
Amplitude (10^5 A/cm^2)	1.98	1.08
Period (ps)	0.413	0.374
Frequency (THz)	2.42	2.67
Time-average (10^5 A/cm^2)	4.18	4.06
Steady-state current (10^5 A/cm^2)	4.50	4.10

would increase). This was indeed the result. The collector current versus time for the 0.24 V simulation is shown in Figure 6.8. The result for 0.25 V was similar. Divergence was very regular, with a decay constant of -0.4/ps at 0.24 V and -0.2/ps at 0.25 V. For the oscillation amplitude at 0.24 V:

$$A(t) \approx 6.31 e^{(0.4t/1\text{ps})} \text{A/cm}^2 . \tag{6.32}$$

Of course, the oscillation amplitude will be bounded, just as it was in Figure 6.7. These results prove that this RTD is inherently unstable at these biases. To model this behavior, Buot and Jensen describe an equivalent circuit model for the RTD [20] that reproduces the bounded instability depicted in Figures 6.7 and 6.8.

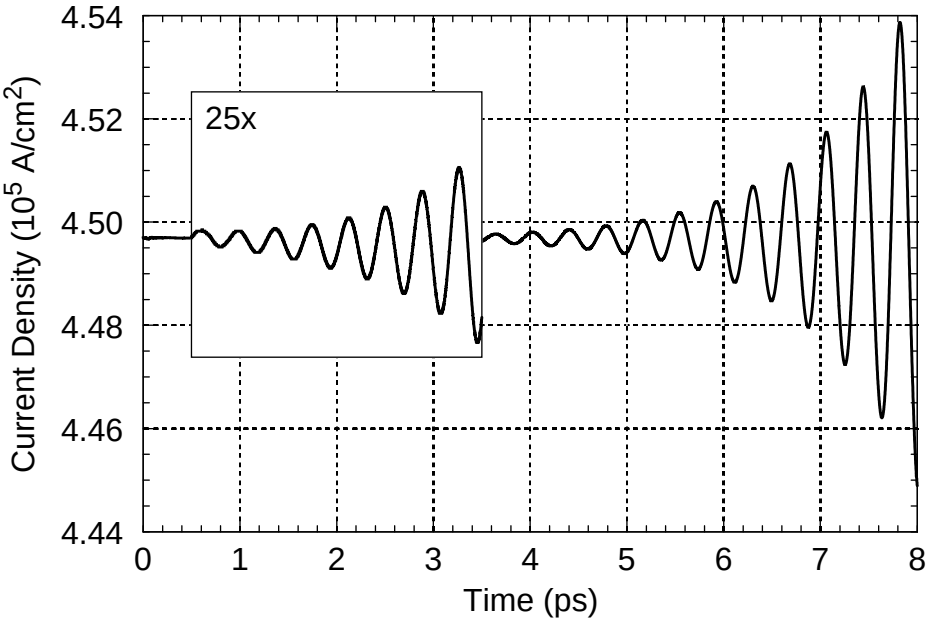


Figure 6.8: Unstable RTD diverging from steady-state

Simulated transient collector current starting from a fully-converged steady-state Gummel iteration simulation at 0.24 V, showing that the RTD is unstable at this bias.

The results at 0.24 V and 0.25 V call into question the conclusion above that the remainder of the plateau is stable. The convergence criterion in (6.30) is perhaps not strict enough to justify this conclusion. In particular, it leaves open the possibility that the RTD might oscillate perpetually with an amplitude of less than $1000\ \text{A/cm}^2$. To verify that the upper portion (0.26 V - 0.31 V) of the plateau is stable, simulations were run at the lower end (0.26 V), middle (0.29 V), and top (0.31 V) of this region with four orders of magnitude stricter convergence criteria. Most importantly, the current variation ΔJ was required to be less than $0.1\ \text{A/cm}^2$ for convergence. Throughout these simulations, the oscillations continued to decay regularly at all three bias points, reaching convergence at 27,906, 10,424, and 7,522 iterations respectively. To illustrate, Figure 6.9 shows a plot of the cur-

rent variation versus time for the 0.26 V simulation.

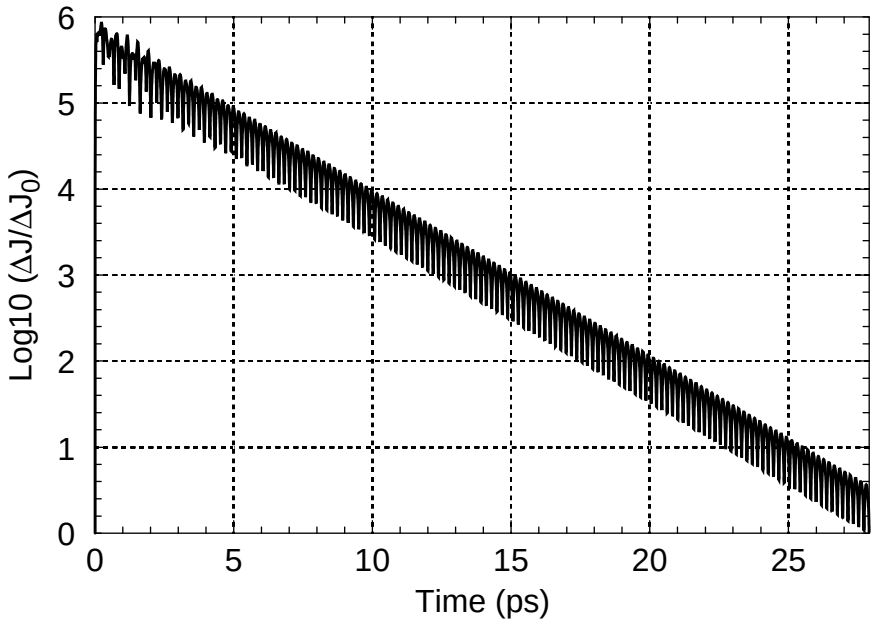


Figure 6.9: Current variation vs. time for marginally-stable operation

Current variation versus time after switching from 0.25 V to 0.26 V. ΔJ is the current variation and ΔJ_0 is the convergence criterion of $0.1\ \mathrm{A/cm}^2$. The simulation converges regularly, showing that the RTD is stable at this bias. The spikes in the curve are due to the decaying oscillations.

Based on the above transient simulations, the plateau in the simulated RTD's I-V curve is composed of two parts: an unstable region (0.24 V - 0.25 V) in which the RTD oscillates forever, and a stable region (0.26 V - 0.31 V) where persistent oscillations eventually die out. Actually, these regions are simply the result of a monotonic increase in the exponential decay constant [see (6.31) and (6.32)] from -0.4/ps at 0.24 V, through 0 at about 0.255 V, and up to about 0.67/ps at 0.31 V. The unstable region agrees with Jensen and Buot's results showing perpetual oscillations in the plateau, while the stable region contradicts their conclusion that these oscillations persist throughout the plateau and are required for the plateau to occur. In fact, these oscillations have only a minor effect on the value of the I-V curve in the unstable region of the plateau (see Table 6.1), and no effect at all elsewhere. Further analysis of Jensen and Buot's work [5] suggests that their incorrect conclusions resulted from the premature termination of transient simulations, the use of an accelerated convergence technique, or both.

The above discussion of transient self-consistency simulations did not mention the

transient Newton iteration method. In fact, only partial I-V curve traces (5 - 10 points in either direction and some plateau region points) were run using this iteration method. These simulations showed that the RTD evolved almost identically with the transient Newton method as with the transient Gummel method. For example, Figure 6.10 compares the collector current from the I-V curve simulations at 0.06 V for the two transient iteration methods. Although the transient Newton method sometimes converged a few iterations faster, for the bias point shown in Figure 6.10, the transient Gummel and Newton methods converged in exactly the same number of iterations (629).

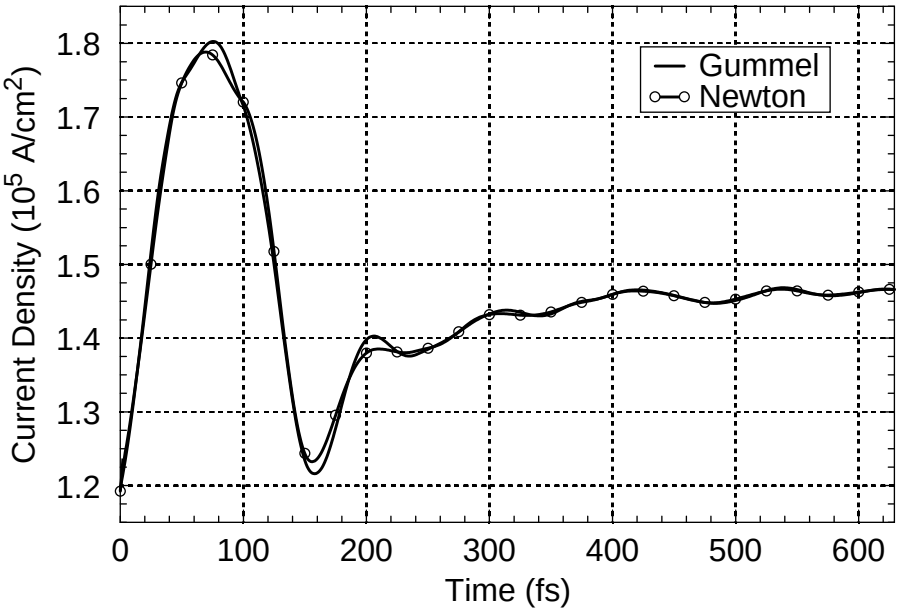


Figure 6.10: Comparison of transient Gummel and Newton results

Simulated collector current for transient Gummel and Newton iteration method simulations after switching from 0.05 V to 0.06 V. This indicates that the Gummel approach is effectively as accurate at the Newton approach for the chosen simulation parameters.

From these observations, it was apparent that performing a full I-V curve trace with the transient Newton method would provide no additional information. Thus, although in theory the transient Newton approach is more accurate than the transient Gummel approach, for the relatively small time step used here, the improvement in accuracy was found to be equally small. The transient Newton I-V curve simulation was also not completed because it would have required an unreasonable amount of CPU time, as discussed in the following section.

6.5.5 Computational Efficiency

The simulations in Sections 6.5.3 and 6.5.4 showed that essentially identical I-V curves are produced for the RTD in Figure 6.4 by all four self-consistency iteration methods. It is reasonable in such a case to use the most efficient iteration method. Thus, the relative efficiencies of the iteration methods is another important point of comparison. As one can surmise from the foregoing discussions, the computational costs of the four iteration methods are vastly disparate. The number of WFTE solves and total CPU time used by each of the iteration methods for the 2-trace I-V curve is summarized in Table 6.2. Data for the non-self-consistent simulation shown in Figure 6.5 is also given for comparison. Also included are data for the standard steady-state Gummel implementation (see Section 6.3), for comparison to the accelerated implementation used in this work.

Table 6.2: Computational cost of self-consistency methods

Number of WFTE solves and total CPU time required for 2-trace I-V curve simulation for each self-consistency iteration method. Data is given for both the standard and accelerated steady-state Gummel approaches. The steady-state Newton simulation required several Gummel loops in some difficult cases. The transient Newton data is estimated. CPU times are for a DEC Alpha 3000/300 LX.

Simulation Type (Iteration Method)	WFTE Solves (*i.e.*, Iterations)	CPU time (hours)
Linear (non-self-consistent)	84	0.28
Steady-state Gummel (std)	4,300	14.3
Steady-state Gummel (acc)	1,450	5.0
Steady-state Newton	410N + 140G	7.2
Transient Gummel	96,500	330
Transient Newton	~96,500	~1,650

Some notes regarding the data in Table 6.2 are in order. The current was simulated at 0.01 V bias increments in both directions over the range 0.0 to 0.4 V, giving a total of 82 bias points plus the equilibrium solution needed for scattering calculations. The 140 steady-state Gummel iterations done during the course of the steady-state Newton simulation were a result of the Newton method's inability in some cases to locate the self-consis-

tent operating point as it entered or exited the plateau region. The transient simulations used 100 fs bias slewing (rather than changing the applied bias in a single time step) to mitigate the "shock" of bias changes and thus to minimize convergence time. The transient simulations further assume that the four oscillating operating points (0.24 V and 0.25 V in both trace directions) were terminated at 8,000 iterations, while all other bias points were run to full convergence. Since a complete transient Newton I-V curve simulation was not conducted, the data in Table 6.2 for this iteration method are estimates, but should be very close, based on the arguments at the end of the previous section.

Note that the simulations for this work were carried out on several platforms. The I-V curves for which data is reported in Table 6.2 were produced on independent processors of an SGI Challenge XL computer and on DEC Alpha 3000/300LX workstations. These platforms were roughly equivalent in performance, requiring about 12 CPU seconds per Gummel loop and 60 seconds per Newton loop. A Cray C-90 supercomputer was used for the longer, single-bias investigations (*e.g.*, the detailed investigations at 0.24 V and 0.25 V). The Cray required only 1.05 CPU seconds per Gummel loop.

Several factors determine the relative computational costs of the self-consistency iteration methods. Considering just the steady-state Gummel simulations, the importance of using the accelerated convergence implementation (see Section 6.3) is clear. In fact, the CPU time advantage of using Fermi-Dirac statistics is often even more dramatic than the roughly 3:1 ratio shown in Table 6.2. Outside the plateau region, the average number of iterations required for convergence to the self-consistent solution was 41 using the standard approach, but only 7 using the accelerated approach. However, for all iteration methods, most of the iterations took place in the challenging plateau region of the I-V curve.[5] For the accelerated Gummel simulation, locating operating points in the plateau often required dropping back to the more reliable standard approach. The result was only a 2.2:1 advantage in CPU time over the standard approach in the plateau region. With its faster convergence, the advantage of the accelerated Gummel implementation increases as convergence criteria become more strict.

A more general factor influencing the relative computational costs of the self-consistency iteration methods is the much greater CPU time required for a Newton loop than a Gummel loop. In this work, the ratio was 5:1. In spite of this, the full steady-state Newton

5. One result of this was that the up-trace always took more CPU time than the down-trace.

simulation required only 44% more CPU time than the accelerated steady-state Gummel simulation, and only half the time of the standard Gummel simulation. This recoup by the steady-state Newton method was a result of yet another factor in the efficiency equation: the Newton method's more sophisticated solution update algorithm (see Section 6.4), meaning that fewer iterations were required for convergence. In spite of the strict convergence criteria used, aside from the plateau region, almost all bias points required only 3 steady-state Newton iterations to meet these criteria.[6] Again, the faster convergence of the steady-state Newton approach improves its favorability in comparison to the steady-state Gummel approach as convergence criteria become more strict.

By far the most significant factor in the computational cost equation is whether the iteration method uses the steady-state or transient approach to finding the self-consistent operating point. The mathematical descriptions, and thus the CPU time per iteration, of the steady-state and transient methods are very similar for each formulation (Gummel or Newton). However, Table 6.2 shows that the transient iteration methods require roughly 2 orders of magnitude more iterations (on average) than the steady-state methods to converge to the self-consistent operating point. The reason for the huge difference is that the transient iteration methods attempt to follow the exact evolution of the device as it relaxes towards steady-state after a bias change, so they must take as long (in simulation time) as a real device would to reach steady-state. Because of the extreme computational cost of the transient iteration method, to complete the transient Gummel simulation in an acceptable amount of real time, several sections of each trace were executed concurrently, using a steady-state Gummel-converged solution for the initial condition (except for the two points on each trace which did not converge).

6.5.6 Discussion

This section discusses the strengths and weaknesses, in terms of efficiency, accuracy, and robustness, of the four self-consistency iteration methods considered in this work. From previous sections, the obvious strength of the steady-state methods is their relative computational efficiency. And as has also been stated, the main strength of the transient methods is their direct physical basis, and their resulting "exact" adherence to the time-

6. The relatively few number of iterations required by both steady-state iteration methods was made possible by the initialization algorithm for u^0, as discussed in Section 6.2.

dependent operation of the device being simulated. These are clearly complementary strengths, so that both the steady-state and transient approaches have important uses. In particular, a steady-state iteration method is recommended for wide-ranging initial investigations (*e.g.*, to trace the I-V curve), thereby gaining the insight necessary to narrow the focus of a more detailed investigation where transient effects are inherent (e.g., switching) or suspected (e.g., oscillations). Strangely, the literature is roughly equally divided between use of transient and steady-state Gummel approaches, with apparently no group simultaneously using the information and advantages provided by both. Hopefully this work will help to end that unnecessary exclusivity.

If a main strength of the steady-state methods is their relative efficiency, their main short-coming, at least in some cases, is accuracy. The inability of the steady-state iteration methods to show the transient oscillations predicted by the transient iteration methods was to be expected: only transient simulations can model time-dependent effects. Much more of a concern was the fact that the steady-state methods offered no concrete indication that an unstable operating condition existed, and thus that a transient simulation should be used. For the simulations in this work, if oscillations hadn't been expected in the plateau, one could easily have assumed that the steady-state simulations told the entire story about this RTD's I-V curve. Admittedly, the actual I-V curve was only slightly different at two points, but the physics underlying those small differences was quite important.

Another short-coming of the steady-state iteration methods is that convergence to a simultaneous solution of the steady-state WFTE and the PE can not be guaranteed. There are several potential causes of this lack of "robustness" or reliability. First, there are almost certainly "pathologic" operating conditions for some quantum devices where the steady-state methods will be unable to converge. Even if a device is stable at a given bias, the operating point may not be found if the previous WFTE and PE solutions are far away from it. Incrementing the bias across a bistable operating point, of which there are three in Figure 6.5, is the usual culprit here. Bistable operating points were, in fact, problematic for both the steady-state Newton method and the (accelerated) steady-state Gummel method. However, SQUADS detects non-convergent behavior during steady-state self-consistent simulations and automatically switches (temporarily) to the standard (and more robust but slower) steady-state Gummel approach. In this way, potential divergence problems of the steady-state iteration methods were completely avoided in this work.

Just as blind faith in the results of steady-state self-consistent simulations is not advisable, so too is complete reliance on transient self-consistent simulations. Admittedly, the basic transient methods are always adequate in terms of reliability and accuracy (*i.e.*, ability to correctly reproduce device physics). However, their extreme computational cost has some harsh consequences. The first is that one can not afford to undertake transient simulations such as those presented in this work without a good reason (and a very fast computer). The problem with this is that often there *is* no concrete reason *a priori* for running a simulation - only a vague notion of how the device might behave. Certainly it is currently completely unfeasible to run multiple week-long transient self-consistent I-V curve simulations to examine the effects of varying simulation or device parameters. In contrast, the decision to run the same steady-state simulations (in a few hours each) hardly merits a second thought.

The opposite side of the tendency for doing too few transient self-consistent simulations is trying to do too many. A good reason to limit reliance on transient simulation where appropriate is that inadequate computing resources invite unnecessary compromises to be made in the implementation of the simulator or in the execution of the simulation. For example, fewer bias points or time steps may be simulated than necessary, the time step or convergence criteria may be larger than accuracy dictates, and so on. One compromise made in this work that seems justified (as discussed in Section 6.5.4) was the use of the transient Gummel method instead of the theoretically more accurate Newton method. On the other hand, the choice of slew rate in this work, based solely on achieving fast convergence rather than modeling reality, is not so easily excused. In fact, investigations using a lower slew rate [21] show that transient current predictions like that in Figure 6.6 may bear little resemblance to what a real RTD would do under test. However, in a circuit of RTD-like devices, 100 fs bias slewing may be reasonable. Since generating the I-V curve was the test-case for this work, the details of the evolution to steady-state could be ignored here. In general, any compromises in implementation or execution should be considered carefully, so that they do not conspire to weaken the direct physical link which is the main advantage of the transient iteration methods over the steady-state approaches. The best defense against these compromises is to focus computing resources on a limited set of transient simulations that are expected to add value to the steady-state results.

The arguments above have advocated using the various self-consistency iteration

methods in a hierarchical manner. An efficient steady-state approach should be used to investigate a broad range of operating conditions, and to narrow the scope for more exacting (and expensive) transient simulations. To implement this approach, one must find clues in steady-state simulations that indicate device operating conditions for which transient simulation might be warranted (*i.e.*, where sustained, significant, or interesting transient effects might occur). Some of these clues are obvious. A negative differential resistance region is a known cause of oscillations, whether intrinsic to the device or a result of the device interacting with the (simulated or real) measurement apparatus or external circuit. Also, any operating point at which the steady-state simulation has significant difficulty converging should raise a red flag. Obviously, if the steady-state iteration method completely fails to converge at a particular bias point, a transient simulation is necessary to determine device operation. Of course, only a transient iteration method can be used for inherently transient self-consistent simulations, such as switching, small-signal, or large-signal investigations.

Finally, the relevance of the discussion and conclusions in this section to the simulation of conventional electronic devices will be assessed. Certainly, the goals of conventional device simulation are identical: to achieve reliably accurate simulations at the least computational cost. Further, the relative costs of the various iteration methods (Gummel or Newton, steady-state or transient) are essentially the same, whether they are used to solve the WFTE for quantum device simulation or, for example, a three-dimensional drift-diffusion equation for conventional device simulation. Thus, in conventional device simulation, steady-state methods will be less costly than transient methods, so a combination of the two should be used to maximize the utility of available computing resources. However, the conclusion that the steady-state Gummel method is preferred for its efficiency and adequate reliability over the Newton method for quantum device simulation is not shared in conventional device simulation. As discussed in [22], when the interaction between the two carrier types is strong (with high concentrations of both carriers in a given region, leading to recombination, generation, scattering, and charge interactions), the Newton method is essential for locating the self-consistent operating point. As long as it is accurate to characterize a quantum device as a single energy band (or multiple non-interacting bands), simulating quantum devices will be simpler in this sense than their classical counterparts.

6.5.7 Other Iteration Methods

As a final note, the Newton and Gummel methods presented above are certainly not the only possible ways to solve the WFTE - PE system and thereby implement self-consistency, although they are perhaps the most basic. Many variations on the Gummel and Newton methods are possible [22], and other non-linear system solving approaches may be used. For example, Jansen *et al.*[23] used the conjugate-gradient (CG) method to compute the self-consistent I-V curve for an RTD. According to their analysis, the CG method is about an order of magnitude faster than the transient Gummel approach, making it about an order of magnitude slower than the steady-state Gummel and Newton iteration methods described herein. However, the CG method has the distinct advantage of a much smaller memory footprint. This would be useful for WFM simulations which are otherwise too large for available hardware. Due to the CG method's limitation to steady-state simulation, its lower speed than the steady-state methods implemented in SQUADS, and the fact that memory usage for solving the WFTE - PE system has been highly optimized in SQUADS, the CG method has not been implemented in SQUADS.

6.6 Self-Consistency and the TMM

6.6.1 Introduction

Up to this point, this chapter has described the implementation and simulation results of self-consistency in the Wigner function method of quantum device simulation. Self-consistency can also be implemented in the transfer matrix method. The main requirement for implementing self-consistency is the ability to calculate the carrier density profile from the quantum calculation. In the TMM, the carrier density is computed as a sum over all of the wavefunctions for electron (or hole) beams from each contact of the device, as described in Section 4.3.4. This section describes in more detail how self-consistency is implemented for TMM simulations, and it also presents the first self-consistent TMM simulations of this dissertation.

The main difference between the WFM and the TMM in terms of self-consistency is related to the nature of their respective state functions. The Wigner function is an *aggregate* state function, containing all of the information about carriers in the system at a given time. In particular, the carrier density is calculated directly from the Wigner function, This

makes the use of the Newton iteration method possible, since the off-diagonal Newton blocks are based on finding such a direct relationship. In contrast, there is *not* a single state function in the TMM, but rather one wavefunction for each energy beam (typically 1000 - 10,000) incident from either contact. During a TMM calculation, one calculates, uses, and discards the amplitudes for each wavefunction, since storing amplitudes for all wavefunctions at every position node would take away the main advantage of the TMM: computational efficiency. But unless *all* wavefunction amplitudes are unknowns in a simultaneous system of equations, a Newton iteration approach to implementing self-consistency in the TMM is impossible. However, a Gummel iteration for self-consistency is still possible, where the Schrödinger equation is solved as usual, and the Poisson equation is updated afterwards. Since the TMM is inherently time-independent, a transient Gummel approach is also not possible, leaving only the steady-state Gummel approach.

6.6.2 Implementation in SQUADS

The steady-state Gummel approach to enforcing self-consistency has been used in TMM simulation by many researchers [24-29] over the past decade and more. The steady-state Gummel approach for the WFM is well described in Section 6.3. This section describes the few changes in the implementation for the TMM. Actually, there is only one significant difference: the calculation of the carrier density as input for the solution of the Poisson equation. Of course, the WFM calculates the carrier density profile from the Wigner function as described in Section 6.2. The standard TMM approach calculates the carrier density profile as described in Section 4.3.4.

The standard TMM calculation does not allow the inclusion of scattering. However, in most real devices, scattering plays an important role in producing the self-consistent energy bands and resulting device operation, as will be illustrated in Section 6.6.3 and Chapter 8. Fortunately, it is possible to approximately incorporate scattering in self-consistent TMM simulations. The “trick” is to use the classical, equilibrium relationship (6.13) to calculate the carrier profile from the energy band profile. More precisely, the classical carrier density calculation is used in the contact regions up to the tunnel barriers, and zero carrier density is assumed in the remaining “active” device region. The classical carrier density inherently includes scattering. Using this carrier density profile requires only minor modifications in setting up the tri-diagonal Poisson equation matrix and RHS

vector of Figure 6.2 (mainly, ignoring the carrier density/feedback terms for position nodes corresponding to the active region). This approach for including scattering in self-consistent TMM simulations gives better accuracy in cases where scattering is high (*e.g.*, high temperature), but it excludes some quantum effects, and it incorporates device-specific code.

6.6.3 Results and Discussion

For comparison to the WFM simulations in Section 6.5, the TMM simulations in this section use the same RTD device structure (see Figure 6.5). Figure 6.11 shows the TMM simulated I-V curve for this RTD for three cases: non-self-consistent (linear potential drop across center 17 nm of RTD, hereafter called TMM0), self-consistent with standard TMM carrier profile computation (hereafter called TMM1), and self-consistent using the classical carrier density in the contact regions (hereafter called TMM2).

Several points are noteworthy in Figure 6.11. First, just as in the WFM simulation (see

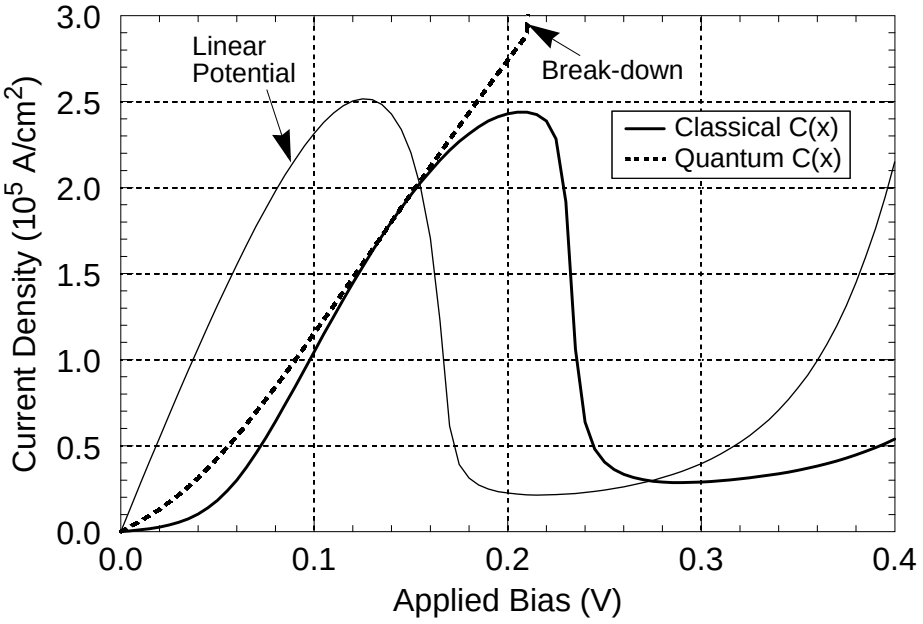


Figure 6.11: Self-consistent TMM I-V curve simulation of RTD

Shown along with the standard TMM simulation (dashed curve) are the non-self-consistent TMM result for a linear potential drop in the active region (thin solid curve) and the self-consistent TMM simulation using the classical carrier density to implement self-consistency (thick solid curve). The standard TMM simulation (with quantum carrier density and no scattering) breaks from expected operation after resonance, where current increases roughly by a factor of 10.

Figure 6.5), enforcing self-consistency moves the peak and valley conditions to higher biases. This is due to the quantum well state not being pulled down (w.r.t. the emitter band minimum) as quickly with increasing applied bias in the self-consistent case since a significant portion of the applied bias is across the collector depletion layer. In fact, both TMM0 and TMM1 gave virtually the same peak and valley applied biases as the analogous WFM simulations. However, currents are a factor of 2 - 3 lower in the TMM I-V curves than in the WFM I-V curves (a factor of 4 when the WFM doesn't include scattering). Test simulations show that this last observation can be traced mainly to inaccuracies in the WFM simulation, as discussed in Section 5.5.7.

Also note that the interesting RTD behavior (I-V plateau, hysteresis, and bistability) seen in the self-consistent WFM I-V curve are not seen here, even in the self-consistent TMM1. The behavior in TMM1 is more in line with expectations, as discussed in Chapter 8. However, the TMM2 simulation (*i.e.*, using the quantum calculation of carrier density) has its own interesting behavior. Before the peak (resonance) condition, the TMM1 and TMM2 follow roughly the same behavior, and the self-consistent energy bands and carrier density profiles are very similar, as indicated by Figure 6.12.

However, in TMM2, when the quantum well state is pulled below the emitter band minimum (*i.e.*, at applied biases greater than 0.21 V), the RTD suffers some kind of breakdown. That is, current increases by about a factor of 10, rather than decreasing by this factor, as in the other two TMM simulations. The cause of this current rise is indicated in Figure 6.13, which shows the carrier density and energy band profiles for TMM2 at 0.22 V applied bias (*i.e.*, just after resonance). Because the standard TMM simulation doesn't include scattering, it is difficult to form the strong accumulation layer in the emitter required to accommodate large biases in the normal way (i.e., with an electric field between the accumulation and depletion layers). The only way for the bias to be accommodated is by depleting the emitter and charging the emitter contact, which situation is indicated in Figure 6.13.

Both TMM1 and TMM2 have drawbacks. TMM1 ignores all quantum effects in the calculation of the carrier profile. As a result, TMM1 can not, for example, simulate intrinsic bistability due to charge storage in the quantum well. On the other hand, TMM2 includes no scattering, making it susceptible to erroneous results at high biases. A simple compromise is to use a classical carrier density calculation in the contact regions and a

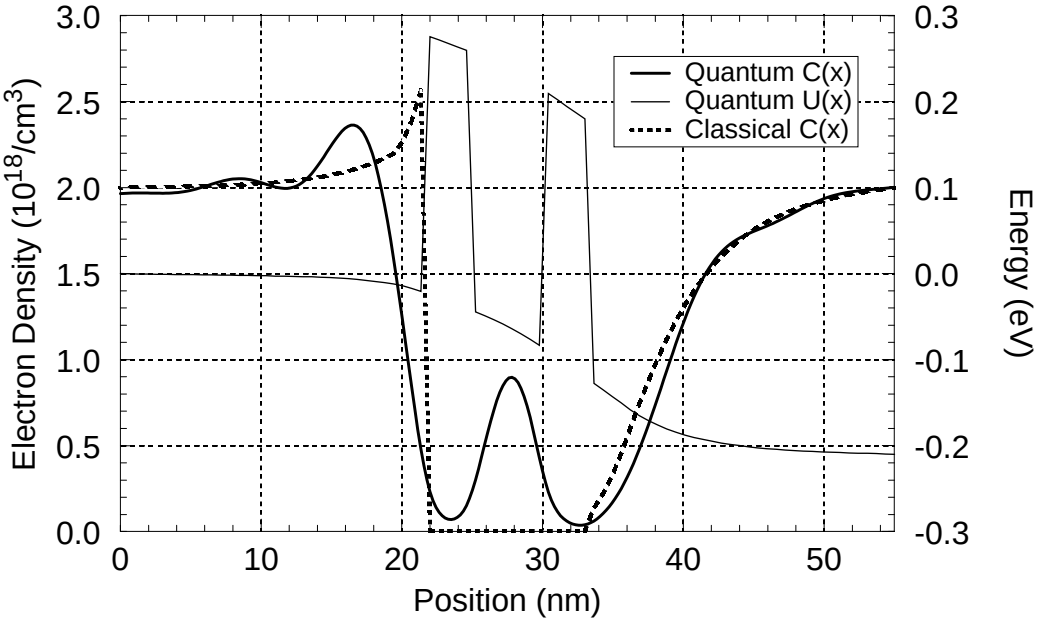


Figure 6.12: Self-consistent TMM-simulated RTD at peak current

At 0.21 V applied bias, both the standard (quantum) and classical carrier density profiles are shown, as well as the energy band profile from the quantum calculation. In spite of the marked difference in carrier density in the quantum well, the energy bands for the two carrier densities are very similar. [Energy bands for classical C(x) simulation are not shown.]

quantum calculation in the "active" (barrier and quantum well) region (see Figure 6.12). This hybrid carrier density model, hereafter called TMM3, is also implemented in SQUADS,[7] and has been implemented by at least one other group [24]. Figure 6.14 shows a plot of the TMM3 I-V curve simulation of the RTD used above. Note that bistability due to charge storage in the quantum well at resonance is predicted by this simulation. Also shown in Figure 6.14 is the I-V curve from Figure 6.11 using the classical charge density model. The two I-V curves differ according to the amount of active region charge in the hybrid case (active region charge is taken to be zero in the classical model simulation).

Other models for the carrier density in a self-consistent TMM simulation are certainly possible, but the lack of an accurate means to include scattering is probably the most serious limitation in self-consistent TMM simulation. Note that if scattering were negligible in an RTD, and the contacts were accurately modeled as ideal ohmic contacts to metal

7. Due to the shared code structure of SQUADS, both the classical and hybrid classical-quantum carrier density calculations can be used in self-consistent WFM simulations as well. However, since the WFM includes scattering in a more accurate way, the reason for using this capability in general is not obvious, in contrast to their manifest usefulness in TMM simulations.

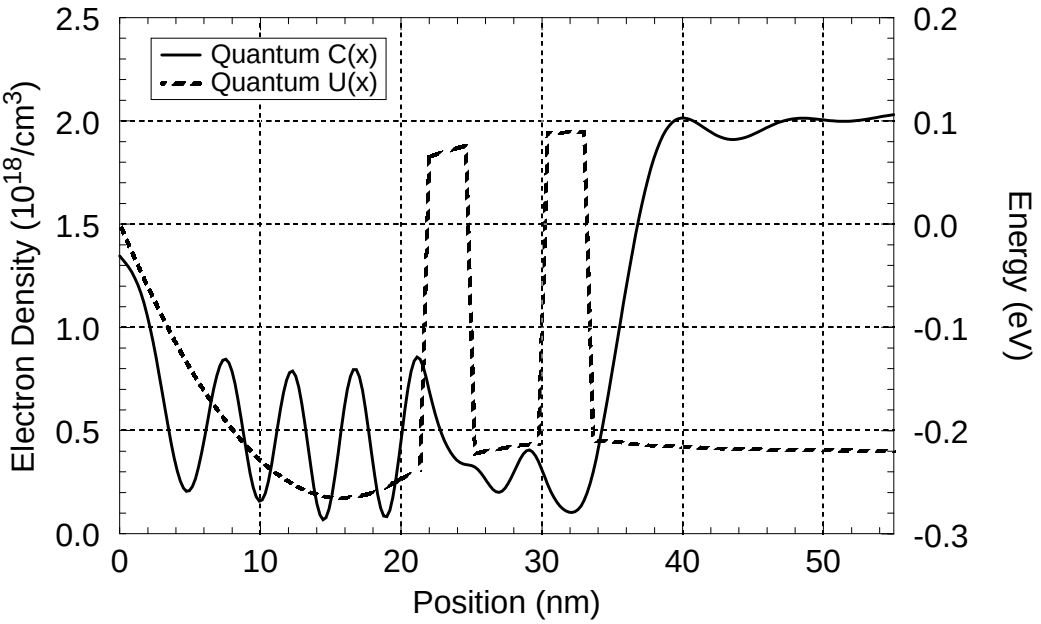


Figure 6.13: Self-consistent TMM-simulated RTD after peak current

When the quantum well state drops below the emitter band minimum, the TMM simulation using the quantum carrier density calculation gives unphysical behavior, indicated by a strongly charged emitter contact. Since scattering is not included, it is not possible to form an accumulation layer in the emitter. Instead, the applied bias is accomodated by a depleted emitter and charged emitter contact.

electrodes, the unexpected energy band profile in Figure 6.13 would occur experimentally, producing the very high current operation simulated in TMM2. However, this behavior has not been observed experimentally, suggesting again that including scattering is important for accurate RTD simulation.

6.7 Summary

This chapter reviewed the theory and numerical implementation of four basic approaches to implementing self-consistency in the Wigner function approach to quantum device simulation. These approaches include steady-state and transient Gummel, and steady-state and transient Newton. This is the first time that all of these approaches have been described in a single mathematical framework and notation. In the process of describing the numerical implementations of these iteration methods, expressions for the off-diagonal Jacobian blocks in the Newton formulation were given, apparently for the first time. Also, an accelerated convergence algorithm was described for the steady-state

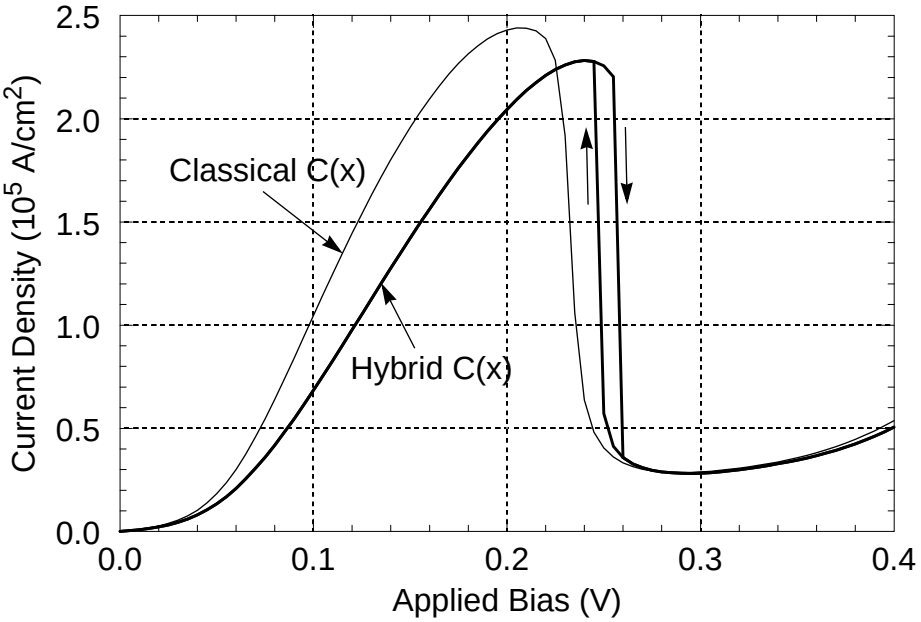


Figure 6.14: Self-consistent I-V curve for RTD with hybrid carrier ddensity

This TMM simulation uses the classical carrier density C(x) in the contact regions of the RTD and the quantum value in the active region. The resulting I-V curve shows bistability and hysteresis, due to significant energy band adjustments when the quantum well charges or discharges in transition between peak and valley operation. This I-V curve diverges from that using the classical C(x) according to the amount of quantum well charge.

Gummel approach which makes it the most efficient means of generating the self-consistent I-V curve for an RTD.

The strengths and weaknesses of the various self-consistency iteration methods were also reviewed. A large part of that analysis concerned relative computational costs. The computational efficiency of the steady-state methods makes them ideal for wide-ranging initial investigations, such as full I-V curve traces. There are undeniable difficulties in using the steady-state iteration methods, such as lack of robustness in the Newton and accelerated Gummel methods, and the relatively slow convergence of the standard Gummel approach. These problems may have discouraged the use of steady-state approaches in the past. This work has demonstrated how these problems can be avoided, and it has shown the excellent results and efficiencies that the steady-state iteration methods can achieve.

This work showed that even if a steady-state iteration method converges to a simultaneous solution of the steady-state WFTE and PE, there is no guarantee that this is a stable

operating point. Transient iteration methods are inherently more accurate and reliable, and are required to treat time-dependent situations (such as unstable oscillations). However, steady-state methods are just as important *in practice* in the investigation of quantum device physics. Efficient steady-state simulations can be used to determine the basic operation of the device (*e.g.*, I-V curve, possible unstable regions), allowing one to narrow the scope of (expensive) transient simulations. Those transient simulations which *are* done can then be implemented and executed without serious compromises so that they will correctly model device physics and add value to the steady-state results.

Finally, the implementation of self-consistency in the transfer-matrix method of quantum device simulation was described and demonstrated. Due to the nature of the TMM, only the steady-state Gummel iteration approach is suitable for enforcing self-=consistency in this case. Implementing this self-consistency iteration in the TMM is almost identical to that in the WFM, the main exception being the calculation of the carrier density. In fact, several different charge density models are implemented for self-consistent TMM simulations in SQUADS. The standard model does not include scattering, and calculates carrier density in the usual TMM manner. In contrast, the classical model includes scattering in the electrodes, while the hybrid classical/quantum model also includes the TMM-calculated carrier density in the active region of the device. A significant result was the conclusion that scattering is required to produce accurate self-consistent TMM simulation results.

References

[1] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer. “Self-consistent study of the resonant-tunneling diode.” *Physical Review B*, 39(11):7720–7735, 1989.

[2] K. K. Gullapalli, D. R. Miller, and D. P. Neikirk. “Simulation of quantum transport in memory-switching double-barrier quantum-well diodes.” *Physical Review B*, 49(4):2622–2628, 1994.

[3] H. Tsuchiya, M. Ogawa, and T. Miyoshi. “Simulation of quantum transport in quantum devices with spatially varying effective mass.” *IEEE Transactions on Electron Devices*, 38(6):1246–1252, 1991.

[4] W. R. Frensley. “Quantum kinetic theory of tunneling devices.” In K. Hess,

J. Leburton, and U. Ravaioli, editors, *Computational Electronics: Semiconductor Transport and Device Simulation*, pages 195–200, Boston, 1991. Kluwer Academic Publishers.

[5] K. L. Jensen and F. A. Buot. “Numerical simulation of intrinsic bistability and high-frequency current oscillations in resonant tunneling structures.” *Physical Review Letters*, 66(8):1078–1081, 1991.

[6] L. L. Chang and R. Tsu. “Resonant tunneling in semiconductor double barriers.” *Applied Physics Letters*, 24(12):593–595, 1974.

[7] E. R. Brown, W. D. Goodhue, and T. C. L. G. Sollner. “Fundamental oscillations up to 200 GHz in resonant tunneling diodes and new estimates of their maximum oscillation frequency from stationary-state tunneling theory.” *Journal of Applied Physics*, 64(3):1519–1529, 1988.

[8] B. Ricco and M. Y. Azbel. “Physics of resonant tunneling. the one-dimensional double-barrier case.” *Physical Review B*, 29(4):1970–1981, 1984.

[9] K. L. Jensen and F. A. Buot. “The methodology of simulating particle trajectories through tunneling structures using a Wigner distribution approach.” *IEEE Transactions on Electron Devices*, 38(10):2337–2347, 1991.

[10] B. A. Biegel. *SQUADS Technical Reference*. (Unpublished), Stanford University, 1996.

[11] M. R. Pinto. *Comprehensive semiconductor device simulation for silicon ULSI*. PhD thesis, Stanford University, Aug. 1990. p. 371.

[12] H. K. Gummel. “A self-consistent iterative scheme for one-dimensional steady-state transistor calculations.” *IEEE Transactions on Electron Devices*, 11(10):455–465, 1964.

[13] F. A. Buot and K. L. Jensen. “Lattice Weyl-Wigner formulation of exact many-body quantum-transport theory and applications to novel solid-state quantum-based devices.” *Physical Review B*, 42(15):9429–9457, 1990.

[14] W. B. Joyce and R. W. Dixon. “Analytic approximations for the fermi energy of an ideal fermi gas.” *Applied Physics Letters*, 31(5):354–356, 1977.

[15] M. R. Pinto. “Comprehensive semiconductor device simulation for silicon ULSI,” Aug. 1990. p. 304.

[16] W. R. Frensley. “Boundary conditions for open quantum systems driven far from

equilibrium." *Reviews of Modern Physics*, 62(3):745–791, 1990.

[17] K. L. Jensen and F. A. Buot. "The effects of scattering on current-voltage characteristics, transient response, and particle trajectories in the numerical simulation of resonant tunneling diodes." *Journal of Applied Physics*, 67(12):7602–7607, 1990.

[18] F. A. Buot and A. K. Rajagopal. "High-frequency behavior of quantum-based devices: Equivalent-circuit, nonperturbative-response, and phase-space analyses." *Physical Review B*, 48(23):17217–17232, 1993.

[19] F. A. Buot and A. K. Rajagopal. "Theory of novel nonlinear quantum transport effects in resonant tunneling structures." *Materials Science and Engineering*, B35(1-3):303–317, 1995.

[20] F. A. Buot and K. L. Jensen. "Intrinsic high-frequency oscillations and equivalent circuit model in the negative differential resistance region of resonant tunneling devices." *COMPEL*, 10(4):241–253, 1991.

[21] B. A. Biegel and J. D. Plummer. "Applied bias slewing in transient wigner function simulation of resonant tunneling diodes." *IEEE Transactions on Electron Devices*, 1997. (to be published).

[22] M. R. Pinto. *Comprehensive semiconductor device simulation for silicon ULSI*. PhD thesis, Stanford University, Aug. 1990. Ch. 5.

[23] R. E. Jansen, B. Farid, and M. J. Kelly. "The steady-state self-consistent solution to the nonlinear Wigner-function equation; a new approach." *Physica B*, 175:49–53, 1991.

[24] H. Ohnishi, T. Inata, S. Muto, N. Yokoyama, and A. Shibatomi. "Self-consistent analysis of resonant tunneling current." *Applied Physics Letters*, 49(19):1248–1250, 1986.

[25] M. Cahay, M. McLennan, S. Datta, and M. S. Lundstrom. "Importance of space-charge effects in resonant tunneling devices." *Applied Physics Letters*, 50(10):612–614, 1987.

[26] B. Zimmermann, E. Marclay, M. Ilegems, and P. Gueret. "Self-consistent calculations of tunneling currents in n+ ga." *Journal of Applied Physics*, pages 36–44, Oct. 1991.

[27] K. K. Gullapalli, A. J. Tsao, and D. P. Neikirk. "Multiple self-consistent solutions at zero bias and multiple conduction curves in quantum tunneling diodes incorpo-

rating N-N+N- spacer layers." *Applied Physics Letters*, 62(23):2971–2973, 1993.

[28] Y. W. Choi and C. R. Wie. "Increased peak current in alas/gaas resonant tunneling structures with gainas emitter spacer." *Journal of Applied Physics*, 71(4):1853–1859, 1992.

[29] Z. Ikonic, V. Milanovic, and D. Tjapkin. "Resonant second harmonic generation by a semiconductor quantum well in electric field." *IEEE Journal of Quantum ELectronics*, 25(1):54–60, 1989.

Chapter 7

Applied Bias Slew Rate

Quantum electronic devices have been proposed as a possible successor to conventional electronic devices in part because the inherently small size of quantum devices makes it possible for them to operate at very high speeds. The transient WFM capability in SQUADS can be used to investigate the transient operation and speed potential of quantum devices. Unfortunately, a typical transient WFM simulation might require thousands of time-stepped solutions of the WFTE. The result is either a very high computational cost or compromises in accuracy (*e.g.*, due to using longer time steps) in an effort to reduce computation time. This chapter investigates a technique for transient WFM simulation that improves both accuracy *and* efficiency: using a finite applied bias slew-rate (the rate at which the applied bias is changed with respect to time). Except for the investigations in Chapter 6, all transient Wigner function simulations of quantum devices to date have used instantaneous changes in the applied bias. Such switching could never be achieved in physical systems. This investigation found that instantaneous switching produces significantly inaccurate quantum device simulations.

To introduce this slew rate investigation, Section 7.1 provides an overview of the physics of transient bias changes in electronic device simulation. Section 7.2 then presents the investigation of whether and to what extent quantum device operation and WFM simulation cost are affected by variations in applied bias slew rate.

7.1 Physics of Transient Bias Changes

Before discussing slew rates directly, the physics of transient bias switching in electronic device simulation will be considered in some detail [1]. In a transient simulation, applied bias changes are completed during a time step (*i.e.*, between consecutive solutions of the system). Changing the applied bias across a device requires a current pulse in the external circuit, essentially "communicating" the new bias to the device. If the applied bias is changed too quickly for free carriers in the device to respond to this current pulse, the bias change is effectively instantaneous, and the external current simply charges the contacts, producing an electrostatic field across the device. In other words, the device acts like a capacitor. For a one-dimensional "capacitor" of area A, width L, and permittivity ε, the external current density J_{ext} necessary to cause a voltage change ΔV in time Δt is

$$J_{\text{ext}} = \frac{I_{\text{ext}}}{A} = \frac{\Delta Q}{A\Delta t} = \frac{C\Delta V}{A\Delta t} = \frac{\varepsilon \Delta V}{L\Delta t}. \tag{7.1}$$

J_{ext} occurs entirely within the time step Δt. Following a bias change, free carriers will respond over time to the electric field as they redistribute, enter, and leave the device to accommodate the changed applied bias.

To maintain physical correctness, SQUADS implements the above described behavior in self-consistent, transient simulations. That is, because time steps are typically on the order of 1 fs, applied bias changes of any magnitude during a time step are effectively instantaneous, and therefore initially appear as electrostatic fields across the entire device. SQUADS computes and prints the average external current J_{ext} required to produce the specified bias change in a single time step, although this current does not appear in any of the simulated internal device currents. Enforcing self-consistency through the Poisson equation naturally causes free carriers in the device to respond appropriately over time to the electric field between the device contacts. Note that slewing the applied bias is accomplished simply by making many small, "instantaneous" bias changes during consecutive small time steps.[1] Given the above description of how transient bias changes should be implemented in a device simulator, the investigation in Section 7.2 demonstrates the significance of the choice of applied bias slew rate in transient Wigner function simulations.

1. In this dissertation, "instantaneous" means "single time step".

7.2 Effect of Slew Rate Variation

7.2.1 Simulated Device and Operation Summary

Resonant tunneling diodes (RTDs) are an excellent quantum device simulation test-bed, for reasons discussed in Section 2.3.4, and will be used in this slew rate investigation. In particular, the RTD used in Chapter 6 (see Figure 6.4) is again used here because of the strong transient effects it displays.[2] Figure 7.1 (a refinement of Figure 6.5) shows the steady-state self-consistent I-V curve for this RTD as simulated by SQUADS. Transient simulations in Section 6.5.4 showed that this RTD is stable at all points on this I-V curve except in the plateau (0.239 V - 0.313 V on the up-trace and 0.254 V - 0.239 V on the down-trace). Where the two traces coincide in the plateau (0.239 V - 0.254 V), perpetual high-frequency (~2.5 THz) current oscillations occur. In the remainder of the plateau on the up-trace, the RTD is only marginally stable (it approaches steady-state in a weakly-damped oscillatory fashion). Since the most interesting transient phenomena occur in the plateau region of the I-V curve, transient simulations of this region offer a very effective means of analyzing the effects and importance of slew rate variation.

All transient simulations in this work used the Cayley transient operator (see Section 5.3.3.5) with a 1 fs time step. The Gummel iteration method (see Section 6.3) was used to implement self-consistency. For I-V curve simulations, operating points were taken every 10 mV. Thee convergence criteria (see Section 6.5.2) used to determine when steady-state had effectively been reached after applied bias changes were: potential change less than 10^{-6} eV, Poisson equation satisfied to less than 10^{-8} eV, and current variation across the device of less than 1000 A/cm^2.

7.2.2 Instantaneous Bias Switching

To determine the effects of slew rate on simulation results, a transient I-V curve simulation was conducted using the standard approach of instantaneous bias switching. Thus, starting from a steady-state solution at one bias, the applied bias was changed to the next bias point in a single time step, and the system was allowed to evolve to steady-state. After each bias switch, a large current pulse of about 1.5×10^5 A/cm^2 peak amplitude and about

2. Obviously, the useful investigation of a transient simulation technique, in this case applied bias slewing, requires a device exhibiting significant transient effects.

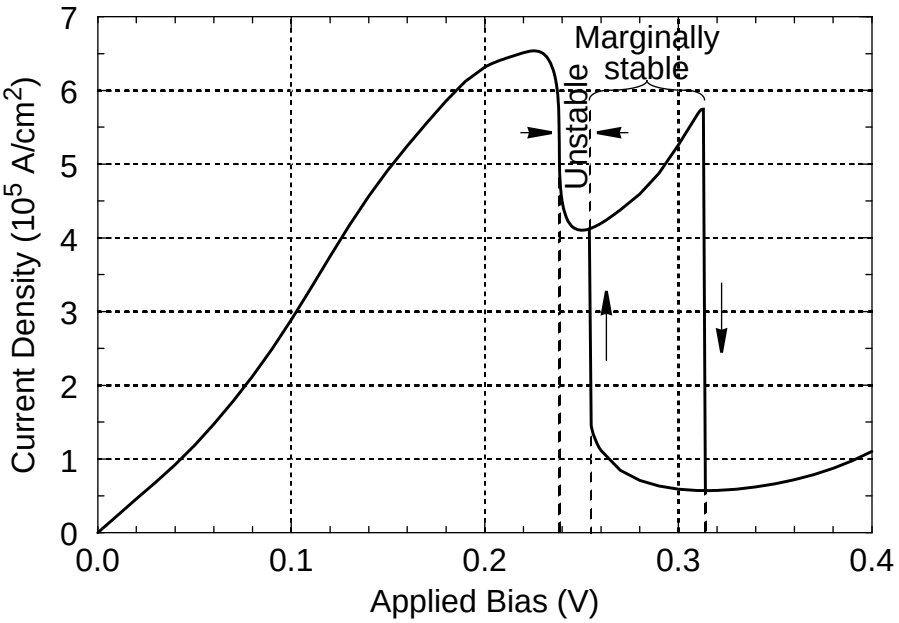


Figure 7.1: Self-consistent, steady-state RTD I-V curve

Effects shown include negative differential resistance, hysteresis, bistability, and a plateau in the I-V curve. The RTD is unstable (oscillates perpetually) in the plateau between 0.239 V and 0.254 V, and it is marginally stable (oscillates with slow damping) in the remainder of the plateau.

50 fs duration occurred. The amplitude of this current pulse often exceeded both the starting and ending currents. For example, Figure 7.2 shows the transient position-averaged current and the collector contact current in the RTD after instantaneous switching from 0 V to 10 mV.[3] Note that the peak current is 7-8 times the final steady-state value. During the down-trace, the current pulse was negative, but with essentially the same amplitude and duration. Based on the consistency of the current pulse in amplitude and duration throughout the I-V curve trace (except at bistable points), simple computations [2] determined that the pulse resulted from charging of the accumulation and depletion layers to accommodate the 10 mV change in applied bias between bias points.

The origin of the current pulse described above has been the source of some consternation in the past. For example, Tsuchiya et al. [3] attempted to explain the current pulse in a transient Wigner function simulation after a bias switch across the negative differential resistance (NDR) region of the I-V curve in terms of the discharging of the quantum well

3. Hereafter, all currents will be position-averaged values, since this is the current induced in the external circuit.

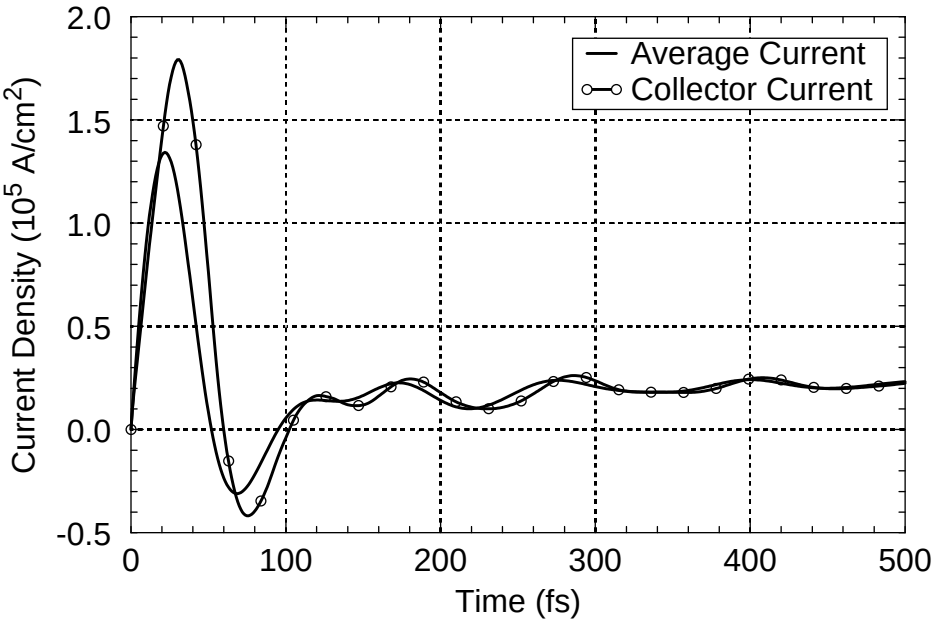


Figure 7.2: Transient current after instantaneous 10 mV bias change

The position-averaged current (plain curve) and collector contact current (circle curve) are shown after an instantaneous bias change from 0.0 V to 10 mV. Note that the peak of the current pulse is 7-8 times the final value.

and the properties of the electrodes (emitter and collector layers). Similarly, Kluksdahl et al. [4] suggested that "The overshoot probably arises from a rapid discharge of the trapped charge in the potential well". Simple calculations [2] show that the quantum well charge in these cases was much too small to produce the observed current pulse, while the required change in accumulation and depletion charge was about right. Use of a finite slew rate in these instances (or switching the bias outside the NDR region) would have shown that the current pulse was largely due to the charging of the accumulation and depletion layers. Thus, instantaneous bias switching in transient Wigner function simulations may obscure device operation to the extent that incorrect conclusions are drawn.[4]

Much worse than the internal current pulse from a practical standpoint is the external circuit current J_{ext} required to change the bias by 10 mV in a single time step (although

4. Note that in the earlier work of Frensley [5], non-self-consistent, transient Wigner function simulations also showed a current pulse after switching an RTD across the NDR region. However, in this case Frensley's conclusion that the pulse was due to the charging or discharging of the quantum well was correct. Since Frensley did not enforce self-consistency, there would be no accumulation and depletion charges.

J_{ext} does not appear in any simulation results). Using (7.1), with L = 55 nm, $\varepsilon = 12.9\varepsilon_o$, ΔV = 10 mV, and Δt = 1 fs, the external current density is $J_{\text{ext}} = 2.1\times10^6$ A/cm^2 for this RTD. This is at least three times larger than any current the external circuit must supply for steady-state device operation anywhere in the simulated bias range (see Figure 7.1). In summary, the use of instantaneous bias switching in self-consistent quantum device simulation produces a huge current pulse within the device, and it would require an even larger current spike from the driving source. Neither of these represent practical quantum device behavior in real measurement or circuit environments.

7.2.3 Realistic Slew Rates

The above simulations show that instantaneous bias switching in transient Wigner function simulations presents a huge “shock” to the quantum device, resulting in a large internal current pulse and damped oscillations thereafter, as shown in Figure 7.2. To more accurately model the operation of real quantum devices and circuits, a finite applied bias slew rate must be used. Based on the simulation results above, RTD-type devices can respond and change state in about 100 fs, so 100 fs bias slewing seems appropriate for studying how an RTD might operate in a circuit of its peers. A second transient I-V curve simulation was therefore conducted with a slew rate of 10 mV/100 fs (100 V/ns). As an example, Figure 7.3 shows the transient current for the slewed switch from 0 V to 10 mV, along with the same plot for instantaneous switching. Note that the accumulation/depletion charging current pulse (which must have the same integral over time, or total charge transfer) of $J_{DA} \approx 4\times10^4$ A/cm^2 is much less severe with slewed switching. Also, (7.1) gives an external current of only $J_{\text{ext}} = 2.1\times10^4$ A/cm^2. Neither of these are large compared to normal operating currents of the device, which confirms that this slew rate could reasonably occur in a quantum circuit.

The use of 100 V/ns slewed switching in transient Wigner function simulations, while improving the accuracy of the simulation, had the ancillary benefit of reducing its very high computational cost. The shock of instantaneous bias switching required a relatively long transient simulation before the convergence criteria were satisfied (*i.e.*, steady-state was reached). Slewed switching lessened the shock, so that, although reaching the next bias took longer, total time to steady-state was much less. For example, for the 0 V to 10 mV switching simulation shown in Figure 7.3, steady-state was reached in 330 fs with

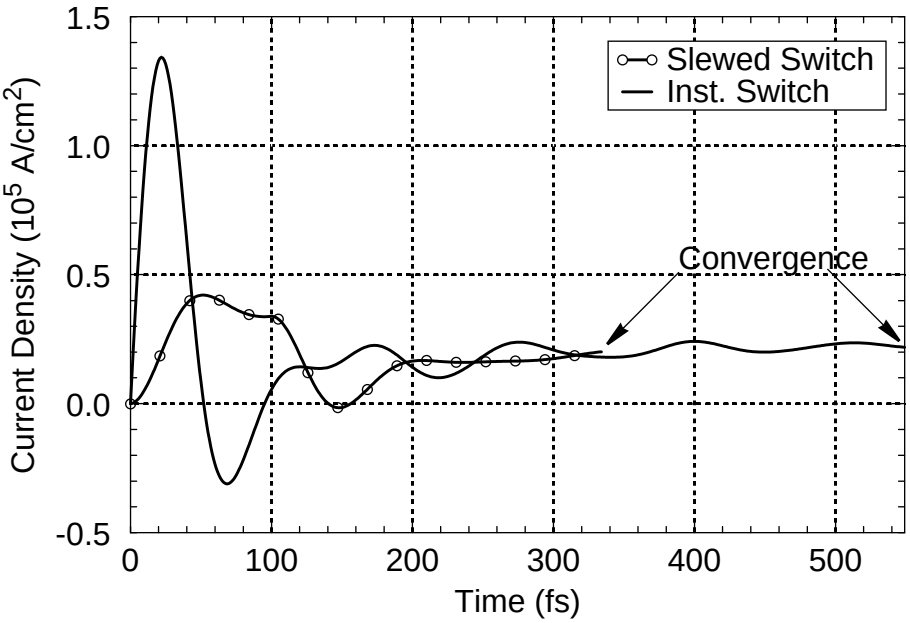


Figure 7.3: Transient current after slewed 10 mV bias change

Transient current after switching from 0.0 V to 10 mV. The plain curve shows the transient current when the bias is switched instantaneously; the circle-curve shows the same when the bias is slewed from 0.0 V up to 10 mV over 100 fs. Note that the slewed switching simulation reaches steady-state significantly faster than the instantaneous switching simulation.

slewing, versus 550 fs without. On average, convergence was reached about 20% faster with slewed switching, even in the critical plateau region (where thousands, rather than hundreds, of femtoseconds were required for convergence). Needless to say, a 20% improvement in computation time is very significant in a several-hundred hour simulation task.[5]

The 100 V/ns slew rate was chosen to model an RTD driven by an equally fast device. If simulation results are to be compared to RTD measurements by a device tester, an even lower slew rate is appropriate. Fast operational amplifiers (which might serve as the front end for a device tester) are capable of perhaps 2000 V/μs slew rates [6]. Taking 1 V/ns as potentially feasible value, this is a factor of 100 lower than the slew rate used above, and would require 10 ps (10,000 time steps) to change the applied bias by 10 mV. The question is, is it necessary to go to such a huge expense in order to simulate experimental device

5. Note that if transient effects are not specifically of interest, the more computationally efficient steady-state simulation methods can be used, as discussed in Chapter 6.

test conditions (*e.g.*, to test simulator accuracy)? In other words, given the instantaneous switching and 100 V/ns slewing simulation results, can lower slew rate effects be estimated by extrapolation or even neglected (*i.e.*, steady-state operation assumed)? In this case, the answer is no. Simulation results in the following section indicate that when the details of *this* device's operation are being investigated, there are cases where even a 1 V/ns slew rate is too high.

7.2.4 Intrinsic Oscillations

As stated earlier, the most interesting transient effects for the chosen RTD occurred in the plateau region of its I-V curve. The effects of slew rate variation in this region of operation were therefore investigated in more detail. In tracing the I-V curve in the plateau, both instantaneous and 100 V/ns slewed switching initiated oscillations that persisted for thousands of femtoseconds (at 1 fs per time step). These oscillations were so persistent that Jensen and Buot [7] concluded that the RTD oscillated perpetually at all biases in the plateau, and that these oscillations were necessary for the plateau's existence. However, Section 6.5.4 showed that this RTD is only truly unstable in the plateau between 0.24 V and 0.25 V.[6] Above this range in the plateau, the oscillations eventually decayed to steady-state. For example, Figure 7.5 shows the current after instantaneous switching from 0.26 V to 0.27 V. Since the RTD is stable (albeit marginally so) at applied biases above 0.254 V, these oscillations were apparently initiated by the abrupt bias changes. A 1 V/ns slew rate (10,000 time steps per 10 mV) simulation from 0.26 V to 0.27 V was conducted to verify this. The result is shown in Figure 7.5. Even this simulation shows very small oscillations after the (abrupt) start and end of slewing. Presumably, even lower slew rates would avoid oscillations entirely. Thus, once again the use of an infinite slew rate (by Jensen and Buot) has been a culprit in producing simulation results which led to invalid conclusions about device operation.

A further set of simulations sought to determine the effect of slewing the applied bias smoothly through the unstable region of operation. This might occur in the simulation of a device with an unstable region that is smaller than the chosen bias step. In this case, the existence of oscillations would probably be missed entirely. Transient simulations starting at 0.23 V and slewing the bias continuously to 0.26 V served to investigate this possibility.

6. Further simulations marked the unstable region more precisely at between 0.239 V and 0.254 V.

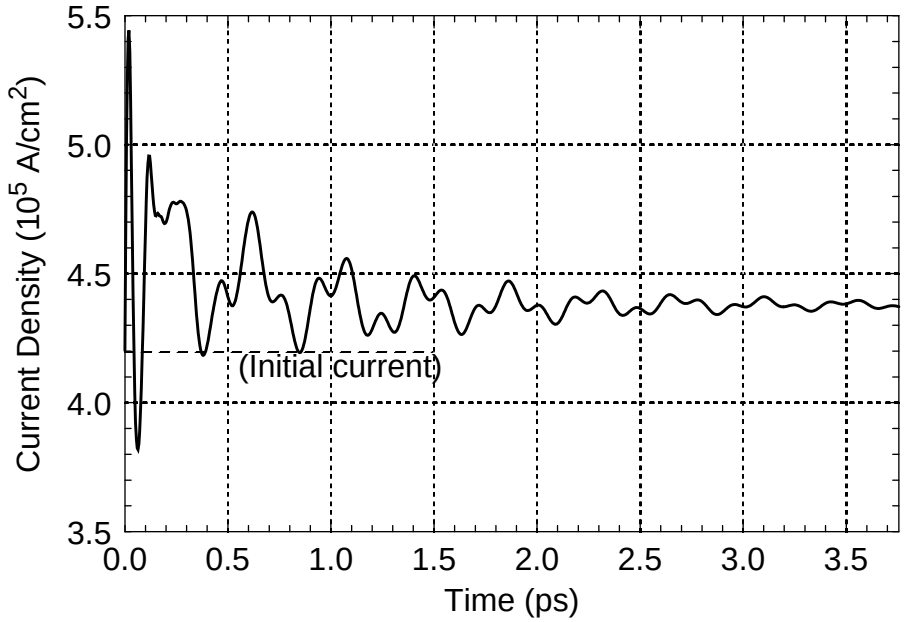


Figure 7.4: Damped oscillations after abrupt switching in plateau

Transient current after instantaneous switching from 0.26 V to 0.27 V in the plateau. Although the difference between initial and final current is less than 2×10^4 A/cm^2, the oscillation amplitude starts at 10 times this value. The oscillations are initiated by the abrupt switching

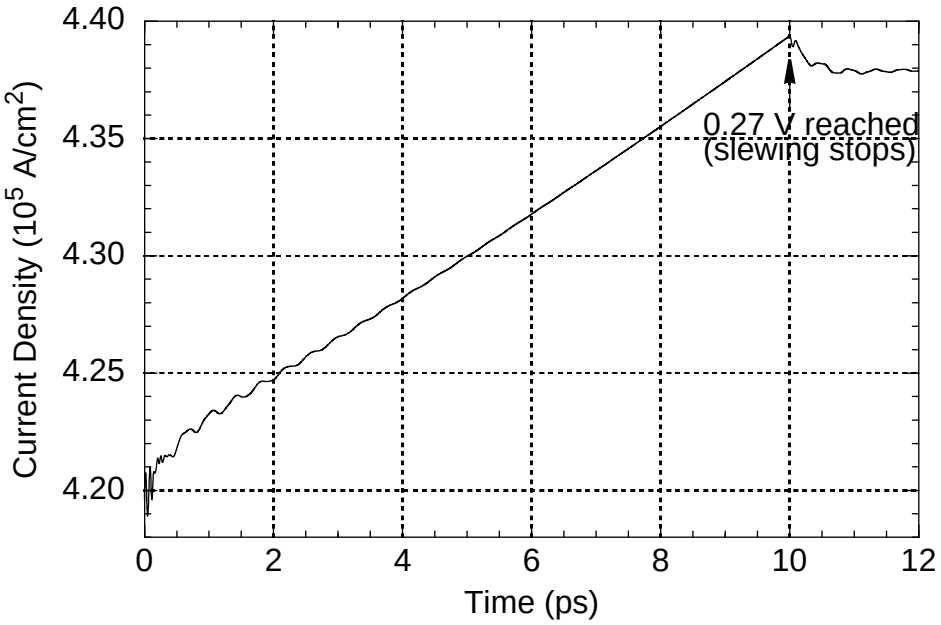


Figure 7.5: Nearly smooth transition with slewed switching in plateau

Transient current during and after 10 ps slewing from 0.26 V to 0.27 V in the plateau (slew rate: 1 V/ns). Although small oscillations occur even at this low slew rate, the oscillation amplitude is less than 1/100th of that in Figure 7.4.

The results are shown in Figure 7.6. At 1 V/ns (10,000 time steps per 10 mV), the device slewed through the unstable region too quickly for oscillations to begin. The results was the same at 0.5 V/ns, (20,000 time steps per 10 mV). Finally, at 0.2 V/ns (50,000 time steps per 10 mV), the RTD was able to achieve the conditions necessary for oscillations (see Chapter 8). The oscillations persisted during the continuous slewing, albeit with decreasing amplitude, until shortly after the unstable region was exited. These results indicate a relatively very slow response time for a device which is otherwise so fast. The lesson is that even the use of a relatively low slew rate may still allow some device physics to be missed.

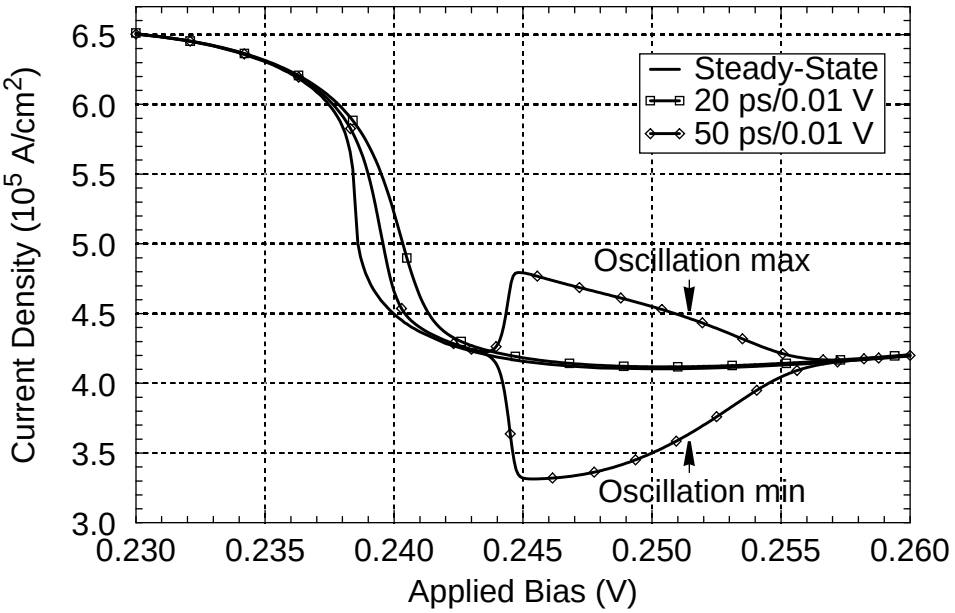


Figure 7.6: Slewing across an unstable region of RTD operation

Trace of unstable region of I-V curve. Continual slewing at 1 V/ns (or 10 ps/10 mV) and 0.5 V/ns (or 20 ps/10 mV) are too fast to allow the RTD to begin oscillating before it leaves the unstable region at 0.254 V. Continual slewing at 0.2 V/ns (or 50 ps/10 mV) is slow enough. Only the peaks of the oscillation are shown, so that the other curves are not obscured.

Since simulation results in the plateau region of the I-V curve depend strongly on the slew rate used, it is again important to consider what might happen in an actual circuit or test environment. Device analyzers (such as the HP 4145) trace an I-V curve by sweeping the applied bias in a step-wise fashion, so that a new bias is established, a delay time (typically a few milliseconds) elapses to allow the device to settle to steady-state, and the cur-

rent is measured. Thus, with an ideal device tester, oscillations would definitely be seen in the unstable region (assuming at least one bias point fell there), since there would be plenty of time for the device to evolve to the unstable conditions. In the marginally stable region, since the slew rate would be less than 1 V/ns and slewing would start and stop more smoothly, no oscillations would occur after 0.254 V. However, device testers are not ideal. External inductance and capacitance in the measuring apparatus could easily change a marginally-stable RTD into an unstable one, causing the RTD to oscillate everywhere in the plateau. Based on the 100 V/ns simulation results, this RTD would certainly oscillate throughout the plateau in a fast-changing RTD circuit.

7.2.5 Bistable Regions

Bistable regions pose yet another hazard for instantaneous bias switching. When a transient simulation is used to trace the steady-state I-V curve (*e.g.*, to search for latent transient effects), for this RTD, instantaneous switching produced the same I-V curve as the steady-state simulation and the slewed-switching, transient simulation. However, it was not difficult to devise a transient simulation that did not follow the steady-state I-V curve. For example, a transient simulation starting from steady-state at 0.23 V and switching instantaneously into the bistable region at 0.26 V, rather than converging to the "correct" higher current state, converged to the lower current state. In contrast, the same simulation with a 10 V/ns slew rate converged to the higher current state. These results are shown in Figure 7.7. The standard slew rate in this work of 100 V/ns was also too fast to converge to the upper I-V curve trace. In general, the shock of instantaneous, or even fast, bias switching may cause a device to "leap the rails" onto another trace in a bistable or multi-stable region of operation. It should be noted that switching to the "wrong" state might be a useful function (*e.g.*, to achieve a higher effective NDR value or produce a multi-state device). By varying the slew rate in simulations, it is possible to investigate how fast the device must be switched in order to produce this type of device operation.

7.2.6 Bias Slewing in Non-Quantum Device Simulation

Virtually all of this slew rate investigation applies equally well to the simulation of conventional (*i.e.*, non-quantum) electronic devices. For example, transient bias changes in conventional device simulation are accomplished in the same single-time-step, incre-

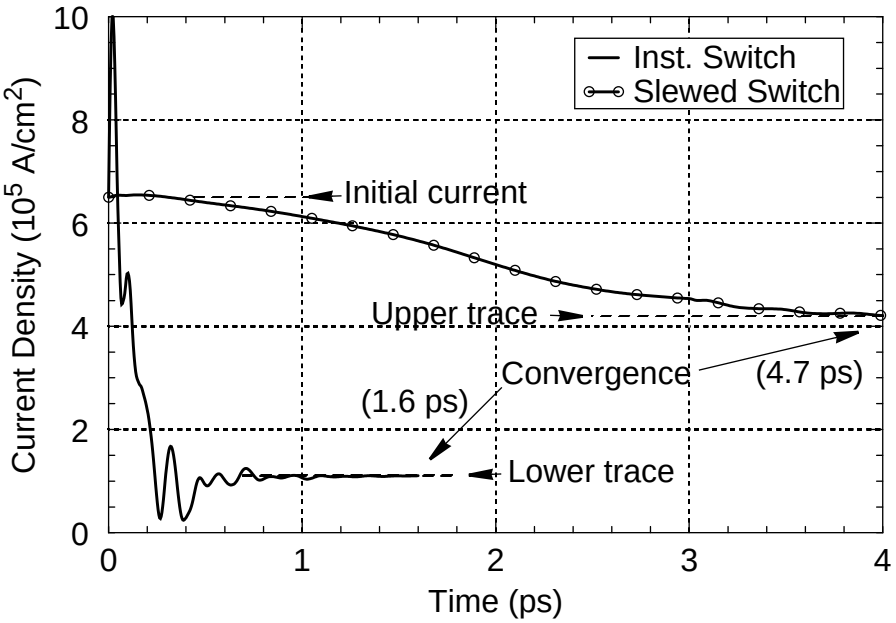


Figure 7.7: Switching or slewing into bistable region of operation

Transient current after switching from 0.23 V to 0.26 V. The instantaneously switched simulation (plain curve) converged to the lower bistable value, while the 10 V/ns slewed simulation (circle-curve) converged to the upper value. This shows that slew rate variation can profoundly affect device function.

mental manner described in Section 7.1. The time step used in conventional device simulations may be orders of magnitude larger than that used in a typical quantum device simulation, but the bias change per time step may also be larger. Further, as indicated in this chapter, the appropriate slew rate depends on the inherent speed of the device being simulated, the intended application, and the desired function. Thus, the critical slew rates where conventional device function may change will be much lower than those found in this chapter for quantum devices.

7.3 Summary

This work has demonstrated the importance of using a finite applied bias slew rate (as opposed to instantaneous switching) to better approximate experimental quantum device conditions, and thus produce more accurate transient Wigner function simulation results. In fact, the proper slew rate for *any* electronic device simulation depends on device speed, intended application conditions, and desired device function. Several instances were high-

lighted where the use of instantaneous switching has led to incorrect conclusions about quantum device operation. As a added benefit, it was shown that the use of slewed switching can also reduce the high computational demands of transient WFM simulation.

References

[1] M. R. Pinto. *Comprehensive semiconductor device simulation for silicon ULSI*. PhD thesis, Stanford University, Aug. 1990. p. 413.

[2] B. A. Biegel and J. D. Plummer. “Origin of the current pulse after a bias change across a resonant tunneling diode.” Unpublished, March. 1996.

[3] H. Tsuchiya, M. Ogawa, and T. Miyoshi. “Static and dynamic electron transport in resonant-tunneling diodes.” *Japanese Journal of Applied Physics*, 31(3):745–750, 1992.

[4] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer. “Self-consistent study of the resonant-tunneling diode.” *Physical Review B*, 39(11):7720–7735, 1989.

[5] W. R. Frensley. “Wigner-function model of a resonant-tunneling semiconductor device.” *Physical Review B*, 36(3):1570–1580, 1987.

[6] F. Moraveji. “Low-power, high-speed, voltage-feedback operational amplifier on a low-cost 40 volt complementary bipolar technology.” In *IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pages 19–22, 1994.

[7] K. L. Jensen and F. A. Buot. “Numerical simulation of intrinsic bistability and high-frequency current oscillations in resonant tunneling structures.” *Physical Review Letters*, 66(8):1078–1081, 1991.

Chapter 8

RTD Device Physics Investigation

In Chapters 6 and 7, the simulation test device was a resonant tunneling diode (RTD) exhibiting several interesting behaviors, including a plateau in the negative differential resistance region of the I-V curve, hysteresis, and intrinsic high-frequency oscillations. While these phenomena were described, their physical causes were not determined. This chapter finally presents a detailed investigation of the physics behind these behaviors. This investigation incorporates and adds to the results of Chapters 6 and 7, producing a fairly comprehensive demonstration of the capabilities of SQUADS for analyzing a quantum device from many different viewpoints until its behaviour is fully understood.

This simulation investigation of RTD physics begins in Section 8.1 with a summary of several remaining controversies about aspects of RTD operation. The RTD physics investigation in this chapter will have implications for each of these open questions. Section 8.2 then analyzes the steady-state operation of this RTD, using both the TMM and WFM capabilities of SQUADS. The most important steady-state effects are the plateau and bistability in the NDR region of the I-V curve. Section 8.3 continues the investigation, with an in-depth look at the transient operation of the RTD, including an analysis of the unstable operation in part of the plateau. Based on this investigation, Section 8.4 significantly revises the interpretation of simulations of this RTD by previous researchers. Finally, Section 8.4 also discusses the remaining discrepancies between experimental RTD measurements and the simulation results for this RTD.

8.1 RTD Controversies

Resonant tunneling diodes [1-3] have undergone intense investigation, both experimental and theoretical, over the past decade and more, due to their potential circuit applications [2, 4] and their status as the prototype quantum electronic device. Even as a detailed understanding of the operation of RTDs has developed, several controversies have persisted, including the cause of an observed plateau in the negative differential resistance (NDR) region of the current-voltage (I-V) curve, whether intrinsic bistability has been observed and how it manifests, the cause of RTD oscillations, the nature of the tunneling process (sequential or resonant), and the correct lumped-parameter equivalent circuit model for an RTD. Although some consensus has developed on most of these issues, additional investigation through accurate numerical simulation of RTDs is needed.

In 1991, Jensen and Buot (JB) [5] published the very interesting results of some ambitious transient Wigner function-based numerical simulations of an RTD (hereafter called the JB RTD), including both self-consistency and scattering. In [5] and subsequent theoretical analysis of these simulation results by Buot et al. [6 -11], a good deal of insight was developed into several of the RTD controversies mentioned above. Unfortunately, more detailed and comprehensive investigations would have been prohibitively expensive at that time (expensive Cray supercomputer time was required) [12]. With the rapid advance in the power of computer workstations, it has recently become feasible to revisit the Wigner function simulation of the JB RTD in greater detail. The investigation of RTD physics in this chapter accomplishes this by using the JB RTD for a case study. Based on these simulation results, the current understanding of the operation physics of this device are extended and revised where necessary. The remainder of this section summarizes the history and current state of the RTD controversies, and indicates how the simulation results of JB and subsequent analysis by Buot et al. impacted these debates.

The device physics underlying the plateau in the NDR region of I-V curve measurements of RTDs has been a source of controversy for over a decade. The plateau is often accompanied by hysteresis/bistability and high-frequency oscillations. For example, Figure 6.1 shows a measured RTD I-V curve where all of these effects were observed [13]. Two opposing explanations were given for the plateau. Some researchers [14-25] advocated extrinsically-induced oscillations caused by reactive elements in the external bias circuit (in concert with intrinsic NDR and capacitance) as the sole cause. Others main-

tained that purely intrinsic RTD operation, such as intrinsic bistability due to charge storage in the quantum well [26-29], or discrete states in the emitter accumulation layer [30], could explain the plateau. Several theoretical calculations and simulations of the RTD [31-37] seemed to support the “intrinsic” case.

The extrinsic-versus-intrinsic plateau debate appeared to be well settled when circuit simulations [20, 22, 25] showed convincingly that all of the observed behavior could be reproduced with extrinsically-induced oscillations. In fact, several researchers [17, 18, 21, 28, 33] argued that keeping an RTD from oscillating in the NDR region, a necessary condition to make the “intrinsic” case, is difficult. Other simulations [13, 24, 38,] suggested that producing an RTD that exhibits intrinsic bistability is also difficult. In fact, many researchers seeking to demonstrate intrinsic bistability used asymmetric barriers [24, 32, 34, 35, 39-41], thick undoped layers near the emitter or collector barrier [39, 41], or a second quantum well just before the emitter barrier [42]. Such structures were not employed in the many experimental measurements where the I-V curve effects were observed. Further, all unambiguous examples of intrinsic bistability produced hysteresis in the main current peak, rather than producing a plateau and hysteresis in the NDR region after the main peak. These facts led to the conclusion that experimentally observed I-V curve effects associated with a plateau in the NDR region were due to, and evidence of, extrinsically-induced oscillations.[1] These efforts also settled the debate as to whether intrinsic bistability could be observed (almost certainly), and how it appears (as a hysteresis loop in the main I-V peak). We also mention that all observed RTD oscillations were assumed to be extrinsic in origin.

In contradiction to the consensus, Jensen and Buot’s simulations [5] provided some convincing evidence for the “intrinsic” explanation of the I-V plateau and related phenomena. Their Wigner-function-based (intrinsic) RTD simulations showed an I-V plateau in the NDR region, hysteresis/bistability in the plateau, and *intrinsic* high-frequency current oscillations at any fixed bias in the plateau. Analysis of these simulations by JB [5] and Buot et al. [6, 8, 9] described the physics that would produce this behavior. In short, they concluded that intrinsic bistability and oscillations conspired to produce the plateau. The mechanism for intrinsic oscillations was the dynamic and self-consistent oscillation of charge in the quantum well and emitter, and the resulting oscillation of the quantum well

1. The equivalent-circuit explanation of these oscillations is given in Section 8.4.2.

state energy. Charge accumulation in the quantum well was responsible for the plateau hysteresis [9]. Buot and Rajagopal [9] also gave a possible explanation for an upward-sloping plateau, which is sometimes observed experimentally, but which is difficult to explain by intrinsic bistability alone.

The controversy [2, 3, 43-47] of whether tunneling is mainly sequential (i.e., tunneling mediated by scattering in the quantum well) or coherent (i.e., tunneling through the entire double-barrier structure without scattering) is perhaps less settled than the I-V plateau and bistability issues. It is not even clear whether the answer makes any practical difference in RTD operation [29, 34-36], although others disagree [2, 3, 44]. The current consensus is that both effects occur in parallel, and either current component can dominate, depending on the scattering rate and device structure [2]. Accurate numerical simulation will be required to further illuminate this issue. JB and Buot et al. did not comment on this issue, but we will return to it briefly in Section 8.2.

Finally, a relatively recent point of controversy concerning RTDs is which lumped-parameter equivalent circuit model is correct for an oscillating RTD. The conventional model [13, 16-22, 25] assumed that the principle source of inductance was extrinsic to the device (*i.e.*, in the biasing circuit), resulting in the series-inductance equivalent circuit in Figure 8.1. This model was not seriously challenged until the numerical simulations of JB [5] showed intrinsic oscillations, requiring an internal inductance model. Two possible circuit models were analyzed in the subsequent analysis of Buot et al. [6, 10, 11]: the series-inductance model (but with an internal inductance), and the parallel-inductance model (see Figure 8.1). As discussed by Buot et al., the location of the inductance has significant implications for stability analysis and oscillation frequency of the RTD. Thus, proper design of RTDs for use as fast oscillators and related applications will depend on use of the correct circuit model. Based on the analysis of many simulation and experimental results, Buot et al. again went against convention in concluding that the parallel-inductance model must be correct. [Note that Gering et al. [48] and Brown and Sollner [49] had previously proposed RTD models with internal inductance, but neither suggested that the RTD could self-oscillate. In fact, Brown and Sollner's negative inductance did not allow self-oscillations.]

As stated earlier, this chapter describes a SQUADS-based investigation of the physics of RTDs in general, and the above RTD controversies in particular, using the JB RTD as a

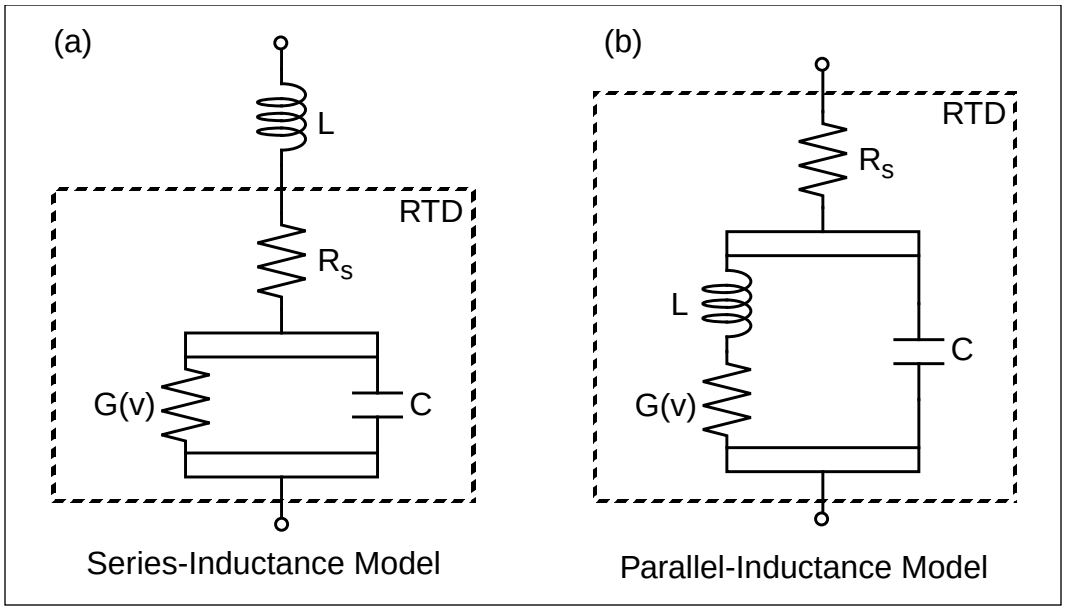


Figure 8.1: Two most common RTD equivalent circuit models

Shown are the series-inductance model and the parallel-inductance model. G is the intrinsic conductance (based on the equilibrium I-V curve), and C is the intrinsic capacitance of the RTD. The series resistance and inductance are often attributed to external causes, such as contact resistance and the measuring apparatus inductance, but in this case of intrinsic oscillations, the cause must be internal. R_S can be attributed to scattering, and L to the delay in current as the QWS charges or discharges after an applied bias change across the DBS.

test case. The results differ from those of JB in some significant conclusions because it was feasible to investigate the operation of the JB RTD in both more detail and more comprehensively that was possible previously.[2] Because the RTD controversies are multifaceted, rather than attempting to demonstrate one position or another on these issues, the approach will be to examine the simulated operation of the JB RTD in sufficient detail to definitively *determine* its underlying physics. Of course, in running a device simulation (as opposed to a circuit simulation), all simulated effects are necessarily intrinsic, so only *inferences* can be made about RTD behavior including measurement circuit parasitics.

This work uses both the Wigner function [50, 51] and transfer-matrix [52, 53] simulation capabilities of SQUADS (see Chapters 4 and 5). To allow direct comparison to the simulations of JB [5], the identical device structure and simulation parameters were used

2. In fact, one of the results of this investigation is a discussion in Section 8.4.3 of the *remaining* inaccuracies in RTD simulation, particularly simulations based on the Wigner function method (used by JB and in SQUADS).

(as was also done in Chapters 6 and 7), and all Wigner function simulations in this chapter include self-consistency and scattering. The JB RTD, shown in Figure 6.4 at equilibrium, was composed of a 5 nm undoped GaAs quantum well between 3 nm undoped $Al_{0.3}Ga_{0.7}As$ tunnel barriers and 3 nm undoped GaAs spacer layers. The GaAs contact layers were 19 nm each, giving a total device width of 55 nm. The electron effective mass was assumed constant at $0.0667m_0$, and the permittivity was also taken as constant at $12.9\varepsilon_0$. The simulations used 86 position points, 72 wavenumber points, a time step of 1 fs, and an effective relaxation time of 525 fs [54] at a simulation temperature of 77 K.

8.2 Steady-State RTD Physics

The fundamental operation characteristic for electronic devices is the current-voltage curve, and this will serve as the starting point in our investigation of the JB RTD. A Wigner function simulation tracing the steady-state I-V curve of this RTD is shown in Figure 8.2. Before investigating the physics behind the more interesting features of this I-V curve, we first describe “normal” RTD operation, the major feature of which is a region of negative differential resistance. The basic cause of NDR is indicated in Figure 8.3, which shows the energy band profile of the JB RTD at both the peak and valley of the I-V curve. At applied biases up to and including the peak current condition (0.23 V in the JB RTD), electrons entering the RTD at the emitter contact can tunnel through the double-barrier structure (DBS) via the quantum well state (QWS). As the bias is increased above the peak condition, the QWS energy drops below the emitter band edge, and current decreases. Current does not drop immediately to its minimum because of the finite width of the QWS and scattering-assisted tunneling into the QWS. Nevertheless, current should normally decrease monotonically from peak to valley (0.32 V in this RTD), as indicated by the dashed curve in Figure 8.2. Current reaches a minimum and eventually increases again because, as the collector barrier is pulled down with increasing applied bias, tunneling through the entire DBS (i.e., not via the QWS) increases. This “normal” RTD behavior is well described in [2] and elsewhere.

Quite obviously from Figure 8.2, the JB RTD does not behave in the simple manner described above in the NDR region of operation. In fact, it was precisely because of this interesting behavior that we used this RTD as the test device in our recent investigations [55, 56] of implementation issues in the Wigner function method of quantum device simu-

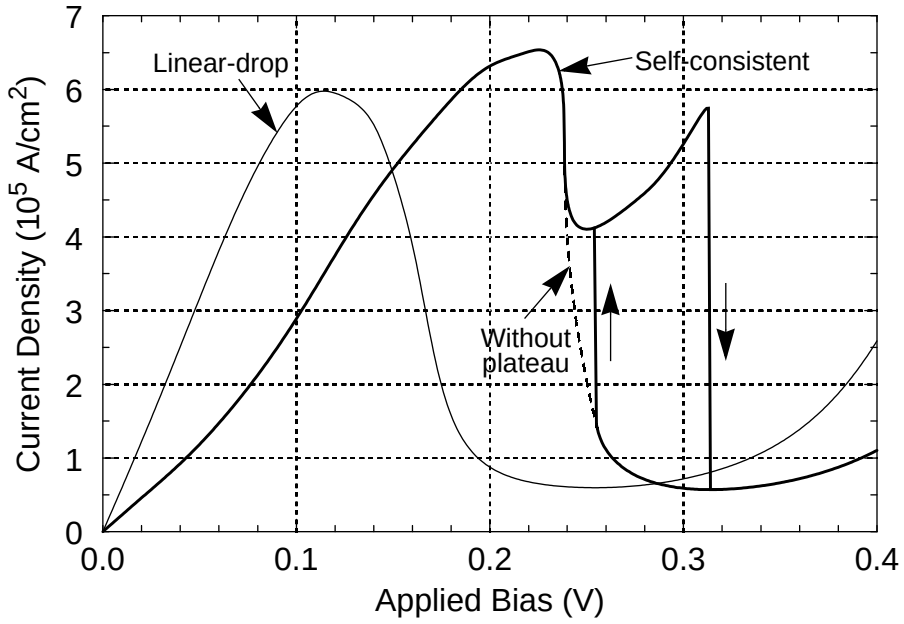


Figure 8.2: Self-consistent, steady-state RTD I-V curve

Note the upward-sloping plateau and hysteresis in the NDR region. "Normal" RTD operation (without the plateau) is indicated by the dashed curve. The linear-drop (non-self-consistent) I-V curve was simulated assuming a linear bias drop across the undoped central region of the RTD.

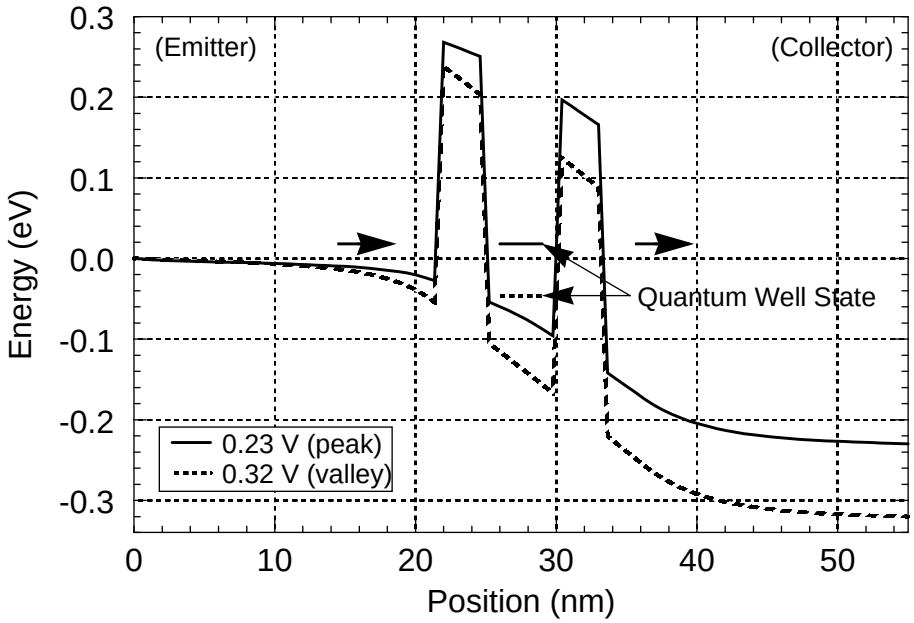


Figure 8.3: RTD energy band profiles at peak and valley operation

Carriers entering from the emitter can tunnel through the quantum well state at the bias for peak current (0.23 V), but not at the bias for valley current (0.32 V).

lation. Instead of the current falling smoothly from peak to valley, a plateau structure occurs in the I-V curve in the NDR region of operation. Apparently a second current path is operative here. The plot in Figure 8.4 of the energy band profile of the RTD at 0.28 V (the center of the plateau) indicates what this new current path may be. Here we see that the QWS is indeed well below the band minimum at the emitter contact, so electrons entering the RTD at the emitter can not tunnel through the QWS directly. However, an extended energy band depression has developed in the emitter layer. This suggests that the plateau current occurs from electrons scattering into the emitter depression and then tunneling through the QWS as usual. Since the emitter depression is narrow (10-20 nm), the quantum mechanically allowed energy levels (below zero) for electrons will be discrete and widely separated, just as in the quantum well. Thus, this explanation for the plateau structure depends on an allowed state in the emitter depression being at roughly the same energy as the QWS so that current can flow from the emitter state to the QWS. Because of the interesting physics involved in this plateau mechanism, a fair amount of discussion below will be devoted to verifying and analyzing it.

It is not straight-forward to verify the above explanation of the plateau. Normally, the transfer-matrix method of quantum device simulation is used to locate discrete energy lev-

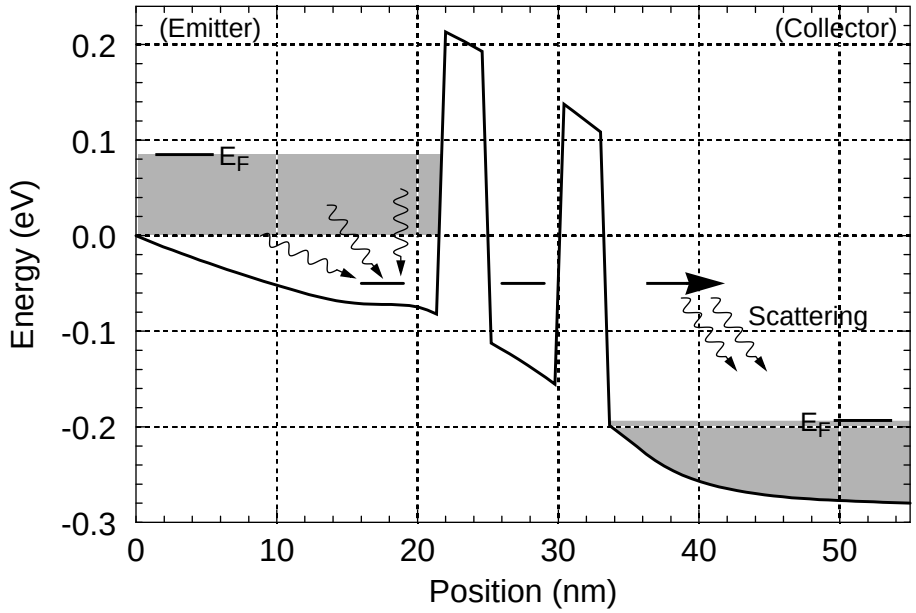


Figure 8.4: RTD energy bands at center of I-V plateau

Self-consistent energy bands at 0.28 V indicate that electrons must scatter into the discrete energy state in the emitter depression to tunnel through the quantum well state.

els by finding the energies at which the transmission of electrons through the device is a maximum. However, for operation of the JB RTD in the plateau, the transmission coefficient in either direction at the QWS energy is necessarily zero, since there are no emitter electrons incident at this energy (because the QWS is below the emitter contact minimum), and all electrons from the collector at the QWS energy will be reflected back (see Figure 8.4). Further, the suspected current path requires scattering, which has not been incorporated accurately into the transfer-matrix method. In spite of these difficulties, it is possible to probe the energy bands in Figure 8.4 for resonant energy states below $E = 0$. The trick is to calculate quantum wavefunctions for mono-energetic electron beams incident from (and reflected back to) the collector at a range of energies. Those wavefunctions which have the highest standing-wave amplitudes in the emitter depression and quantum well are said to "fit", or resonate, there. The computation of single-energy electron wavefunctions is a simple extension of the transfer-matrix method of quantum system simulation (see Chapter 4).

By default, SQUADS computes wavefunctions at energies relative to the incident contact minimum. In this case, wavefunctions are incident from the collector, which will effectively shift the reference energy down by the applied bias. Using the approach discussed above, Figure 8.5 shows the resulting energy spectrum (normalized wavefunction amplitude versus energy) of carriers in the emitter depression (solid curve) and quantum well (dashed curve) for the energy bands of the JB RTD biased at the center of the plateau (0.28 V). The first discrete emitter state (DES) energy is only about 5 meV below the QWS energy, which is close enough for them to interact. Note the constructive interference at the respective resonant energies and destructive interference between.[3]

Note that the DES and QWS energies are separated by only 5 meV when the RTD is biased at the center of the plateau, yet the plateau extends over about 75 mV of applied bias. This requires that the two energy levels must stay essentially "locked" together via some (as-yet undetermined) mechanism during this portion of the I-V curve: any changes in energy of the two states must be virtually equal. If the energy levels became widely separated, the plateau current path would be broken, and current flow would decrease.[4] Figure 8.6 shows the variation of the two energy states versus applied bias in the plateau. As

3. The presentation of these results was refined based on similar work by another researcher [57].
4. In fact, this is what happens at the end of the plateau, as discussed later.

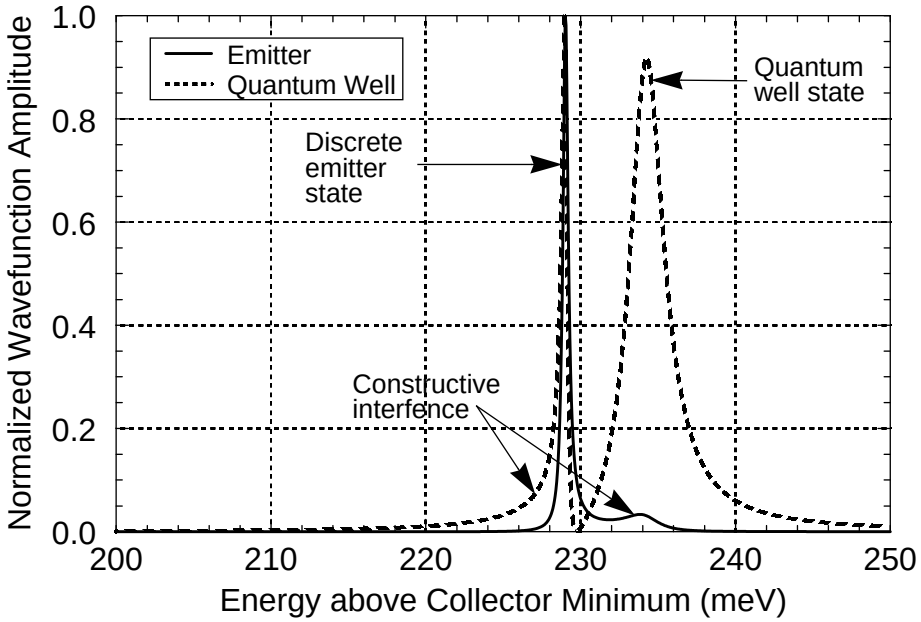


Figure 8.5: Energy occupation spectrum in emitter and quantum well

Energy occupation spectrum (normalized wavefunction amplitude versus energy) of carriers in the emitter depression (solid curve) and quantum well (dashed curve) for the band diagram of Figure 8.4. The first emitter energy level is only about 5 meV below the quantum well state. Constructive interference is apparent near the respective resonant energies, and destructive interference between.

expected, the two energy states do remain very close (within 10 meV) throughout the plateau. Collecting data for this plot (from curves like that in Figure 8.5) is somewhat complicated due to the interference of the two resonances. However, a couple of general trends are evident: the QWS energy is nearly constant, while the DES energy rises gradually, until at the end of the plateau the two are equal. Each of these observations is significant, as discussed below.

The fact that the QWS energy does not rise with respect to the collector band minimum indicates that the energy band profile in the collector and quantum well does not change appreciably through the plateau. Therefore, all increases in applied bias must be accommodated by band-bending in the emitter. To check this definitively, energy band profiles were plotted for consecutive biases in the plateau, with the collector electrode grounded and biases applied to the emitter (rather than vice versa, as in Figure 8.4). The energy bands in the collector and quantum well should line up very closely, and then diverge in the emitter. This is exactly what occurs, as shown in Figure 8.7. Thus, all of the additional band-bending in the plateau is accomplished by charging the emitter contact

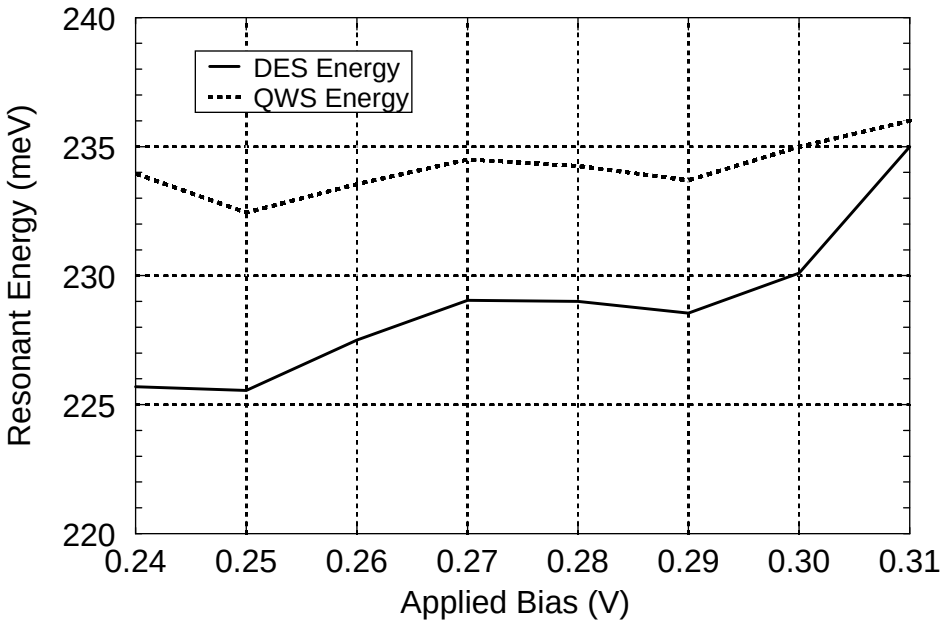


Figure 8.6: Emitter and quantum well energy levels in plateau

Resonant energy (relative to the collector band minimum) versus applied bias in the plateau. The solid curve shows the energy of the lowest state in the emitter depression, while the dashed curve shows that in the quantum well. Both energies are relatively constant relative to the applied bias change. However, the emitter state energy rises until it reaches the quantum well state energy at the end of the plateau.

and discharging of the emitter itself, so that the e-field at the emitter barrier, and thus the energy band profile in the rest of the RTD, remains unchanged.

Another conclusion from the fact that the quantum well and collector energy band profiles do not change in the plateau region is that the total charge in the quantum well and collector must remain constant throughout the plateau. If the charges changed appreciably, then the electric fields in the device would also be modified, as would the potential profile. On the other hand, as mentioned above, the emitter charge (absolute magnitude) must decrease to screen the charging of the emitter contact from the rest of the RTD. To check these conclusions, Figure 8.8 shows the displaced charge (difference in charge from equilibrium) in the emitter, quantum well, and collector versus applied bias. As expected, the quantum well and collector charges are effectively constant during the plateau, while the magnitude of the emitter charge decreases significantly. Also shown in Figure 8.8 is the change in total charge in the RTD. This curve indicates that the RTD does not remain net charge neutral, reflecting the charging of the emitter contact to restore charge neutrality.

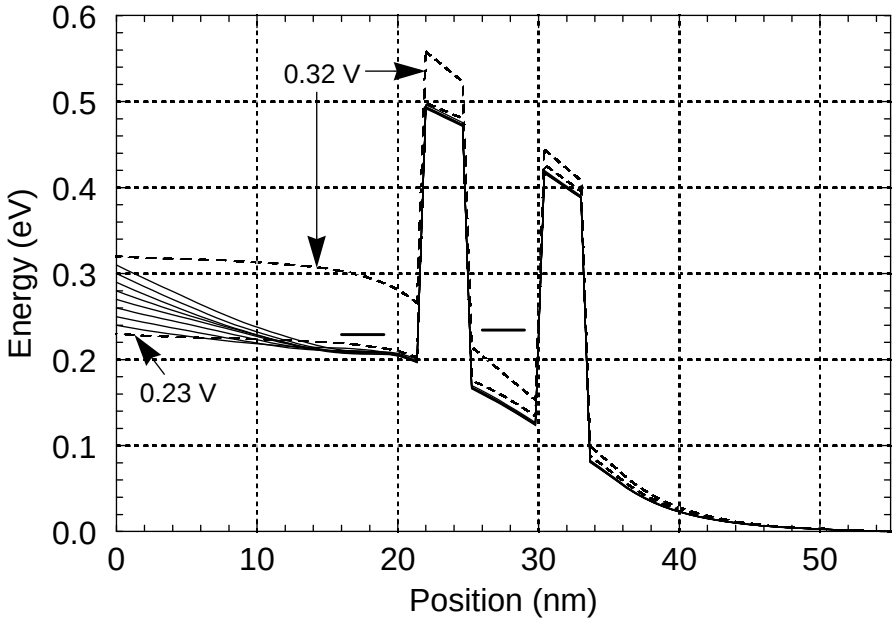


Figure 8.7: RTD Energy band profiles in plateau operation

Self-consistent energy band profile for the plateau (solid curves) and adjacent biases (dashed curves). All applied bias changes in the plateau are accomodated by charging of the emitter contact and discharging of the emitter itself. The resonant states at the center of the plateau (0.28 V) are shown in the emitter depression and quantum well.

The significance of the charged emitter contact is examined in Section 8.4.3.

Comparison of the I-V curve in Figure 8.2 and the quantum well charge in Figure 8.8 suggests that current flow is proportional to the quantum well charge Q_{QW}. To understand why this is true, recall the basic current density equation (ignoring sign conventions):

$$J = qnv \tag{8.1}$$

where q is the elementary charge, n is the electron density, and v is the average electron velocity. Suppose that the probability of an electron tunneling through the collector barrier, and thus contributing to n, is fixed. Then J is exactly proportional to Q_{QW}. In typical RTD operation, the height of the collector barrier above the QWS energy decreases gradually with increasing applied bias, so Q_{QW} is not exactly proportional to n over wide applied bias ranges. However, when the JB RTD operates in the plateau, where the collector barrier does remain essentially unchanged, the proportionality should hold very well. Further, since electric fields in the collector remain unchanged in the plateau, the average velocity v of the carriers in the collector should remain constant throughout plateau operation. This suggests that current should be constant in the plateau, and this is far from the

case (see Figure 8.2).

The only reasonable conclusion from the above analysis is that (once again) a parallel current path must be operating. Figure 8.4 shows what this additional current path is: tunneling through the entire double-barrier structure by carriers in the emitter that do not scatter into the DES. Considering the plateau energy band diagrams in Figure 8.7, as the applied bias increases, carriers entering at the emitter contact which do not scatter will see both barrier heights reduced by approximately $V_a - 0.23$ V (*i.e.*, the difference between the applied bias and the bias at the current peak). Since tunneling probability (the transmission coefficient) varies exponentially with barrier height, this non-resonant current component (from unscattered electrons in the emitter) should increase exponentially versus applied bias. This expectation is consistent with the current increase seen in the plateau of Figure 8.2. The positive slope of the plateau is again an issue in Section 8.4.3.

The other observations from Figure 8.6 are now investigated: that the discrete emitter state energy rises relative to the collector energy band during the plateau until it equals the QWS energy at the end. The fact that the DES energy rises makes sense: the emitter

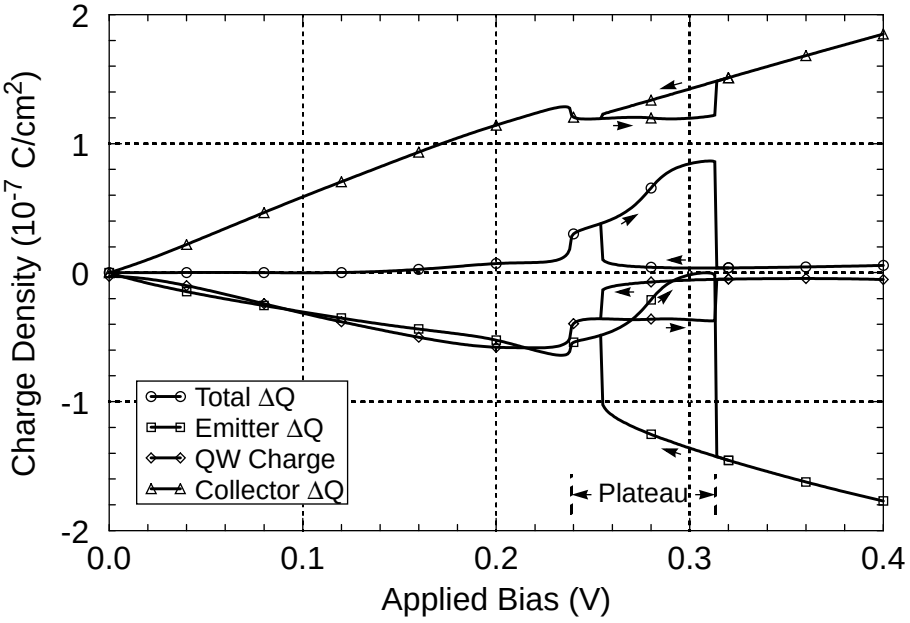


Figure 8.8: Integrated charge versus applied bias in RTD

For the total, emitter, and collector charges, only the change from the equilibrium is shown. In the plateau region, the quantum well and collector charges are essentially constant, while the magnitude of the emitter charge decreases significantly to screen the rest of the RTD from the charged emitter contact. The charge on the emitter contact can be inferred from the non-zero change in total charge.

depression gets narrower and steeper, so the resonant quantum wavelength will be shorter, and the resonant energy is therefore higher. The following argument shows that as long as the DES energy is below the QWS energy, the plateau current path is maintained. With the QWS above the DES, if the QWS charge density increases, the e-field in the collector barrier increases while that in the emitter barrier decreases, so the potential of the QWS rises further above the DES. This reduces the current flow from DES to QWS, reducing the QW charge and the QWS energy. By symmetry, as the QWS charge decreases, the potential of the QWS decreases towards the DES energy, so the supply of electrons from DES to QWS increases, as does the QWS charge and energy, and the cycle repeats. Thus, a negative feedback mechanism due to charge storage in the quantum well keeps the QWS slightly above the DES, and maintains the DES-QWS current path. However, if the DES ever rises above the QWS, we argue that the plateau current path is no longer stable, and the RTD switches to the lower I-V curve. For example, as 0.313 V is approached on the up-trace, exact alignment of the DES and QWS produces maximum current from DES to QWS. When the DES rises just slightly above the QWS, the supply of electrons to the QWS decreases, and the QWS begins to deplete. The e-field in the collector barrier decreases while that in the emitter barrier increases, so the potential of the QWS drops further below the DES. This further reduces the supply of electrons to the QWS. A run-away conditions ensues, which ends when the lower I-V curve operating conditions are reached.

Since plateau operation is apparently so fragile, it should not be difficult to switch the RTD out of plateau operation without actually biasing the RTD above 0.313 V. This expectation was proven correct in simulations described in Chapter 7 and [56], where simply slewing the bias too quickly from the peak into the plateau region caused the RTD to switch to the lower I-V curve. Another point should be made that although the non-resonant current (i.e., current that tunnels through the DBS, *not* via the DES/QWS) is quite significant in the plateau, this current component contributes negligibly to the quantum well charge, because the wavelengths of electrons at these energies do not “fit” (resonate) in the quantum well. Thus, when the resonant current component can not support the quantum well charge at the end of the plateau, the non-resonant carriers can not make up the difference.

Given the above description of the two simultaneous current paths in this RTD, it may seem that these simulations can shed some light on the controversy of whether tunneling

in RTDs is dominantly resonant or sequential. However, in this RTD, the important scattering which makes the plateau current path possible takes place *before* electrons tunnel through the double-barrier structure, rather than in the quantum well. Actually, since scattering in these simulations is constant throughout the device, both electron current paths involve components of both sequential and coherent tunneling. Thus, once again, differentiating between the two, even in simulations, will require more ingenuity. Nevertheless, with its ability to include scattering in a meaningful way, the Wigner function approach should be well suited to such an investigation.

Having done the transfer-matrix analysis of the DES and QWS energy levels, the above conclusions are further verified with a few additional Wigner function simulation results. Note that the Wigner function $W(x, k)$ gives the density of carriers at each position and wavenumber in the simulation domain. Therefore, a cross-section of the Wigner function at a particular position gives the number of carriers at each wavenumber (and thus energy) at that position. Consider two such cross-sections of the Wigner function at 0.31 V applied bias: one at the collector contact ($x = 55$ nm), and one in the emitter depression ($x = 18$ nm). In the collector contact cross-section (Figure 8.9), there should be two peaks of carriers (besides the large equilibrium distribution centered at $k = 0$) at positive velocity (wavenumber), corresponding to carriers which tunneled through the QWS, and those which tunneled directly through the DBS from the emitter contact. Wavenumber is related to energy by:

$$k = \frac{\sqrt{2m^*E}}{\tilde{h}}, \tag{8.2}$$

where m^* is the electron effective mass, E is its kinetic energy, and $\tilde{h}$ is the reduced Planck constant. Using Figure 8.7 and neglecting scattering, at 0.31 V, transmitted carrier peaks should appear at about 0.24 eV above the collector minimum for the QWS carriers, and between 0.31 eV (the emitter minimum) and about 0.4 eV (the emitter Fermi level) for the direct tunneling carriers. This translates to wavenumbers of $k_{\mathrm{QWS}} \approx 0.65/\mathrm{nm}$ and $0.74/\mathrm{nm} < k_{\mathrm{dir}} < 0.84/\mathrm{nm}$. These values match the two peaks in Figure 8.9 quite well.

In the Wigner function cross-section in Figure 8.10, three carrier peaks are evident. The two outter peaks are due to carriers travelling from the emitter contact towards the emitter barrier (positive k) and reflected off the emitter barrier back towards the emitter contact (negative k). The middle peak is due to carriers in the DES, which should be about

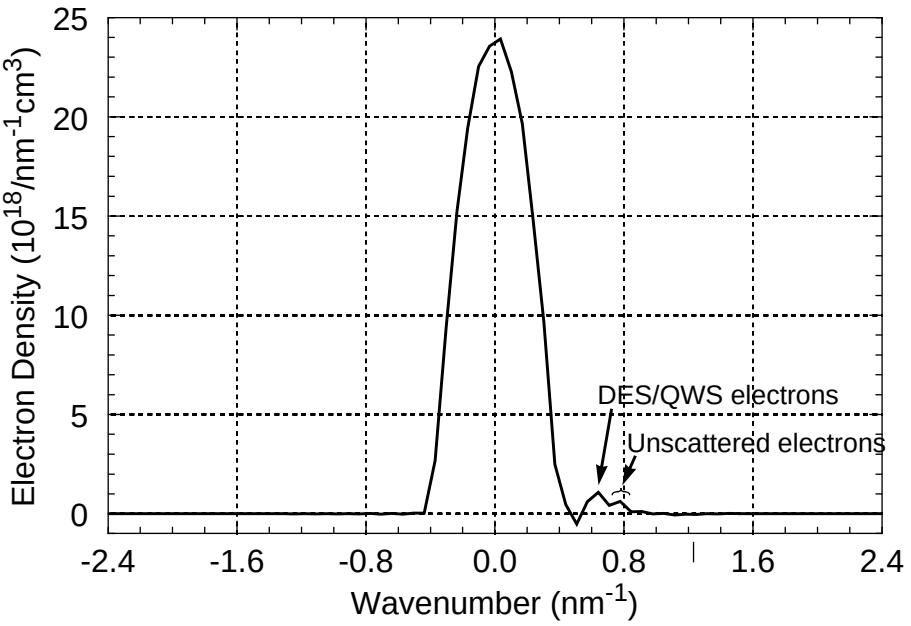


Figure 8.9: Wavevector spectrum of carriers at collector contact

Wigner function cross-section at the collector contact (x = 55 nm) at 0.31 V applied bias. The large peak is the (largely undisturbed) equilibrium carrier distribution, while the two small peaks at positive wavenumber (velocity) account for the RTD current. The peak at k = 0.65/nm is due to electrons which tunneled through the DES and QWS, while the peak near k = 0.8/nm is due to carriers tunneling directly from the emitter without scattering into the DES.

0.026 eV above the band edge at $x = 18$ nm, or $k_{DES} \approx 0.21/\text{nm}$, which is again a reasonable match. Finally, note that the DES carriers are largely not reflected, since there is no corresponding carrier peak at $k \approx -0.21/\text{nm}$. This is because the DBS is nearly transparent at the QWS energy, so the DES electrons tunnel through the DBS, rather than reflecting back like the majority of the electrons not in the DES. Admittedly, because of the relatively low number of wavenumber points used in these simulations, the peaks in both plots discussed above are not resolved very well, even though we used 0.31 V biasing for maximum energy separation. It is worthy of note that the Wigner function method can discern discrete energy level effects without a highly refined wavenumber grid. However, the relative accuracy of such simulations will be considered in Section 8.4.3.

One minor issue remains in the analysis of the basic I-V curve of this RTD. The physics of the transition at 0.313 V as given above is also described by Goldman et al. [26], but the physics of the transition point at 0.254 V was considered a mystery. Actually, the physics of the transition point on the down-trace is even simpler. As 0.254 V is approached

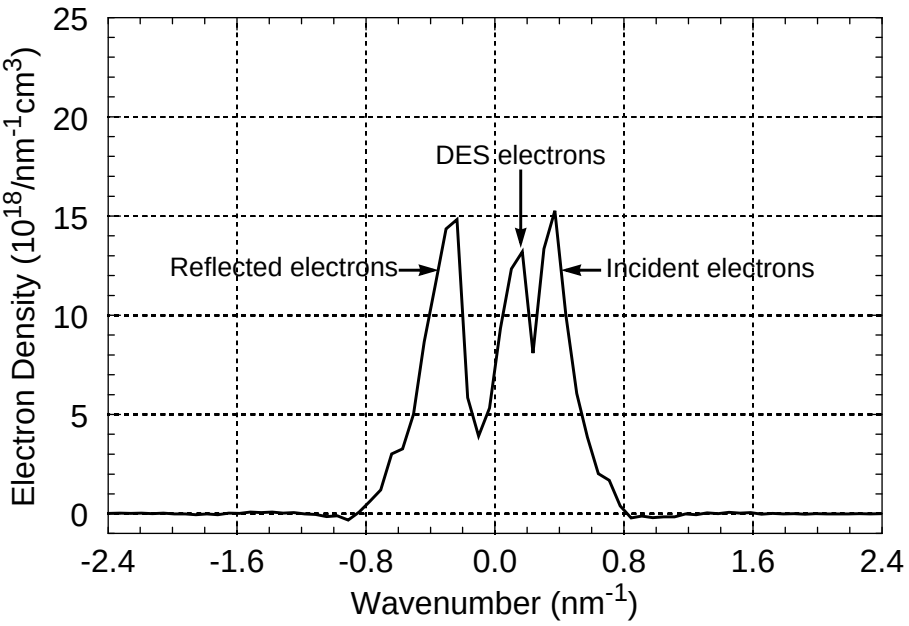


Figure 8.10: Wavevector spectrum of carriers in emitter depression

Wigner function cross-section in the emitter depression (x = 18 nm) at 0.31 V applied bias. The outter peaks are carriers incident on the DBS (positive k) and reflected from the DBS (negative k) which have experienced minimal inelastic scattering. The middle peak shows electrons in the DES and travelling towards the DBS. There is no reflected peak at the DES energy because the DBS is largely transparent at the QWS energy, so the DES electrons tunnel through, rather than reflecting.

from above, the emitter energy band (without a depression) is not far above the QWS. Scattering-assisted tunneling allows the (empty) QWS to begin to fill, raising its potential. This brings the emitter and QWS closer together, and both the scattering-assisted and resonant tunneling currents increase further. As with the 0.313 V transition, a run-away condition ensues that ends when the emitter depression develops and the lower trace reaches the plateau operating conditions. Buot [7] also described the bistable transitions in some detail.

To summarize the steady-state investigation of the JB RTD in this section, the physics of the I-V plateau and associated hysteresis can be described as as an interaction of several phenomena: scattering, the development of a potential depression in the emitter, the alignment of a discrete emitter state with the quantum well state, and charge storage in the quantum well. On the up-trace, as the QWS drops below 0 (*i.e.*, the emitter contact energy band minimum) after the peak condition, the quantum well begins to deplete. Normally,

the emitter charge increases to compensate, but in the plateau, another means of accommodating the applied bias comes into play: the development of an energy depression in the emitter. A discrete emitter state develops in this emitter depression which electrons scatter into, and which provides a current path through the slightly higher energy QWS. A negative feedback mechanism due to quantum well charge keeps the QWS slightly above the DES as the bias increases. Thus, current through the DES-QWS current path remains essentially constant throughout the plateau. However, current due to electrons which do not scatter into the DES increases with bias as the height of the DBS decreases. Also with increasing bias, the DES is slowly pushed up towards the QWS. When the two states cross, the electron supply from DES to QWS decreases, the QWS energy drops as it depletes, and the plateau ends abruptly. On the down-trace, the QWS is initially empty, and the bias must be decreased to the point where the QWS is just below the emitter energy before electrons from the emitter began to scatter into the QWS, raising its potential, and returning the RTD to plateau operation. Thus, for the JB RTD, quantum well charge is solely responsible for the plateau's hysteresis, as determined by JB [5].

8.3 Transient RTD Physics

Given the description in the previous section of the basic physics of the plateau in the I-V characteristic of the JB RTD, it is now possible to menaingfully investigate the transient physics of the plateau. This begins with a transient Wigner function simulation trace of the I-V curve[5] similar to that of Jensen and Buot [5], which simulation brought attention to this device and to the WFM simulation method. Like JB's, these simulations showed high-frequency current oscillations at fixed biases throughout the plateau after switching from one applied bias to the next. However, for biases above 0.25 V, the plateau is actually stable, since the oscillations decayed and the device eventually reached steady-state, while JB concluded that the plateau was unstable throughout. Section 8.4 will return to this important discrepancy between these simulation results and those of JB. According to this transient I-V simulation, the JB RTD is unstable in the plateau only at biases of 0.25 V and below. For example, Figure 8.11 shows the current oscillations at 0.24 V after they have converged to a steady waveform and amplitude after about 20 ps. These are quite sig-

5. Note that Figure 8.2 is a steady-state I-V curve, which traces the (stable or unstable) equilibrium operating point.

nificant oscillations, with a frequency of about 2.5 THz and an amplitude of 1.8×10^5 A/cm^2, which is over 40% of the time-average current.

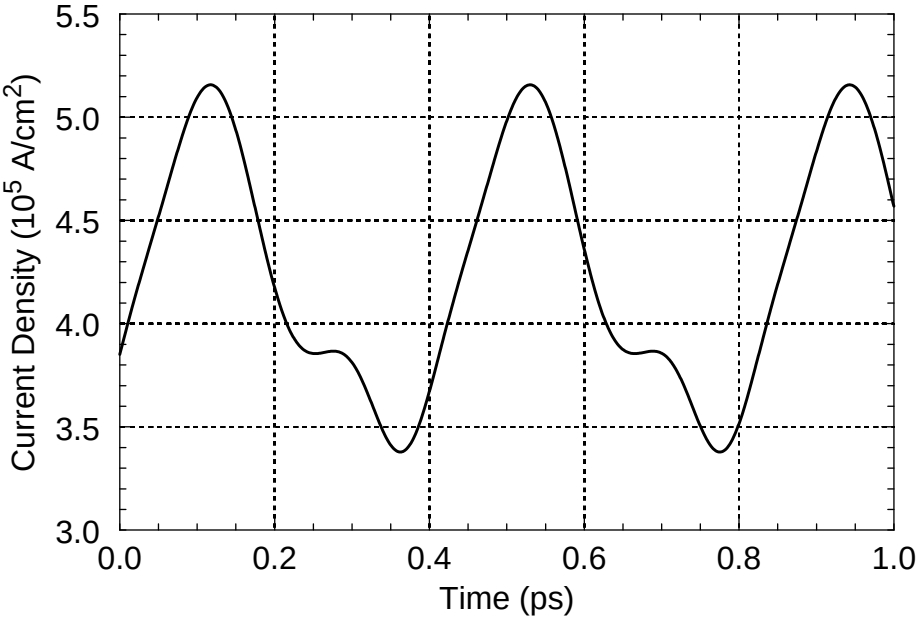


Figure 8.11: Intrinsic current oscillations of unstable RTD

2.5 THz intrinsic current oscillations at V_a = 0.24 V in the narrow-emitter RTD. The oscillations result from the changing relative positions of the quasi-bound states in the emitter depression and the quantum well.

The transient I-V curve was identical to the steady-state curve where the RTD was stable. However, in the small range of biases where there were perpetual oscillations (*i.e.*, where the transient simulations did not converge to steady-state), the time-average current was not equal to the (unstable) equilibrium value found by the steady-state simulation. In such cases, since the transient simulation follows the actual evolution of the device, and since experiments typically measure time-average current, the transient I-V curve is the physically correct one. Figure 8.12 shows a detail view of the unstable region of the equilibrium steady-state and the time-average transient I-V curves. In following the down-trace of the transient I-V curve, a second hysteresis loop not seen in the steady-state simulation was discovered. Thus, there are already three features of the transient operation of the JB RTD to investigate: the cause of the oscillations, the physical difference between the lower (unstable) and upper (stable) portions of the plateau, and the cause of the second hysteresis.

Considering the first issue, the discussion of negative feedback in the previous section

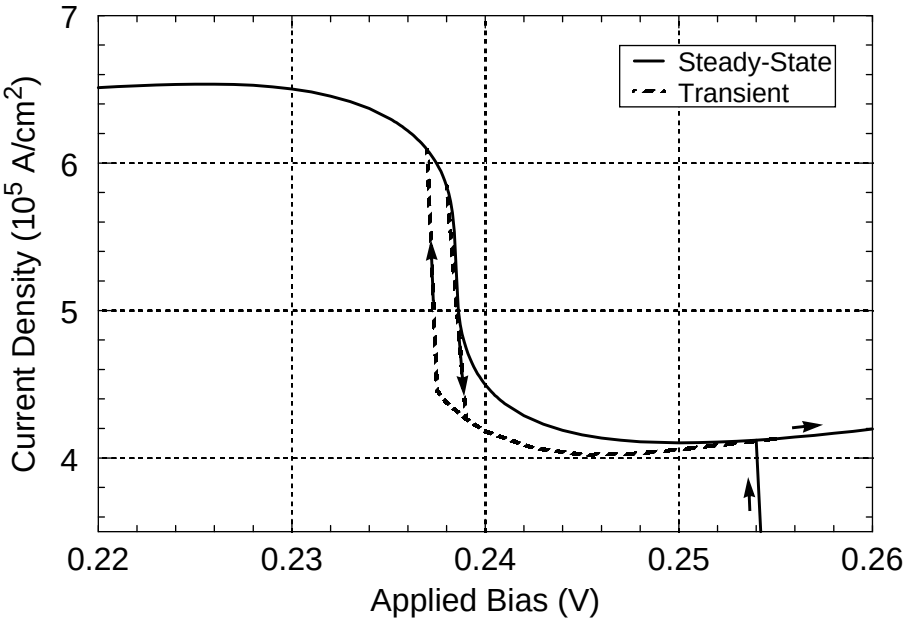


Figure 8.12: Transient hysteresis below main I-V current peak

Steady-state (equilibrium) and transient (time-averaged) I-V curve detail near the upper transition to the plateau. The transient I-V curve has a second hysteresis loop near the main peak of the I-V curve.

suggests that variations in the alignment of the DES and QWS due to charge density variations might produce the plateau oscillations. To investigate further, Figure 8.13 shows charge density and energy band profiles for the minimum and maximum current conditions of Figure 8.11. To see charge variation more clearly, Figure 8.14 shows integrated charge in the emitter and quantum well versus time (the collector charge variation is much smaller). Thus, during oscillations, the emitter and quantum well charges oscillate essentially 180 degrees out of phase with each other. As described in Section 8.2, as the quantum well charge increases, the electric field in the collector barrier increases, and that in the emitter barrier decreases, raising the potential of the QWS further above the DES. The misalignment of the DES and QWS lowers the current from DES to QWS, so that the QWS discharges and the emitter recharges. This lowers the QWS with respect to the DES again, and the cycle repeats. Thus, the plateau oscillations in the JB RTD result from the self-consistent interplay between the charge in the quantum well and that in the accumulation region, and the resulting variation in the alignment of the discrete energy states in these two regions.

The analysis above largely agrees with the detailed description of the physics of pla-

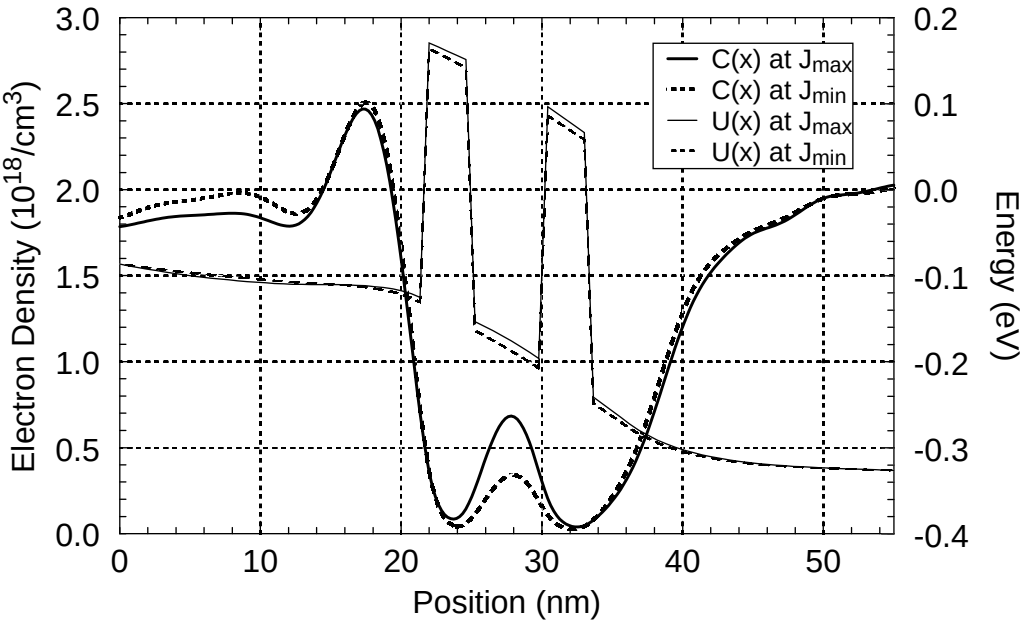


Figure 8.13: Carrier and energy band profiles during oscillations

Self-consistent energy band and electron density during oscillations at 0.24 V. The solid curves correspond to the maximum current; the dashed curves to the minimum. the quantum well potential only varies by about 12 meV.

teau oscillations given by Buot and Rajagopal [8, 9], but differs on one important point. From Figure 8.13, the variation in alignment of the QWS and DES is only about 10 meV. Thus, the occurrence of plateau oscillations requires a discrete energy state in the emitter, since only a narrow energy state would produce, with only a small variation in the QWS energy, a large variation in current into the QWS and corresponding large variation in quantum well charge. Buot and Rajagopal used the Fermi level, rather than the DES energy, as the relevant emitter energy. The Fermi level in this device is 86.4 meV above the emitter contact minimum, making the incoming electron distribution much too wide to produce the observed quantum well charge variation with such small changes in the QWS energy. More importantly, Figure 8.4 shows that the emitter Fermi level is nowhere near the QWS energy while this RTD is operating in the plateau. Even if Buot and Rajagopal meant to describe a quasi-Fermi level in the emitter depression, these oscillations definitely require a discrete emitter state.

The possibility of oscillations occurring in an RTD where a discrete emitter state charges the QWS was first predicted by Ricco and Azbel [57]. However, they did not foresee that the DES must remain below the QWS for a negative-feedback mechanism (in this

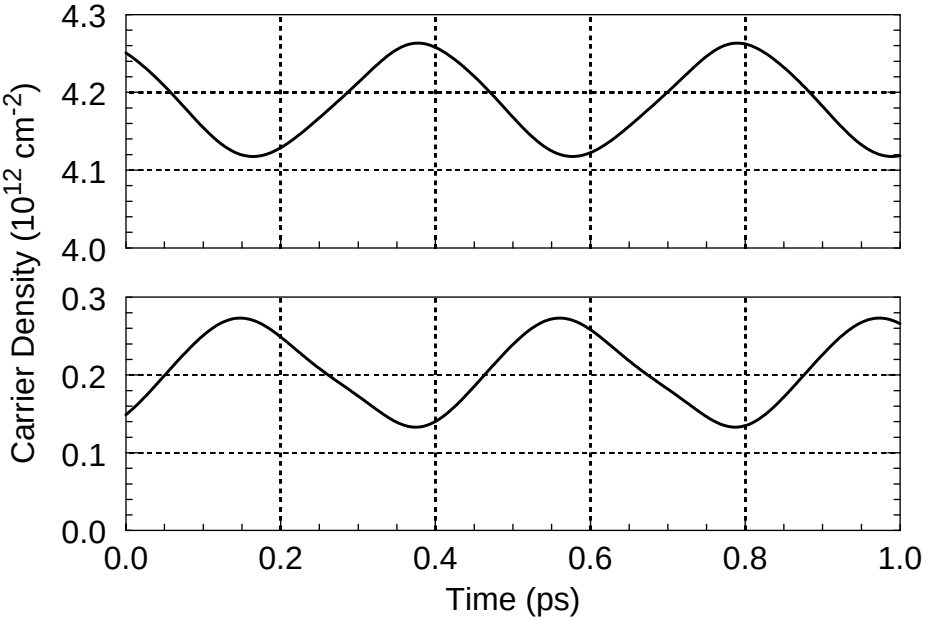


Figure 8.14: Emitter and quantum well charge during oscillations

Total (integrated) charge in the emitter layer (top) and quantum well (bottom) versus time during oscillations at 0.24 V bias. Self-consistency tries to maintain a constant net charge in the device, so a decrease in one region causes an increase in the other, and vice-versa.

case, self-consistency) to maintain this current path. They also suggested that an RTD would never reach steady-state under these circumstances. Simulations in this section (and in Chapter 6) showed conclusively that this RTD *is* stable in the upper portion of the plateau, even though it displays damped oscillations. Explaining why the plateau is partly stable and partly unstable is the second "mystery" concerning transient plateau physics. The answer is quite obvious: one of the requirements of unstable operation of either of the equivalent circuits in Figure 8.1 is that the differential conductance G must be negative. The plateau will only be unstable where the RTD exhibits NDR.[6] Thus, for plateau operation at biases of 0.25 V and below, the RTD will be unstable, while above 0.25 V it will be stable. Since the JB RTD is unstable in the NDR portion of the plateau, it may seem odd that the RTD is stable in NDR portion of the lower I-V curve on the down-trace (0.31 V - 0.255 V). The reason is, of course, that the negative feedback mechanism of the plateau (variation in alignment of the DES and QWS) is not operational except in the plateau. In

6. This statement will be modified slightly in Section 8.4.2.

fact, there is essentially no resonant charge in the QWS at all, so the emitter charge, which is not in a discrete state, has nothing to oscillate out of phase with.Tthis difference between NDR regions will be important in the equivalent circuit analysis of Section 8.4.2.

Finally, this section concludes with an analysis of the third "mystery" concerning the transient operation of this RTD: the cause of the narrow hysteresis loop just below 0.24 V (see Figure 8.12). Since there is no hysteresis in the steady-state I-V curve in the main current peak, the usual causes of hysteresis must be ruled out here: load-line hysteresis due a series resistance greater than the RTDs intrinsic NDR [21, 23, 24], and bistability due to charge storage in the quantum well (see Section 8.1). The cause of this hysteresis must be a dynamic effect. Indeed, the RTD is still oscillating here on the transient down-trace. Figure 8.15 shows the position-averaged current after the RTD is switched from 0.238 V to 0.2375 V (the end of the plateau on the transient down-trace). Since the maximum current during the oscillation is $J = 5.36\times10^5$ A/cm^2, it is clear that the RTD is not oscillating around the equilibrium operating point ($J = 6.0\times10^5$ A/cm^2) found by the steady-state I-V trace. The oscillations somehow cause the RTD to remain in plateau operation (i.e., with an emitter depression and DES/QWS current path) longer than a non-oscillating RTD would. Actually, this I-V curve feature has been shown in RTD equivalent circuit simulations and experimental measurements previously [15, 17, 20, 58]. Sollner [15] used a kind of momentum argument to explain this form of hysteresis: "it is necessary to bias the diode nearer the region of maximum negative conductance to begin oscillations...than to suppress oscillations after they have begun...." Wallis and Teitsworth [25, 58] use the term "subcritical Hopf bifurcation" for this effect. Buot and Rajagopal [59] reported the effects of this as a double hysteresis in the original JB simulations [5], although at that bias (0.24 V), both transient I-V traces do eventually converge to the same time-average current. This transient I-V curve simulation is therefore apparently the first to demonstrate dynamic hysteresis definitively in intrinsic RTD simulations.

8.4 Discussion

Sections 8.2 and 8.3 described the steady-state and transient physics behind the operation of the JB RTD in some detail. This section discusses the significance of these results. Section 8.4.1, points out the main discrepancy between the simulation results and conclusions in this chapter and those of JB [5], and explains the reason for the incorrect simula-

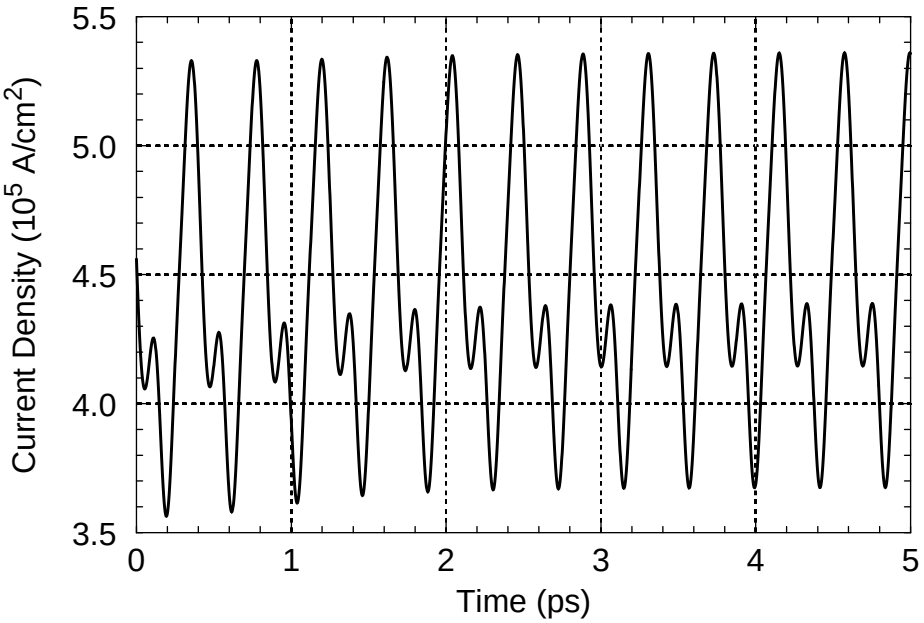


Figure 8.15: Oscillating current in lower state of dynamic bistability

Position-averaged current after the RTD is switched from 0.238 V to 0.2375 V (the end of the plateau on the transient down-trace). Since the maximum current during oscillation is J = 5.36x10^5 A/cm^2, the RTD is not oscillating around the equilibrium operating point (j = 6.0x10^5 A/cm^2) found by the steady-state I-V trace. This indicates a dynamic bistability.

tion results and conclusions of JB. Section 8.4.2 then discusses the revisions that are necessary in other researchers' equivalent circuit analysis based on the JB simulations. Finally, Section 8.4.3 attempts to determine, through further analysis and more accurate simulations, whether and to what extent the foregoing simulations (and those of JB) correctly model the physics and operation of real RTDs.

8.4.1 Plateau Interpretation Error

As mentioned in Section 8.3, the main discrepancy between the transient Wigner function method (WFM) simulation results in this chapter and the nominally identical ones of Jensen and Buot [5] is the fact that the simulations in this work showed the positive differential resistance (PDR) portion of the plateau to be stable, while JB concluded that the entire plateau was unstable. With their conclusion, JB's results matched three related experimental observations very well: high-frequency oscillations throughout a plateau in the NDR region, bistability and hysteresis in the plateau, and the end of the plateau where

the lower I-V curve returns to PDR. Subsequent analysis by Buot et al. [5, 6, 8-11, 60] concluded that the JB simulation results resolved the plateau controversy (see Section 8.1), claiming that the I-V plateau and it associated oscillations were intrinsic in origin. Indeed, the plateau controversy appeared to require exactly such a resolution: oscillations to produce the observed plateau, and charge-bistability to produce the observed hysteresis. As discussed in Section 8.1, it is now clear that extrinsically-induced oscillations can produce all of the observed plateau features. Further, the results in Section 6.5.4 (and [55]) proved conclusively that the JB RTD operating in the PDR portion of the plateau is stable (*i.e.*, does not oscillate). Thus, the claimed resolution of the plateau controversy by Buot et al. must be incorrect, since it does not match observations (oscillations throughout the plateau). Further evidence against the claim of Buot et al. is given in Section 8.4.3.

Since the transient WFM simulations in Section 8.3 were nominally identical to those of JB, it is instructive to determine why JB found the PDR region of the plateau to be unstable. The most importance difference between the transient WFM simulations in this work and that of JB was their use of an "accelerated convergence technique". The idea behind this technique was that if the potential profile is not updated at a given time step, a much less expensive (non-self-consistent) Wigner function update computation can be used. The accelerated convergence technique therefore allowed the potential to remain fixed for up to 50 time steps (*i.e.*, 50 fs) between updates. It is not difficult to see how this might produce oscillations in a region of operation which is marginally stable. However, JB apparently recognized this problem, and switched to the "natural time-evolution approach" (potential updated at each time step) in the plateau. Therefore, the most likely cause of their incorrect conclusion about plateau stability is that only 1600 time steps (*i.e.*, 1600 fs) were allowed per bias point.[7] Indeed, from the instantaneous switching simulations of the JB RTD described in Chapter 6 and [55], the oscillations decayed so slowly after instantaneous switching in the PDR region of the plateau that the RTD did not converge to steady-state in 1600 fs at any bias point in this region of operation. Convergence times for instantaneous switching ranged from over 7500 fs at 0.26 V down to just under 2000 fs at 0.31 V. Thus, terminating the transient simulation prematurely in the PDR portion of the plateau caused JB to incorrectly conclude that the RTD was unstable in this

7. Bias points in JB's plateau simulations were 0.01 V apart, with instantaneous bias switching between.

region of the plateau.

Given the root cause of JB's simulation error, a few related points are worth mentioning. Transient WFM simulations are inherently CPU intensive, as they seek to follow the exact evolution of the quantum device being investigated. Each time step (typically 1 fs) requires the solution of a huge system of equations (assuming "natural", or physically-based, time evolution is used). As a result, both the accelerated convergence technique and the limitation on time steps by JB were a practical response to the limited and expensive computer resources required. True instability in the NDR region of the plateau may have resulted in reduced vigilance in the remainder of the plateau, especially when a fully unstable plateau was expected, based on experimental observations. Indeed, many Cray C90 supercomputer CPU hours were used in the transient WFM simulations of Chapter 6 to verify that the PDR portion of the plateau was, in fact, stable.

In the above discussion, the high computational cost of transient WFM simulations has become an issue once again. As discussed in Chapter 6, for maximum effect, a WFM investigation should make use of the complementary advantages of both steady-state and transient simulations where appropriate. In particular, (efficient) steady-state simulations are appropriate for wide-ranging initial investigations (e.g., to trace the I-V curve or to determine the effects of varying simulation parameters). These results will provide the insight necessary to narrow the focus of a more detailed (and expensive) transient investigation to those cases where dynamic effects are inherent (e.g., switching) or suspected (e.g., oscillations). In this way, the basic operation of the device is known, and adequate computer resources can be applied to a few critical transient simulations, without making significant compromises in implementation or execution. In this case, note that although the transient I-V trace for the JB RTD required roughly 100 times as much CPU time as the steady-state trace, the two I-V curves were almost identical, even though significant dynamic and bistable effects occurred in the NDR region. This confirms the reliability of the steady-state WFM for investigating the basic (i.e., non-transient) operation of quantum devices, without diminishing the importance of properly conducted transient WFM simulations, when necessary.

As another example, transient WFM simulations are *not* generally required, as Buot and Rajagopal claim [59], to trace I-V curves simply because of bistable transition points. This is a strong claim to make without first determining the capabilities of steady-state

WFM simulations. Although only transient simulations can show the transition *process*, this process is generally not shown in an I-V plot. The simulations of the JB RTD in this chapter have demonstrated that a properly implemented steady-state WFM simulation (see Chapter 6 for self-consistency implementation challenges and solutions) can locate bistable transition points accurately when the initial operating point is stable. Finally, even though only transient simulations can model the "exact" evolution of a device, and even if computer resources were infinite and infinitely fast, steady-state simulations can still provide information that a transient simulation can not: the (unstable) equilibrium operating point in an unstable region of operation. The significance of such knowledge was demonstrated in the determination in Section 8.1 that the I-V plateau is largely *not* a dynamic phenomenon, and in the discovery and analysis in Section 8.3 of an actual dynamic bistability and hysteresis loop in the JB RTD.

8.4.2 Equivalent Circuit Analysis

One main investigative thrust based on the JB simulations was an extensive RTD equivalent circuit analysis by Buot et al. [6-9] and Woolard et al. [10, 11] (hereafter referred to as BW). The determination in Section 8.2 that the plateau is a steady-state effect will require some of this work to be substantially revised, at least in relation to the JB RTD. Based on the initial conclusion by JB that experimentally-observed plateau effects and the JB simulated plateau were one and the same, BW drew further analogies with analysis of RTD measurements. They assumed that, as in the experimental case, the plateau was not the "real" I-V curve, but was a purely dynamic effect: the time-average of an oscillating current. They assumed that the intrinsic I-V curve followed the simple RTD behavior described in Section 8.2. Their equivalent circuit analysis work thus centered on trying to produce the JB transient simulation results by adding circuit elements to a bias-dependent conductance $G(V)$ which gave the linear-drop I-V curve (see Figure 8.2). Both the series-inductance model and parallel-inductance model in Figure 8.1 were considered.

Given the initial assumption that the JB RTD was behaving like experimental RTDs, BW followed a very reasonable equivalent circuit analysis. They used a series resistance R_s to shift the linear-drop (*i.e.*, non-self-consistent) I-V curve out to the "normal" RTD behavior without plateau effects (the dashed curve and lower curve in Figure 8.2). The intrinsic capacitance C and an inductance L due to a quantum well charging delay [5]

were proposed to cause the oscillations in the NDR region.[8] The time-average of these oscillations was expected to produce the plateau. The main challenge was to achieve a plateau with positive slope in an oscillating RTD. Admittedly, RLC circuits have been shown to produce all of the simulated effects (oscillations throughout a positive-slope plateau and hysteresis with the down-trace), as we discussed in Section 8.1. However, BW were constrained to use circuit elements which corresponded to the JB simulation results, a constraint which previous circuit analyses did not have. Several attempts were made to explain the positive slope and unstable plateau, which explanations relied on either a complicated interaction of effects [9] or a complicated equivalent circuit with unspecified nonlinear elements [10, 11]. Since BW's initial assumption about the correct steady-state I-V curve was incorrect, these attempts were not completely successful.

Section 8.2 showed that in the JB RTD simulations, the lower curve has nothing to do with plateau operation, so the "normal" I-V curve can not be used in a circuit element to analyze plateau operation. Indeed, the plateau is not purely a dynamic effect (*i.e.*, the average of oscillations), but is a fundamental part of the equilibrium I-V curve (the solid curve in Figure 8.2). Of course, the true equilibrium I-V curve must be used as the starting point for equivalent circuit analysis. The simplest DC equivalent circuit model for this RTD consists of a variable conductance $G_{\mathrm{DC}}(V)$ which produces this I-V curve in parallel with the RTD's DC capacitance $C_{\mathrm{DC}}(V)$, as shown in Figure 8.16. The DC capacitance versus bias can be computed as the displaced charge, Q(V), divided by the applied bias. In Figure 8.16, the positive displaced collector charge is used for Q, since the negative "plate" includes three components in different locations: quantum well, emitter, and emitter contact (see Figure 8.8).

The remainder of this section is outlined as follows. First, starting with the DC equivalent circuit model in Figure 8.16, the elements of the transient equivalent circuit models in in Section 8.1 are described in more detail. Then equivalent circuit analysis is used to explain why the NDR portion of the plateau displays sustained oscillations, while the NDR portion of the lower curve does not. Based on this analysis, some observations and conclusions are made concerning the correct transient equivalent circuit model from the candidates in Figure 8.1. Finally, a slightly modified equivalent circuit model for the JB

8. See Sections 8.2 and 8.3 for a more detailed description of how self-consistency creates a feedback mechanism that produces oscillations in the JB RTD.

RTD is made based on the analysis of the operation of this RTD in the previous sections.

To begin this equivalent circuit analysis of the JB RTD transient I-V curve trace, the origins of the four transient model circuit elements for the two RTD equivalent circuit models in Figure 8.1 will be described in more detail. Extending Figure 8.16, the equilibrium I-V curve $V_{DC}(I)$ is now across the series combination of R_s and $G(V)$. Therefore, $G(V)$ is computed from the hypothetical I-V curve

$$V(I) = V_{DC}(I) - IR_s, \tag{8.3}$$

which essentially skews the equilibrium I-V curve towards lower voltages. The series resistance R_s and inductance L are often attributed to external causes, but since intrinsic device simulations do not include any external effects, R_s and L must have internal causes in this work. For simplicity and comparison to the work of Buot and Jensen [6], a fixed R_s is used here. As mentioned above, inductance L is usually attributed to the delay in current as the QWS charges after an applied bias change across the DBS [9, 48, 49]. However, even when the quantum well charge is negligible, enforcing self-consistency automatically

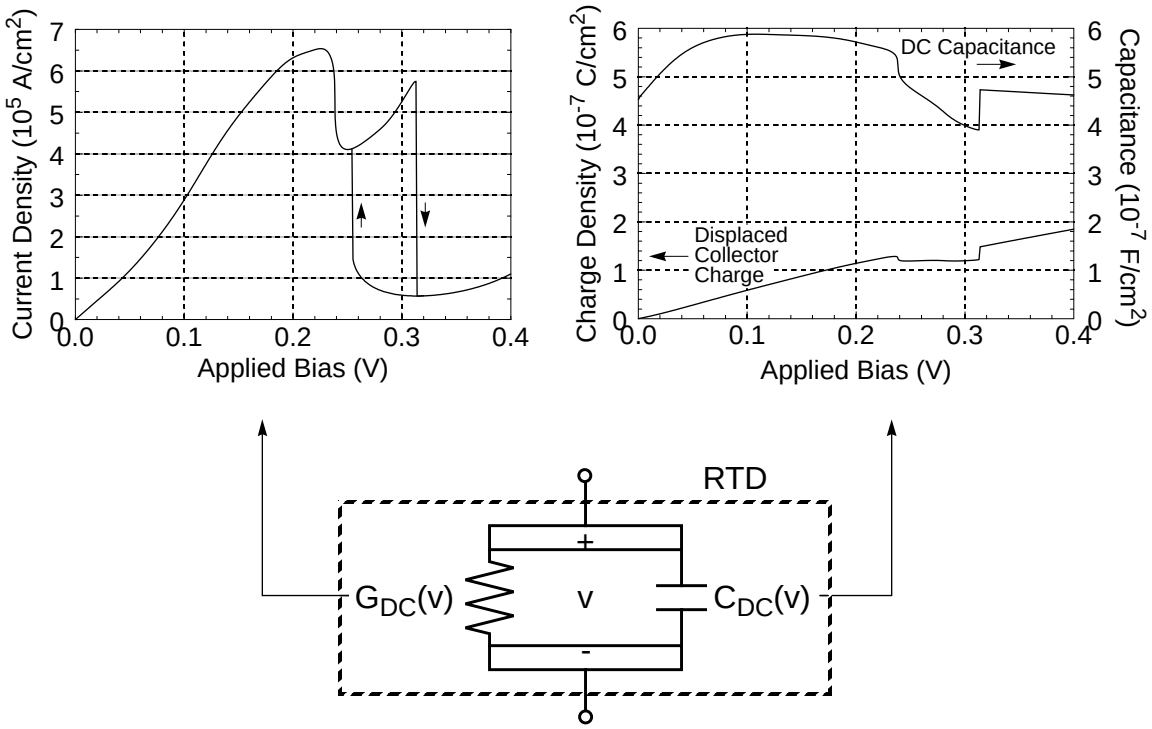


Figure 8.16: DC equivalent circuit for simulated RTD

The bias-dependent resistance is based on the equilibrium I-V curve, and the bias-dependent capacitance is the equilibrium displaced collector charge divided by the applied bias.

results in an LC "ringing" effect, as the device oscillates around a new equilibrium after a quick bias change. Finally, C is the dynamic capacitance of the RTD, which is, in general, not equal to the DC capacitance, as discussed by Buot and Jensen [6].

Even before choosing specific values for, R_s, L, and C, stability criteria will allow the correct circuit model (parallel or series inductance) to be identified, based on circuit stability arguments. In general, Woolard et al. [11] noted that two energy storage elements (in this case L and C) are required for circuit oscillations of any kind. Self-oscillations under DC bias further require an element exhibiting NDR, while R_s serves to damp oscillations. One condition given by Buot and Jensen [6] for sustained oscillation of each circuit is:

$$\text{Series-inductance: } R_s C < L|G|, \tag{8.4}$$

$$\text{Parallel-inductance: } R_s C > L|G|. \tag{8.5}$$

As stated in Section 8.3, note that a negative differential resistance G is required to make either RTD circuit model in Figure 8.1 unstable (*i.e.*, to initiate osciallations at a given operating point).[9]

Now consider the difference in stability of the two NDR regions of the I-V curve from an equivalent circuit viewpoint.[10] The determination of the correct RTD equivalent circuit hinges on the fact that both L and $|G|$ (the local slope of the I-V curve) are smaller in the lower NDR region. It is clear from Figure 8.2 that the slope of the I-V curve is greater at the beginning of the plateau than in the lower NDR region, so $|G|$ is certainly smaller in the lower NDR region. Two independent arguments also show that the inductance L in the lower NDR region is smaller. First, Section 8.3 found that the essential difference between the plateau and lower I-V curve is that quantum well charge is negligible in the lower curve, effectively eliminating the inductive delay due to quantum well charging. Second, it is apparent that the remaining inductance due to self-consistency (*i.e.*, the self-consistent interplay of charge density and energy bands) is significantly smaller, based on oscillation frequency, which generally increases with decreasing inductance. In the NDR portion of the plateau, the (sustained) oscillation frequency is about 2.5 THz, while outside the plateau the (damped) oscillation frequency is roughly 10 THz. With both L and $|G|$ smaller in the lower NDR region, (8.4) would tend to become invalid (no oscillations), while (8.5)

9. However, if the circuit is already oscillating, it may continue to do so *near* an NDR region if external circuit elements allow the applied bias across $G(V)$ to be in the NDR region over at least part of the oscillation cycle.

10. The physics of this behavior was described in Section 8.3.

would be more strongly satisfied (oscillations would occur). Since oscillations do *not* occur in the lower NDR region, (8.4) and the series-inductance RTD circuit model must be correct in this case. This result is contrary to that of Buot and Jensen [6], who concluded that the parallel-inductance model was correct.

Having chosen an RTD circuit model, the task returns to choosing reasonable equivalent circuit elements. Admittedly, it is perhaps futile to attempt to accurately model the complex physics occurring in the JB RTD (as seen in the transient WFM simulations) using a simple lumped-parameter equivalent circuit. However, the gross features of the WFM simulations are not too difficult to reproduce. In contrast to the rather complicated circuit elements proposed by Woolard et al. [10, 11], a simple series-inductance circuit model with constant R_s, L, and C will produce the basic behavior predicted by the transient WFM simulations. Circuit simulations using HSPICE and the RTD equivalent circuit model shown in Figure 8.17 were used to demonstrate this. The device area was assumed to be 1 μm^2, to keep currents in the reasonable range of a few milliamps. These circuit simulations also used C = 5 fF (the approximate average DC capacitance; see Figure 8.16), while R_s = 5 Ω and L = 600 fH were chosen to approximately match circuit and WFM simulations (especially oscillation frequency and amplitude, and the width of RTD instability). The NDR element $G(V)$ was computed directly from the DC I-V curve using (8.3), so the DC I-V curve trace was exactly as simulated by the WFM. The transition from high to low current was accomplished by switching one $G(V)$ element out and simultaneously switching the other in at the correct bias (0.314 V on the up-trace; 0.254 V on the down-trace).

A complete transient I-V up-trace (using continuous bias slewing at 10 mV/ps) for the circuit in Figure 8.17 is shown in Figure 8.18. Note that only the NDR region of the plateau is unstable, as expected. Figure 8.19 shows more detailed simulations of this simple RTD circuit model (Figure 8.17) in the unstable region, including transient bias slewing (at 1 mV/ps) in both directions. Note the dynamic bistability in both directions, due to the fact that oscillations started later in one direction than they ended in the other, for reasons discussed in Section 8.3. It was unnecessary to change the inductance to keep the circuit from oscillating in the lower NDR region - the decrease in negative differential conductance was sufficient.

Simulations of the parallel-inductance circuit (created by connecting C as shown by

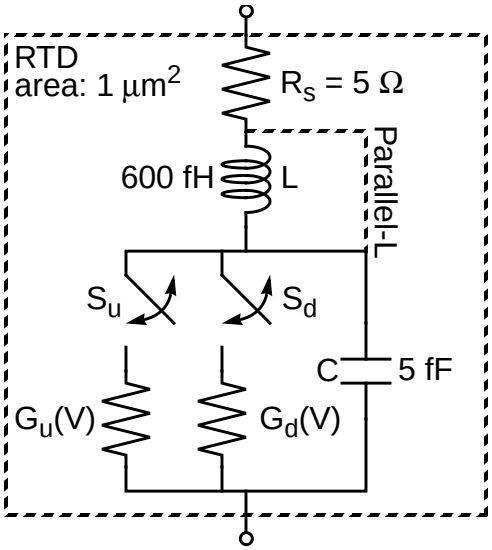


Figure 8.17: RTD equivalent circuit used in HSPICE simulations

RTD series-inductance equivalent circuit model used in HSPICE simulations. The device area is taken as 1 μm^2. The average DC capacitance was used for C, while R_s and L were chosen to approximately match circuit simulations to WFM simulations (frequency and bias range of oscillations). S_u and S_d switch in and out the two I-V curves (upper and lower) at the proper applied bias (0.314 V on the up-trace; 0.254 V on the down-trace). One switch closes as the other opens. The parallel-inductance model used by JB is formed by simply moving the capacitor terminal to the other side of the inductor, as shown by the dashed line.

the dashed-line in Figure 8.17) were also attempted. Although DC simulations gave the correct I-V curve, transient HSPICE simulations were unable to converge in either NDR region. As a result, it was not possible to meaningfully investigate the behavior of the parallel-inductance model. Returning to the series-inductance model, a more elaborate equivalent circuit model could be developed in an attempt to more accurately match WFM simulations of the JB RTD. However, as discussed in the next section, this effort is perhaps not worthwhile, since the WFM simulation results themselves may be substantially inaccurate.

8.4.3 Simulation Accuracy

This chapter has shown that the connection is merely visceral between the transient Wigner function simulations of the JB RTD and experimental RTD measurements showing oscillations resulting in a plateau in the NDR region of operation. Further, these WFM simulation results seem quite suspicious, considering that the effects predicted in the RTD

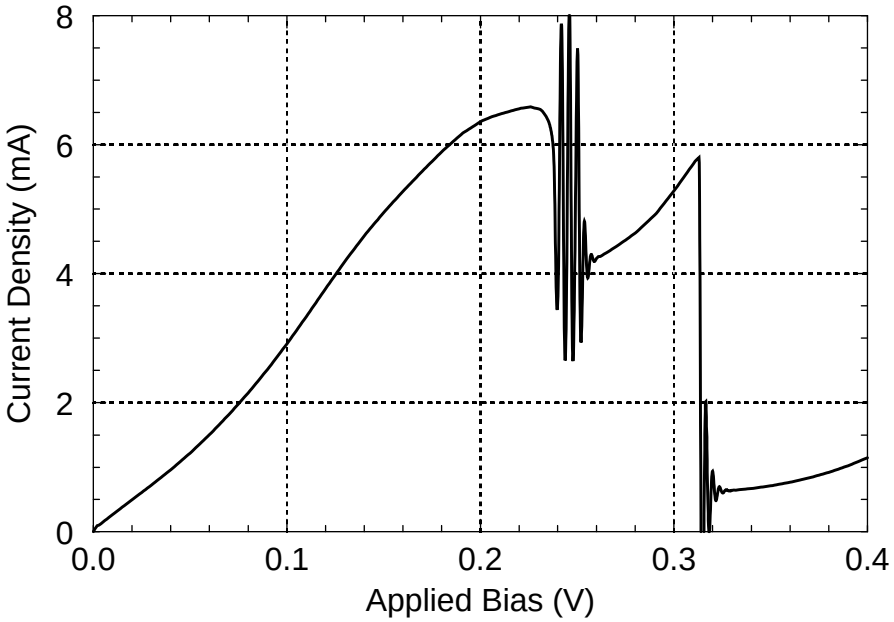


Figure 8.18: HSPICE simulated I-V curve trace

Transient I-V curve (continuous bias slewing at 10 mV/ps) for the series-inductance RTD circuit model shown in Figure 8.17. Only the NDR portion of the plateau is unstable. Elsewhere, the transient I-V curve follows the DC I-V curve. At the bistable transition, switching from the high to low curve causes a brief and highly damped oscillation.

simulations have apparently not been observed experimentally. In particular, in the main current peak where intrinsic bistability due to charge storage in the quantum well might be expected experimentally, the JB RTD instead showed a hysteresis loop due to unstable oscillations. And in the NDR region where experimental results often show an I-V plateau and hysteresis due to unstable oscillations, the JB RTD simulations instead predicted a plateau due to a potential depression and discrete energy state in the emitter, and hysteresis due to intrinsic bistability.

The question considered in this section is whether the effects predicted by the Wigner function simulations should occur in measurements of the JB RTD, or whether they are simply artifacts of inaccurate simulations. If the effects are real, then they can undoubtedly be used in quantum functional devices. If they are not, then more attention should be paid in the future to improving and verifying the accuracy of WFM simulation results. This section attempts to asses the accuracy and reliability of the foregoing WFM simulation results by executing more accurate (and correspondingly more expensive) Wigner function simulations. The determination will hinge on whether and to what degree these simu-

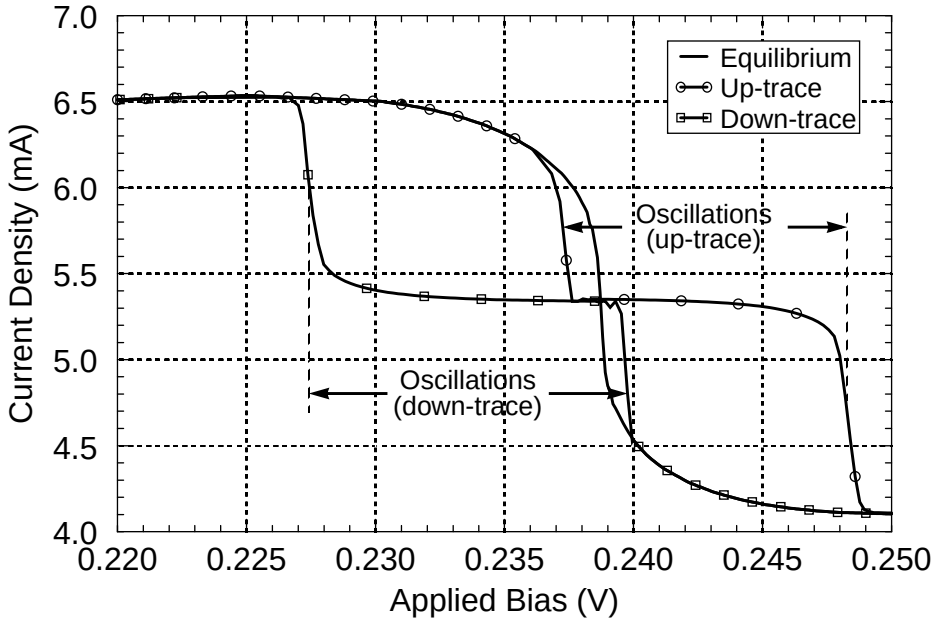


Figure 8.19: Detail of HSPICE I-V curve showing dynamic bistability

Transient I-V curve (continuous bias slewing at 1 mV/ps) for the series-inductance RTD circuit model shown in Figure 8.17. The unstable portion of the I-V curve is detailed, including both the up-trace and down-trace. In each trace, oscillations started later than they ended on the opposite trace, resulting in dynamic bistabilities. The maximum and minimum of the oscillations for both traces were about 8.1 mA and 2.5 mA, respectively. The oscillation frequency was 2.5 THz, as in the WFM simulations.

lation results differ from the previous ones.

The most obvious possible source of inaccuracy of the WFM simulations above is indicated by the high electric field at the emitter contact during plateau operation. A high e-field at a contact indicates (in addition to a charged contact) that the simulation results will not be independent of the simulation region boundary location. To accurately model experimental devices which are wider, a wider simulation width should be used. Even for RTDs which are this narrow, non-equilibrium boundary conditions (e.g., drifted Fermi-Dirac) boundary conditions [61] should be used to improve simulation accuracy. Most experimental RTDs are many times the 55 nm simulation width used by JB and in this work, so the previous simulation results may say little about the operation of most experimental RTDs. In particular, since the emitter contact e-field was significant for plateau operation, the interesting physics (which all occurred in the plateau) could be entirely a result of choosing too small of simulation width. To determine this, steady-state and tran-

sientWFM simulations were run with a wider emitter to determine the effect of emitter width on simulated RTD operation. To prevent any effect from position grid spacing changes, the grid spacing was maintained by simply increasing the number of position points in the simulation.

In short, the emitter width did indeed have a significant influence on the I-V plateau. For example, Figure 8.20 shows the equilibrium I-V curve for a 63 nm emitter layer. The narrow (19 nm) emitter I-V curve is shown for comparison. Clearly, using a narrow emitter forces the RTD into plateau operation at lower biases and more abruptly, and prolongs plateau operation to higher biases, as compared to the wide emitter RTD. Nevertheless, the I-V plateau still occurs in the wide emitter RTD. Examination of other simulation results showed that the plateau is caused in the same manner, and displays all of the same features, as in the narrow emitter RTD. Figure 8.21 shows the energy band profile for the wide emitter RTD operating in the plateau at 0.28 V. Note that the emitter contact e-field is small, as intended. This indicates that the formation of an emitter depression and the resulting I-V plateau are not simply the result of inaccurate boundary conditions in the narrow emitter RTD simulations.

The wide-emitter RTD also self-oscillates in the plateau. For example, at 0.27 V, the final oscillation amplitude was almost 10^5 A/cm^2 around an average value of 4.4×10^5 A/cm^2, and at a frequency of just under 2.5 THz. Other wide-emitter simulations showed that the effect of emitter width changes was minimal, even in plateau operation, for emitter widths above about 50 nm (versus 19 nm in our previous simulations, not including the 3 nm buffer layer). As expected, the emitter width made little difference for RTD operation outside the I-V plateau, since the emitter contact e-field was low here, even in the narrow-emitter RTD. For the same reason, increasing the collector width made negligible difference in the equilibrium I-V curve under any conditions.

Figure 8.20 also shows that hysteresis and bistability still occur in the wide emitter RTD, but note that the hysteresis no longer has the visceral connection to the plateau of experimental I-V traces. In particular, the up-trace transition point no longer occurs where the lower I-V curve returns to PDR, in contrast to experiments. Also the transition from lower to upper trace is direct, since it occurs at biases below the plateau. In fact, except for the small plateau, the hysteresis now appears much more like experimental observations of intrinsic bistability, where the hysteresis loop appears in the main I-V peak. This suggests

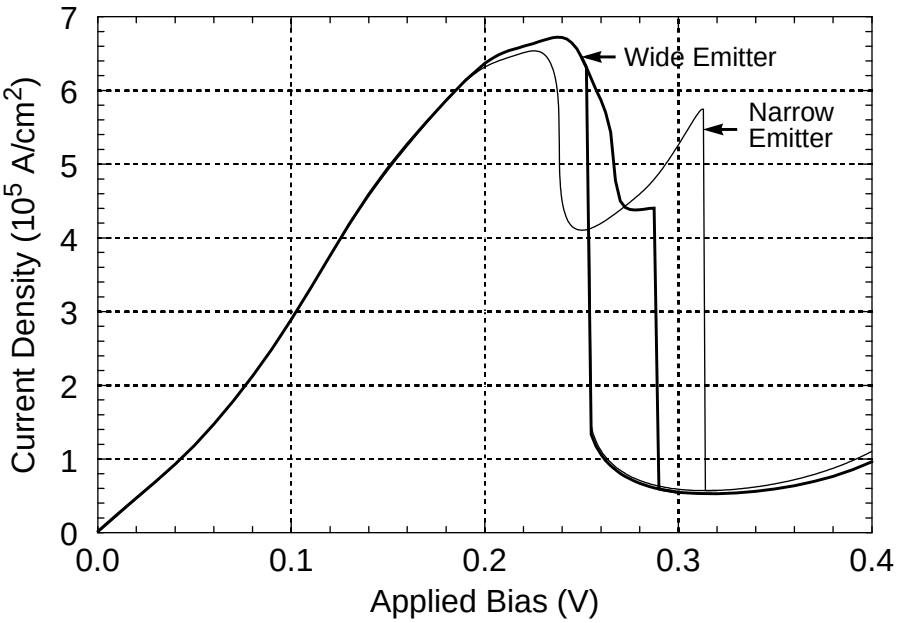


Figure 8.20: Steady-state I-V curve for wide-emitter RTD

Wide (63 nm) emitter equilibrium I-V curve. The narrow (19 nm) emitter I-V curve is shown for comparison. Using a narrow emitter causes the RTD to begin plateau operation at lower biases and more abruptly, and prolongs it to higher biases, than the wide emitter RTD. However, the I-V plateau still occurs in the wide emitter RTD, and is caused in the same manner as in the narrow emitter RTD.

that further improvements in the accuracy of the simulation may remove the plateau (and the associated emitter depression and oscillations) entirely. This would bring the WFM simulation results into parity with experimental results: hysteresis due to intrinsic bistability appears as an I-V loop in the main current peak, and any plateau would be due to externally-induced oscillations. The WFM simulations would then support the existing consensus in these RTD operation controversies, as discussed in Section 8.1.

One other obvious source of concern relating to the accuracy of WFM simulations is the need to use a relatively small number N_k of wavenumber grid points. As reported by Frensley [50], memory usage in WFM simulations is proportional to $N_x N_k^2$, and computation increases with $N_x N_k^3$, where N_x is the number of position points. Thus, while transfer-matrix method simulations typically use thousands of energy values, it is very costly to use even 100 wavenumber points in WFM simulations. However, as computing power increases, it will be feasible to refine the energy (wavenumber) spectrum of WFM simulations, even with the inherently costly transient simulations. Some ambitious steady-state

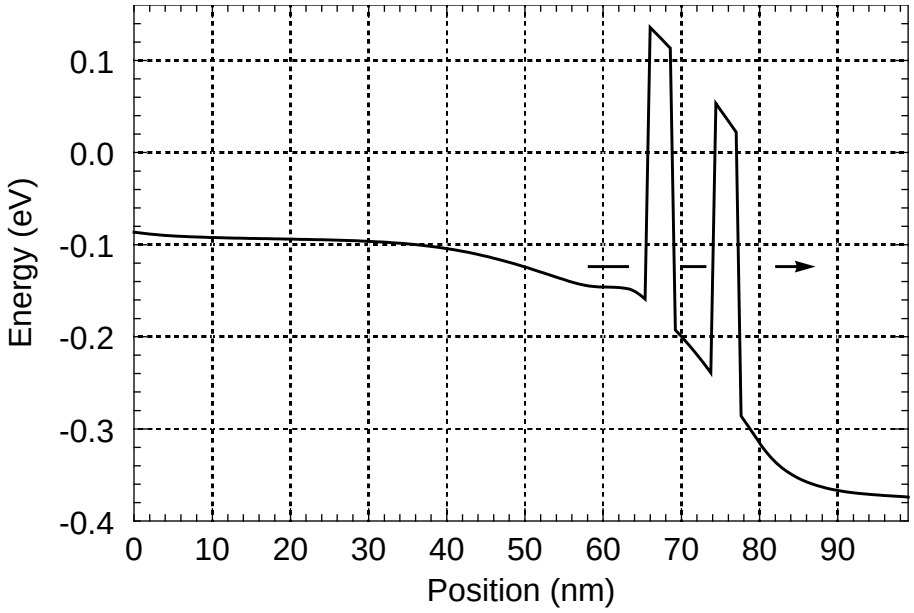


Figure 8.21: Energy band profile for wide emitter RTD in plateau

Wide emitter energy band profile during plateau operation at 0.28 V. Also indicated are the positions of the DES and QWS (found using transfer-matrix analysis). Note that the emitter contact e-field is small, as intended. This shows that the formation of an emitter depression and the resulting I-V plateau are not simply the result of inaccurate boundary conditions in the narrow emitter RTD simulations.

simulations have been reported by Gullapalli et al. [62] using $N_x = 200$ and $N_k = 144$. However, the JB RTD has not been investigated in this much detail, and no transient WFM simulations approaching this magnitude have been reported to date. Such large simulations remain in the future of WFM research. A final and less obvious reason that the JB and similar WFM simulations may be inaccurate was discussed in Section 5.5.7: the possibility that the standard implementation of the discrete WFTE may be inaccurate.

Based on the wide emitter simulation results, it appears that much of the interesting behavior of the JB RTD may be an artifact of inaccurate WFM simulations. The obvious question is whether this would render worthless the results of previous sections, as well as work by other researchers based on the JB simulations. The answer has several parts. The first is that these WFM simulations were very useful in the study of WFM simulation. In fact, more is usually learned from imperfection than from perfection. Of course, in recognizing the WFM simulation errors using hindsight, there is no intention to suggest that JB should have used different simulation parameters or device size, since they were at the limits of available computing power with their pioneering work. JB produced the first

credible transient WFM simulations including both self-consistency and scattering. Further, previous sections gave an analysis of the JB simulation results, not experimental results, so in this sense, the simulation parameters were necessary, and the analysis was entirely accurate.

The second part of the question of whether simulations of the JB RTD to date have been worth while focuses on the narrowest implications of these simulations: whether they produced an accurate description of the operation of the JB RTD itself. While most of the I-V curve is qualitatively correct, the NDR portion, in which most of the interesting "physics" occurred, and around which most subsequent investigations focused, was incorrect to some degree. In this sense, the simulation results have been somewhat misleading.

However, the third part of the question is whether these investigations have been worth while for the study of quantum devices in general. Here the answer is yes. The JB RTD may not exhibit these interesting behaviors experimentally, but a similar device which *did* have a potential depression in the emitter would. Indeed, experimental RTD measurements showing effects due to emitter potential wells and resulting discrete emitter states have been claimed or demonstrated [26, 30, 42, 63], although the structures are usually specially designed for these results, unlike the very conventional JB RTD. All of the phenomena described (quantum and otherwise, including tunneling, interference, scattering, self-consistency, carrier transport, etc.) were self-consistent, and thus reasonably describe a quantum system tot he accuracy ofthe model. Therefore most of the analysis of the JB simulations (in this work and in that of others) remains relevant to the investigation of quantum devices in general, although not all of it is relevant to the JB RTD specifically.

The discussion of this section shows that even after more than a decade of active development, producing accurate WFM simulations is still very difficult. More care must be taken in the future to check the accuracy of WFM simulation results before using them to draw conclusions about experimental observations. Due to this difficulty, it is not surprising that, in a survey of over 40 papers by numerous groups (including our own) describing Wigner function simulations of RTDs, only a single paper [64] showed experimental I-V measurements of the same device. The reason for the quantitative disconnect between experiment and Wigner function simulation is simply that the typically large disparity between the results, whether due to inaccurate simulations or poor device quality, has generally made such comparisons useless. However, qualitative agreement continues to

improve greatly. As a result, Wigner function simulation is already a useful tool for the investigation of quantum devices, as this chapter has demonstrated. In general, dynamic carrier transport, scattering, self-consistency, quantum confinement, and open boundaries are all important features of real RTDs and other quantum devices, and the WFM can do a good job modeling these. In fact, all of these effects were in play in the simulated plateau operation of the JB RTD.

8.4.4 RTD Physics Controversies

Before concluding, the RTD operation controversies that started this work will now be considered in light of the foregoing simulations an analysis. The wide-emitter simulations in the previous section have changed some of the conclusions regarding these issues. Until that point, it appeared that neither of the two main explanations for the plateau, extrinsic oscillations or intrinsic bistability, were correct. Instead, a third possibility described in Section 8.1, current flow through a discrete emitter state, was shown to be the cause of the I-V plateau. Based on the more accurate simulations and lack of experimental corroboration, the plateau may eventually be shown to be an artifact of inaccurate WFM simulations. Thus, the consensus explanation of the plateau, extrinsically-induced oscillations, has not been weakened by JB RTD simulations, contrary to previous claims. The related controversy concerning the origin of observed oscillations in the I-V plateau is resolved similarly (as far as the JB RTD simulations are concerned): without an intrinsic plateau, RTD oscillations must be extrinsically-induced, rather than purely intrinsic.

Although the simulated operation of the JB RTD in the plateau does involve two current paths, one of which requires scattering, Section 8.2 concluded that these WFM simulations can not speak to the resonant-versus-sequential tunneling controversy. However, these simulations do have repercussions for the other two RTD operation controversies discussed. The first concerned the correct RTD equivalent circuit model. The conclusion in Section 8.4.2 was that the consensus view was again supported, that the series-inductance model is correct for the JB RTD. Note that these results were based on the behavior of the JB RTD in the I-V plateau. It is unclear which circuit model would be supported if the plateau does not occur in more accurate WFM simulations. The final controversy discussed was whether and how RTDs demonstrate intrinsic bistability. Once again, more accurate (*i.e.*, wide-emitter) WFM simulations indicated that the consensus is correct:

intrinsic bistability manifests as a hysteresis loop in the main I-V peak, not as a hysteresis loop in the NDR region, as shown in previous JB RTD simulations.

8.5 Summary

This chapter has revisited the very intriguing transient Wigner function method simulations of a resonant tunneling diode published about 5 years ago by Jensen and Buot. The advancement of available computing power in the interim made it possible for this more detailed yet more comprehensive investigation of the JB RTD. The simulation results in this chapter differ from those of JB on some key points. First, steady-state WFM simulations showed that the I-V curve plateau, which was previously ascribed to dynamic effects, is actually an equilibrium phenomenon. Detailed analysis of both WFM and TMM simulation results revealed the origin of the plateau and related effects. In short, during plateau operation, two parallel current paths are operating. Along the first current path, electrons scatter into a discrete quantum state in a potential depression that develops in the emitter, and then tunnel through the resonant state in the quantum well. Because of the emitter depression, the height of the tunnel barriers is greatly reduced for electrons which do not scatter into the depression. The second current path is due to electrons which tunnel directly through the lowered double barrier structure. It is this second current path that is responsible for the positive slope of the plateau in these simulations, a feature which has been the focus of some speculation and analysis for both the JB RTD and similar devices.

Transient Wigner function simulations of the JB RTD in this work also differed in several respects with previous results. First, the I-V plateau was shown to be only partly unstable, while previous results concluded that it was unstable throughout. In fact, because the plateau was an equilibrium feature, only the NDR portion of the plateau *could* be unstable, while the PDR portion (the majority) must ultimately be stable. Previous descriptions of the cause of the plateau oscillations were largely confirmed: self-consistent interaction of the charge in the emitter and quantum well, resulting in out-of-phase oscillations of these charges. One new discovery was that a discrete energy state in the emitter is required to produce the oscillations and the abrupt termination of the plateau. It was the oscillation in alignment of the discrete states in the emitter and quantum well that modulated the current. Finally, the discovery was made of a second, albeit smaller, hysteresis loop below the main current peak on the I-V curve. While the main hysteresis loop (below

the plateau) is not a dynamic bistability, contrary to previous conclusions, this second hysteresis loop *is* dynamic, since only the transient simulations (with oscillating current) showed this feature.

The main difference between this work and previous work of JB was the respective conclusions about the stability of the plateau. The probable reason for the error by JB was found to be inadequate iterations of the transient simulation, so that the RTD did not have time to reach steady-state, and thus appeared to be unstable. This incorrect conclusion was very attractive, since it matched perfectly the symptoms of experimentally-observed I-V plateaus. Given the high computational cost of WFM simulations, the best way to avoid such errors is to use the complimentary advantages of both steady-state and transient simulations for best effect. In the simulations described here, steady-state simulation results uncovered the physics behind the plateau, while transient simulations were needed to detail its stability characteristics.

The appropriate equivalent circuit model for the JB RTD was also investigated. Only the series-inductance circuit model could match the behavior of the JB RTD in the NDR region of operation. By contrast, previous analysis of JB RTD simulations had concluded that a parallel-inductance model was correct. Based on the determination that the plateau is an equilibrium effect, and that it is only unstable in the NDR portion, the values of the circuit elements necessary to reproduce the essential features of the transient I-V curve trace were significantly simplified compared to previous analysis. A slightly more detailed equivalent circuit model was also proposed, which explicitly included the two current paths discovered in the steady-state simulations.

This work concluded that the simulated behavior of the JB RTD, although viscerally similar to experimental results, was of a fundamentally different origin. In particular, the NDR region effects seen in these WFM simulations were caused by a potential depression in the emitter, rather than high-frequency oscillations. Since the emitter depression is not seen experimentally in simple RTDs such as that investigated herein, the accuracy of these WFM simulations was called into question. Three possible sources for this inaccuracy were identified: the simulated RTD was narrower than experimental devices, resulting in suspicious boundary conditions; too few wavenumber points were used to accurately resolve phenomena in the energy dimension; and the implementation of the WFM (used by all researchers to date) may be inaccurate. All of these potential sources of inaccuracy

reflect an attempt to mitigate the inherently high cost of WFM simulations. The latter two items remain to be investigated in future work. However, simulations using a wide emitter layer with the JB RTD were described. This was intended to make the equilibrium boundary conditions more self-consistent (i.e., have a low e-field at the contact). This brought the RTD simulations in much closer agreement with experiment. For example, the equilibrium plateau (not seen in experiment) nearly disappeared. Similarly, the intrinsic bistability loop then appeared nearly as it does experimentally - as a bistability loop beneath the main current peak.

In spite of the remaining innacuracies in Wigner function simulations, this chapter has demonstrated that the WFM can produce and self-consistently model all of the phenomena that occur in real quantum devices. These include dynamic carrier transport, scattering, self-consistency, quantum confinement, tunneling, and open boundaries. No other quantum device simulation method yet devised has shown this range of capabilities. Unfortunately, this work also demonstrated that although WFM simulation of RTDs has advanced swiftly over the past decade, it is still experiencing growing pains as the amount of computing resources required to produce accurate results with it becomes apparent. Clearly, researchers (ourselves included) need to make more certain in the future that their WFM simulations accurately model real systems before drawing conclusions about the physics or operation of real systems. However, note that the wide-emitter simulations supported, either implicitly or explicitly, each of the fairly well established consensus views (described at the beginning of the chapter) on the RTD operation controversies, in contrast to previous work. By this measure alone, this work and this quantum device simulation tool have made a significant contribution to advancing the accuracy of quantum device analysis through simulation.

References

[1] L. L. Chang and R. Tsu. “Resonant tunneling in semiconductor double barriers.” *Applied Physics Letters*, 24(12):593–595, 1974.

[2] F. Capasso, K. Mohammed, and A. Y. Cho. “Resonant tunneling through double barriers, perpendicular quantum transport phenomena in superlattices, and their device applications.” *Journal of Quantum Electronics*, 22(9):1853–1868, 1986.

[3] S. Luryi. “Observation of intrinsic bistability in resonant tunneling diode model-

ing." In F. Capasso and G. Margaritondo, editors, *Heterojunction Band Discontinuities: Physics and Device Applications*, chapter 12. Elsevier Science Publishers, 1987.

[4] F. Capasso, S. Sen, F. Beltram, L. M. Lunardi, A. S. Vengurlekar, P. R. Smith, N. J. Shah, R. K. Malik, and A. Y. Cho. "Quantum functional devices: Resonant-tunneling transistors, circuits with reduced complexity, and multiple-valued logic." *IEEE Transactions on Electron Devices*, 36(10):2065–2082, 1989.

[5] K. L. Jensen and F. A. Buot. "Numerical simulation of intrinsic bistability and high-frequency current oscillations in resonant tunneling structures." *Physical Review Letters*, 66(8):1078–1081, 1991.

[6] F. A. Buot and K. L. Jensen. "Intrinsic high-frequency oscillations and equivalent circuit model in the negative differential resistance region of resonant tunneling devices." *COMPEL*, 10(4):241–253, 1991.

[7] F. A. Buot. "Mesoscopic physics and nanoelectronics: Nanoscience and nanotechnology." *Physics Reports*, 234(2-3):73–174, 1993.

[8] F. A. Buot and A. K. Rajagopal. "High-frequency behavior of quantum-based devices: Equivalent-circuit, nonperturbative-response, and phase-space analyses." *Physical Review B*, 48(23):17217–17232, 1993.

[9] F. A. Buot and A. K. Rajagopal. "Theory of novel nonlinear quantum transport effects in resonant tunneling structures." *Materials Science and Engineering*, B35(1-3):303–317, 1995.

[10] D. L. Woolard, F. A. Buot, D. L. Rhodes, X. Lu, and B. S. Perlman. "An assessment of potential nonlinear circuit models for the characterization of resonant tunneling diodes." *IEEE Transactions on Electron Devices*, 43(2):332–341, 1996.

[11] D. L. Woolard, F. A. Buot, D. L. Rhodes, X. J. Lu, R. A. Lux, and B. S. Perlman. "On the different physical roles of hysteresis and intrinsic oscillations in resonant tunneling structures." *Journal of Applied Physics*, 79(3):1515–1525, 1996.

[12] Private conversation with K. L. Jensen.

[13] E. R. Brown, W. D. Goodhue, and T. C. L. G. Sollner. "Fundamental oscillations up to 200 GHz in resonant tunneling diodes and new estimates of their maximum oscillation frequency from stationary-state tunneling theory." *Journal of Applied Physics*, 64(3):1519–1529, 1988.

[14] T. J. Shewchuk, J. M. Gering, P. C. Chapin, D. P. Coleman, W. Kopp, C. K. Peng, and H. Morkoc. “Stable and unstable current-voltage measurements of a resonant tunneling heterostructure oscillator.” *Applied Physics Letters*, 47(9):986–988, 1985.

[15] T. C. L. G. Sollner. “Comment on ’observations of intrinsic bistability in resonant-tunneling structures’.” *Physical Review Letters*, 59(14):1622, 1987.

[16] J. F. Young, B. M. Wood, H. C. Liu, M. Buchanan, and D. Landheer. “Effect of circuit oscillations on the dc current-voltage characteristics of double barrier resonant tunneling structures.” *Applied Physics Letters*, 52(17):1398–1400, 1988.

[17] H. C. Liu. “Simulation of extrinsic bistability of resonant tunneling structures.” *Applied Physics Letters*, 53(6):485–486, 1988.

[18] E. S. Hellman, K. L. Lear, and J. S. Harris. “Limit cycle oscillation in negative differential resistance devices.” *Journal of Applied Physics*, 64(5):2798–2800, 1988.

[19] H. C. Liu. “Circuit simulation of resonant tunneling double-barrier diode.” *Journal of Applied Physics*, 64(9):4792–4794, 1988.

[20] C. Y. Belhadji, K. P. Martin, J. J. L. Rascol, R. J. Higgins, R. C. Potter, H. Hier, and E. Hempfling. “Bias circuit effects on the current-voltage characteristic of double-barrier tunneling structures: Experimental and theoretical results.” *Applied Physics Letters*, 57(1):58–60, 1990.

[21] C. Kidner, I. Mehdi, J. R. East, and G. I. Haddad. “Bias circuit instabilities and their effect on the D.C. current-voltage characteristic of double-barrier resonant tunneling diodes.” *Solid State Electronics*, 34(2):149–156, 1991.

[22] B. Jogai and E. T. Koenig. “A parametric study of extrinsic bistability in the current-voltage curves of resonant-tunneling diodes.” *Journal of Applied Physics*, 69(5):3381–3383, 1991.

[23] J. Chen, J. G. Chen, C. H. Yang, and R. A. Wilson. “The I-V characteristics of double-barrier resonant tunneling diodes: Observation and calculation of their temperature dependence and asymmetry.” *Journal of Applied Physics*, 70(6):3131–3136, 1991.

[24] A. D. Martin, M. L. F. Lerch, P. E. Simmonds, and L. Eaves. “Observation of intrinsic tristability in a resonant tunneling structure.” *Applied Physics Letters*, 64(10):1248–1250, 1994.

[25] C. R. Wallis and S. W. Teitsworth. “Hopf bifurcations and hysteresis in resonsant tunneling diode circuits.” *Journal of Applied Physics*, 76(7):4443–4445, 1994.

[26] V. J. Goldman, D. C. Tsui, and J. E. Cunningham. “Observation of intrinsic bistability in resonant-tunneling structures.” *Physical Review Letters*, 58(12):1256–1259, 1987.

[27] V. J. Goldman, D. C. Tsui, and J. E. Cunningham. “Reply to comment on ’observation of intrinsic bistability in resonant-tunneling structures’.” *Physical Review Letters*, 59(14):1623, 1988.

[28] F. W. Sheard and G. A. Toombs. “Space-charge buildup and bistability in resonant-tunneling double-barrier structures.” *Applied Physics Letters*, 52(15):1228, 1988.

[29] Y. Fu and M. Willander. “Charge accumulation and band edge in the double barrier tunneling structure.” *Journal of Applied Physics*, 71(8):3877–3882, 1992.

[30] N. Tabatabaie and M. C. Tamargo. “Determination of elastic tunneling traversal times.” In *International Elestron Devices Meeting (IEDM) Technical Digest*, pages 80–83, 1986.

[31] H. L. Berkowitz and R. A. Lux. “Hysteresis predicted in I-V curve of heterojunction resonant tunneling diodes simulated by a self-consistent quantum method.” *Journal of Vacuum Science and Technology B*, 5(4):967–970, 1987.

[32] T. Baba and M. Mizuta. “Simulation of intrinsic bistability in resonant tunneling diodes.” *Japanese Journal of Applied Physics*, 28(8):L1322–L1325, 1989.

[33] D. Landheer and G. C. Aers. “Role of carrier equilibrium in self-consistent calculations for double barrier resonant diodes.” *Superlattices and Microstructures*, 7(1):17–21, 1990.

[34] M. Rahman and J. H. Davies. “Theory of intrinsic bistability in a resonant tunneling diode.” *Semiconductor Science and Technology*, 5:168–176, 1990.

[35] S. M. Booker, F. W. Sheard, and G. A. Toombs. “Current bistability in resonant tunneling double barrier diodes.” *Superlattices and Microstructures*, 9(1):111–114, 1991.

[36] F. W. Sheard and G. A. Toombs. “Theory of resonant tunneling between 2-D electron systems.” *Semiconductor Science and Technology*, 7:B460–B461, 1992.

[37] Y. Abe. “Bifurcation of resonant tunnelling current due to the accumulated electrons in a well.” *Semiconductor Science and Technology*, 7:B498–B501, 1992.

[38] Y. Hu and S. Stapleton. “Self-consistent model of a double-barrier resonant tunneling diode: Dependence of intrinsic bistability on structural parameters.” *Journal of Applied Physics*, 73(10):5254–5263, 1993.

[39] E. S. Alves, L. Eaves, M. Henini, O. H. Hughes, M. L. Leadbeater, F. W. Sheard, G. A. Toombs, G. Hill, and M. A. Pate. “Observation of intrinsic bistability in resonant tunneling devices.” *Electronics Letters*, 24(18):1190–1191, 1988.

[40] A. Zaslavsky, V. J. Goldman, and D. C. Tsui. “Resonant tunneling and intrinsic bistability in asymmetric double-barrier heterostructures.” *Applied Physics Letters*, 53(15):1408–1410, 1988.

[41] M. L. Leadbeater, E. S. Alves, L. Eaves, M. Henini, O. H. Hughes, F. W. Sheard, and G. A. Toombs. “Charge build-up and intrinsic bistability in an asymmetric resonant-tunnelling structure.” *Semiconductor Science and Technology*, 3:1060–1062, 1988.

[42] C. R. Wie and Y. W. Choi. “Designing resonsant tunneling diode structures for increased peak current density.” *Applied Physics Letters*, 58(10):1077–1079, 1991.

[43] S. Luryi. “Mechanism of operation of double-barrier resonant-tunneling oscillators.” In *International Elestron Devices Meeting (IEDM) Technical Digest*, pages 666–669, 1985.

[44] T. Weil and B. Vinter. “Equivalence between resonant tunneling and sequential tunneling in double-barrier diodes.” *Applied Physics Letters*, 50(18):1281–1283, 1987.

[45] M. Buttiker. “Coherent and sequential tunneling in series barriers.” *IBM Journal of Research and Development*, 32(1):63–75, 1988.

[46] R. Gupta and B. K. Ridley. “The effect of level broadening on the tunneling of electrons through semiconductor double-barrier quantum-well structures.” *Journal of Applied Physics*, 64(6):3089–3097, 1988.

[47] S. Luryi. “Coherent versus incoherent resonant tunneling and its implications for fast devices.” *Superlattices and Microstructures*, 5(3):375–382, 1989.

[48] J. M. Gering, D. A. Crim, D. G. Morganb, P. D. Coleman, W. Kopp, and H. Morkoc. “A small-signal equivalent-circuit model for gaas-algaas resonant tunneling heterostructures at microwave frequencies.” *Journal of Applied Physics*, 61(1):271–276, 1987.

[49] E. R. Brown and T. C. L. G. Sollner. “Resonant-tunneling: Physics and modeling.” In *7th International Workshop on Future Electron Devices, Superlattices, and Quantum Functional Devices*, 1989.

[50] W. R. Frensley. “Wigner-function model of a resonant-tunneling semiconductor device.” *Physical Review B*, 36(3):1570–1580, 1987.

[51] F. A. Buot and K. L. Jensen. “Lattice Weyl-Wigner formulation of exact many-body quantum-transport theory and applications to novel solid-state quantum-based devices.” *Physical Review B*, 42(15):9429–9457, 1990.

[52] R. Tsu and L. Esaki. “Tunneling in a finite superlattice.” *Applied Physics Letters*, 22(11):562–564, 1973.

[53] C. M. Tan, J. Xu, and S. Zukotynski. “Study of resonant tunneling structures: A hybrid incremental Airy function plane-wave approach.” *Journal of Applied Physics*, 67(6):3011–3017, 1990.

[54] K. L. Jensen and F. A. Buot. “The effects of scattering on current-voltage characteristics, transient response, and particle trajectories in the numerical simulation of resonant tunneling diodes.” *Journal of Applied Physics*, 67(12):7602–7607, 1990.

[55] B. A. Biegel and J. D. Plummer. “Comparison of self-consistency iteration options for the Wigner function method of quantum device simulation.” *Physical Review B*, 54(11):8070–8082, 1996.

[56] B. A. Biegel and J. D. Plummer. “Applied bias slewing in transient wigner function simulation of resonant tunneling diodes.” *IEEE Transactions on Electron Devices*, 1997. (to be published).

[57] B. Ricco and M. Y. Azbel. “Physics of resonant tunneling. the one-dimensional double-barrier case.” *Physical Review B*, 29(4):1970–1981, 1984.

[58] C. R. Wallis and S. W. Teitsworth. “Conditions for current oscillations and hysteresis in resonant tunneling diode circuits.” In *Proceedings of the International Semiconductor Device Research Symposium*, pages 487–490, University of Virginia, Charlottesville, 1993. Engineering Academic Outreach.

[59] F. A. Buot and A. K. Rajagopal. “Binary information storage at zero bias in quantum-well diodes.” *Journal of Applied Physics*, 76(9):5552–5560, 1994.

[60] F. A. Buot and K. L. Jensen. “Quantum transport: Novel approaches in the formulation and applications to quantum-based solid-dtate devices.” *COMPEL*,

10(4):509–524, 1991.

[61] W. R. Frensley. “Effect of inelastic processes on the self-consistent potential in the resonant-tunneling diode.” *Solid State Electronics*, 32(12):1235–1239, 1989.

[62] K. K. Gullapalli, D. R. Miller, and D. P. Neikirk. “Simulation of quantum transport in memory-switching double-barrier quantum-well diodes.” *Physical Review B*, 49(4):2622–2628, 1994.

[63] J. S. Wu, C. Y. Chang, C. P. Lee, K. H. Chang, D. G. Liu, and D. C. Liou. “Resonant tunneling of electrons from quantized levels in the accumulation layer of double-barrier heterostructures.” *Applied Physics Letters*, 57(22):2311–2312, 1990.

[64] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer. “Self-consistent study of the resonant-tunneling diode.” *Physical Review B*, 39(11):7720–7735, 1989.

Chapter 9

Conclusion

This chapter summarizes the contents of this dissertation (Section 9.1), lists the specific contributions of this work to the field of quantum device simulation (Section 9.2), and makes recommendations for future work in this field (Section 9.3). Finally, a list of recommended techniques is given for the successful and efficient development of large software projects such as SQUADS (Section 9.4).

9.1 Summary

As conventional electronic devices shrink and their energy dissipation decreases, these devices are be increasingly antagonized by quantum effects. When these effects no longer permit further scaling of conventional electronic devices while maintaining reliable operation, the only solution will be to use quantum effects to control the operation of electronic devices. Quantum electronics is the concept of producing useful computing (analog or digital signal processing) with quantum devices. If it achieves its goals, quantum electronics will produce not only quantum scale devices, but also integration levels, computing efficiencies, and system functionality well beyond that of ULSI. Of the three possible approaches (conceptual, computational, experimental) to pursuing quantum electronics research, this work took the computational approach, and this dissertation described an effort to develop a numerical simulation tool called SQUADS for modeling quantum electronic devices. This approach leverages the advancing power of computers to provide more efficient investigation than experiment, and more detail than theory.

To determine the quantum device characteristics that SQUADS must handle, it was necessary to have a conceptual understanding of quantum electronics, from relevant quantum phenomena to complete quantum computing systems and likely computation approaches. At the smallest scale, the optical analogy to quantum wave systems implies that the phenomena, structures, and device functions of quantum electronics at the smallest scale will be similar to those of optical systems. At the large scale, analysis of the requirements of digital computing systems concluded that far-from-equilibrium quantum devices show the most promise for quantum electronics in the near term. In general, far-from-equilibrium quantum devices are unipolar, heterojunction-based, support high biases, include scattering, and are very high speed. Due to its simplicity and strong quantum effects, and the availability of experimental results, the resonant tunneling diode (RTD) was chosen as the prototype quantum device for testing the accuracy and capabilities of SQUADS.

Given these quantum device characteristics, the optimal formulation of quantum mechanics was chosen for the development of SQUADS. This formulation must also provide an internal view of device operation, be suitable for both steady-state and transient simulation, be independent of any particular device or material system, and be feasible to execute on existing hardware. Based on these requirements, the Wigner function method (WFM) of quantum device simulation was chosen as the primary basis for SQUADS. The simulator was limited to 1-D quantum device simulation due to limitations of existing computing hardware. Since the WFM is a computationally expensive and its accuracy is unknown, the more established transfer matrix method (TMM) was also implemented in SQUADS to provide an efficient check on WFM results. SQUADS was designed to be flexible and easily extensible, enabling the efficient investigation of both quantum device *simulation*[1] and quantum device *operation*.

The TMM is well-suited to efficient simulation of steady-state quantum device operation, such as tracing the I-V curve. The TMM calculates current flow through a quantum system by adding the transmission of many independent, mono-energetic electron beams over a distribution of incident energies represented by the boundary conditions. The calculation of the transmission coefficient is facilitated by division of the position domain into

1. Note that because the TMM is well-established and can not handle scattering or transient simulations, most of quantum *simulator* research was applied to investigation of the WFM.

many small regions. Each region and interface then has a simple transfer matrix, as determined by solution of the Schrödinger equation. Extensions of the TMM in SQUADS for the calculation of the wavefunction, energy spectrum, carrier density profile, and Wigner function were also described. Within its limitations (no scattering or transient simulation), the TMM proved to be a very reliable basis for quantum device simulation.

When scattering or transient effects are to be investigated in quantum devices, the WFM is required. The WFM solves the Wigner function transport equation (WFTE) at a discrete set of points to predict the evolution of the Wigner function $f_w(x, k, t)$, which describes the action of charge carriers in the quantum device. The discretization of the WFTE has many complications and alternative implementations, all correctly handled by SQUADS. The resulting computation requires the solution of typically 5000 to 50,000 simultaneous equations, so it is essential to reduce extraneous memory usage and computation as much as much as possible. SQUADS uses optimized matrix storage and computation schemes. The WFM simulation of a Gaussian wave packet in free space, which also has an analytic solution, allowed the accuracy and cost of various discretization schemes for the WFTE to be compared. The discrepancy between WFM and TMM simulations of an RTD was rather large, presumably due to inaccuracy in the standard WFM implementation. Clearly, the WFM is still in a development phase, although it can give qualitative results about quantum device operation. The remainder of this work described three WFM-based investigations of RTDs and the WFM itself.

The first investigation examined the implementation of self-consistency in the WFM. Enforcing self-consistency requires an iterative solution of the WFTE and the Poisson equation (PE) to achieve a simultaneous solution to both equations. Four self-consistency iteration methods are implemented in SQUADS: steady-state and transient Gummel, and steady-state and transient Newton.[2] The Gummel approaches alternately solve the WFTE and PE, while the Newton approaches solve the two equations simultaneously. Due to their computational efficiency, the steady-state methods are recommended for wide-ranging initial investigations, such as I-V curve traces. For transient operation, or to verify device operation in critical regions, expensive transient simulations are required. The expense of the Newton approaches is not worth the small additional accuracy they afford.

---

2. Most other quantum device simulators implement only a single self-consistency iteration approach, to simplify the programmer's job, rather than to maximize usefulness of the simulator.

The proper selection of convergence criteria (*i.e.*, when to terminate the iteration) is also important. The implementation of self-consistency in the TMM is much simpler, since the TMM can only use the steady-state Gummel iteration method.

The second investigation studied the effect on device function of using a finite applied bias slew rate in transient, self-consistent WFM simulations of an RTD. The conventional approach of instantaneously changing the applied bias in simulations causes huge current pulses both within the RTD and in the external biasing circuit. Using lower slew rates, the current pulse amplitudes decreased to more tolerable values, and the internal pulse was then shown to be due to charging of the depletion and accumulation layers to accommodate the new applied bias. Other instances where RTD function depended on slew rate were demonstrated. For example, fast slewing initiated oscillations in a marginally stable region of operation, while low slew rates did not. Also, very low slew rates were necessary to produce some modes of RTD operation. Finally, the slew rate could determine which state an RTD ended up in when switched into a bistable region of operation. Thus, slew rate *does* affect device function, and should be chosen in simulations based on the intended application for the device.

The final SQUADS investigation was an in-depth study of the very intriguing physics of the RTD that was used in the two investigations summarized above. The interesting effects included a plateau in the negative differential resistance of the I-V curve, hysteresis, and high-frequency oscillations. Based on the visceral similarity between these phenomena and experimental observations, other researchers concluded that they had reproduced the experimental effects. On the basis of this conclusion, they broke with the established consensus view on several controversies about the nature of RTD operation. The investigation in this work showed conclusively that the simulated RTD physics was *not* of the same origin as the apparently similar experimental effects. In fact, further investigation indicated that more accurate WFM simulations would probably support the consensus view on each of the controversial issues. Although the WFM appears to require further refinements, this investigation showed that SQUADS can provide a rich array of information from which to draw in the study of quantum device physics.

9.2 Contributions

This section lists the principle contributions of this work. Roughly in the order they are

presented in this dissertation, these contributions include:

- One of the most comprehensive conceptual reviews and analyses of the theory of quantum electronics to date, including discussions of
  - the array of available quantum effects and basic quantum device structures,
  - the types of quantum devices (quasi-equilibrium and far-from-equilibrium),
  - the architectures that may be utilized (cellular automaton or quantum filter),
  - the probable nature of computing with quantum electronic circuits, and
  - the ultimate future of quantum electronics in true quantum computers.
- An analysis of quantum device simulation approaches, including discussions of
  - the goals of quantum device simulation and its relation to theoretical and experimental quantum device research,
  - the lessons learned from conventional electronic device simulation, and
  - the strengths and weaknesses of all significant quantum device simulation approaches in use and proposed.
- Most importantly, the development of a numerical simulation called SQUADS (Stanford QUAntum Device Simulator) tool for modeling 1-D quantum devices. Important features of SQUADS include:
  - Highly functional: TMM and WFM simulation approaches are implemented. Self-consistency, scattering (WFM), steady-state analysis, transient analysis (WFM), and Gaussian wave packet simulations (WFM) are all available.
  - Extensible: Modular structure makes enhancements and alternative implementations easily added. Also, general enhancements to SQUADS are immediately available, if appropriate, to both the TMM and WFM.
  - Efficient: a great deal of effort was applied to the effort of producing accurate simulations with a minimum of computation.
  - Portable: Compiled and executed without modification on at least ten platforms, ranging from a 386 PC to a Cray C-90 supercomputer.
  - Publication-quality graphical output, including line and surface plots.
- TMM simulation contributions:
  - Most comprehensive mathematical derivation and description of the TMM to date, including discussions (apparently for the first time) of
    - how to take advantage of contact flat-band regions for improved compu-

tational efficiency,
    - how quantum turning points must be treated for robust TMM simulation,
    - when numerical overflow may occur and how to protect against it,
    - handling classically neutral regions, and
    - correctly handling classically-forbidden T-contacts.
  - Alternative transmission matrix computation algorithms were found to be significantly more efficient for some TMM simulation tasks.
  - Using a piece-wise linear (as opposed to piece-wise constant) potential was found to have three times the computational cost, with little accuracy benefit.
  - The use of a position-dependent effective mass was shown to significantly alter quantum device simulation results.
  - Calculation of the energy spectrum of carriers and the Wigner function were added to the standard TMM simulator capabilities.
- WFM simulation contributions:
  - Most comprehensive mathematical derivation and description of the WFM to date, including
    - implementation of numerous discretization options for the diffusion term and three for the drift term of the WFTE,
    - five alternative implementations for transient simulations,
    - a Gaussian wave packet simulation capability, and
    - efficient storage and solution schemes for solving the compute-intensive WFTE.
  - Optimal discretization approaches for the diffusion and transient terms were determined, both for efficiency and accuracy.
  - Inaccuracy in the standard implementation of the WFM was demonstrated, and possible solutions to this short-coming were proposed.
- Self-consistency contributions:
  - Implemented the four basic self-consistency iteration methods for the WFM, allowing direct comparison of accuracy, efficiency, and robustness.
  - Demonstrated the complementary roles of steady-state and transient self-consistency iteration approaches.
  - Described the Newton iteration method for the WFTE.

- Derived and implemented a modified steady-state Gummel iteration method that achieves much faster convergence that the standard Gummel approach.
- Implemented a sophisticated steady-state self-consistency iteration algorithm, which is both efficient and robust.
- Described necessary and sufficient criteria to gauge convergence in self-consistency simulations.
- Demonstrated that converged steady-state self-consistency simulations is only an *equilibrium* operating point, and may be stable or unstable.
- Implemented, as an alternative to Gummel self-consistency, a quasi-classical self-consistency algorithm (to imitate scattering) for the TMM.

- Slew rate contributions:
  - Showed that the conventional approach of using instantaneous switching in transient WFM simulations has led to substantially inaccurate simulations of quantum devices and resulting incorrect conclusions about device operation.
  - Showed that use of different applied bias slew rates in quantum device simulation can dramatically change device function: the slew rate used in simulations should match that of the intended device application.
- Conducted the most detailed simulation investigation to date of RTD physics. Some results included:
  - Showed the rich physics that RTDs are capable of, and the depth of simulation investigation that may be needed to uncover all significant aspects of this physics.
  - Showed that an RTD having a discrete emitter state can produce interesting and useful operation, including a plateau in the NDR region, unstable oscillations, bistability, and hysteresis.
  - Uncovered and corrected errors in previous interpretation of RTD simulation results and experimental measurements.
  - Achieved improved agreement between simulation and experiment for a particularly interesting RTD, with which improvement, the simulations effectively support the consensus conclusions on several RTD physics controversies, in contrast to previous analysis of simulations of this RTD.
  - Demonstrated that the standard WFM implementation used for over ten years

still requires accuracy improvements.

- Demonstrated dynamic hysteresis in RTD simulations for the first time.
- Showed the importance of including self-consistency and scattering in simulations to demonstrate some very interesting and potentially useful modes of RTD operation.
- For accurate simulations, showed the significance of using a simulation region which naturally accommodates the entire applied bias.

9.3 Recommendations for Future Work

As the contributions above indicate, this work encompasses a significant advance in the field of quantum device simulation. However, this work also demonstrates that there are still many significant questions to answer, investigations to pursue, and advances to make in this endeavor. Most of the resulting recommendations for future work center on the WFM of quantum device simulation, since this approach was shown to have the most potential for accuracy and capabilities. The following capabilities should be added to, and investigated with, the WFM:

- Improved discrete WFTE implementation. One of the contributions of this work was a determination that the standard WFM implementation (used by all WFM researchers to date) apparently has a fundamental flaw, resulting in compromised accuracy. Other methods of implementing the discrete WFTE should be investigated in an attempt to clarify and avoid this flaw. Mains and Haddad [1] have proposed one such implementation.
- Higher N_k simulations. TMM simulations typically require about 1000 energy points to achieve acceptable accuracy. In contrast, typical WFM simulations only use 100 wavenumber points (related to energy). Further, a single energy value in a TMM simulation incorporates both forward-travelling (positive wavenumber) and backward-travelling (negative wavenumber) wavefunctions. Thus, accurate WFM simulations may require $N_k = 2000$ wavenumber points.
- Interface to a classical device simulator. Although quantum devices are a long-term prospect, understanding quantum effects in conventional electronic devices is becoming increasing important as these devices are down-scaled. Because the WFM uses classical boundary conditions, it can interface to non-quantum device

simulators. This would allow quantum regions of a conventional device to be simulated with a quantum simulator, with the remainder being treated with the conventional device simulator. Such multi-modal simulation has been used with other quantum simulation approaches [2, 3], but this essentially requires that a computationally expensive Monte Carlo approach be used for the classical portion of the simulation.

- Interband interactions. Some TMM simulators have added interband coupling [4-9] including a limited scattering model [10], only one WFM simulator [11] has attempted to include these effects. The importance of including interband coupling for the accurate simulation of many tunneling devices has been widely demonstrated [4, 7, 12, 13]. Interacting bands would make multi-band WFM simulations much more costly, but a suitable re-ordering of the unknowns in the WFTE matrix equation should make the computation feasible on existing hardware.
- Bipolar and optical capability. By far the most widely used quantum effects in electronics today are optoelectronic in nature, in devices such as the quantum well laser diode. Adding a bipolar capability (oppositely-charged carriers, recombination/generation) and optical effects (photogeneration/absorption) would enable the simulation of these important optoelectronic devices. Some proposed and demonstrated resonant tunneling transistors are bipolar devices, and would also be accessible to simulation with the bipolar capability. Even with a 1-D simulator, a quasi-three-terminal capability could be produced by maintaining by the potential at an internal device node with the introduction or depletion of the requisite number of carriers.
- Detailed energy bands. Quantum device simulators almost universally use parabolic energy bands (effective mass independent of energy).[3] This is not very accurate in far-from-equilibrium quantum devices, where charge carriers may accelerate to several hundred meV above the energy band minimum. It should not be difficult to implement more accurate energy bands in the WFM [11], and some TMM simulators [5, 7] and other quantum device simulators [14, 15] already have this feature.

3. In E-M wave systems, this would be equivalent to a dispersionless material.

- Dynamic boundary conditions. Quantum devices will eventually be small enough, and have low enough scattering, that classical boundary conditions such as those used in the WFM are not appropriate. Carriers will maintain some phase-coherence into the contact, and into the next device. This means that neighboring devices will have a more complicated and dynamic effect on each other. These effect could be studied using dynamic boundary conditions.
- Greater variety in quantum device investigations. Investigations to date with WFM-based quantum device simulators, including those in this work, have centered on the GaAs RTD. Besides a lack of time in this relatively new endeavor, this narrow focus suggests a lack of imagination in the field of quantum device simulation. The devices which may make quantum electronics a success are almost certainly not even be known at this point. Thus, a wider array of devices, material systems, and simulation parameters should be investigated with the WFM in an effort to discover more promising quantum devices.
- 2-D and 3-D simulation. Few quantum devices can be accurately modeled as 1-D structures, so implementing multi-dimensional WFM simulation is essential. Some Schrödinger equation quantum device simulators already provide 2-D [16-19] and 3-D [20, 21] modeling. Unfortunately, it is unlikely that direct 2-D or 3-D WFM simulation will be feasible in the near future, since the required computing power would be immense, as discussed in Section 5.2.1. However, the WFM could be used for one dimension, and another quantum or classical approach for the other(s). Another option is to use the WFM-based Monte-Carlo approach [22], which should be feasible on existing hardware in 2-D or 3-D. Finally, approximate quantum corrections to classical simulators could be effective for some quantum effects.

The following capabilities have been investigated to some degree in at least one other WFM simulator, but should be implemented in the WFM simulator in SQUADS:

- Position-dependent effective mass. TMM simulations in Section 4.5.3 showed that using a position-dependent effective mass significantly affects quantum device simulation results. Therefore, accurate WFM simulations require the inclusion of a position-dependent effective mass as well. Several researchers [1, 23, 24] have described how this may be accomplished.

- Small-signal WFM simulations. Adding his capability to the existing steady-state and transient WFM capabilities would make SQUADS a fully general quantum device simulator. Frensley [25] has demonstrated a WFM-based small-signal capability.

Quantum device simulation research to date has been rather disorganized and uncoordinated. In particular, there is no software package that provides a base-line of functionality which researchers can use and enhance (for example, as PISCES [26] does for 2-D conventional device simulation). Instead, every research team must essentially implement the same base-line functionality before the enhancements of interest can be added. Of course, this new functionality is not available to anyone else, since the various quantum device simulation tools have independently-evolved structures and interfaces. Thus, a final and more general recommendation for future quantum device simulation work is to create a generally available, highly functional, highly usable, and robust quantum device simulation package. In addition to alleviating the problems described above, this would also greatly increase the number of researchers to whom quantum device simulation would be available as a research tool. Note that SQUADS is currently being evaluated as a candidate for this "base-line" quantum device simulation tool.

9.4 Recommendations for Software Development

Finally, for those attempting to create a large software package like SQUADS, this section contains a few suggestions (learned during the development of SQUADS) for coding large software projects, in the interest of maintaining programmer sanity (and producing increasingly functional code). These points may also be of use to those trying to decipher code in SQUADS (for example, for the purpose of adding enhancements).

General issues:

- Choice of programming language. For numerical simulation tools, speed of execution is a key issue, as are the availability of programing tools, portability of the source code to multiple computing platforms (*i.e.*, level of language standardization, and widespread existence of compilers), widespread use (in case others will look at or work on the code), and ability to structure the code in separate subroutines and files. Based on these considerations, the C programming language was chosen for SQUADS, and GCC is used as the compiler on most platforms.

- Platform independence. For a large programming project, it is unlikely that the entire programming and execution of the project will be conduced on a single computing platform. Efforts should be taken to make sure the project has minimal platform-dependence. SQUADS has been run under OS/2, Ultrix, Irix, OSF1, SunOS4, Solaris, SystemVr4, Unicos (Cray). To cope with the necessary differences between platforms (usually, different function library files), SQUADS has a run script which determines the operating system on the current platform, modify a copy of the makefile for this platform, compile SQUADS in a directory for that OS, and execute.

Before writing any code:

- Have at least one generation full backup of working source files. Refresh the backup only after major changes have been fully tested.
- Always have a working executable. There will always be occasions to run it on short notice, whether for demonstration or to get last-minute results.
- For mathematically complex functions, do not begin coding until the theoretical derivation is complete.

When writing and debugging code:

- Use a hierarchical structure. In other words, divide and conquer. Use reasonably-sized subroutines and multiple files to break a problem down into small parts, each of which can be completely understood as a whole.
- Document new code as soon as possible. Although non-obvious code sections should be commented within the source file, the most important code documentation is the interface (header) files. These files are all someone (including the programmer) should have to read to understand what the associated subroutines do. A header file is incomplete if someone must look at the actual code.
- Upgrade gracefully. Don’t make all planned changes at once. Make enhancements in small steps whenever possible, and verify proper program function after each step.
- Use appropriate tools to make programming tasks easier. For example, during SQUADS development, *grep* was used to find all occurrences of a variable whose name or definition was about to change, *diff* was used to find all differences between an old (working, backup) source file and a new (non-working)

one, an execution profiler was used to tune critical sections code for efficiency, and a debugger was used to quickly locate fatal run-time errors.

- Implement new functionality without destroying working code. Add features in separate subroutines, rather than directly modifying existing code. Alternatively, keep a copy of previous (working) lines, subroutines, or files, for comparison the (almost invariably) non-working ones after modifications.

When running the program:

- Automate busy work. For example, much of SQUADS' execution script was described above. Other automation scripts create the directory structure used by SQUADS, pack and unpack all SQUADS files for transport to other machines, clean out old object files or simulation results, rotate surface plot files, filter plot file data, convert plot files between various formats, or scan all library files for needed functions. In general, try to automate anything that must be done more than twice.
- User interface. If possible, add a GUI (graphical user interface) only *after* the code (or at least the input to the program) is essentially unchanging, and only if the GUI is faster and more straightforward to use than a text interface. For the programmer, the GUI represents an additional level of complexity to be maintained and upgraded with the remainder of the code. Therefore, it must be very easy to upgrade, or it will eventually be abandoned for a text interface.

References

[1] R. K. Mains and G. I. Haddad. "An accurate re-formulation of the wigner function method for quantum transport modeling." *Journal of Computational Physics*, 112:149–161, 1994.

[2] H. Ohnishi, N. Yokoyama, and A. Shibatomi. "Modeling electron transport in ingaas-based resonant-tunneling hot-electron transistors." *IEEE Transactions on Electron Devices*, 36(10):2335–2339, 1989.

[3] K. K. Gullapalli, D. R. Miller, and D. P. Neikirk. "Hybrid boltzmann transport - schrödinger equation model for quantum well injection transit (qwitt) diodes." In *International Electron Devices Meeting*, pages 511–513, 1991.

[4] J. R. Söderström, E. T. Yu, M. K. Jackson, Y. Rajakarunanayake, and T. C. McGill.

"Two-band modeling of narrow band gap and interband tunneling devices." *Journal of Applied Physics*, 68(3):1372–1375, 1990.

[5] K. V. Rousseau and K. L. Wang. "Gamma- and X-state influences on resonant tunneling current in single- and double-barrier GaAs/AlAs structures." *Applied Physics Letters*, 54(14):1341–1343, 1989.

[6] J. C. Chiang and Y.-C. Chang. "Resonant tunneling of electrons in si/ge strained-layer double-barrier tunneling structures." *Applied Physics Letters*, 61(12):1405–1408, 1992.

[7] D. Mui, M. Patil, J. Chen, S. Agarwala, N. S. Kumar, and H. Morkoc. "Modelling of the i-v characteristic of single and double barrier tunneling diodes using a k-p band model." *Solid State Electronics*, 32(11):1025–1031, 1989.

[8] K. Fobelets, R. Vounckx, and G. Borghs. "Matrix formalism for the triple-band effective-mass equation." *Semiconductor Science and Technology*, 8:1815–1821, 1993.

[9] Y. Fu, Q. Chen, and M. Willander. "Resonant tunneling of holes in Si/GeSi." *Journal of Applied Physics*, 70(12):7468–7473, 1991.

[10] G. Y. Wu and T. C. McGill. "Effects of barrier phonons on the tunneling current in a double-barrier structure." *Physical Review B*, 40(14):9969–9972, 1989.

[11] D. R. Miller and D. P. Neikirk. "Simulation of intervalley mixing in double-barrier diodes using the lattice Wigner function." *Applied Physics Letters*, 58(24):2803–2805, 1991.

[12] G. D. Sanders and Y. C. Chang. "Optical properties in modulation-doped gaas-ga_1-xal_xas quantum wells." *Physical Review B*, 31(10):6892–6895, 1985.

[13] D. Landheer, H. C. Liu, M. Buchanan, and R. Stoner. "Tunneling through alas barriers: Gamma-x transfer current." *Applied Physics Letters*, 54(18):1784–1786, 1989.

[14] D. Z.-Y. Ting, E. T. Yu, and T. C. McGill. "Band structure effects in interband tunnel devices." *Journal of Vacuum Science and Technology B*, 9(4):2405–2410, 1991.

[15] M. D. J. andd Y. Sekiguchi, J. S. Y. J. Chan, D. Pavlidis, and M. Quillec. "A study of charge control in n- and p-type lattice matched and strained channel modfets with gaas and inp substrates." In *IEEE/Cornell Conference on Advanced Concepts*

in High Speed Semiconductor Devices and Circuits, pages 70–79, 1987.

[16] C. S. Lent and D. J. Kirkner. “The quantum transmitting boundary method.” *Journal of Applied Physics*, 67(10):6353–6359, 1990.

[17] S. Bhobe, W. Porod, and S. Bandyopadhyay. “Modulation of impurity scattering rates by wavefunction engineering in quasi 2-D systems and its device applications.” *Solid State Electronics*, 32(12):1083–1087, 1989.

[18] L. F. Register, U. Ravaioli, and K. Hess. “Numerical simulation of mesoscopic systems with open boundaries using the multidimensional time-dependent Schrödinger equation.” *Journal of Applied Physics*, 69(10):7153–7158, 1991.

[19] H. Taniyama, M. Tomizawa, and A. Yoshi. “Two-dimensional analysis of resonant tunneling using the time-dependent schr"odinger equation.” *Japan Journal of Applied Physics*, 33, Part 1(4A):1781–1786, 1994.

[20] G. Neofotistos and K. Diff. “Time-dependent modeling of resonant tunneling structures using the 3-dimensional schrodinger equation: Investigation of the intrinsic time characteristics of a zero-dimensional semiconductor nanostructure.” In M. A. Reed and W. P. Kirk, editors, *Nanostructure Physics and Fabrication*, pages 135–139, New York, 1989. Academic Press. Proceedings of the International Symposium, College Station, TX, March 13-15.

[21] T. Palm. “Self-consistent calculations of an electron y-branch switch.” *Journal of Applied Physics*, 74(5):3551–3557, 1993.

[22] K. L. Jensen and A. K. Ganguly. “Quantum transport simulations of electron field emission.” *Applied Physics Letters*, 55(7):669–671, 1989.

[23] H. Tsuchiya, M. Ogawa, and T. Miyoshi. “Simulation of quantum transport in quantum devices with spatially varying effective mass.” *IEEE Transactions on Electron Devices*, 38(6):1246–1252, 1991.

[24] K. K. Gullapalli and D. P. Neikirk. “Incorporating spatially sarying effective-mass in the Wigner-Poisson model for AlAs/GaAs resonant-tunneling diodes.” In *Proceedings of the 3rd International Workshop on Computational Electronics*, pages 171–174, 1994.

[25] W. R. Frensley. “Boundary conditions for open quantum systems driven far from equilibrium.” *Reviews of Modern Physics*, 62(3):745–791, 1990.

[26] M. R. Pinto, C. S. Rafferty, and R. W. Dutton. *PISCES-II - Poisson and Continuity*

Equation Solver. Stanford University, 1984.