THREE-DIMENSIONAL SUPERCELL SIMULATION OF NOVEL SEMICONDUCTOR NANOSTRUCTURES

Thesis by

Shaun Kirby

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1994

(Submitted August 25, 1993)

to My Family

Acknowledgements

It is a great pleasure to acknowledge my advisor, Professor Thomas C. McGill, whose keen administration of our large research group has made my experience at Caltech both stimulating and rewarding. In addition to providing technical leadership, Tom imparts to his students an invaluable perspective on how science, economics and politics interact in the evolution of technology. Practically speaking, his generous support affords students the opportunity to work with state-of-the-art facilities and to attend numerous conferences and workshops. In spite of the size of the group, Tom manages to interact with each of his students on a personal level.

I am also grateful to Dr. David Ting, who has served as my mentor throughout the tenure of the thesis research. David's careful, systematic analysis is an example of the highest caliber of scientific research. His deep physical insights and creativity have more than a few times unraveled mysteries which had thrown me into confusion. His enthusiasm for simulation and modeling are inspirational and highly contagious.

I owe much of my theoretical foundation to Yixin Liu, with whom I worked during my first year in the group. Yixin helped me bridge the gap between first year courses in relativity and field theory and state-of-the-art solid state device physics. Our lengthy discussions on Chinese politics, economics, culture and history serve as an education in themselves. In addition I have found Yixin to be a most enthusiastic travel companion and a spirited tennis, squash and racketball player as well!

Were it not for Marcia Hudson's graceful facilitation of travel arrangements, correspondence, documents and official records, the McGill group would come to a grinding halt. I have also found the endless assortment of snacks on her desk the perfect cure for the mid-afternoon munchies. Sandy Brooks also deserves great credit for top-notch secretarial support.

I could not have found a more exciting, inspiring and amiable group of colleagues than those I have met in the McGill group. Dr. Ed Croke, Dr. Ed Yu, Dr. Doug Collins, Dr. Mark Phillips, Harold Levy, Ron Marquardt, Rob Miles, P. O. Pettersson, David Reich, Chris Springfield, Johanes Swenberg, and Mike Wang have greatly enriched my experiences with their diverse research interests and piqued my curiosity in everything from cinematography to snorkeling.

I am very grateful for the support of an Office of Naval Research Fellowship, and for the funding from the Office of Naval Research under Grant No. N00014-89-J-1141 that made this research possible.

Finally, I thank my parents, my grandparents and my brother for their generous support and unflagging encouragement over many years.

List of Publications

Work related to this thesis has been, or will be, published under the following titles:

Atomic Scale Imperfections and Fluctuations in the Transmission Properties of a Quantum Dot,
S. K. Kirby, D. Z.-Y. Ting, and T. C. McGill, to be submitted to *Phys. Rev. B* (1993).

Fluctuations in the Transmission Properties of a Quantum Dot with Interface Roughness and Impurities,
S. K. Kirby, D. Z.-Y. Ting, and T. C. McGill, to be presented at the Eight International Conference on Hot Carriers in Semiconductors, Oxford, U.K. (1993).

Planar Supercell Simulations of 3D Quantum Transport in Semiconductor Nanostructures,
D. Z.-Y. Ting, S. K. Kirby, and T. C. McGill, to be presented at the International Workshop on Computational Electronics, Leeds, England (1993).

Neutral Impurities in Tunneling Structures,
S. K. Kirby, D. Z.-Y. Ting, and T. C. McGill, submitted to *Phys. Rev. B* (1993).

Three-dimensional Simulations of Quantum Transport in Semiconductor Nanostructures,
D. Z.-Y. Ting, S. K. Kirby, and T. C. McGill, presented at the Twentieth Conference on the Physics and Chemistry of Semiconductor Interfaces, Williamsburg, VA (1993).

Three-dimensional Supercell Simulations of Quantum Transport,
S. K. Kirby, D. Z.-Y. Ting, and T. C. McGill, *Proceedings of the International Workshop on Computational Electronics, Urbana–Champaign*, 289, (1992).

Exciton Transfer and Excitonic Band Structure in Semiconductor Quantum Structure Systems,
Y. X. Liu, S. K. Kirby, and T. C. McGill, to be submitted to *Phys. Rev. B* (1993).

Abstract

In this thesis we investigate atomic scale imperfections and fluctuations in the quantum transport properties of novel semiconductor nanostructures. For this purpose, we have developed a numerically efficient supercell model of quantum transport capable of representing potential variations in three dimensions. This flexibility allows us to examine new quantum device structures made possible through state-of-the-art semiconductor fabrication techniques such as molecular beam epitaxy and nanolithography. These structures, with characteristic dimensions on the order of a few nanometers, hold promise for much smaller, faster and more efficient devices than those in present operation, yet they are highly sensitive to structural and compositional variations such as defect impurities, interface roughness and alloy disorder. If these quantum structures are to serve as components of reliable, mass-produced devices, these issues must be addressed.

In Chapter 1 we discuss some of the important issues in resonant tunneling devices and mention some of thier applications. In Chapters 2 and 3, we describe our supercell model of quantum transport and an efficient numerical implementation. In the remaining chapters, we present applications.

In Chapter 4, we examine transport in single and double barrier tunneling structures with neutral impurities. We find that an isolated attractive impurity in a single barrier can produce a transmission resonance whose position and strength are sensitive to the location of the impurity within the barrier. Multiple impurities can lead to a complex resonance structure that fluctuates widely with impurity configuration. In addition, impurity resonances can give rise to negative differential resistance. In Chapter 5, we study interface roughness and alloy disorder in double barrier structures. We find that interface roughness and alloy disorder can shift and broaden the $n = 1$ transmission resonance and give rise to new resonance peaks, especially in the presence of clusters comparable in size to the electron deBroglie wavelength. In Chapter 6 we examine the effects of interface roughness and impurities on transmission in a quantum dot electron waveguide. We find that variation in the configuration and stoichiometry of the interface roughness leads to substantial fluctuations in the transmission properties. These fluctuations are reduced by an attractive impurity placed near the center of the dot.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview and Motivation

Novel semiconductor nanostructures, such as quantum wells, quantum wires and quantum dots, have given rise to a wealth of new physics and offer promise for new devices. With dimensions small compared to the electron mean free path, these structures provide laboratories in which to study quantum confinement, quantum interference and low-dimensional systems. One of the most widely studied nanostructures is the double barrier resonant tunneling structure, consisting of a quantum well composed of narrower band gap material such as GaAs, surrounded by barriers composed of wider band gap material such as AlAs, and sandwiched between doped electrodes. A quasibound level in the quantum well gives rise to a transmission resonance [1], leading to negative differential resistance as originally proposed by Tsu and Esaki [2] and demonstrated by Chang et al. [3]. Since the initial demonstration, fabrication techniques have improved dramatically, leading to better performance, and the double barrier has stimulated much interest in potential applications such as very high frequency microwave devices, logic elements with new functionality, and novel neural networks [4, 5, 6]. Lately quantum wires and quantum dots have also attracted attention, both for their interesting new properties and for their potential device applications [7, 8, 9].

Qualitative features of the current-voltage characteristics of the double barrier resonant tunneling structure are understood [2, 10], but good quantitative agreement with experiment is still lacking. One of the most important outstanding problems is that calculated peak-to-valley current ratios are much higher than those observed experimentally, causing considerable concern, as a good peak-to-valley ratio is crucial to device performance. Likewise, transport in real quantum wires and quantum dots is far from fully understood.

Discrepancies between experiment and theory are thought to be the result of many complex physical phenomena. Among those which can play a role in the operation of the double barrier and other nanostructures are electron-electron interactions, electron-phonon interactions, band structure effects, and structural and compositional imperfections. Much effort has been devoted to the understanding of electron-electron interactions, as well as electron-phonon interactions and band structure effects in double barriers. However, the treatment of structural and compositional imperfections, such as interface roughness, alloy disorder and impurities, has been lacking, and these issues are believed to be a major source of remaining differences between theory and experiment.

There is a great deal of experimental evidence that structural and compositional imperfections can play a major role in transport in quantum structures. Improvements in interface quality in double barriers since the first observation of negative differential resistance by Chang et al. [3] have led to dramatic improvement in the peak-to-valley current ratio [11, 12] yet there is substantial evidence that interface roughness is still a dominant contributor to valley current. Gueret et al. [13, 14] have given convincing evidence that interface roughness may account for most of the disagreement of more than an order of magnitude between their measured and calculated peak-to-valley ratios. (Their calculations assumed smooth interfaces.) There is also considerable experimental evidence that defect impurities can dramatically alter transport in quantum structures. An isolated conductance peak observed below the turn-on of the first transverse mode in a narrow constriction has been attributed to resonant tunneling via a single impurity[15]. Degradation in the quantized conductance steps of a dual electron waveguide has been seen when the conductance channel is electrostatically steered into a scatterer[16].

Thus far, models of structural and compositional imperfections [17, 18, 19, 20, 21] have relied on first order perturbation treatments in essentially one-dimensional simulations, limiting them to the weak scattering, weak localization regime and preventing a realistic description of device imperfections and multiple scattering in three dimensions. In addition, these one-dimensional models are incapable of treating low-dimensional structures such as quantum wires and quantum dots. If nanostructures are to serve as building blocks for reproducible circuits in future technologies, atomic scale imperfections and fluctuations in their transmission properties must be understood. Indeed, understanding the effect of structural imperfections offers the best hope for continued improvement in the characteristics of double barriers as well as quantum wires and quantum dots—as proficiency in fabricating and manipulating atomic structures improves, structural and compositional imperfections may be reduced or controlled.

In this thesis we study the effects of structural and compositional imperfections in quantum structures. For this purpose, we have developed a supercell model of quantum transport in three dimensions, capable of representing three-dimensional potential variations on an atomic scale. This flexibility permits not only a more accurate description of imperfections in double barriers, but it also allows us to study novel geometries and material configurations and to understand quantum wires, quantum dots and other low-dimensional structures. We find that interface roughness, alloy disorder, impurities and other structural and compositional imperfections can dramatically alter device transport properties in ways that can only be understood properly in terms of a three-dimensional model in which quantum transport can be calculated exactly.

Although we do not treat the effects of electron-electron interactions, electron-phonon interactions and detailed band structure, we will give a brief overview of some of the important results in each of these areas and discuss how they could affect the properties of the structures we consider. We find that, for our purposes, a one-band, nearest neighbor, tight-binding Hamiltonian serves admirably to address the effects of interface roughness, alloy disorder and impurities, yielding important new insight in these areas.

The remainder of this chapter is organized as follows: we first present some background on the double barrier resonant tunneling structure, describing early theoretical and experimental work and some considerations of electron-electron interactions, electron-phonon interactions, band structure and work thus far on interface roughness, alloy disorder and impurity scattering. We then describe our supercell model and how it can be used to study not only double barriers but novel geometries and low-dimensional structures such as quantum wires and quantum dots as well. We present a brief overview of some of the important issues in one- and zero-dimensional structures, and we conclude with a summary of the thesis and our results.

1.2 Background

1.2.1 Early Investigations

We begin with a discussion of early theoretical and experimental efforts on the double barrier resonant tunneling structure. The basic operation of the double barrier can be described as follows. A typical double barrier structure consists of a quantum well of narrower band gap material (such as GaAs) in between two barrier layers of wider band gap material (such as AlAs). The structure is sandwiched between two heavily doped electrodes for carrier injection (see Figure 1.1). Confinement along the growth direction gives rise to a quasibound state in the well, since the barriers are neither infinitely thick nor infinitely high. The transmission coefficient for the double barrier, as shown in Figure 1.1, therefore exhibits a resonance of finite width centered on the quasibound level. At low bias, the quasibound level lies above the Fermi energy, and little current flows through the structure. As the bias is increased, the quasibound level is lowered below the Fermi energy, and a substantial number of electrons can tunnel resonantly through the double barrier, increasing the current. As bias is further increased, the quasibound level drops below the conduction band edge in the emitter, and electrons can no longer tunnel resonantly, leading to a reduction in current and negative differential resistance (see Figure 1.1).

The first calculation of current-voltage characteristics in the double barrier resonant tunneling structure is due to Tsu and Esaki [2]. In their model, the double barrier was assumed to possess perfect translational symmetry in the plane normal to the growth direction, and the transmission coefficient was calculated by solving a one-dimensional effective mass Schrödinger equation. An applied bias was assumed to produce a linear drop across the double barrier region, similar to that shown in Figure 1.1. The transmission coefficient was integrated over the Fermi distributions in the electrodes and over the in-plane momenta to yield the current. In this model, the current exhibits a peak when the applied bias is approximately twice the quasibound energy in the well (since the well band edge is assumed to drop by an amount equal to half the applied bias), and at higher bias the current drops sharply, leading to a large peak-to-valley current ratio and a narrow region of negative differential resistance, even at room temperature.

Shortly after this theory was presented, negative differential resistance was first observed experimentally [3] in GaAs/$Ga_{0.3}Al_{0.7}As$ double barriers. At 77 K, structures with 80 Å barriers and a 50 Å well showed rounded peaks in the current and negative differential resistance, but the peak-to-valley ratio was much lower than predicted by the theory of Tsu and Esaki. When the temperature was lowered to 4.2 K, the structure in the negative differential resistance did not sharpen, and the authors attributed this and the discrepancy from predictions in

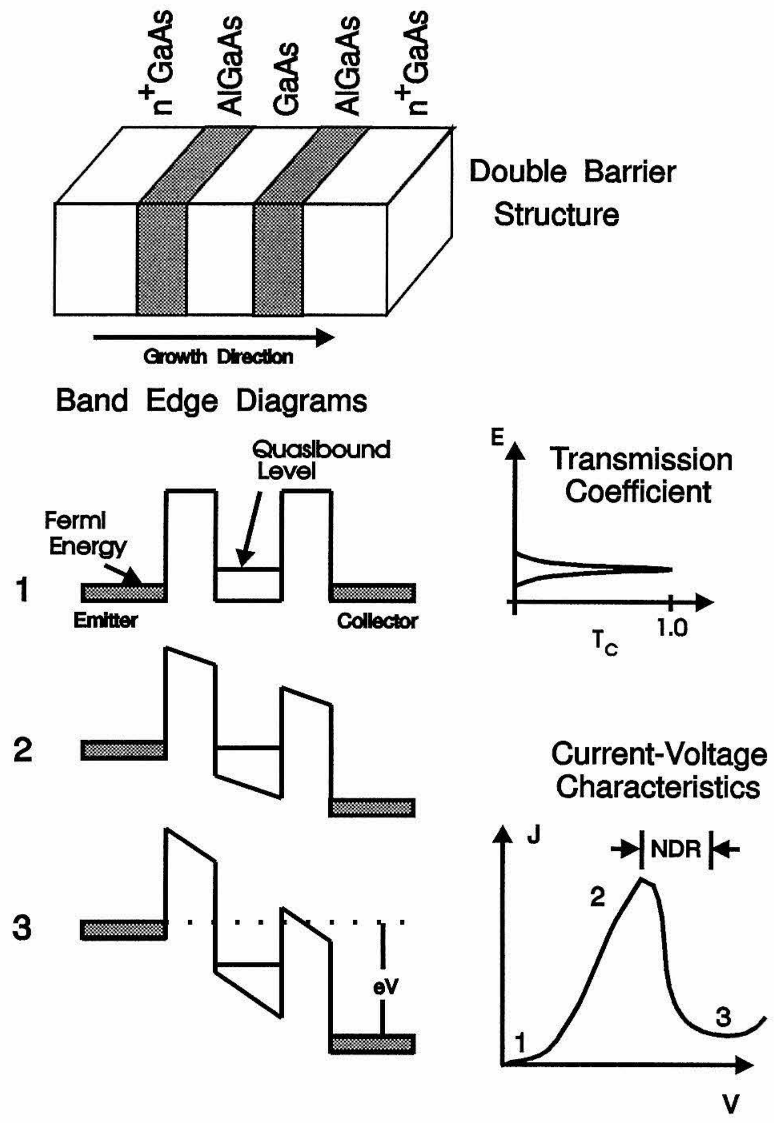

Figure 1.1: The basic operation of a double barrier resonant tunneling structure is as follows. At low bias, the quasibound level lies above the Fermi energy, and little current flows through the structure (stage 1). As the bias is increased, the quasibound level is lowered below the Fermi energy, and a substantial number of electrons can tunnel resonantly through the double barrier, increasing the current (stage 2). As bias is further increased, the quasibound level drops below the conduction band edge in the emitter, and electrons can no longer tunnel resonantly, leading to a reduction in current and negative differential resistance (stage 3).

the Tsu-Esaki model to structural fluctuations and impurity scattering in the samples used. Subsequent improvements in fabrication techniques led to improvements in performance of the double barrier resonant tunneling structure [11, 22, 23, 24], yet substantial departures from theory persisted. Most notably the peak-to-valley ratio remained much higher in theory than in experiment, owing to substantial valley current in real devices. In addition, early theories did not predict the intrinsic tristability [25] measured in real double barriers, as we shall discuss. Several phenomena are thought to be responsible for these effects. Band bending and space charge in the well, longitudinal optical phonon emission, real band structure effects and elastic scattering due to structural and compositional imperfections have been shown to impact the physics of resonant tunneling. In what follows, we give a brief overview of results in each of these areas.

1.2.2 Electron-Electron Interactions

We first consider some of the effects of electron-electron interactions. In semiconductor quantum structures, many effects of electron-electron interactions have been successfully treated with Thomas-Fermi theory, wherein electrons collectively give rise to a charge distribution leading to changes in the potential each electron senses. The accumulation of space charge in the electrodes and in the well causes screening of the applied bias and can lead to hysteresis and tristability, which we discuss below. This accumulation of charge density in double barriers has, in fact, been observed experimentally [26, 27, 28, 29].

Theoretical analysis of space charge in double barriers must account for the fact that, as bias is applied, the conduction band profile changes, leading to modified transmission characteristics and to a new charge distribution which in turn alters the conduction band profile, etc. A proper treatment calls for a simultaneous solution of the Schrödinger and Poisson equations at each applied bias [30, 31, 32, 33]. Schrödinger-Poisson self-consistent calculations begin with a guess for the conduction band potential profile, and then a transmission spectrum is calculated from which a charge distribution is determined. This is then used as a source term in the Poisson equation, from which a new potential profile is calculated. The procedure is repeated until the potential profile has converged to the desired degree of accuracy.

The resulting potential profiles exhibit band bending. In most double barrier structures, the barriers and the well are intrinsic, while the electrodes are heavily doped (typically between $10^{17}/cm^3$ and $2 \times 10^{18}/cm^3$). Therefore, at zero bias, the conduction band edge in the well lies above the conduction band edge in the emitter (see Figure 1.2). The details of the bending will depend upon whether or not undoped spacers are included between the electrodes and the double barrier. (Spacers are often included to reduce the number of ionized impurities in the barriers and in the well.) In general, the band bending increases with increasing electrode doping and with increasing temperature [33].

As bias is applied, an accumulation layer forms between the emitter and the left barrier, and a depletion layer forms between the collector and the right barrier (see Figure 1.2). The effect is a decrease in the bias across the double barrier region, although the position of the well resonance is not changed relative to the emitter Fermi energy unless the band bending in the left and right electrodes is unequal (as might be the case when the electrodes are doped with different concentrations, for example). The general impact of band bending, therefore, is that the rise in the conduction band edge in the well due to electrode doping leads to a higher threshold voltage for resonant tunneling, since the quasibound level in the well must be lowered more to establish resonance.

Another interesting effect involves accumulation of space charge in the well of a double barrier. Transport calculations accounting for space charge [34, 35, 36, 37, 38] have predicted intrinsic tristability (see Figure 1.3) and a region where the current is a triple-valued function of bias in double barriers on account of the direct Coulomb interaction. The extent of the region in which the current is

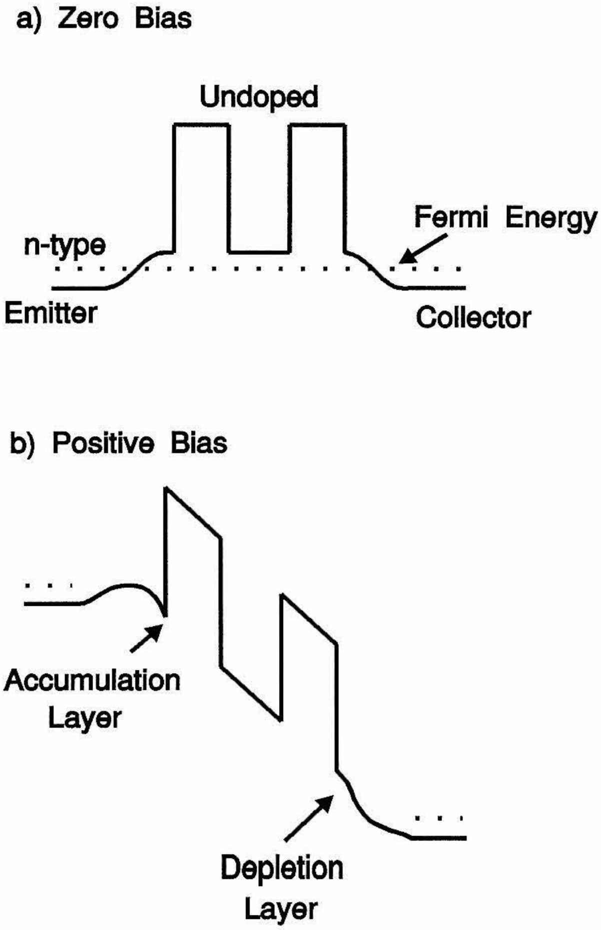

Figure 1.2: At zero bias, the conduction band edge adjusts to accommodate the difference in doping concentrations in the various regions of the double barrier. When a bias is applied, an accumulation layer forms between the emitter and the left barrier, and a depletion layer forms between the right barrier and the collector.

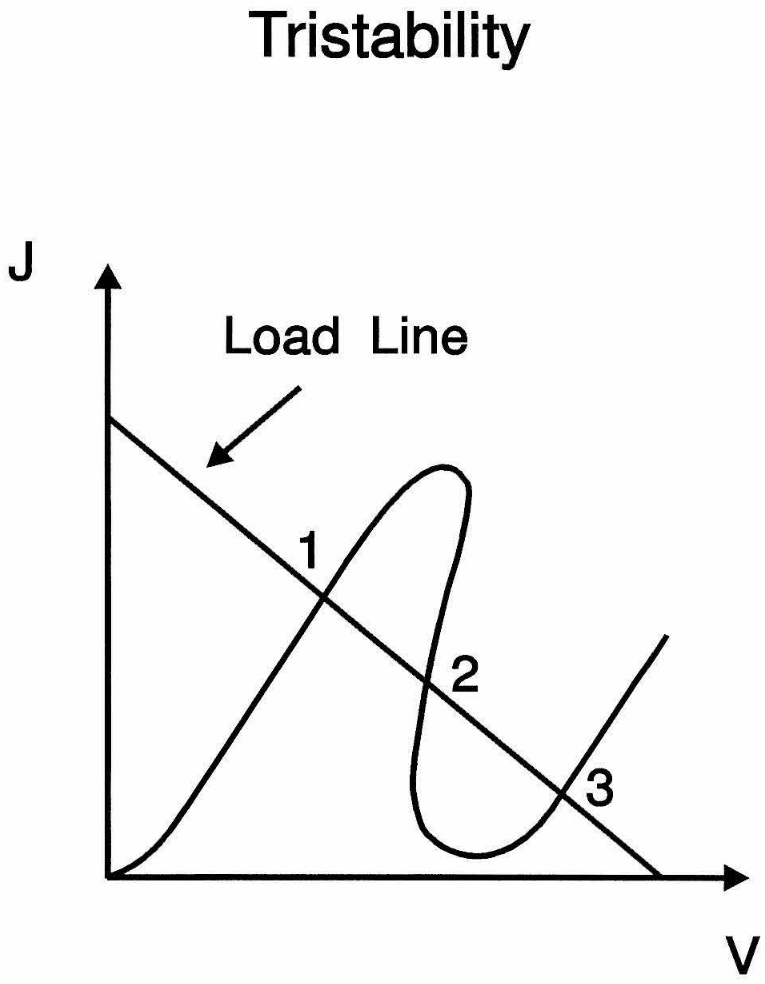

Figure 1.3: Due to the accumulation of space charge in the well, the current-voltage characteristic of a double barrier can exhibit a region where the current is a triple-valued function of the applied bias. A conventional load line analysis reveals three stable operating points.

a triple-valued function of bias is slightly reduced by exchange interactions [35]. Although tristability has only very recently been observed [25], Goldman et al. [39, 40, 41] first observed its consequences, reporting hysteresis in the current-voltage characteristics measured while sweeping the bias up and down in the region of peak current. This phenomenon depends on a feedback mechanism due to the buildup of significant charge density in the well as bias is applied, and it is hence much more prevalent in asymmetric double barriers where the collector barrier is higher and/or thicker than the emitter barrier, allowing charge to tunnel easily into the well near the resonance, but making escape difficult.

An intuitive explanation of hysteresis is as follows. At low bias, the quasibound state in the well is above the Fermi energy, little current flows, and there is little charge density in the well (region 1 in Figure 1.4). As bias increases, the quasibound level approaches the Fermi energy, significant current begins to flow, and charge begins to build up in the well. As the charge density increases with increasing bias, the resulting electric field causes the portion of the potential drop occurring across the left barrier to be less than that across the right barrier, and thus additional bias is required to reach peak current (region 2). When enough bias is applied so that the quasibound level finally drops below the emitter conduction band edge, the current drops (region 3), the space charge leaks out of the well, and the quasibound level drops to well below the conduction band edge in the emitter. Next the bias is lowered. Throughout region 4, the quasibound level remains below the emitter conduction band edge, and little current flows. When the bias is decreased to where the quasibound level again rises above the emitter conduction band edge, resonance is again established, and current increases (region 5).

Since hysteresis depends on the accumulation of significant charge density in the well, it is most prevalent in asymmetric structures, as described earlier. Since we do not consider structures where, at zero bias, the collector barrier is substantially higher or thicker than the emitter barrier, the effect of space charge in the well

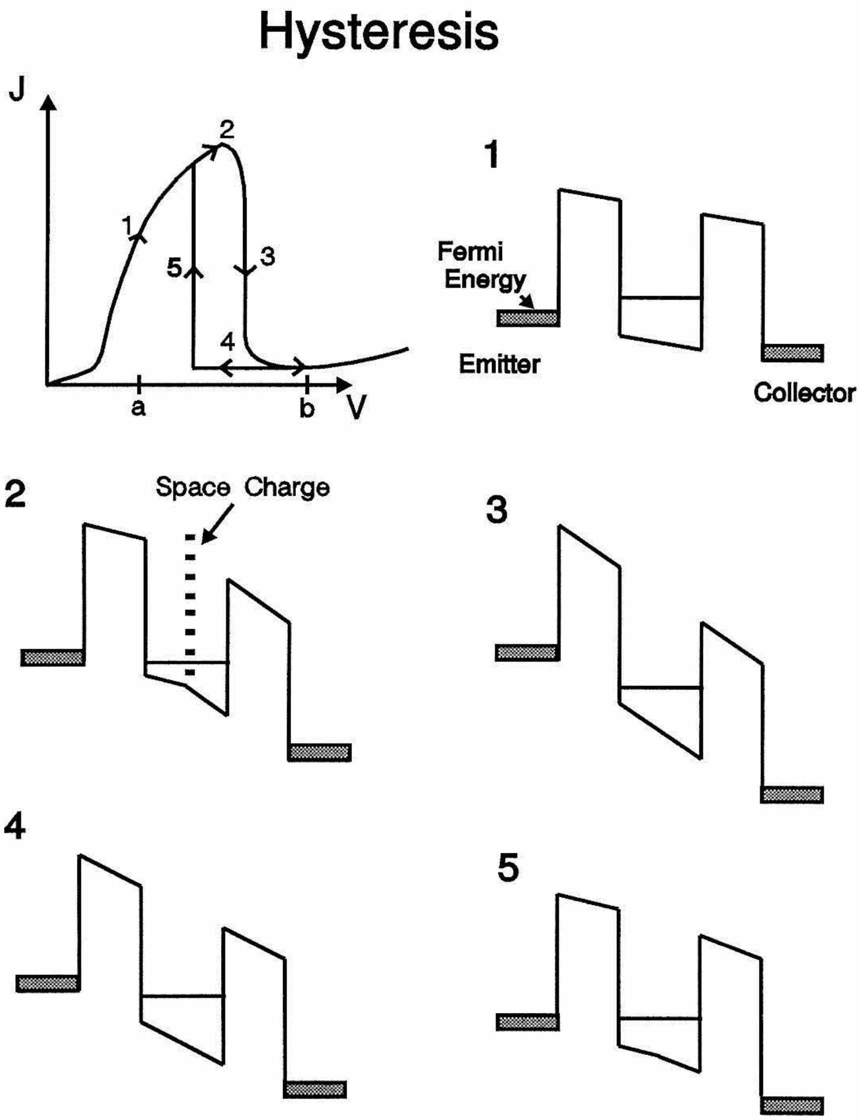

Figure 1.4: Conduction band edge diagrams of a double barrier at key points labeled on the schematic current-voltage characteristic. The bias is swept slowly up and down in the region from *a* to *b*. When bias is increasing from below resonance, charge builds up in the well, raising the band edge (part 2) and causing the peak current to occur at higher bias than when bias is decreasing from above resonance (parts 4 and 5).

should be minimal. For instance, in a typical example of a double barrier that we consider with a peak current density of $J \approx 10^5 A/cm^2$ and a quasibound state lifetime of $\tau = \hbar/\Delta E \approx 10^{-13}s$ (where ΔE is the resonance width), the areal charge density, σ, in the well should be $\sigma \approx J\tau \approx 10^{11}e^-/cm^2$ [42]. (For thicker barriers, the increase in τ is compensated by a decrease in peak current [13], so this estimate for σ should be representative for a range of double barriers.) This charge density gives rise to an electric field of $E = \sigma/2\epsilon_0\epsilon_r \approx 10^6 V/m$, where $\epsilon_r \approx 10$ is the barrier dielectric constant. This leads to an increase in the well band edge of only on the order of 1 meV in a double barrier with 10 Å thick barriers. This shift is negligible compared to the 60 meV shift in peak position due to alloy clustering as calculated in Chapter 5 (see Figure 5.6).

1.2.3 Electron-Phonon Interactions

Another factor which can influence the operation of the double barrier resonant tunneling structure is electron-phonon interactions. Electrons in the double barrier can interact with acoustic phonons (deformation potential) and with optical phonons (polarization field). Although acoustic phonons in double barriers have been considered [43, 44], their effect on transport is substantially less than that of optical phonons [43] due to the weaker electron-phonon coupling of the deformation potential. Electrons can, however, interact strongly with longitudinal optical phonons through the electric field of the polarization wave, which gives rise to a long range Coulomb interaction, different from the deformation potential interaction. Transverse optical phonons generally interact less strongly on account of their smaller electric field. Absorption and emission of phonons allows electrons to change energy and momentum en route through the double barrier enabling resonant tunneling from energies in the emitter different from the quasibound level in the well. This leads to replica peaks in the current-voltage characteristics as we discuss below. At low temperature phonon absorption is minimized, but phonon emission (notably longitudinal optical phonon emission) can still impact transmission [45].

The example of longitudinal optical phonon emission serves well to illustrate the effect of electron-phonon interactions. An intuitive description of this process, which can result in a peak or shoulder in the region of the valley current, follows. When the applied bias is high enough that the quasibound level in the well lies below the emitter conduction band edge, electrons cannot tunnel resonantly through the double barrier directly from the Fermi sea in the emitter. When the quasibound level lies within $\hbar\omega_{LO}$ of the band edge, however, electrons from the emitter can create a longitudinal optical phonon, losing energy $\hbar\omega_{LO}$, and tunnel resonantly through the well (see Figure 1.5). This leads to a replica peak in the valley current at a bias approximately $2\hbar\omega_{LO}/e$ above the bias required for peak current.

A number of quantitative theoretical models [43, 46, 47, 48, 49, 50, 51, 52, 53, 54] have investigated inelastic scattering by phonons. Phenomenological models [46, 47] have predicted sidebands of the main resonance at multiples of the phonon energy (corresponding to the emission and absorption of multiple phonons). An exactly solvable, one-dimensional model due to Wingreen et al. [48] predicts a downshifting and diminishing of the elastic transmission resonance peak in addition to the appearance of sidebands. Calculations based on Fermi's golden rule have shown [43, 49] that the inelastic current (due to phonon assisted tunneling) can be several orders of magnitude higher than the elastic current (without phonon assisted tunneling) in certain ranges of applied bias. Calculations involving the Wigner distribution function [52] and real-time path integral techniques [53, 54] have also been used to account for electron-phonon interactions.

Effects of electron-phonon interactions have also been observed experimentally, both in single barriers [55] and in double barriers [45, 56]. In a GaAs/$\mathrm{Al}_{0.4}\mathrm{Ga}_{0.6}\mathrm{As}$ double barrier, for example, there are three energies to consider: that of the longitudinal optical phonon in the pure GaAs well (36 meV), that of the GaAs-like

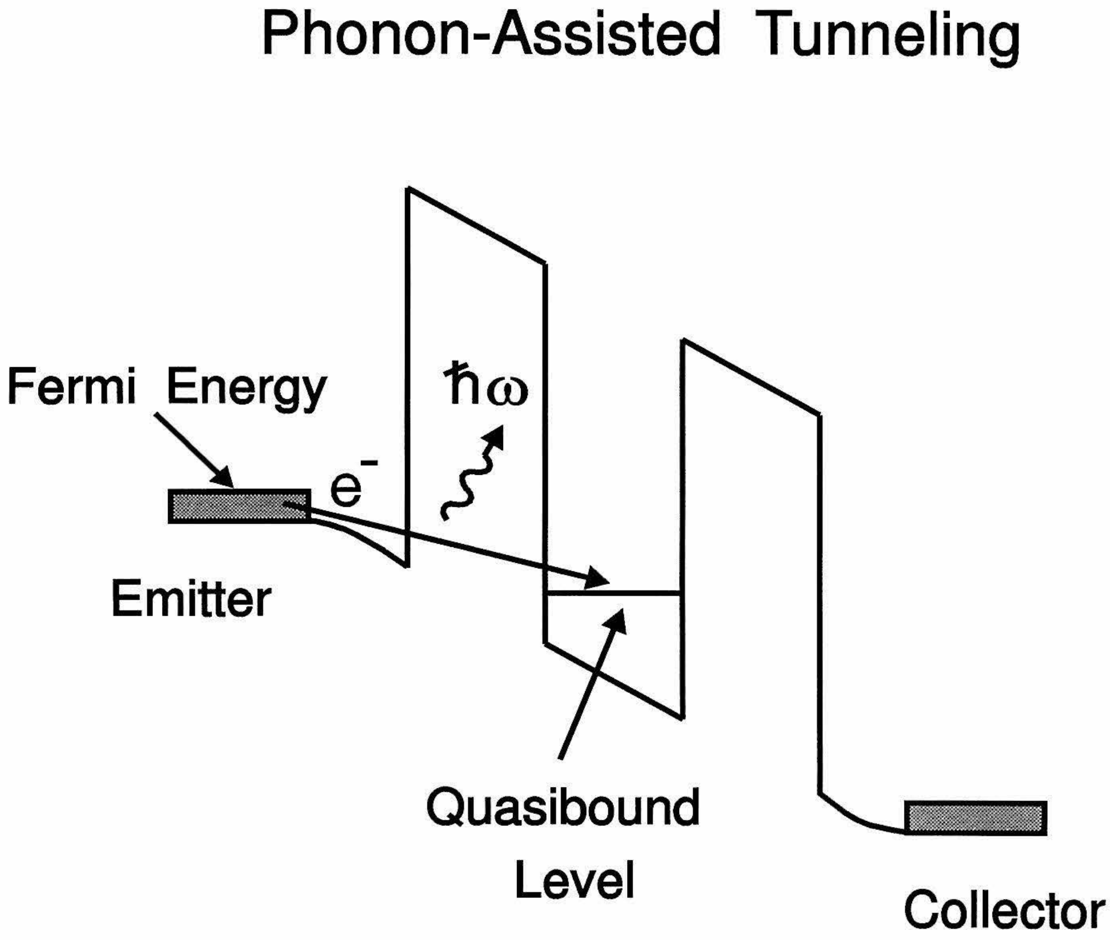

Figure 1.5: Electrons can emit a phonon of energy $\hbar\omega$ en route through a double barrier, allowing resonant tunneling at a bias above that for which elastic resonant tunneling is possible. This can result in the appearance of a replica peak in the valley current due to longitudinal optical phonon emission, for example. The electron-acoustic phonon coupling is much weaker and does not substantially alter the current-voltage characteristics.

phonon in the alloy barriers (35 meV) and that of the AlAs-like phonon in the barriers (47 meV) [57]. In a structure with a 56 Å thick well and 85 Å thick barriers, an AlAs-like longitudinal optical phonon emission peak in the valley current with a magnitude of 4% of that of the main elastic current peak was observed in the current-voltage characteristics at 4.2 K [45]. In a similar structure, Leadbeater et al. [56] observed interaction with both GaAs longitudinal optical phonons and with AlAs-like longitudinal optical phonons via studies involving a magnetic field.

As far as current-voltage characteristics are concerned, electron-phonon interactions mainly affect the operation of double barrier resonant tunneling structures through longitudinal optical phonon emission, contributing to increased valley current and a broadened negative differential resistance region which persists even as zero temperature is approached. As these effects are already fairly well understood, we shall be concerned with elastic scattering from structural and compositional imperfections such as interface roughness, alloy disorder and impurities.

1.2.4 Band Structure

The model of Tsu and Esaki, based on an effective mass Schrödinger equation, describes transport with a single conduction band having a single minimum at $\mathbf{k} = \mathbf{0}$. Detailed semiconductor band structures are actually much more complex. In Figure 1.6, we show the band structure of GaAs, which has the zinc-blende crystal structure, whose underlying Bravais lattice is the face centered cubic lattice. The Brillouin zone for the face centered cubic lattice is also shown in the figure with high symmetry points labeled. Local minima in the conduction band of GaAs occur at the $\Gamma-$point ($\mathbf{k} = \mathbf{0}$) as well as at the X- and L-points. The Γ-point minimum is the lowest, roughly 250 meV below the L-point minimum and 400 meV below the X-point minimum.

These details of the real band structure can lead to interesting effects. Mixing between states in different valleys in the conduction band (such as Γ-X mixing)

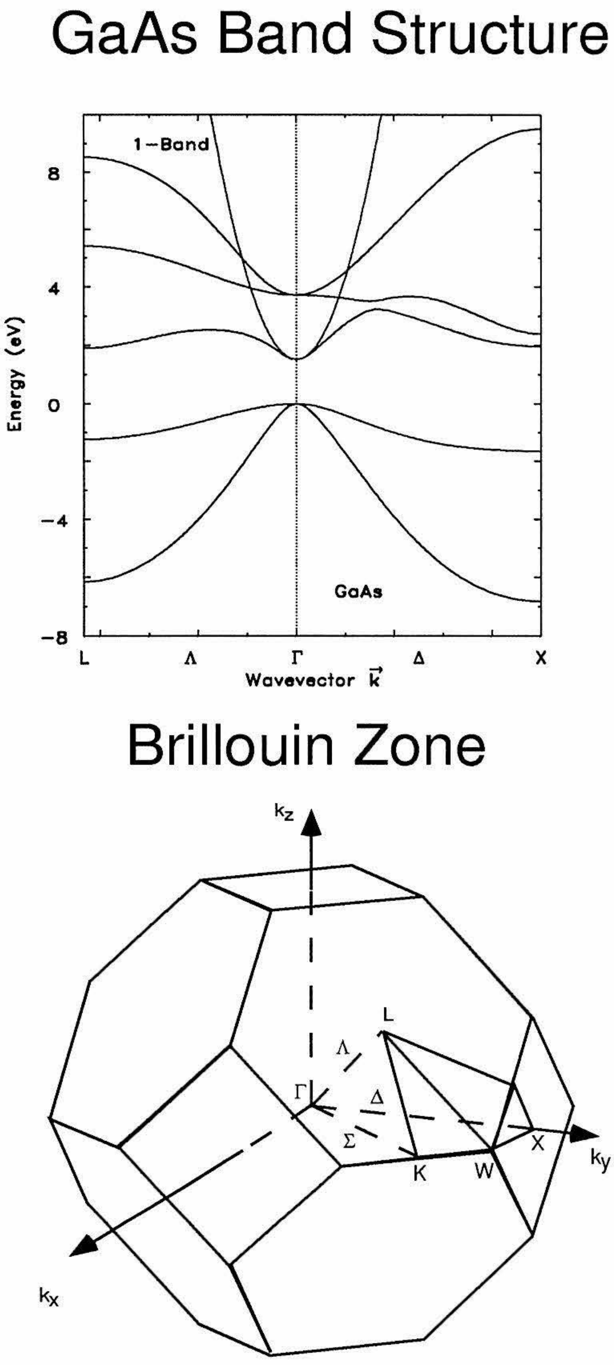

Figure 1.6: Band structure for GaAs. Symmetry points are labeled on the Brillouin zone diagram.

can impact tunneling [58, 59, 60]. Mixing of light holes and heavy holes in the valence bands can substantially affect hole-tunneling times in double barrier heterostructures [61]. Since we consider n-type devices, however, the majority carriers are electrons, so we will not be concerned with the valence bands. In addition, since the Fermi energy in the GaAs electrodes in our structures rarely exceeds 50 meV (corresponding to a doping of $2 \times 10^{18}/cm^3$), transport in our calculations is well described by the parabolic region near the Γ-point minimum (see Figure 1.6). In addition, it has been shown that Γ-X mixing is not critical in double barrier structures with wide wells and $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ barriers where $x < 0.4$ [59, 60], so a one-band, nearest neighbor tight-binding Hamiltonian serves admirably for our purposes.

1.2.5 Elastic Scattering

In addition to electron-electron interactions, electron-phonon interactions and real band structure effects, structural and compositional imperfections can play a vital role in transmission in nanostructures. Interface roughness [12] is thought to be a leading contributor to the valley current measured experimentally in double barrier structures [13, 14] at low temperature. Localized states due to defect impurities are believed to provide preferential current paths and to give rise to resonant tunneling in a variety of nanostructures [62, 63, 64, 65, 66]. Alloy disorder should also play a role in transport, especially when substantial clustering exists, as we demonstrate in Chapter 5.

Nevertheless, theoretical treatments [17, 18, 19, 20, 21] of these topics seem unsatisfactory, and quantitative understanding of the effects of elastic scattering is lacking. Models proposed thus far rely on perturbation theory or are essentially one-dimensional in nature, imposing restrictions on the effects which they can treat. For example, perturbation theory allows only investigation of the weak scattering limit, and important effects such as multiple scattering and virtual transitions are excluded from the analysis. In addition, correlations in the imperfections, such as clustering and ordering, are neglected, and the models cannot adequately address fluctuations. Limitation to one dimension also imposes restrictions. One-dimensional simulations are inherently unphysical and do not make a realistic account of scattering. In addition, they exaggerate disorder and structural imperfections. More importantly, the models to date are not capable of treating transport in low-dimensional structures such as quantum wires and dots.

1.3 Supercell Model

In order to understand the effects of structural and compositional imperfections in a variety of nanostructures, we propose a supercell model of quantum transport in three dimensions, capable of representing three-dimensional potential variations on an atomic scale. This flexibility allows us to treat elastic scattering due to interface roughness, alloy disorder and impurities in a physically realistic, three-dimensional setting. In addition, we can address strong scattering and correlation effects due to alloy clustering and interface island formation or impurity clustering. An added advantage of the model is the capability to investigate novel geometries and low-dimensional structures, such as quantum wires and quantum dots, with structural and compositional imperfections as well.

A basic description of the model is as follows. We model a three-dimensional device structure as a series of monolayer planes normal to the z-direction. Each plane consists of an infinite periodic array of identical rectangular supercells n_x sites in the x-direction and n_y sites in the y-direction, as in Figure 1.7. The sites for the supercell in a particular plane are chosen to reflect the properties of that plane. For example, if the plane represents a region of bulk material, the sites are identical. To represent a cross-sectional plane of a quantum dot with interface roughness and an impurity in the center we configure the supercell as in Figure 1.7. Three materials are represented: one for the impurity, and one each for the interior

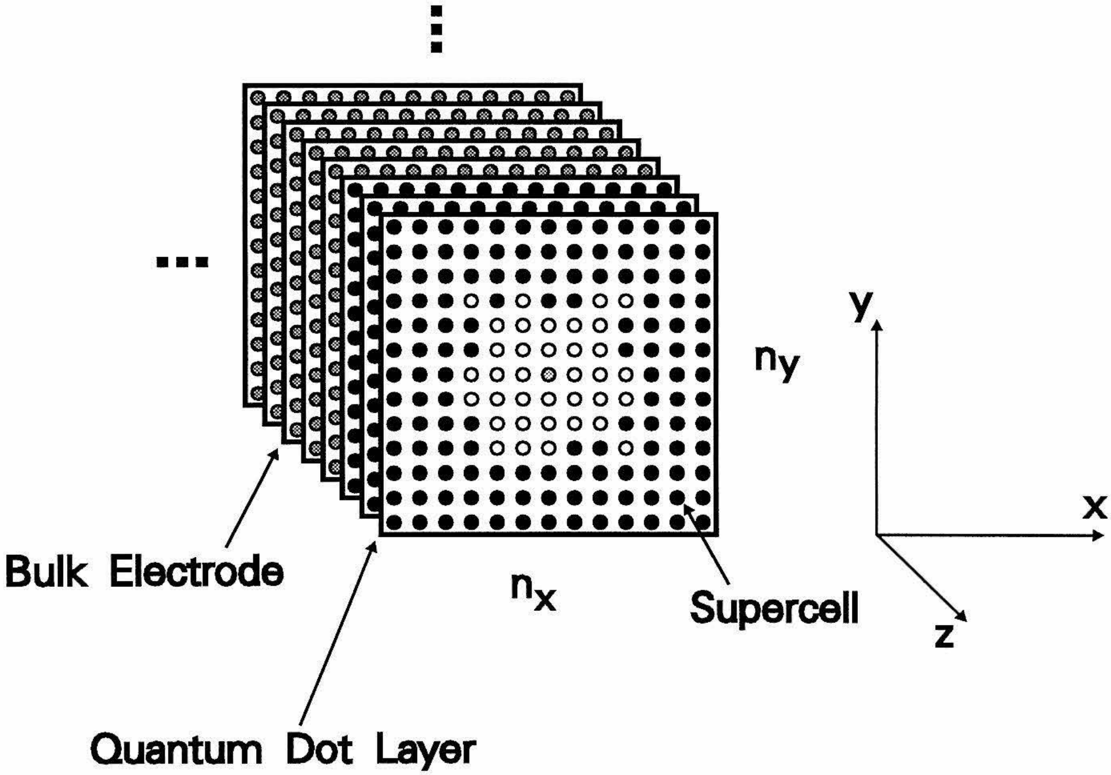

Figure 1.7: Supercell representation of a quantum dot resonator with rough walls and an impurity in the cavity. The supercells repeat in the planes normal to the z-direction. The darkly shaded sites represent the electrode, the solid sites represent the confining walls of the dot, the unshaded sites represent the well-type material of the quantum dot, and the lightly shaded site in the center represents an impurity.

of the dot and the confining region, which meet at a rough interface. Thus in the supercell method, the infinite layers normal to the z-direction are modeled by a finite supercell, and a device structure is specified by a finite series of supercells normal to the z-direction.

A drawback to this model is the fact that the supercells repeat in the x- and y-directions in the planes normal to the z-direction, imposing somewhat artificial periodic boundary conditions. This repetition of supercells can lead to artifacts in the transmission coefficient curves (see Chapters 4 and 5). To fully represent macroscopic cross sections, we would need to employ supercells with a computationally prohibitively large number of sites. We have generally found, however, that a 25×25 supercell is adequate for the issues we consider. In any event, our model is particularly well suited to simulating local probing over an area of a few nanometers on an edge, such as with scanning tunneling microscopy. As we shall see in Chapter 4, local probing of a single barrier with impurities can lead to a detailed resonance structure in the transmission coefficient. Whether or not this fine structure would be observed in a macroscopic sample would depend on the details of the impurity distribution. At low temperature, we might observe resonant transmission through impurities in a single barrier of macroscopic cross-section for a high concentration of well-isolated impurities confined to the middle barrier layer, for example (see Chapter 4).

A major advantage of our approach is that it allows us to study novel geometries such as quantum wires and quantum dots. These structures have stimulated great interest, offering both new physics and promise for new technologies. Just as the double barrier, however, these structures exhibit imperfections. Interface roughness over the scale of a few monolayers is currently unavoidable in etched quantum wires. In addition, compositional variation, particularly due to impurities, is difficult to eliminate. These structural and compositional imperfections play a vital role in the transport properties of one-dimensional and zero-dimensional structures. A small width increase in one place in a quantum wire has been shown to produce dips in the step-like conductance structure[67]. An isolated conductance peak observed below the turn-on of the first transverse mode in a narrow constriction has been attributed to resonant tunneling via a single impurity[15]. Degradation in the quantized conductance steps of a dual electron waveguide has been seen when the conductance channel is electrostatically steered into a scatterer[16]. In Chapter 6 we examine the impact of interface roughness and impurities on the transport properties of a quantum dot. We find that interface roughness over a single monolayer leads to substantial fluctuations in the transmission coefficient and that neutral impurities can dramatically alter the resonance modes of the dot. For background we present a brief overview of one- and zero-dimensional systems in the next section.

1.4 1D and 0D Systems

Laterally restricting a quasi-two-dimensional system, such as a quantum well, produces a quasi-one-dimensional system, where motion is free in only one direction and limited in the other two. This leads to the interesting and useful property of quantized conductance. When a small bias is applied along a quantum wire, the conductance as a function of the Fermi energy in the electrodes, E_F, is quantized in multiples of $2e^2/h$. Here we give a short derivation of the conductance in a quasi-one-dimensional system following Weisbuch [68].

The Schrödinger equation for an electron in a quasi-one-dimensional wire of size $L_x \times L_y \times L_z$ oriented along the z-direction can be written

$$[\frac{p_x^2 + p_y^2 + p_z^2}{2m^*} + V(x,y)]\psi(x,y,z) = E\psi(x,y,z), \tag{1.1}$$

where m^* is the effective mass of the electron in the wire, and $V(x,y)$ is the lateral confining potential. Since the motion is free along the z-direction, we can write

$$\psi(x,y,z) = \frac{1}{\sqrt{L_z}}\zeta_i(x,y)e^{ik_z z}, \tag{1.2}$$

for the ith subband, where $\zeta_i(x,y)$ satisfies

$$[\frac{p_x^2+p_y^2}{2m^*}+V(x,y)]\zeta_i(x,y)=E_i\zeta_i(x,y). \tag{1.3}$$

The subband dispersion relation is

$$E_{i,k_z}=E_i+\frac{\hbar^2k_z^2}{2m^*} \tag{1.4}$$

where

$$E_i=E_{n_x,n_y}=\frac{\hbar^2\pi^2}{2m^*}(\frac{n_x^2}{L_x^2}+\frac{n_y^2}{L_y^2}) \tag{1.5}$$

in the case of an infinitely high confining potential, for example. Each state contributes $e\hbar k_z/m^*L_z$ to the current. When a small bias V is applied, the chemical potential for states with $k_z>0$ in the left electrode lies eV above that for the states with $k_z<0$ in the right electrode, so the current from each subband for which $E_i<E_F$ is

$$I_i=\frac{e\hbar k_i(E_F)}{m^*L_z}\frac{D_i(E_F)}{2}eV, \tag{1.6}$$

where

$$k_i(E_F)=\frac{\sqrt{2m^*(E_F-E_i)}}{\hbar}, \tag{1.7}$$

and

$$D_i(E)=2\frac{g_sL_z}{\pi\hbar}\sqrt{\frac{m^*}{2(E-E_i)}} \tag{1.8}$$

is the familiar one-dimensional density of states assuming periodic boundary conditions of period L_z along the z-direction, and $g_s=2$ to account for electron spin. Thus the conductance from each subband is

$$I_i/V=\frac{2e^2}{h}. \tag{1.9}$$

The higher the Fermi energy in the electrodes, the more subbands there are available to carry current. In the case in which $L_x<<L_y$, for example, the conductance versus Fermi energy has a staircase-like structure, as we plot in Figure 1.8.

Whether or not the effects of this quantized conductance can be observed experimentally depends on the deviation of real quantum wires from ideality. Roughness

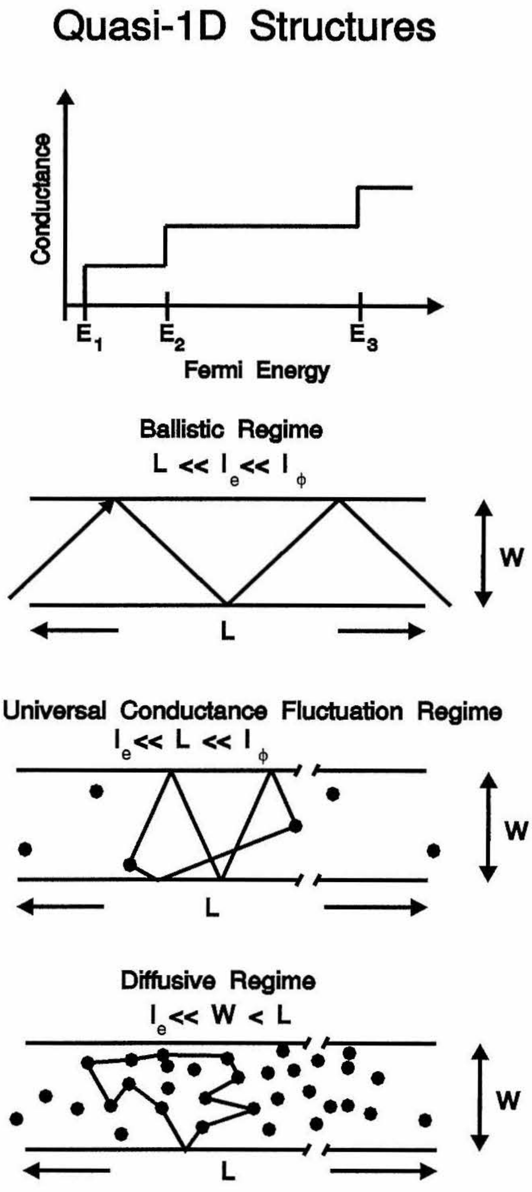

Figure 1.8: In the top panel, the staircase-like conductance versus Fermi energy for an ideal quasi-one-dimensional quantum wire with $L_x << L_y$, for example, is shown. The one-dimensional subband edges are labeled E_i. In the bottom panel, different transport regimes are shown. L is the length of the wire, W is a characteristic cross-sectional dimension, l_e is the elastic mean free path between scattering processes involving structural and compositional imperfections, and l_ϕ is the inelastic or phase breaking mean free path between phonon scattering events.

in the walls of the wire, impurities and phonons can all play a role in transport. (Thermal broadening can also smooth out the sharp step-like structures observed at low temperatures.) Elastic and inelastic scattering will contribute to the deterioration of ideal characteristics to different degrees in different regimes. To discuss the different regimes, it is convenient to define two length scales: l_e, the elastic mean free path between scattering processes involving structural and compositional imperfections, and l_ϕ, the inelastic or phase breaking mean free path between phonon scattering events. The relation between each of these lengths and the longitudinal and lateral dimensions, L and W, of the wire will determine what effects are important. In the ballistic regime (see Figure 1.8), $L, W, << l_e << l_\phi$, electrons sense only the confining potential of the structure, and the wire behaves ideally, giving quantized conductance. In the universal conductance fluctuation regime, $W << l_e << L << l_\phi$, and there are a few defects along the wire (see Figure 1.8) which can cause mixing of different wire modes, increasing the reflection probability for electrons entering the wire. Multiple scattering from impurities and the walls of the wire can lead to trapped states, localized on the length scale of l_e. These states no longer contribute to current. The behavior of the wire in this regime depends strongly on the particular configuration of the impurities. In the diffusive regime at low temperature, $l_e << W < L << l_\phi$, and impurity scattering dominates, so wire modes no longer have meaning. States are localized on the scale of l_e (see Figure 1.8) and no longer sense the confining potential of the structure. No states exist that extend from one end of the structure to the other, and at low temperatures, there will be no conductivity. Transport could, however, take place via inelastic scattering between localized states at higher temperatures. In the classical Boltzmann regime, $L, W >> l_\phi$, electrons diffuse through the wire, effectively averaging over impurity positions. Thus the temperature and dimensions of the wire and characteristics of structural and compositional imperfections will determine different regimes corresponding to quite different behavior.

One-dimensional structures have been fabricated using a number of techniques.

Lateral confinement of a quantum well has been achieved by deep mesa etch [69], electrostatic confinement [70], shallow etch [71], ion beam exposure [72, 73], and selective growth on a patterned substrate [74, 75, 76, 77]. Several device applications for quantum wires, such as the quantum modulated transistor [7] and the split-gate dual electron waveguide [8] with voltage tunable conductance properties have been suggested. Nonetheless, measured properties of these devices deviate substantially from predictions for ideal structures. Interface roughness over the scale of a few monolayers is currently unavoidable in etched quantum wires. In addition, compositional variation, particularly due to impurities, is difficult to eliminate. As a consequence, the effects of these variations on device performance have drawn considerable attention.

Theoretical studies of interface roughness in quantum wires have revealed alterations of the transmission spectra. A small width increase in one place in a quantum wire has been shown to produce dips in the step-like conductance structure[67]. It has also been shown that cross-sectional area variations along a wire lead to a smearing of the peak-like structure of the average density of states plotted as a function of carrier energy[78].

Impurities in quantum wires have been studied both experimentally and theoretically. An isolated conductance peak observed below the turn-on of the first transverse mode in a narrow constriction has been attributed to resonant tunneling via a single impurity[15]. Degradation in the quantized conductance steps of a dual electron waveguide has been seen when the conductance channel is electrostatically steered into a scatterer[16]. Theoretical studies of an impurity in a narrow channel have revealed the ways in which scattering alters the transmission properties[79, 80, 81]. In these papers, dips, peaks, and shifts in the conductance and transmission coefficient curve features as a function of impurity location and strength have been calculated. Calculations involving a T-shaped quantum wire junction have shown that a repulsive impurity can either enhance or suppress transmission[82]. Impurities near the aperture of an electron waveguide have been shown to destroy quantized conductance[83], and ionized donors have been shown to affect the quantized conductance of point contacts in a way that reflects the detailed configuration of the impurities[84].

Adding another degree of confinement, we come to quasi-one-dimensional structures, where the electron is confined in all dimensions, giving rise to a set of discrete levels. Since early work on tunneling in systems with small metal particles [85, 86, 87, 88], these structures have drawn much attention for their novel transport properties. Recently, periodic, two-dimensional arrays of quantum dots with an effective diameter on the order of 100 nm have been fabricated [9] using holographic lithography and deep mesa etch. A similar field effect array has been fabricated by depositing a metal gate over a photoresist mask on an n-AlGaAs/GaAs heterojunction [9]. With discrete levels, these dots act like artificial atoms, and periodic arrays of dots are suggestive of crystal lattices, stimulating renewed interest in band structure engineering [9, 89]. Quantum dots have also been proposed in applications such as cellular automata [68].

Tunneling through a quantum dot isolated from electrodes by thin barriers has attracted a great deal of attention. In the absence of a magnetic field, two main phenomena play a role in transport through a zero-dimensional structure: electron charging and energy quantization in the structure. In large metallic particles, the lowest empty electron energy levels are closely spaced, almost forming a continuum, and electron charging plays the dominant role, leading to the Coulomb blockade effect [90]. In small, semiconducting quantum dots, where only a few electrons are present, the lowest available levels are spaced further apart. When the level spacing is comparable to the single electron charging energy, both energy quantization and electron charging play a role in tunneling [91]. We will focus on the effects of structural and compositional imperfections on the quasibound levels in quantum dots. We shall see that imperfections can substantially impact the transmission properties of quantum dots.

Indeed, one of the main challenges in engineering quantum dots into useful devices will lie in achieving reproducibility and uniformity. Atomic scale variations in the structure of quantum dots lead to fluctuations in their properties. In this thesis, we examine fluctuations in the transmission coefficient of a quantum dot with interface roughness. We find that variations in both the stoichiometry and configuration of the roughness lead to fluctuations in the transmission resonance positions, widths and maxima. If these novel quantum structures are to find use in future technologies, these fluctuations must be understood.

1.5 Summary of Thesis and Results

The remainder of this thesis is organized as follows. In Chapter 2, we develop the formalism of our supercell model. Expressions for the transmission coefficient, electron wave function, probability current density and current-voltage characteristics are derived. We conclude with an indication of how our model could be adapted to incorporate more extensive band structure, which would extend the range of applicability to include interband transport and hole transport, for example.

In chapter 3 we develop the numerical tools for calculating transport in the supercell model. It is only by way of highly efficient numerical techniques that we are able to implement our exact three-dimensional model on presently available computers. Our numerical technique relies on a new method [92, 93, 94] for calculating quantum transport in the tight-binding model. The method formulates the quantum transport problem into a linear system of equations, overcoming instability problems which plague the transfer matrix method [95, 96] in structures with active regions longer than a few tens of Å [97]. For a typical device structure, calculation of a single transmission coefficient at a given energy requires solving a $40,000 \times 40,000$ system of equations. This presents a formidable challenge, both in terms of execution time and storage requirements. We give an overview of the various methods we have considered for solving large, sparse linear systems and for storing sparse matrices. We then describe the particulars of our implementation and present storage and execution time benchmarks for some typical calculations. In certain cases our calculations are highly amenable to parallel computing, and thus we conclude the section on numerics with a discussion of various topics in concurrent computing.

In Chapter 4, we present our results on transport in single and double barrier tunneling structures with neutral impurities. We find that an isolated attractive impurity in a single barrier can produce a transmission resonance whose position and strength are sensitive to the location of the impurity within the barrier. We also study transmission in the presence of two closely spaced impurities as a function of their separation and orientation relative to the incident plane wave. Multiple impurities can lead to a complex resonance structure that fluctuates widely with impurity configuration. In addition, impurity resonances can give rise to negative differential resistance.

In Chapter 5, we study interface roughness and alloy disorder in double barrier structures. We find that interface roughness can affect transmission in two ways: in-plane momentum ($\mathbf{k}_{||}$) scattering produces a transmission enhancement just above the $n = 1$ resonance, and wave function localization broadens and reduces the energy of the $n = 1$ resonance. We also find that the degree of disorder and clustering in the alloy barriers of a double barrier structure has a dramatic impact on transmission. An analysis of the transmission coefficient curve for different cluster sizes reveals that as the cluster size increases, the barriers grow less confining, broadening resonances and shifting them to lower energy. In addition, localized states arise, leading to new transmission resonance structure.

In Chapter 6 we examine the effects of atomic scale imperfections on the transmission properties of a quantum dot resonator. We find that variation in the surface roughness of quantum dots leads to substantial fluctuations in the transmission properties. Impurities in a quantum dot are studied as a function of impurity strength and location, and it is found that an attractive impurity near the center of a dot can reduce fluctuations caused by surface roughness. Nevertheless, the presence of more than a single impurity can give rise to a complex resonance structure that varies with impurity configuration.

Bibliography

[1] D. Bohm, *Quantum Theory* (Prentice-Hall, Englewood Cliffs, NJ, 1951), p. 283.

[2] R. Tsu and L. Esaki, Appl. Phys. Lett. **22**, 562 (1973).

[3] L. L. Chang, L. Esaki, and R. Tsu, Appl. Phys. Lett. **24**, 593 (1974).

[4] S. Luryi, in *Heterojunction and Band Discontinuities: Physics and Device Applications*, edited by F. Capasso and G. Margaritondo (North-Holland, Amsterdam, 1987), p. 489.

[5] F. Capasso, *Physics of Quantum Electron Devices* (Springer-Verlag, Berlin, 1990), p. 283.

[6] H. J. Levy and T. C. McGill, IEEE Trans. Neural Netw. **4**, 427 (1993).

[7] F. Sols, M. Macucci, U. Ravaioli, and K. Hess, Appl. Phys. Lett. **54**, 350 (1989).

[8] C. C. Eugster, J. A. del Alamo, M. J. Rooks, and M. R. Melloch, Appl. Phys. Lett. **60**, 642 (1992).

[9] D. Heitmann, and J. P. Kotthaus, Phys. Today **46**, 56 (1993).

[10] B. Ricco and M. Ya. Azbel, Phys. Rev. B **29**, 1970 (1984).

[11] T. C. L. G. Sollner, W. D. Goodhue, P. E. Tannenwald, C. D. Parker, and D. D. Peck, Appl. Phys. Lett. **43**, 588 (1983).

[12] H. Sakaki, T. Noda, K. Hirakawa, M. Tanaka, and T. Matsusue, Appl. Phys. Lett. **51**, 1934 (1987).

[13] P. Gueret, C. Rossel, E. Marclay, and H. Meier, J. Appl. Phys. **66**, 278 (1989).

[14] P. Gueret, C. Rossel, W. Schlup, and H. P. Meier, J. Appl. Phys. **66**, 4312 (1989).

[15] P. L. Mceuen, B. W. Alphenaar, and R. G. Wheeler, Surf. Sci. **229**, 312, (1990).

[16] C. C. Eugster, J. A. del Alamo, M. R. Melloch, and M. J. Rooks, Phys. Rev. B **46**, 10,146 (1992).

[17] P. A. Schulz, and C. E. T. Goncalves da Silva, Phys. Rev. B **38**, 10718 (1988).

[18] M. Tabe and M. Tanimoto, J. Appl. Phys. **67**, 593 (1990); H. C. Liu and D. D. Coon, J. Appl. Phys. **64**, 6785 (1988).

[19] J. Leo and A. H. MacDonald, Phys. Rev. Lett. **64**, 817 (1990); J. Leo and A. H. MacDonald, Phys. Rev. B **43**, 9763 (1991).

[20] P. Roblin and W.-R. Liou, Phys. Rev. B **47**, 2146 (1993).

[21] F. Chevoir and B. Vinter, Phys. Rev. B **47**, 7260 (1993).

[22] T. C. L. G. Sollner, P. E. Tannenwald, D. D. Peck, and W. D. Goodhue, Appl. Phys. Lett. **45**, 1319 (1984).

[23] M. Tsuchiya, H. Sakaki, and J. Yoshino, Jpn. J. Appl. Phys. **24**, L466 (1985).

[24] V. J. Goldman, D. C. Tsui, J. E. Cunningham, and W. T. Tsang, J. Appl. Phys. **61**, 2693 (1987).

[25] A. D. Martin, M. L. F. Lerch, P. E. Simmonds, L. Eaves, and M. L. Leadbeater, presented at the 8th International Conference on Hot Carriers in Semiconductors, Oxford, UK (1993).

[26] V. J. Goldman, D. C. Tsui, and J. E. Cunningham, Phys. Rev. B **35**, 9387 (1987).

[27] M. L. Leadbeater, E. S. Alves, F. W. Sheard, L. Eaves, M. Henini, O. H. Hughes, and G. A. Toombs, J. Phys. Condens. Matter **1**, 10605 (1989).

[28] I. Bar-Joseph, T. K. Woodward, D. S. Chemla, Y. Gedalyahu, A. Yacoby, D. Sivco, and A. Y. Cho, Superlattices and Microstructures **8**, 409 (1990).

[29] T. K. Woodward, D. S. Chemla, I. Bar-Joseph, H. U. Baranger, D. L. Sivco, A. Y. Cho, Phys. Rev. B **44**, 1353 (1991).

[30] Y. Rajakarunanayake and T. C. McGill, J. Vac. Sci. Tech. B **5**, 1288 (1991).

[31] H. Ohnishi, T. Inata, S. Muto, N. Yokoyama, and A. Shibatomi, Appl. Phys. Lett. **49**, 1248 (1986).

[32] M. Cahay, M. McLennan, S. Datta, and M. S. Lundstrom, Appl. Phys. Lett. **50**, 612 (1987).

[33] K. F. Brennan, J. Appl. Phys. **62**, 2392 (1987).

[34] H. L. Berkowitz and R. A. Lux, J. Vac. Sci. Technol. B **5**, 967 (1987).

[35] K. M. S. V. Bandara and D. D. Coon, Appl. Phys. Lett. **53**, 1865 (1988).

[36] D. D. Coon, K. M. S. V. Bandara, and H. Zhao, Appl. Phys. Lett. **54**, 2115 (1989).

[37] F. W. Sheard and G. A. Toombs, Appl. Phys. Lett. **52**, 1228 (1988).

[38] F. W. Sheard and G. A. Toombs, Semicond. Sci. and Technol. **7**, B460 (1992).

[39] V. J. Goldman, D. C. Tsui, and J. E. Cunningham, Phys. Rev. Lett. **58**, 1256 (1987); *ibid.*, p. 1623.

[40] V. J. Goldman, D. C. Tsui, and J. E. Cunningham, Journal de Physique **C5**, 463 (1987).

[41] A. Zaslavsky, V. J. Goldman, and D. C. Tsui, Appl. Phys. Lett. **53**, 1408 (1988).

[42] Strictly speaking, the τ we should use here is not equal to the lifetime [29]. However, in symmetric double barriers, it is almost identical to $\hbar/\Delta E$.

[43] G. Y. Wu and T. C. McGill, Phys. Rev. B **40**, 9969 (1989).

[44] L. I. Glazman and R. I. Shekter, Solid State Commun. **66**, 65 (1988).

[45] V. J. Goldman, D. C. Tsui, and J. E. Cunningham, Phys. Rev. B **36**, 7635 (1987).

[46] A. D. Stone, M. Ya. Azbel, and P. A. Lee, Phys. Rev. B **31**, 1707 (1985).

[47] A.-P. Jauho, Phys. Rev. B **41**, 12327 (1990).

[48] N. S. Wingreen, K. W. Jacobsen, and J. W. Wilkins, Phys. Rev. Lett. **61**, 1396 (1988); N. S. Wingreen, K. W. Jacobsen, and J. W. Wilkins, Phys. Rev. B **40**, 11834 (1989).

[49] F. Chevoir and B. Vinter, Appl. Phys. Lett. **55**, 1859 (1989).

[50] W. Cai, T. F. Zheng, P. Hu, B. Yudanin, and M. Lax, Phys. Rev. Lett. **63**, 418 (1989).

[51] B. G. R. Rudberg, Semicond. Sci. and Technol. **5**, 328 (1990).

[52] W. R. Frensley, Solid-State Electronics **31**, 739 (1988).

[53] B. A. Mason, K. Hess, R. E. Cline, and P. G. Wolynes, Superlattices and Microstructures **3**, 421 (1987).

[54] B. A. Mason and K. Hess, Phys. Rev. B **39**, 5051 (1989).

[55] R. T. Collins, J. Lambe, T. C. McGill, and R. D. Burnham, Appl. Phys. Lett. **44**, 532 (1984).

[56] M. L. Leadbeater, E. S. Alves, L. Eaves, M. Henini, O. H. Hughes, A. Celeste, J. C. Portal, G. Hill, and M. A. Pate, Phys. Rev. B **39**, 3438 (1989).

[57] R. Tsu, H. Kawamura, and L. Esaki, *Proceedings of the 11th International Conference on the Physics of Semiconductors, Warsaw, Poland, 1972* (PWN, Warszawa, 1972), p. 1135.

[58] D. Landheer, H. C. Liu, M. Buchanan, and R. Stoner, Appl. Phys. Lett. **54**, 1784 (1989).

[59] D. Z.-Y. Ting and T. C. McGill, J. Vac. Sci. Technol. B **10**, 1980 (1992).

[60] D. Z.-Y. Ting and T. C. McGill, Phys. Rev. B **47**, 7281 (1993).

[61] D. Z.-Y. Ting, E. T. Yu, and T. C. McGill, Phys. Rev. B **45**, 3576 (1992).

[62] R. H. Koch and A. Hartstein, Phys. Rev. B **54**, 1848 (1985).

[63] S. J. Bending and M. R. Beasley, Phys. Rev. Lett. **55**, 324 (1985).

[64] F. Capasso, K. Mohammed, and A. Y. Cho, Phys. Rev. Lett. **57**, 2303 (1986).

[65] M. Tabe and M. Tanimoto, Appl. Phys. Lett. **58**, 2105 (1991).

[66] M. W. Dellow, P. H. Beton, C. J. G. M. Langerak, T. J. Foster, P. C. Main, L. Eaves, M. Henini, S. P. Beaumont, and C. D. W. Wilkinson, Phys. Rev. Lett. **68**, 1754 (1992).

[67] T. Itoh, S. Nobuyuki, and A. Yoshii, Phys. Rev. B **45**, 14,131 (1992).

[68] C. Weisbuch and B. Vinter, *Quantum Semiconductor Structures* (Academic Press Inc., San Diego, 1991), p. 189.

[69] K. K. Choi, D. C. Tsui, and K. Alavi, Appl. Phys. Lett. **50**, 110 (1987).

[70] T. J. Thornton, M. Pepper, H. Ahmed, D. Andrews, and G. J. Davies, Phys. Rev. Lett. **56**, 1198 (1986).

[71] H. van Houten, B. J. van Wees, M. G. J. Heijman, and J. P. Andre, Appl. Phys. Lett. **49**, 1781 (1986).

[72] A. Scherer, M. L. Roukes, H. G. Craighead, R. M. Ruthen, E. D. Beebe, and J. P. Harbison, Appl. Phys. Lett. **51**, 2133 (1987).

[73] T. L. Cheeks, M. L. Roukes, A. Scherer, and H. G. Craighead, Appl. Phys. Lett. **53**, 1964 (1988).

[74] Y. Qian, J. M. Zhang, and J. Y. Xu, Superlatt. Microstruct. **13**, 241 (1993).

[75] K. Inoue, K. Kimura, and K. Maehashi, J. Cryst. Growth **127**, 1041 (1993).

[76] X. Q. Shen, M. Tanaka, and T. Nishinaga, J. Cryst. Growth **127**, 932 (1993).

[77] S. Tsukamoto, Y. Nagamune, and M. Nishioka, Appl. Phys. Lett. **62**, 49 (1993).

[78] V. V. Mitin, Superlatt. Microstruct. **8**, 413 (1990).

[79] E. Tekman and S. Ciraci, Phys. Rev. B **42**, 9098 (1990).

[80] P. F. Bagwell, Phys. Rev. B **41**, 10,354 (1990); A. Kumar and P. F. Bagwell, Phys. Rev. B **43**, 9012 (1991); P. F. Bagwell, T. P. Orlando, and A. Kumar, in *Resonant Tunneling in Semiconductors*, edited by L. L. Chang et al., (Plenum Press, New York, 1991). p. 417.

[81] C. S. Chu and R. S. Sorbello, Phys. Rev. B **40**, 5941 (1989).

[82] Y. Takagaki and D. K. Ferry, Phys. Rev. B **45**, 6715 (1992).

[83] D. van der Marel and E. G. Haanappel, Phys. Rev. B **39**, 7811 (1989); E. G. Haanappel and D. van der Marel, Phys. Rev. B **39**, 5484 (1989).

[84] J. A. Nixon, J. H. Davies, and H. U. Baranger, Phys. Rev. B **43**, 12,638 (1991).

[85] I. Giaever and H. R. Zeller, Phys. Rev. Lett. **20**, 1504 (1968).

[86] H. R. Zeller and I. Giaever, Phys. Rev. **181**, 789 (1969).

[87] J. Lambe and R. C. Jaklevic, Phys. Rev. Lett. **22**, 1371 (1969).

[88] J. A. A. J. Perenboom, P. Wyder, and F. Meier, Phys. Reports **78**, 173 (1981).

[89] L. Esaki and R. Tsu, IBM J. Res. Dev. **14**, 61 (1970).

[90] M. A. Kastner, Physics Today **46**, 24 (1993).

[91] H. van Houten, C. W. J. Beenakker, and A. A. M. Staring, in *Single Charge Tunneling*, edited by H. Grabert and M. H. Devoret (Plenum, New York, 1992), p. 167.

[92] D. Z.-Y. Ting, E. T. Yu, and T. C. McGill, Phys. Rev. B **45**, 3583 (1992).

[93] C. S. Lent and D. J. Kirkner, J. Appl. Phys. **67**, 6353 (1990).

[94] W. R. Frensley (private communication).

[95] E. O. Kane, in *Tunneling Phenomena in Solids*, edited by E. Burstein and S. Lundqvist (Plenum, New York, 1969), p. 1.

[96] J. N. Schulman and Y. C. Chang, Phys. Rev. B **27**, 2346 (1983).

[97] C. Mailhiot and D. L. Smith, Phys. Rev. B **33**, 8360 (1986).

Chapter 2

The Supercell Model

2.1 Formalism

A three-dimensional, one-band, nearest neighbor, rectangular lattice tight-binding model forms the basis for all calculations in this thesis. A solid is represented with a rectangular lattice, each site of which is assigned a material type, specified by a band edge and an effective mass. This translates into assigning an onsite energy to each site in the lattice and a hopping matrix element to the bond between each nearest neighbor pair of sites. A uniform bulk region, for example, is represented by assigning the same onsite energy to each site and the same hopping matrix element to each nearest neighbor bond in the region. This yields a cosine-shaped band structure, as shown below. In a disordered alloy region, by contrast, the onsite energies and hopping matrix elements vary throughout. The model thus accounts for potential variations in three dimensions.

Representing a macroscopic sample in this manner would require on the order of 10^{23} sites, a prohibitively large number for present-day computers. We therefore apply a planar supercell method to the model, implementing periodic boundary conditions. We model a device structure as a set of monolayer planes along the $z-$direction. (The $z-$axis is chosen along the direction separating the electrodes by which the structure is probed.) Each plane consists of an infinite periodic array of identical rectangular supercells n_x sites in the x-direction and n_y sites in the y-direction, as in Figure 2.1. The sites for the supercell in a particular plane are chosen to reflect the properties of that plane. For example, if the plane represents a region of bulk material, the sites are identical. To represent an impurity in a particular layer, we choose the supercell for that layer to contain a site representing the impurity, and we assign to the other sites the appropriate type of surrounding material. To represent the binary alloy A_xB_{1-x}, we assign material of type A to a fraction x of the sites, and material of type B to the remaining sites. Thus the infinite layers along the $z-$direction are modeled by a finite supercell, and a device structure is specified by a finite set of supercells along the $z-$direction.

The nearest neighbor tight-binding Hamiltonian for a structure can be written

$$H = \sum_{\mathbf{n}} \epsilon_{\mathbf{n}} |\mathbf{n}\rangle\langle\mathbf{n}| + \sum_{<\mathbf{nm}>} t_{\mathbf{nm}} |\mathbf{n}\rangle\langle\mathbf{m}|. \tag{2.1}$$

The $\{|\mathbf{n}\rangle\}$ are orbitals localized at the lattice sites, the $\{\epsilon_{\mathbf{n}}\}$ are the onsite energies, and the $\{t_{\mathbf{n},\mathbf{m}}\}$ are hopping matrix elements. The second sum extends over all nearest neighbor pairs of sites in the lattice. As stated above, site $|\mathbf{n}\rangle$ is characterized by a particular type of material with band edge $E_{\mathbf{n}}$ and effective mass $m_{\mathbf{n}}$. In terms of these material parameters, the onsite and hopping matrix elements are

$$\begin{aligned} \epsilon_{\mathbf{n}} &= E_{\mathbf{n}} - \sum_{\mathbf{m}} t_{\mathbf{nm}}, \\ t_{\mathbf{nm}} &= -\frac{1}{2{d_{\mathbf{n},\mathbf{m}}}^2}(\frac{\hbar^2}{2m_{\mathbf{n}}} + \frac{\hbar^2}{2m_{\mathbf{m}}}). \end{aligned} \tag{2.2}$$

The sum in the first line above runs over all nearest neighbors $\mathbf{m}$ of $\mathbf{n}$. The parameter $d_{\mathbf{n},\mathbf{m}}$ is the distance between sites $\mathbf{n}$ and $\mathbf{m}$.

It should be noted that Eq. (2.2) implies that the hopping matrix element between sites of different materials is taken as the arithmetic mean of the hopping matrix elements of bulk samples of each of the materials. This, along with the dependence of the onsite energy on the hopping matrix elements to the nearest

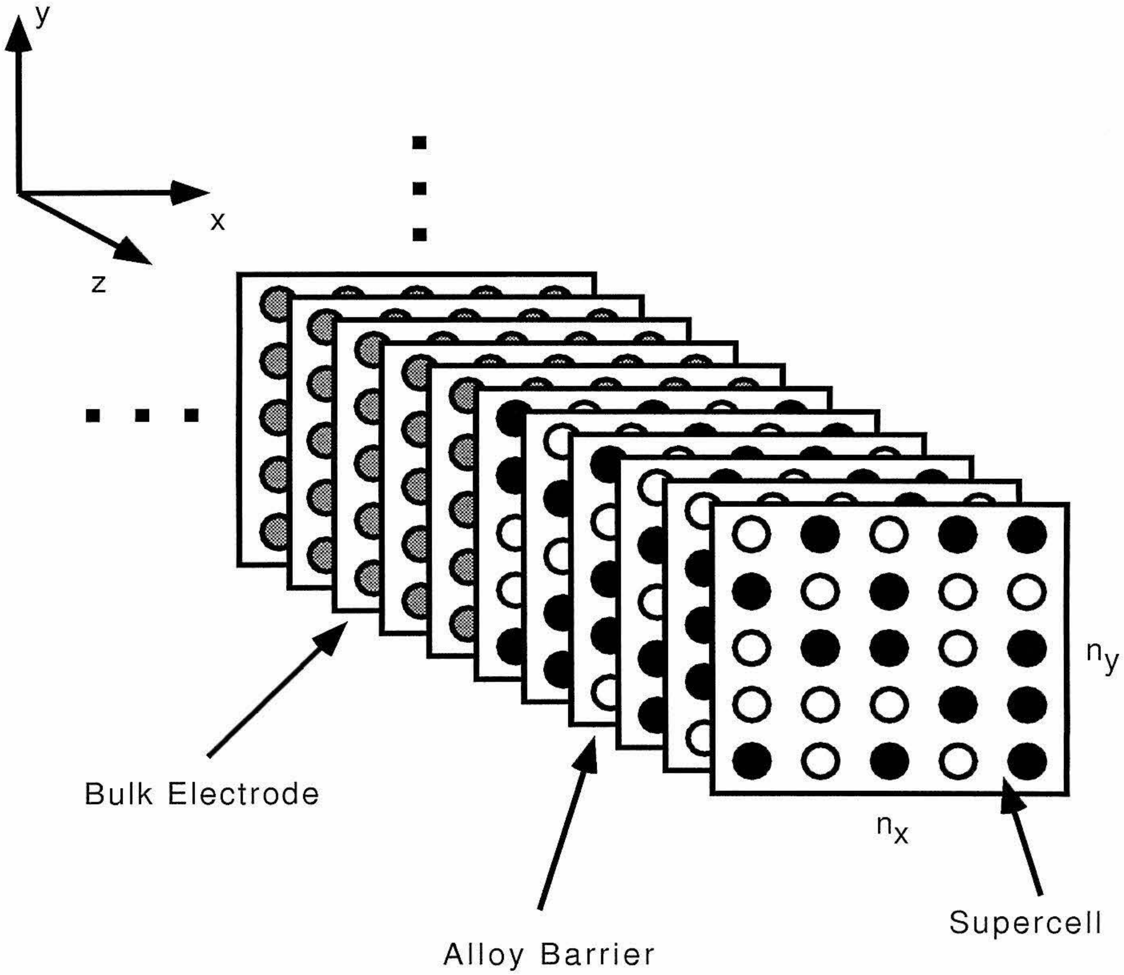

Figure 2.1: 5 × 5 supercell representation of an electrode followed by an alloy region. The supercells repeat in the $x-$ and $y-$directions. In the tight-binding model, an onsite energy corresponds to each site, and a hopping matrix element corresponds to each nearest neighbor pair of sites.

neighbors, stems from a discretization [1, 2] of the simplest manifestly Hermitian Hamiltonian incorporating a varying effective mass, namely

$$H = -\frac{\hbar^2}{2}\nabla \cdot \frac{1}{m^*(\mathbf{r})}\nabla + V(\mathbf{r}). \tag{2.3}$$

The $\{d_{\mathbf{n},\mathbf{m}}\}$ in Eq. (2.2) are the discretization lengths, chosen based on typical dimensions and the rate of change of the potential in a particular problem. More complicated formulations have been proposed [3], but the above serves well when the variation in the effective mass is not too large [1].

The definitions in Eq. (2.2) are familiar in the case of a bulk region of uniform onsite energy ϵ and effective mass m. The Hamiltonian then becomes

$$H = \epsilon \sum_{\mathbf{n}} |\mathbf{n}\rangle\langle\mathbf{n}| + t \sum_{<\mathbf{nm}>} |\mathbf{n}\rangle\langle\mathbf{m}|, \tag{2.4}$$

and, due to complete translational symmetry by any direct lattice vector, the eigenstates can be chosen with definite crystal momentum $\mathbf{k}$:

$$|\mathbf{k}\rangle = \frac{1}{\sqrt{N}} \sum_{\mathbf{n}\in\Gamma} e^{i\mathbf{k}\cdot\mathbf{n}}|\mathbf{n}\rangle. \tag{2.5}$$

Here N is the number of sites in the lattice Γ, representing the discretization of the Schrödinger equation. The energy band structure as a function of crystal momentum $\mathbf{k}$ is thus

$$\langle\mathbf{k}|H|\mathbf{k}\rangle = \epsilon + 2t(\cos k_x d_x + \cos k_y d_y + \cos k_z d_z) \tag{2.6}$$

where d_x, d_y and d_z are the discretization lengths along the $x-$, $y-$ and $z-$directions.

In the supercell method, the in-plane translational symmetry is reduced (having the period of the supercell), and we must choose a new basis. In addition, there may be no translational symmetry along the $z-$direction, as in the case of most epitaxially grown structures. Thus we choose eigenstates of definite in-plane momentum, $\mathbf{k}_{||}$:

$$|\mathbf{k}_{||}, \sigma, \alpha\rangle = \sum_{\mathbf{n}\in\Gamma_{\alpha,\sigma}} e^{i\mathbf{k}_{||}\cdot\mathbf{n}}|\mathbf{n}\rangle. \tag{2.7}$$

Here σ indexes the plane along the $z-$direction, and α indexes the supercell sites. The sum is over all sites in the lattice $\Gamma_{\alpha,\sigma}$ (see Figure 2.2) comprised of site α in each supercell in plane σ. Due to the reduced in-plane translational symmetry, $\mathbf{k}_{||}$ ranges over the reduced Brillouin zone, shown in Figure 2.3. In this basis, the Hamiltonian is block diagonal in $\mathbf{k}_{||}$. With only nearest neighbor interactions involved, only matrix elements between supercell basis states in the same plane and in adjacent planes need be considered.

We may write the electron wave function

$$\psi = \sum_{\sigma,\alpha} C_{\sigma,\alpha} |\mathbf{k}_{||}, \sigma, \alpha\rangle \tag{2.8}$$

as a linear combination of the supercell basis. In this representation, the Schrödinger equation, $(H - E)\psi = 0$, becomes

$$\mathbf{H}_{\sigma,\sigma-1}\mathbf{C}_{\sigma-1} + \bar{\mathbf{H}}_{\sigma,\sigma}\mathbf{C}_{\sigma} + \mathbf{H}_{\sigma,\sigma+1}\mathbf{C}_{\sigma+1} = 0, \tag{2.9}$$

where

$$\mathbf{C}_{\sigma} = \begin{bmatrix} C_{\sigma 1} \\ C_{\sigma 2} \\ \vdots \\ C_{\sigma M} \end{bmatrix}, \tag{2.10}$$

$$[\bar{\mathbf{H}}_{\sigma,\sigma}]_{\alpha,\alpha'} = \langle \sigma, \alpha, \mathbf{k}_{||} | (H - E) | \sigma, \alpha', \mathbf{k}_{||} \rangle, \tag{2.11}$$

$$[\mathbf{H}_{\sigma,\sigma'}]_{\alpha,\alpha'} = \langle \sigma, \alpha, \mathbf{k}_{||} | H | \sigma', \alpha', \mathbf{k}_{||} \rangle, \tag{2.12}$$

and $M = n_x n_y$ is the number of sites in a supercell. The significance of the matrices $\mathbf{H}_{\sigma,\sigma\pm1}$ and $\bar{\mathbf{H}}_{\sigma,\sigma}$ is illustrated schematically in Figure 2.4. $\bar{\mathbf{H}}_{\sigma,\sigma}$ contains information about the electron energy and the hopping matrix elements and onsite energies in plane σ, and $\mathbf{H}_{\sigma,\sigma+1}$ describes the hopping matrix elements between planes σ and $\sigma + 1$.

In order to solve for the wave function ψ, we need to specify the boundary conditions. All the devices we consider are bounded by bulk material electrodes

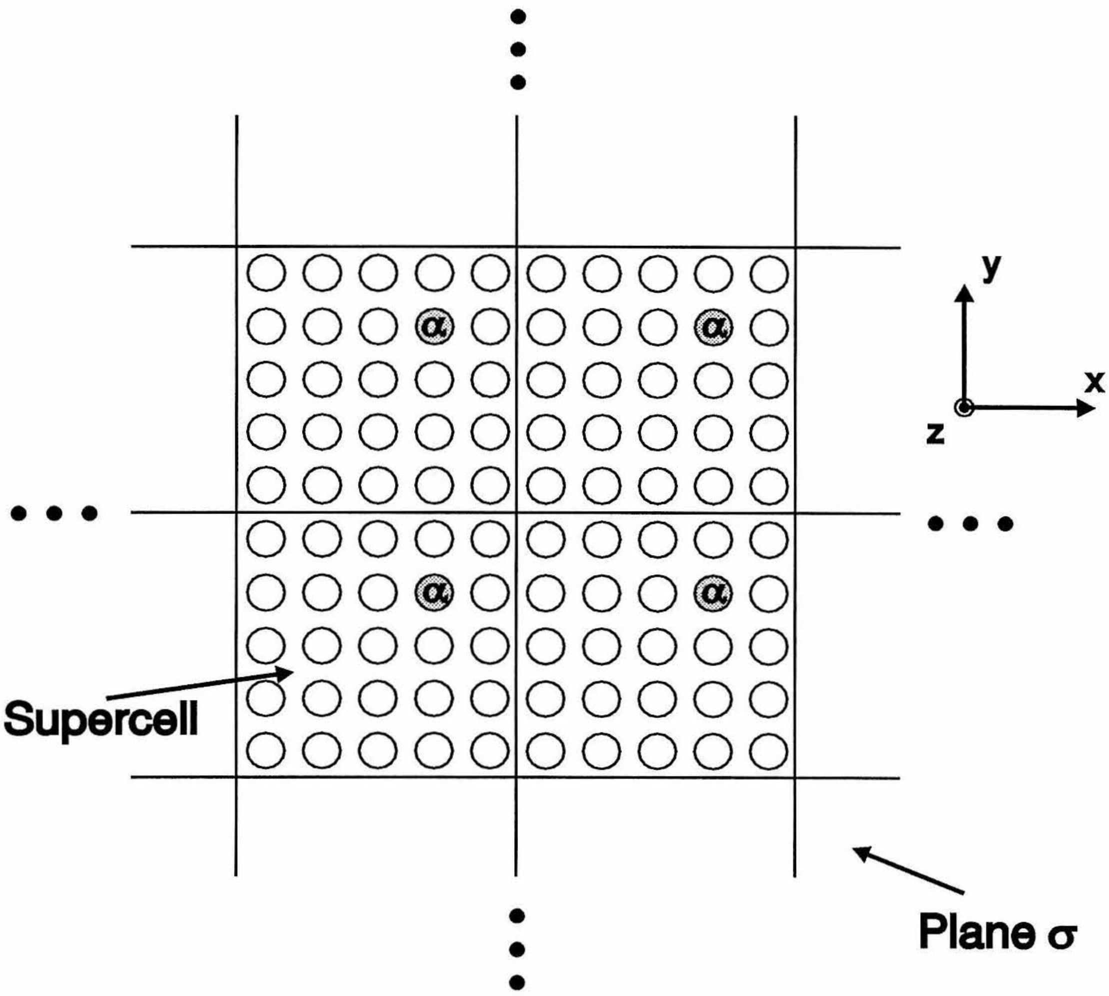

Figure 2.2: The lattice $\Gamma_{\alpha,\sigma}$ consists of site α in each supercell in plane σ. The supercells are identical and repeat in the $x-$ and $y-$directions.

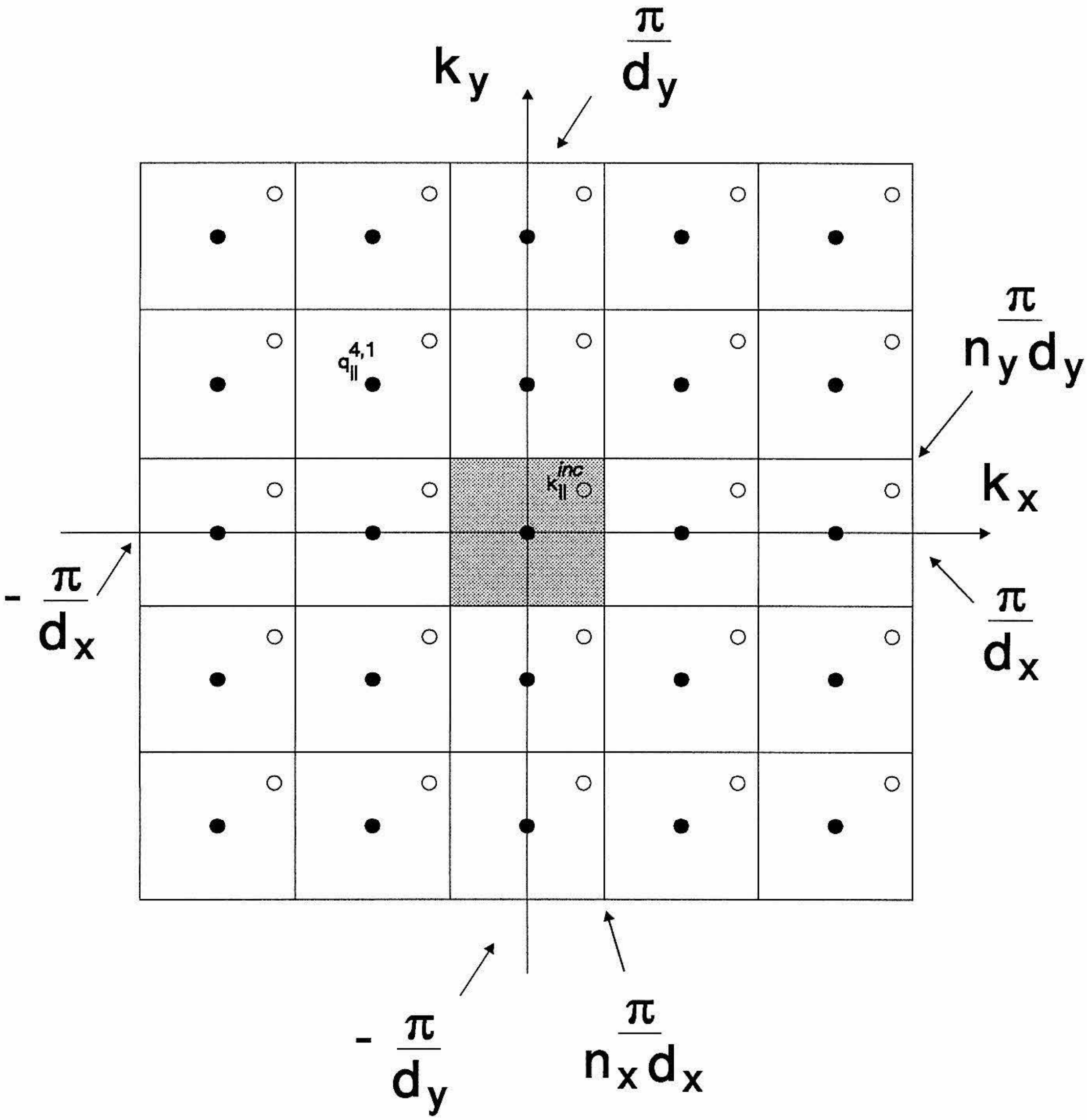

Figure 2.3: The reduced Brillouin Zone, shown shaded above, corresponding to a 5×5 supercell. The $\{\mathbf{q}_{\|}^{l,m}\}$ used in (2.15) are given by the solid circles. Electrons incident with in-plane momentum $\mathbf{k}_{\|}^{inc}$ can scatter only into states with $\mathbf{k}_{\|}$ given by the set of open circles.

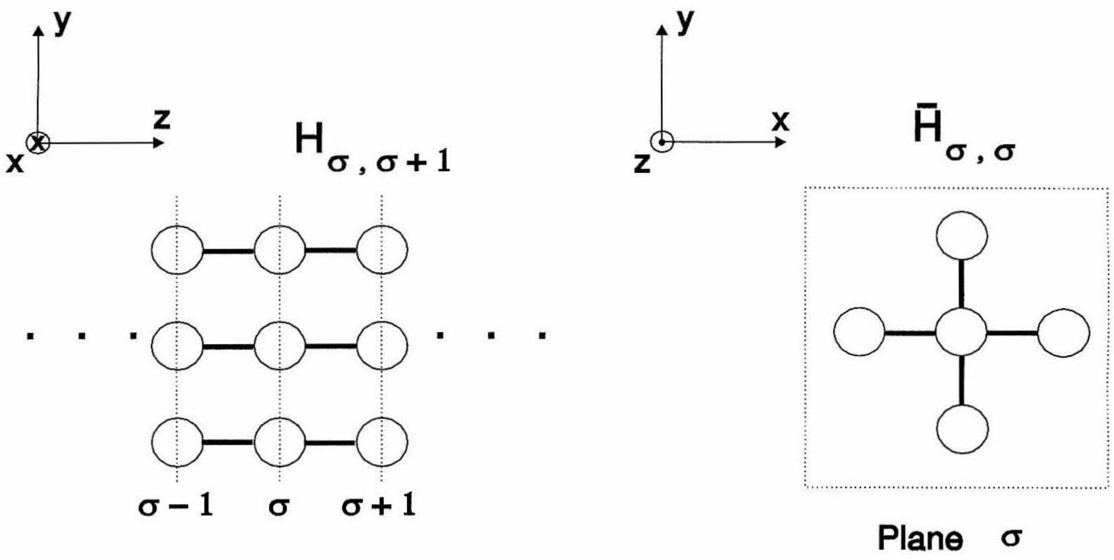

Figure 2.4: Schematic for the two types of sparse blocks appearing in the matrix in Eq. (2.19). Blocks of the form $\mathbf{H}_{\sigma,\sigma\pm1}$ describe the hopping matrix elements between adjacent planes, and blocks of the form $\bar{\mathbf{H}}_{\sigma,\sigma}$ contain information about the electron energy and the hopping matrix elements and onsite energies in plane σ.

on each end along the z−direction. The boundary conditions we specify are that in the emitter we have an incident plane wave characterized by an energy E and by in-plane momentum $\mathbf{k}_{||}^{inc}$, along with reflected plane waves, and that in the collector we have only transmitted plane waves. Thus the boundary conditions are

$$\begin{aligned} \psi_e &= |\mathbf{k}_{||}^{inc}, k_{z,e}^{inc}\rangle + \sum_{l,m} r_{l,m}|\mathbf{k}_{||}^{inc} + \mathbf{q}_{||}^{l,m}, -k_{z,e}^{l,m}\rangle, \\ \psi_c &= \sum_{l,m} t_{l,m}|\mathbf{k}_{||}^{inc} + \mathbf{q}_{||}^{l,m}, k_{z,c}^{l,m}\rangle \end{aligned} \tag{2.13}$$

where ψ_e and ψ_c are the wave function in the emitter and collector, respectively. Here $\mathbf{q}_{||}^{l,m} = (\frac{2\pi l}{N_x d_x}, \frac{2\pi m}{N_y d_y})$. Due to the reduced in-plane symmetry of the supercells, a plane wave with in-plane momentum $\mathbf{k}_{||}^{inc}$ can scatter only into a state with $\mathbf{k}_{||} = \mathbf{k}_{||}^{inc} + \mathbf{q}_{||}^{l,m}$ (see Figure 2.3). Once the electron energy and the in-plane momentum, $\mathbf{k}_{||}$, are specified, k_z is determined, depending on the local band structure. Thus $k_{z,c}(E, \mathbf{k}_{||})$ and $k_{z,e}(E, \mathbf{k}_{||})$ may be different functions, such as when a positive bias is applied to the device, lowering the collector band edge relative to that of the emitter.

We need to translate these boundary conditions into the supercell basis set $\{|\mathbf{k}_{||}, \sigma, \alpha\rangle\}$ used in expressing Schrödinger's equation. We do this by writing the plane wave basis $\{|\mathbf{k}_{||}, k_z\rangle\}$ in terms of the supercell basis $\{|\mathbf{k}_{||}, \sigma, \alpha\rangle\}$. Since $|\mathbf{k}_{||}, \sigma, \alpha\rangle$ is a state localized on the sublattice $\Gamma_{\alpha,\sigma}$, $\mathbf{C}_1$ and $\mathbf{C}_2$ for the state $\sum_{l,m} I_{l,m}|\mathbf{k}_{||}^{inc} + \mathbf{q}_{||}^{l,m}, k_{z,e}^{l,m}\rangle + \sum_{l,m} r_{l,m}|\mathbf{k}_{||}^{inc} + \mathbf{q}_{||}^{l,m}, -k_{z,e}^{l,m}\rangle$ in the emitter are given by

$$\begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{U} & \mathbf{U}\mathbf{V}^e \\ \mathbf{U}\mathbf{V}^e & \mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{r} \end{bmatrix} \tag{2.14}$$

where

$$\begin{aligned} [U]_{\alpha,\beta} &= e^{i\mathbf{n}_{||}^{\alpha}\cdot\mathbf{q}_{||}^{\beta}}, \\ [V^e]_{\alpha,\beta} &= \delta_{\alpha,\beta}e^{ik_{z,e}^{\alpha}d_z}, \end{aligned} \tag{2.15}$$

$$\mathbf{r} \;=\; \begin{bmatrix} r_{1,1} \\ r_{1,2} \\ \vdots \\ r_{n_x,n_y} \end{bmatrix},$$

and

$$\mathbf{I} = \begin{bmatrix} I_{1,1} \\ I_{1,2} \\ \vdots \\ I_{n_x,n_y} \end{bmatrix}.$$

In the above α and β index supercell sites, and $\mathbf{k}_{||}^{\alpha} = \mathbf{k}^{inc} + \mathbf{q}_{||}^{\alpha}$. Likewise, in the collector we have

$$\begin{bmatrix} \mathbf{C}_{n_z-1} \\ \mathbf{C}_{n_z} \end{bmatrix} = \begin{bmatrix} \mathbf{U} & \mathbf{U}\mathbf{V}^c \\ \mathbf{U}\mathbf{V}^c & \mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{t} \\ \mathbf{0} \end{bmatrix} \tag{2.16}$$

for the state $\sum_{l,m} t_{l,m}|\mathbf{k}_{||}^{inc} + \mathbf{q}_{||}^{l,m}, k_{z,c}^{l,m}\rangle$, where $\mathbf{t}$ is analogous to $\mathbf{r}$. From this we have

$$\begin{aligned} \mathbf{C}_1 &= \mathbf{U}\mathbf{I} + \mathbf{U}\mathbf{V}^e\mathbf{r}, \\ \mathbf{C}_2 &= \mathbf{U}\mathbf{V}^e\mathbf{I} + \mathbf{U}\mathbf{r}, \\ \mathbf{C}_{n_z-1} &= \mathbf{U}\mathbf{t}, \\ \mathbf{C}_{n_z} &= \mathbf{U}\mathbf{V}^c\mathbf{t}. \end{aligned} \tag{2.17}$$

Eliminating $\mathbf{r}$ and $\mathbf{t}$ from the above gives

$$\begin{aligned} \mathbf{C}_1 - \mathbf{U}\mathbf{V}^e\mathbf{U}^{\dagger}\mathbf{C}_2 = \mathbf{U}\mathbf{I} - \mathbf{U}(\mathbf{V}^e)^2\mathbf{I}, \\ \mathbf{C}_{n_z} - \mathbf{U}\mathbf{V}^c\mathbf{U}^{\dagger}\mathbf{C}_{n_z-1} = 0, \end{aligned} \tag{2.18}$$

where we have invoked the unitarity of $\mathbf{U}$.

These equations, together with the Schrödinger Eq. (2.9), can be formulated into the following linear system:

$$\mathbf{A}\begin{bmatrix} \mathbf{C_1} \\ \mathbf{C_2} \\ \mathbf{C_3} \\ \vdots \\ \mathbf{C_{n_z-1}} \\ \mathbf{C_{n_z}} \end{bmatrix} = \begin{bmatrix} \mathbf{UI} - \mathbf{U}(\mathbf{V}^e)^2\mathbf{I} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{2.19}$$

where $\mathbf{A}$ is the $n_x n_y n_z \times n_x n_y n_z$ matrix

$$\begin{bmatrix} \mathbf{1} & -\mathbf{U}\mathbf{V}^e\mathbf{U}^\dagger & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{H}_{2,1} & \bar{\mathbf{H}}_{2,2} & \mathbf{H}_{2,3} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{3,2} & \bar{\mathbf{H}}_{3,3} & \mathbf{H}_{3,4} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \ddots & \ddots & \vdots & \\ \mathbf{0} & & & \mathbf{0} & \mathbf{H}_{N-1,N-2} & \bar{\mathbf{H}}_{N-1,N-1} & \mathbf{H}_{N-1,N} \\ \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} & -\mathbf{U}\mathbf{V}^c\mathbf{U}^\dagger & \mathbf{1} \end{bmatrix}$$

The quantum transport problem is thus formulated into a linear system of equations. This method is just as efficient and easy to implement as the transfer matrix method[4], but it has the advantage of numerical stability. For devices with long active regions, exponentially increasing modes can "blow up" causing the transfer matrix method to fail. The above linear system, however, always has a stable solution.

2.2 Physical Quantities

2.2.1 Transmission Coefficient

Having solved for the coefficients $C_{\alpha,\sigma}$, the electronic wave function in the device is known, and we can calculate many quantities of physical interest. For example, the transmission coefficient is given by

$$T(E,\mathbf{k}_{||}^{inc}) = \sum_{l,m} |t_{l,m}(E,\mathbf{k}_{||}^{inc})|^2 \frac{|v_z^c(E,\mathbf{k}_{||}^{l,m})|}{|v_z^e(E,\mathbf{k}_{||}^{inc})|} \tag{2.20}$$

where v_z^e and v_z^c are the group velocities along the $z-$direction in the emitter and collector, respectively. They are given by

$$\begin{aligned} v_z^c(E,\mathbf{k}_{||}^{l,m}) &= \frac{\hbar}{md_z}\sin{(k_{z,c}^{l,m}d_z)}, \\ v_z^e(E,\mathbf{k}_{||}^{inc}) &= \frac{\hbar}{md_z}\sin{(k_{z,e}^{inc}d_z)}. \end{aligned} \tag{2.21}$$

$$\tag{2.22}$$

The $\{t_{l,m}\}$ are determined from (2.16):

$$\mathbf{t} = \mathbf{U}^{\dagger}\mathbf{C}_{n_z-1}. \tag{2.23}$$

By calculating the transmission coefficient at different incident energies and at different incident in-plane momenta, one can gain insight into the characteristics of a device. This will form the basis of much discussion in subsequent chapters, wherein the effects of imperfections on transmission are studied.

Examining the transmitted and reflected state amplitudes in a bulk region other than the electrodes (in a quantum well, for example) can also yield interesting information. For layers σ and $\sigma+1$ in a region where both transmitted and reflected states exist, we have

$$\begin{bmatrix} \mathbf{C}_{\sigma} \\ \mathbf{C}_{\sigma+1} \end{bmatrix} = \begin{bmatrix} \mathbf{U} & \mathbf{UV} \\ \mathbf{UV} & \mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{t} \\ \mathbf{r} \end{bmatrix} \equiv \mathbf{D} \begin{bmatrix} \mathbf{t} \\ \mathbf{r} \end{bmatrix}. \tag{2.24}$$

So

$$\begin{bmatrix} \mathbf{t} \\ \mathbf{r} \end{bmatrix} = \mathbf{D}^{-1} \begin{bmatrix} \mathbf{C}_{\sigma} \\ \mathbf{C}_{\sigma+1} \end{bmatrix}, \tag{2.25}$$

where the quantities are as defined earlier. Motivated by the form of $\mathbf{D}$ when $\mathbf{U}$ and $\mathbf{V}$ are scalars, and paying attention to the order of the matrices in products, we can calculate

$$\mathbf{D}^{-1} = \left[\begin{array}{cc} (1-\mathbf{V}^2)^{-1}\mathbf{U}^\dagger & -(1-\mathbf{V}^2)^{-1}\mathbf{V}\mathbf{U}^\dagger \\ -(1-\mathbf{V}^2)^{-1}\mathbf{V}\mathbf{U}^\dagger & (1-\mathbf{V}^2)^{-1}\mathbf{U}^\dagger \end{array} \right] \tag{2.26}$$

where

$$[(1-\mathbf{V}^2)^{-1}]_{\alpha,\beta} = \delta_{\alpha,\beta}\frac{1}{1-e^{2ik_z^\alpha d_z}}. \tag{2.27}$$

From this the transmitted components $\{t_{l,m}\}$ and the reflected components $\{r_{l,m}\}$ can be computed.

2.2.2 Electronic Wave Function

The electronic wave function can shed much light on quantum transport phenomena. The probability density at site (σ, α) is just $|C_{\sigma,\alpha}|^2$, and the phase of the wave function is $Arg[C_{\sigma,\alpha}]$. To help provide a stronger intuitive grasp of the physics, plots of the electronic wave function probability density are included in parts of this thesis.

2.2.3 Probability Current Density

Another useful construct is the probability current density, which, for a particle of mass m with continuous wave function ψ, is given by

$$\mathbf{J} = \frac{\hbar}{2im}(\psi^*\nabla\psi - \psi\nabla\psi^*). \tag{2.28}$$

In our model, the mass varies in space according to the material configuration, so we need to reformulate $\mathbf{J}$ slightly. In addition, our model is based on a discrete tight-binding Hamiltonian, so we need to use some care. We derive below an expression for $\mathbf{J}$ which represents the flow of probability density from site to site. The derivation is motivated by the traditional one in which a quantity $\mathbf{J}$ is sought, such that

$$\nabla \cdot \mathbf{J} = \frac{\partial|\psi|^2}{\partial t}. \tag{2.29}$$

A solution to the Schrödinger equation can be written

$$\psi = \sum_{\mathbf{n}} a_{\mathbf{n}}(t)|\mathbf{n}\rangle. \tag{2.30}$$

The probability density at site $\mathbf{n}$ is $|a_{\mathbf{n}}|^2$. We seek a quantity $\mathbf{J}$ on the lattice such that

$$\frac{\partial}{\partial t}|a_{\mathbf{n}}(t)|^2 = \nabla \cdot \mathbf{J}_{\mathbf{n}}, \tag{2.31}$$

where $\nabla\cdot$ is taken as

$$\nabla \cdot \mathbf{J}_{\mathbf{n}} \equiv \frac{\mathbf{J}^x_{\mathbf{n}+d_x\hat{\mathbf{x}}} - \mathbf{J}^x_{\mathbf{n}}}{d_x} + \frac{\mathbf{J}^y_{\mathbf{n}+d_y\hat{\mathbf{y}}} - \mathbf{J}^y_{\mathbf{n}}}{d_y} + \frac{\mathbf{J}^z_{\mathbf{n}+d_z\hat{\mathbf{z}}} - \mathbf{J}^z_{\mathbf{n}}}{d_z}. \tag{2.32}$$

$\mathbf{J}$ is then a vector field which represents the flow of probability density from site to site. Now,

$$\begin{aligned}
\langle \mathbf{n}|H|\psi\rangle &= i\hbar\frac{\partial}{\partial t}a_{\mathbf{n}}(t) \\
&= \sum_{\mathbf{m}} t_{\mathbf{n},\mathbf{m}}\langle \mathbf{m}|\psi\rangle + \epsilon_{\mathbf{n}}\langle \mathbf{n}|\psi\rangle \\
&= \sum_{\mathbf{m}} t_{\mathbf{n},\mathbf{m}} a_{\mathbf{m}}(t) + \epsilon_{\mathbf{n}} a_{\mathbf{n}}(t) = i\hbar\frac{\partial}{\partial t}a_{\mathbf{n}}(t).
\end{aligned} \tag{2.33}$$

The sums in the above are over the nearest neighbors $\mathbf{m}$ of site $\mathbf{n}$. Multiplying the above equation by $a^*_{\mathbf{n}}(t)$ and then subtracting the complex conjugate, we have

$$\frac{\partial}{\partial t}|a_{\mathbf{n}}|^2 = \frac{-i}{\hbar}\sum_{\mathbf{m}} t_{\mathbf{n},\mathbf{m}}(a^*_{\mathbf{n}}a_{\mathbf{m}} - a_{\mathbf{n}}a^*_{\mathbf{m}}) \equiv \nabla \cdot \mathbf{J}_{\mathbf{n}}. \tag{2.34}$$

We make the Ansatz (motivated by the derivation of $\mathbf{J}$ in the continuum case with a constant effective mass) that

$$\begin{aligned}
\mathbf{J}_{\mathbf{n}} &= \frac{-i}{\hbar}\hat{\mathbf{x}}\; t_{\mathbf{n},\mathbf{n}-d_x\hat{\mathbf{x}}} d_x[a^*_{\mathbf{n}}(a_{\mathbf{n}} - a_{\mathbf{n}-d_x\hat{\mathbf{x}}}) - a_{\mathbf{n}}(a^*_{\mathbf{n}} - a^*_{\mathbf{n}-d_x\hat{\mathbf{x}}})] \\
&+ \frac{-i}{\hbar}\hat{\mathbf{y}}\; t_{\mathbf{n},\mathbf{n}-d_y\hat{\mathbf{y}}} d_y[a^*_{\mathbf{n}}(a_{\mathbf{n}} - a_{\mathbf{n}-d_y\hat{\mathbf{y}}}) - a_{\mathbf{n}}(a^*_{\mathbf{n}} - a^*_{\mathbf{n}-d_y\hat{\mathbf{y}}})] \\
&+ \frac{-i}{\hbar}\hat{\mathbf{z}}\; t_{\mathbf{n},\mathbf{n}-d_z\hat{\mathbf{z}}} d_z[a^*_{\mathbf{n}}(a_{\mathbf{n}} - a_{\mathbf{n}-d_z\hat{\mathbf{z}}}) - a_{\mathbf{n}}(a^*_{\mathbf{n}} - a^*_{\mathbf{n}-d_z\hat{\mathbf{z}}})] \\
&= \hat{\mathbf{x}}\;\frac{2t_{\mathbf{n},\mathbf{n}-d_x\hat{\mathbf{x}}} d_x}{\hbar}(a^I_{\mathbf{n}}a^R_{\mathbf{n}-d_x\hat{\mathbf{x}}} - a^R_{\mathbf{n}}a^I_{\mathbf{n}-d_x\hat{\mathbf{x}}}) \\
&+ \hat{\mathbf{y}}\;\frac{2t_{\mathbf{n},\mathbf{n}-d_y\hat{\mathbf{y}}} d_y}{\hbar}(a^I_{\mathbf{n}}a^R_{\mathbf{n}-d_y\hat{\mathbf{y}}} - a^R_{\mathbf{n}}a^I_{\mathbf{n}-d_y\hat{\mathbf{y}}}) \\
&+ \hat{\mathbf{z}}\;\frac{2t_{\mathbf{n},\mathbf{n}-d_z\hat{\mathbf{z}}} d_z}{\hbar}(a^I_{\mathbf{n}}a^R_{\mathbf{n}-d_z\hat{\mathbf{z}}} - a^R_{\mathbf{n}}a^I_{\mathbf{n}-d_z\hat{\mathbf{z}}}),
\end{aligned} \tag{2.35}$$

where $a^R_{\mathbf{n}} = \Re\{a_{\mathbf{n}}\}$, and $a^I_{\mathbf{n}} = \Im\{a_{\mathbf{n}}\}$.

2.2.4 Current-Voltage Characteristics

Once the transmission coefficient for a device has been calculated, the current-voltage characteristics can be determined. To calculate current density J at a specified bias V, the transmission coefficient is integrated over the in-plane momentum and the Fermi distributions of electrons in the electrodes, including the appropriate velocity factors:

$$\begin{aligned} J &= \frac{e}{4\pi^3}\{\int dk_{z,e}d^2k_{\|}T_{\rightarrow}(E,\mathbf{k}_{\|})f(E)[1-f(E+eV)]\frac{1}{\hbar}\left(\frac{\partial E}{\partial k_{z,e}}\right)_{z=0} \\ &- \int dk_{z,c}d^2k_{\|}T_{\leftarrow}(E,\mathbf{k}_{\|})f(E+eV)[1-f(E)]\frac{1}{\hbar}\left(\frac{\partial E}{\partial k_{z,c}}\right)_{z=n_zd_z}\}, \quad (2.36) \end{aligned}$$

where $f(E)$ and $f(E+eV)$ are the Fermi distributions in the emitter and in the collector, and $(\frac{\partial E}{\partial k_z})$ is the group velocity along the $z-$direction. $T_{\rightarrow}$ and $T_{\leftarrow}$ refer to the transmission coefficients for electrons traversing the device from emitter to collector and from collector to emitter respectively.

The above integral may be simplified substantially in special cases. We shall describe two. The first case is that of a device at $0K$ for which the transmission coefficient may be approximated as independent of the direction of $\mathbf{k}_{\|}$. In this case, we may integrate over the direction of $\mathbf{k}_{\|}$ analytically. In addition, at $0K$, the second integral vanishes in forward bias, since there are no empty states available in the emitter to be filled by those tunneling from the collector. The expression for current then reduces to

$$J = \frac{e}{2\pi^2\hbar}\int k_{\|}dk_{\|}\int dE_zT_{\rightarrow}(E,\mathbf{k}_{\|})f(E)[1-f(E+eV)], \quad (2.37)$$

where E_z is the energy corresponding to k_z in the emitter. This integral requires considerably less computational effort than the general form (2.36). We shall attempt to provide some justification for this approximation when we invoke it in Section 4.2.5. The second case involves approximating $T(E,\mathbf{k}_{\|})$ as independent of $\mathbf{k}_{\|}$, and $T_{\leftarrow}(E) \approx T_{\rightarrow}(E)$. In this situation,

$$J = \frac{e}{4\pi^3\hbar}\int dE_z 2\pi k_{\|}dk_{\|}T(E)[f(E)-f(E+eV)]$$

$$= \frac{emk_BT}{2\pi^2\hbar^3}\int_{E_0}^{\infty} dE_zT(E)ln(\frac{1+e^{-(E_z-\mu)/k_BT}}{1+e^{-(E_z-eV-\mu)/k_BT}}), \tag{2.38}$$

where m is the effective mass in the electrodes, E_0 is the conduction band edge in the emitter, and μ is the Fermi level in the emitter. The approximations behind this formula serve well for devices with mild deviations from full translational symmetry in the $x-y$ plane, such as for a double barrier with interface roughness. This integral requires even less computation than Eq. (2.37) and could therefore be used in simulations where calculating transmission coefficients is particularly expensive.

2.3 Extensions

The above development has assumed a one-band, nearest neighbor, rectangular lattice tight-binding Hamiltonian. It is, however, straightforward to extend the model to include elements such as a multiband band structure, next-nearest neighbor interactions, and new lattice topologies such as those of the face-centered cubic crystal or the diamond lattice. Instead of a basis of one orbital localized around each site, a set of orbitals such as those used in the multiband analysis of Ting et al. [5] can be associated with each lattice site. The lattice topology, together with symmetry considerations and the number of neighbors included in coupling, then determines the sparsity pattern of the linear system representing the Schrödinger equation and boundary conditions.

Expressing the boundary conditions would require solving an eigenvalue problem [5, 6]. We illustrate this in the case of a nearest neighbor rectangular lattice multiband model for which the supercell basis consists of several orbitals localized at each site in the lattice. In the bulk electrodes, we require the supercell basis coefficients to obey $\mathbf{C}_{\sigma+1} = e^{ik_zd_z}\mathbf{C}_\sigma$. This, together with the Schrödinger equation

$$\mathbf{H}_{\sigma,\sigma-1}\mathbf{C}_{\sigma-1} + \bar{\mathbf{H}}_{\sigma,\sigma}\mathbf{C}_\sigma + \mathbf{H}_{\sigma,\sigma+1}\mathbf{C}_{\sigma+1} = 0 \tag{2.39}$$

is equivalent to

$$\begin{bmatrix} -\mathbf{H}_{\sigma,\sigma-1}^{-1}\bar{\mathbf{H}}_{\sigma,\sigma} & -\mathbf{H}_{\sigma,\sigma-1}^{-1}\mathbf{H}_{\sigma,\sigma+1} \\ \mathbf{1} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{\sigma} \\ \mathbf{C}_{\sigma+1} \end{bmatrix} = e^{-ik_z d_z} \begin{bmatrix} \mathbf{C}_{\sigma} \\ \mathbf{C}_{\sigma+1} \end{bmatrix}. \qquad (2.40)$$

In the above, $\mathbf{H}_{\sigma,\sigma-1}$ describes the Hamiltonian matrix elements between orbitals in planes σ and $\sigma-1$, and $\bar{\mathbf{H}}_{\sigma,\sigma}$ describes the overlap between orbitals in plane σ, as earlier, only in this case the matrices would be much larger, owing to the multiband nature. The eigenvectors of the above problem express the boundary conditions in terms of the supercell basis coefficients, allowing the quantum transport problem to be formulated into a sparse linear system of equations.

Although straightforward in principle, these extensions present a formidable numerical challenge. As such, they are out of reach on all but the largest supercomputers presently available. In addition, the one band nearest neighbor model used in this thesis serves admirably in an enormous variety of fascinating problems. Exploration of extensions and their consequences is therefore left to progeny.

Bibliography

[1] W. R. Frensley, Rev. Mod. Phys. **62**, 745 (1990).

[2] R. K. Mains, I. Mehdi and G. I. Haddad, Appl. Phys. Lett. **55**, 2631 (1989).

[3] R. A. Morrow and K. R. Brownstein, Phys. Rev. B **30**, 678 (1984).

[4] E. O. Kane, in *Tunneling Phenomena in Solids*, edited by E. Burstein and S. Lundqvist (Plenum, New York, 1969), p. 1; J. N. Schulman and Y. C. Chang, Phys. Rev. B **27**, 2346 (1983).

[5] D. Z.-Y. Ting, E. T. Yu, and T. C. McGill, Phys. Rev. B **45**, 3583 (1992).

[6] J. N. Schulman and Y. C. Chang, Phys. Rev. B **27**, 2346 (1983).

Chapter 3

Numerics

3.1 Overview

As formulated in the previous section, quantum transport calculations in the supercell model depend on the solution of a large, sparse linear system of equations. To get an idea of the size of the problem, we note that each of the blocks in the matrix in (2.19) are $M \times M$ matrices, where $M = n_x n_y$ is the number of orbitals in a supercell. There are as many block rows as there are layers, n_z, along the $z-$direction in the structure. Thus, for example, using a 20×20 supercell to represent a structure 100 layers thick would result in a $40,000 \times 40,000$ complex linear system. Such a system presents a formidable numerical challenge, both in terms of storage requirements and execution time.

Fortunately, the system is very sparse. The densest portions are the $M \times M$ blocks of the form $\mathbf{UVU}^\dagger$, which have no non-zero elements (c.f. Eq. (2.15)). The blocks of the form $\mathbf{H}_{\sigma,\sigma\pm1}$ are diagonal, as they represent the nearest neighbor overlap between two different planes. Blocks of the form $\bar{\mathbf{H}}_{\sigma,\sigma}$ have precisely nine non-zero complex diagonals, as explained in detail in Section 3.3.6. There is thus hope of solving such a system on present-day computers.

Having acquired a feel for the problem, we are now ready to pursue an efficient numerical solution. The rest of this section is organized as follows: First, we attack the problem of execution time, discussing different approaches to solving linear systems, giving an overview of various direct and iterative methods. Next, we attack the storage problem, describing several sparse matrix storage modes and their advantages and disadvantages. We then present a benchmark comparison of some of the methods. The quasi-minimum residual iterative method using the compressed diagonal storage mode, which we have used for all the simulations in this thesis, is benchmarked for various problem sizes. We conclude with some observations about concurrent computing and sketch how quantum transport in the supercell model could be calculated on a parallel machine.

3.2 Solving Linear Systems

3.2.1 Introduction

The problem of solving a linear system of equations specified by a coefficient matrix $\mathbf{A}$ and a right hand side $\mathbf{b}$ is to find a vector $\mathbf{x}$ such that $\mathbf{Ax} = \mathbf{b}$. Numerically speaking, there are two broad classes of methods for solving the problem. The first is termed direct solvers, and the algorithms in this class terminate in a predetermined, fixed number of steps, depending on the size of the problem. The second consists of iterative solvers, which begin with a trial solution and iterate until the solution $\mathbf{x}$ is found to within an acceptable error.

Each type of method has advantages and disadvantages[1]. Direct solvers terminate in a predictable number of steps, but they often require substantially more memory than iterative solvers for a given problem. Two of the more popular direct methods, Gauss-Jordan elimination and LU decomposition, are described below. Iterative solvers usually require less memory than direct solvers and are more robust against loss of significance, especially for large systems. The tradeoff is that they may require many more iterations than expected and thus take longer in arriving at a solution than direct methods. Nonetheless, they are particularly well suited to large sparse linear systems on account of their storage efficiency.

3.2.2 Direct Methods

Gauss-Jordan Elimination

Gauss-Jordan elimination constructs $\mathbf{A}^{-1}$. This makes the method useful when the solution for several right hand sides using the same $\mathbf{A}$ is sought. The method constructs $\mathbf{A}^{-1}$ starting from the identity matrix, $\mathbf{1}$, and performing the same operations on $\mathbf{1}$ as on $\mathbf{A}$ to transform $\mathbf{A}$ into the identity matrix. $\mathbf{A}$ is transformed into the identity matrix one column at a time as follows: The first row of $\mathbf{A}$ is multiplied by $1/A_{11}$, and then the appropriate multiple of the first row is subtracted from the remaining rows so as to eliminate their first entries. The procedure then moves to the next row of $\mathbf{A}$, multiplies that by the current value of $1/A_{22}$, and so on.

When preparing to work on row n of $\mathbf{A}$, the number A_{nn} is known as the pivot. Before working on row n, a suitable pivot should be chosen by interchanging row n with a row below. This procedure, known as partial pivoting, is essential to the numerical stability of the procedure. The optimal choice of a pivot is not completely known theoretically, but the largest available element below and to the right of the last pivot is usually a good choice. Sometimes the equations are normalized so that the largest element in each row of $\mathbf{A}$ is 1 prior to determining $\mathbf{A}^{-1}$. This produces what is known as implicit pivoting. The same row interchange involved in pivoting on $\mathbf{A}$ must be performed on $\mathbf{1}$. Full pivoting involves interchanging columns as well as rows, and in this case the permutation of the rows of $\mathbf{A}^{-1}$ must be recorded and undone in the end. Once $\mathbf{A}^{-1}$ is found, the solution for any right hand side $\mathbf{b}$ can be computed as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, although it is better from the standpoint of numerical stability to work on the right hand sides along with $\mathbf{1}$ in reducing $\mathbf{A}$ to $\mathbf{1}$[1].

When applied to our problem, Gauss-Jordan elimination bears little fruit in that we are usually only interested in solving a particular system for one right hand side. In addition, Gauss-Jordan elimination has no execution speed advantage over iterative methods for our problem.

LU Decomposition

Another direct method, one of a group of matrix factorization approaches, is known as LU decomposition[1]. The idea is to factor the matrix $\mathbf{A}$ into a lower triangular matrix $\mathbf{L}$ and an upper triangular matrix $\mathbf{U}$. The system then becomes $\mathbf{Ax} = \mathbf{L}(\mathbf{Ux}) = \mathbf{b}$, which can be solved in two steps: $\mathbf{Ly} = \mathbf{b}$, which is readily solved by forward substitution:

$$y_1 = \frac{b_1}{L_{11}}; \quad y_i = \frac{1}{L_{ii}}[b_i - \sum_{j=1}^{i-1} L_{ij} y_j], \tag{3.1}$$

and $\mathbf{Ux} = \mathbf{y}$, which is solved by backward substitution:

$$x_N = \frac{y_N}{U_{NN}}; \quad x_i = \frac{1}{U_{ii}}[y_i - \sum_{j=i+1}^{N} U_{ij} x_j]. \tag{3.2}$$

$\mathbf{A}$ is factored by a method known as Crout's algorithm. Again, pivoting is crucial to the stability of this technique[1].

This method is handy for solving a given system $\mathbf{A}$ for many right hand sides since once the factorization has been made, only the forward and backward substitutions need be performed for each different right hand side. A major drawback in our situation, however, is the large amount of "fill-in" generated. "Fill-in" is the extra storage required for $\mathbf{L}$ and $\mathbf{U}$; depending on the sparsity pattern of $\mathbf{A}$, $\mathbf{L}$ and $\mathbf{U}$ may wind up dense. We were able to apply a sparse matrix version of this technique[2] to problems involving structures with a supercell size up to about 7×7 and up to about $30 - 35$ layers thick. Beyond this size, however, fill-in becomes unmanageable—in the case of a 100 layer structure with 20×20 supercells, a dense matrix would require $(20 \times 20 \times 100)^2 \times 8$ bytes $= 12.8$ Gbytes of storage. Nonetheless, the method proved reasonably fast for small problems. A comparison of storage requirements and execution times for typical problems solved by various direct and iterative methods is given in Table 3.2 in Section 3.4.

3.2.3 Iterative Methods

Overview

In contrast with direct methods, iterative methods begin with a trial solution and iterate, each time refining the trial solution until the desired degree of accuracy is reached. The aim of most iterative methods is to minimize $|\mathbf{Ax}-\mathbf{b}|$ over $\mathbf{R}^N$, where N is the dimension of $\mathbf{A}$. This is usually accomplished by extending the dimension of the subspace over which $|\mathbf{Ax}-\mathbf{b}|$ is minimized[3] with each iteration. Thus, in the absence of roundoff error, the solution is guaranteed within N iterations. With any luck, an acceptably accurate solution can usually be found within well fewer than N iterations. We have applied two iterative methods to the solution of our model, the conjugate gradient method, and the quasi-minimum residual method. Both performed far better in terms of storage than the direct methods discussed earlier without any sacrifice in execution time. The state-of-the-art quasi-minimum residual method provided the fastest solutions of any algorithm by more than a factor of two. This is the method used in all the simulations presented in this thesis.

To impart some of the flavor of iterative methods, we describe the conjugate gradient method in detail. The related bi-conjugate gradient and generalized minimum residual methods are mentioned briefly, and an indication of the differences between the conjugate gradient and the quasi-minimum residual method is given, though details are left to the references.

Conjugate Gradient

Our description of the conjugate gradient method follows that of E. F. Van de Velde[3]. The construction presented finds the solution for $\mathbf{Ax}=\mathbf{b}$ provided that

$\mathbf{A}$ is symmetric and positive definite. However, since $\mathbf{Ax} = \mathbf{b} \Leftrightarrow \mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$, one can always convert a linear system with an $\mathbf{A}$ which is not symmetric and positive definite into one which is by replacing $\mathbf{A}$ with $\mathbf{A}^T\mathbf{A}$ and $\mathbf{b}$ with $\mathbf{A}^T\mathbf{b}$. This results in extra matrix-vector products in the final algorithm.

The method works by minimizing $F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{b}$, since the minimum $\mathbf{x}^*$ of F occurs when $\nabla F = 0 = \mathbf{Ax}^* - \mathbf{b}$. The starting point of the conjugate gradient method is an initial guess, $\mathbf{x}_0$. The remainder $\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0$ is computed, and then refined trial solutions $\mathbf{x}_i$ are computed. At iteration i, $\mathbf{x}_i$ is chosen so as to minimize F over the space $\mathbf{x}_0 + K_{i-1}(\mathbf{r}_0, \mathbf{A})$. ($K_i(\mathbf{r}_0, \mathbf{A})$ is the degree$-i$ Krylov Space of $\mathbf{A}$ at $\mathbf{r}_0$: $K_i(\mathbf{r}_0, \mathbf{A}) = span(\mathbf{r}_0, \mathbf{Ar}_0, \ldots, \mathbf{A}^i\mathbf{r}_0)$.) It will be convenient in what follows to have an $\mathbf{A}-$orthogonal basis $S_i \equiv \{\mathbf{p}_0, \ldots, \mathbf{p}_{i-1}\}$ of $K_{i-1}(\mathbf{r}_0, \mathbf{A})$ with $\mathbf{p}_0 = \mathbf{r}_0$. (By an $\mathbf{A}-$orthogonal basis, we mean a basis such that $\mathbf{p}_i^T\mathbf{Ap}_j = 0$ unless $i = j$.) It can be shown that the basis defined by the three-term recursion relation

$$\begin{aligned} \mathbf{p}_{-1} &= 0; \quad \mathbf{p}_0 = \mathbf{r}_0 \\ \mathbf{p}_{i+1} &= \lambda_i(\mathbf{Ap}_i - \mu_i\mathbf{p}_i - \nu_i\mathbf{p}_{i-1}) \end{aligned} \tag{3.3}$$

with

$$\begin{aligned} \mu_i &= (\mathbf{Ap}_i)^T(\mathbf{Ap}_i)/\mathbf{p}_i^T\mathbf{Ap}_i \\ \nu_i &= (\mathbf{Ap}_i)^T(\mathbf{Ap}_{i-1})/\mathbf{p}_{i-1}^T\mathbf{Ap}_{i-1} \end{aligned} \tag{3.4}$$

and λ_i a scaling factor to normalize $\mathbf{p}_{i+1}$, is such a basis.

Now, if $\mathbf{x}_i$ is the minimum of F over $\mathbf{x}_0 + S_i$, then

$$\begin{aligned} \mathbf{x}_i &= \mathbf{x}_0 + \sum_{j=0}^{i-1} \xi_j\mathbf{p}_j, \\ \mathbf{r}_i &= \mathbf{r}_0 - \sum_{j=0}^{i-1} \xi_j\mathbf{Ap}_j. \end{aligned} \tag{3.5}$$

A lemma[3] shows how the ξ_j are determined: $\mathbf{x}_i$ is the minimum of F over $\mathbf{x}_0+S_i \Leftrightarrow$

$$\forall j \in 0, \ldots, i-1 \quad \mathbf{p}_j^T \mathbf{r}_i = 0. \tag{3.6}$$

The proof is as follows: $\mathbf{x}_i$ is the minimum of F over $\mathbf{x}_0 + S_i \Leftrightarrow$

$$\begin{aligned} F(\mathbf{x}_i) &< F(\mathbf{x}_i + \epsilon \mathbf{v}) \\ &= F(\mathbf{x}_i) + \epsilon \mathbf{v}^T(\mathbf{A}\mathbf{x}_i - \mathbf{b}) + \frac{1}{2}\epsilon^2 \mathbf{v}^T \mathbf{A} \mathbf{v} \quad \forall \epsilon > 0, \forall \mathbf{v} \in S_i. \end{aligned} \tag{3.7}$$

Since $\mathbf{A}$ is positive definite, $\frac{1}{2}\epsilon^2 \mathbf{v}^T \mathbf{A} \mathbf{v} > 0$, so

$$\mathbf{v}^T(\mathbf{A}\mathbf{x}_i - \mathbf{b}) = -\mathbf{v}^T \mathbf{r}_i = 0 \quad \forall \mathbf{v} \in S_i. \tag{3.8}$$

(If $\mathbf{v}^T \mathbf{r}_i \neq 0$ for some $\mathbf{v} \in S_i$, we could always choose an $\epsilon > 0$ and give $\mathbf{v}$ the proper sign so as to violate the inequality in (3.7)). Since this is true for all $\mathbf{v} \in S_i$, it is, *a fortiori*, true for the basis $\{\mathbf{p}_j\}$ of S_i. Thus the lemma is proved. This indicates how to choose the ξ_i:

$$\forall j \in 0, \ldots, i-1 \quad \mathbf{p}_j^T \mathbf{r}_i = 0 = \mathbf{p}_j^T \mathbf{r}_0 - \sum_{l=0}^{i-1} \mathbf{p}_j^T \mathbf{A} \mathbf{p}_l \, \xi_l. \tag{3.9}$$

But since we have constructed the $\mathbf{p}_i$ to be $\mathbf{A}-$orthogonal, we see that

$$\xi_j = \frac{\mathbf{p}_j^T \mathbf{r}_0}{\mathbf{p}_j^T \mathbf{A} \mathbf{p}_j}. \tag{3.10}$$

At each iteration a new $\mathbf{p}_i$ is calculated from Eq. (3.3), giving a new ξ_i term to add on to the solution estimate, $\mathbf{x}_i$. At each step $\mathbf{r}_i$ is usually calculated and used in an error estimate to determine when the procedure should be terminated.

Variants of the conjugate gradient method have been proposed, such as the bi-conjugate gradient method[4] and the generalized minimum residual method[5]. These have some advantages over the conjugate gradient method, though the generalized minimum residual method is plagued by slow convergence, and the bi-conjugate gradient method usually exhibits an irregular convergence pattern. The recently proposed quasi-minimum residual method[6] overcomes many of the difficulties of the bi-conjugate gradient and generalized minimum residual methods, converging smoothly and quickly. The method is similar in flavor to the conjugate gradient method presented above, though the Krylov space basis is chosen differently (via a look-ahead Lanczos algorithm), and the iterates are chosen on the basis of a quasi-minimum principle. The details are left to the references[3, 6].

The implementation of the quasi-minimum residual method we are using is adapted from the netlib directory on netlib@ornl.gov. We find that this method provides the fastest and most efficient method of solving the sparse linear system that arises from the supercell formulation of quantum transport. Despite its success, quasi-minimum residual is susceptible to breakdown when, in the course of the calculation, a divide by zero (or a very small number) is called for. Exact breakdown almost never occurs in practice, although we have often needed to restart the calculation with a new initial guess (usually the last iterate) in order to proceed closer to the solution. (In the quasi-minimum residual method, an iterate depends on a few of the previous iterates, so restarting the algorithm with the previous iterate actually sets out on a new course toward the solution.)

We have also investigated the choice of the initial guess, $\mathbf{x}_0$. We find that simply starting from $\mathbf{x}_0 = 0$ is as good as most any other scheme. We have, however, tried two other schemes for picking $\mathbf{x}_0$. Since we are usually interested in solving Eq. (2.19) at a number of closely spaced incident electron energies (to produce a transmission coefficient curve, for example), once we have the solution $\mathbf{x}_E^*$ at some particular energy E, we may choose $\mathbf{x}_0 = \mathbf{x}_E^*$ when solving at a slightly different energy. This scheme offers only a marginal improvement over simply picking $\mathbf{x}_0 = 0$. The second scheme we tried, a secant approach, takes the previous two solutions into account in order to predict $\mathbf{x}_0$ when repeatedly solving (2.19) over a range of closely spaced monotonically increasing energies $E_{i-2}, E_{i-1}, E_i, \ldots$:

$$\mathbf{x}_{0,E_i} = \mathbf{x}^*_{E_{i-1}} + \beta(\mathbf{x}^*_{E_{i-1}} - \mathbf{x}^*_{E_{i-2}}). \tag{3.11}$$

We let β vary from 0 (which is equivalent to choosing the previous solution as a starting point) to about 10%. The result was nearly independent of β, so we simply chose the previous solution as a starting point in most of the simulations in this thesis.

Preconditioning

For most iterative methods, preconditioning can make a substantial difference in the number of iterations required for convergence. Preconditioning consists of finding a matrix $\mathbf{M}$ which, in some sense, approximates $\mathbf{A}$ and can be decomposed into $\mathbf{M} = \mathbf{M}_1\mathbf{M}_2$ such that $\mathbf{M}_1$ and $\mathbf{M}_2$ are easily invertible. The system $\mathbf{Ax} = \mathbf{b}$ is then transformed into $\mathbf{A}'\mathbf{y} = \mathbf{b}'$ where $\mathbf{A}' = \mathbf{M}_1^{-1}\mathbf{A}\mathbf{M}_2^{-1}$, $\mathbf{b}' = \mathbf{M}_1^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}_0)$ and $\mathbf{y} = \mathbf{M}_2(\mathbf{x} - \mathbf{x}_0)$. The solution to $\mathbf{Ax} = \mathbf{b}$ is then $\mathbf{x}^* = \mathbf{x}_0 + \mathbf{M}_2^{-1}\mathbf{y}^*$, where $\mathbf{y}^*$ solves $\mathbf{A}'\mathbf{y} = \mathbf{b}'$. Although calculating $\mathbf{M}_1^{-1}$ and $\mathbf{M}_2^{-1}$ requires overhead, and each iteration takes longer (since $\mathbf{A}' = \mathbf{M}_1^{-1}\mathbf{A}\mathbf{M}_2^{-1}$ necessitates three matrix vector products as opposed to just one with preconditioning), the savings in the number of iterations often more than compensates, making preconditioning a powerful technique for speeding up the calculation.

The choice of an appropriate $\mathbf{M}$ is the key to the problem. If we could choose $\mathbf{M}_1$ and $\mathbf{M}_2$ such that $\mathbf{A}' = \mathbf{M}_1^{-1}\mathbf{A}\mathbf{M}_2^{-1}$ were diagonal, the solution to $\mathbf{A}'\mathbf{y} = \mathbf{b}'$ would be trivial. This suggests one way of picking $\mathbf{M}_1$ and $\mathbf{M}_2$: let $\mathbf{M}_1$ approximate $\mathbf{A}$, and let $\mathbf{M}_2 = \mathbf{I}$; this is known as "left preconditioning." We have experimented with this type of preconditioning, letting $\mathbf{M}_1$ be the diagonal part of $\mathbf{A}$. Thus $\mathbf{M}_1^{-1}$ is easy to compute and requires little storage. The result was, however, only about a 10% reduction in the number of iterations, each of which took roughly 10% longer than without preconditioning. The benefit was therefore marginal. More sophisticated preconditioning schemes have been explored[6], though we have not applied them to our problem.

3.3 Storage Modes

3.3.1 Overview

We have thus far addressed the issue of choosing the best algorithm for solving (2.19). We now search for the best implementation of this algorithm, the quasi-minimum residual method. We shall strive for two goals: minimizing storage required and maximizing execution speed. As it turns out, the two goals are, for the problem at hand, unusually compatible. As we can see from the foregoing discussion, the brunt of the numerical effort in solving (2.19) by the quasi-minimum residual method lies in calculating matrix vector products ($\mathbf{A}\mathbf{x}_i$) and transposed matrix-vector products ($\mathbf{A}^T\mathbf{x}_i$). We should therefore choose a matrix storage mode that will allow for fast matrix-vector products. In order to choose such a mode, we must first understand the basics of sparse matrices.

Sparse matrices may be characterized as having relatively few non-zero entries. The sparsity pattern, or arrangement of the non-zero elements in the matrix, may vary from well-ordered (such as when all non-zero elements are concentrated along a few diagonals) to random. In any event it is usually advantageous to adopt a scheme by which the matrix can be stored without all of its zeroes. There are several schemes in widespread use; determining which to adopt depends on the sparsity pattern of the matrix involved and upon the way in which the matrix will be used. In the remainder of this section, we examine some sparse matrix storage modes and their strengths and weaknesses. (The discussion follows that found in the manual for the IBM Engineering and Scientific Subroutine Library, Release 4, from which the examples are taken.) We then give the details of how we store the matrix in (2.19) to conserve space and to allow efficient matrix-vector products.

3.3.2 Compressed Matrix Storage Mode

The first mode we discuss is known as the compressed matrix storage mode, which essentially "compresses" the non-zero elements in each row to the left. The mode uses two matrices to store the sparse $m \times n$ matrix $\mathbf{A}$: an $m \times l$ matrix $\mathbf{AS}$, and an $m \times l$ matrix $\mathbf{K}$, where l is the maximum number of non-zero elements in a row of $\mathbf{A}$. $\mathbf{AS}$ contains the non-zero elements of $\mathbf{A}$, row by row, padding with 0's each row of $\mathbf{A}$ containing fewer than l non-zero elements. $\mathbf{K}$ contains the corresponding column indices of each non-zero element in $\mathbf{A}$. For example, to store the 6×6 matrix

$$\mathbf{A} = \begin{bmatrix} 11 & 0 & 13 & 0 & 0 & 0 \\ 21 & 22 & 0 & 24 & 0 & 0 \\ 0 & 32 & 33 & 0 & 35 & 0 \\ 0 & 0 & 43 & 44 & 0 & 46 \\ 51 & 0 & 0 & 54 & 55 & 0 \\ 61 & 62 & 0 & 0 & 65 & 66 \end{bmatrix}, \tag{3.12}$$

$$\mathbf{AS} = \begin{bmatrix} 11 & 13 & 0 & 0 \\ 22 & 21 & 24 & 0 \\ 33 & 32 & 35 & 0 \\ 44 & 43 & 46 & 0 \\ 55 & 51 & 54 & 0 \\ 66 & 61 & 62 & 65 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} 1 & 3 & * & * \\ 2 & 1 & 4 & * \\ 3 & 2 & 5 & * \\ 4 & 3 & 6 & * \\ 5 & 1 & 4 & * \\ 6 & 1 & 2 & 5 \end{bmatrix}. \tag{3.13}$$

It is easy to see that this method is most effective for sparse matrices with approximately the same number of non-zero elements in each row.

When $\mathbf{A}$ is to be used in matrix-vector products, the compressed matrix storage mode brings up the problem of random gather (with its close cousin, random scatter[7]). Random gather involves "randomly accessing" data stored in computer memory. For example, in the compressed matrix mode, the matrix-vector product $\mathbf{v} \leftarrow \mathbf{Aw}$ would be coded as

```
do i = 1, m
    v(i) = 0
    do j= 1, l
        v(i) = v(i) + AS(i,j) * w(K(i,j))
    enddo
enddo
```

If the non-zero elements of **A** occur in random positions in each row of **AS**, the above code will randomly jump around in memory, gathering the correct elements of **w** to form the product. (Usually memory access is fastest when successively accessed elements are distributed in memory with unit stride, i.e., at consecutive addresses. Thus, in the above example, even if the successive row elements of **A** are not completely randomly distributed, the algorithm will be bogged down if they are not arranged contiguously in memory.) For most present computer architectures, random gather is a very inefficient process. Sparse matrices with an irregular sparsity pattern therefore present a special challenge in terms of storage and processing.

3.3.3 Storage by Indices

Another mode, well suited to matrices with a random sparsity pattern, is storage by indices. In this mode, **A** is stored in three data structures: **AS**, **IA**, and **JA**, all vectors of length l, where l is the number of non-zero elements of **A**. **AS** contains the non-zero elements of **A**, in any order, and **IA** and **JA** contain the row and column indices, respectively, of the corresponding elements of **AS**. For example, the matrix

$$\mathbf{A} = \begin{bmatrix} 11 & 0 & 13 & 0 & 0 & 0 \\ 21 & 22 & 0 & 24 & 0 & 0 \\ 0 & 32 & 33 & 0 & 35 & 0 \\ 0 & 0 & 43 & 44 & 0 & 46 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 61 & 62 & 0 & 0 & 65 & 66 \end{bmatrix} \qquad (3.14)$$

might be stored as

$$\begin{aligned} \mathbf{AS} &= (11, 22, 32, 33, 13, 21, 43, 24, 66, 46, 35, 62, 61, 65, 44), \\ \mathbf{IA} &= (1, 2, 3, 3, 1, 2, 4, 2, 6, 4, 3, 6, 6, 6, 4), \\ \mathbf{JA} &= (1, 2, 2, 3, 3, 1, 3, 4, 6, 6, 5, 2, 1, 5, 4). \end{aligned}$$

Storage by indices is intuitive and simple to implement, yet it is not the most storage efficient for matrices that are not particularly sparse. For example, if **A** is an integer matrix with more than 1/3 of its entries non-zero, this mode actually consumes more memory than storing all of **A**!

3.3.4 Storage by Columns or Rows

Another mode is storage by columns (or, analogously, by rows). Storage by columns stores the $m \times n$ matrix **A** in three data structures: two arrays **AS** and **IA** of length l, and an array **JA** of length $n+1$, where l is the number of non-zero elements of **A**. **AS** contains the non-zero elements of **A**, column by column, left to right. **IA** contains the row indices of the corresponding elements in **AS**, and **JA** lists the positions in **AS** at which each new column of **A** begins. (The last element of $\mathbf{JA}$ is $l+1$.) For example, the matrix

$$\mathbf{A} = \begin{bmatrix} 11 & 0 & 13 & 0 & 0 & 0 \\ 21 & 22 & 0 & 24 & 0 & 0 \\ 0 & 32 & 33 & 0 & 0 & 0 \\ 0 & 0 & 43 & 44 & 0 & 46 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 61 & 62 & 0 & 0 & 0 & 66 \end{bmatrix} \tag{3.15}$$

would be stored as

$$\begin{aligned} \mathbf{AS} &= (11, 61, 21, 62, 32, 22, 13, 33, 43, 44, 24, 46, 66), \\ \mathbf{IA} &= (1, 6, 2, 6, 3, 2, 1, 3, 4, 4, 2, 4, 6), \\ \mathbf{JA} &= (1, 4, 7, 10, 12, 12, 14). \end{aligned}$$

3.3.5 Compressed Diagonal Mode

The final mode which we describe in detail is the compressed diagonal storage mode. This is the mode we use for storing most of the matrix in (2.19), as it is designed for square sparse matrices whose elements are concentrated along a few diagonals. As we shall see, this mode also lends itself well to fast matrix-vector products. The compressed diagonal storage mode stores the $m \times m$ matrix $\mathbf{A}$ in two data structures: an $m \times l$ matrix $\mathbf{AS}$ and a vector $\mathbf{LA}$ of length l, where l is the number of non-zero diagonals in $\mathbf{A}$. The elements of $\mathbf{LA}$ give the positions of the non-zero diagonals relative to the major diagonal, and the columns of $\mathbf{AS}$ give the diagonals, padded with leading zeroes for diagonals below the major diagonal and with trailing zeroes for diagonals above the major diagonal. For example, the

6 × 6 matrix

$$\mathbf{A} = \begin{bmatrix} 11 & 0 & 13 & 0 & 0 & 0 \\ 21 & 22 & 0 & 24 & 0 & 0 \\ 0 & 32 & 33 & 0 & 35 & 0 \\ 0 & 0 & 43 & 44 & 0 & 46 \\ 51 & 0 & 0 & 54 & 0 & 0 \\ 61 & 62 & 0 & 0 & 65 & 66 \end{bmatrix} \tag{3.16}$$

would be stored as

$$\mathbf{AS} = \begin{bmatrix} 11 & 13 & 0 & 0 & 0 \\ 22 & 24 & 21 & 0 & 0 \\ 33 & 35 & 32 & 0 & 0 \\ 44 & 46 & 43 & 0 & 0 \\ 55 & 0 & 54 & 51 & 0 \\ 66 & 0 & 65 & 62 & 61 \end{bmatrix}, \tag{3.17}$$

and

$$\mathbf{LA} = (0, 2, -1, -4, -5). \tag{3.18}$$

The compressed diagonal method is storage efficient for sparse matrices whose elements are concentrated along a few diagonals. In addition, it is well suited to matrix-vector products. One can verify that to calculate the matrix-vector product $\mathbf{v} \leftarrow \mathbf{Aw}$, one simply accumulates the dot products of the columns of $\mathbf{AS}$ with $\mathbf{w}$ (properly aligned according to $\mathbf{LA}$) into $\mathbf{v}$. Dot products run very quickly on most architectures, as they require accessing contiguous pieces of memory, one after another. The dot product accumulations run particularly fast on the IBM RS/6000's we used for the simulations in this thesis on account of their superscalar capability, allowing a multiply and an add instruction to be performed in one cycle. The compressed diagonal method is thus extremely well suited to our application, both from the point of view of storage and of speed.

3.3.6 Supercell Application

Having explored some sparse matrix storage modes, we now present our method for storing the matrix in (2.19). Since we have found the quasi-minimum residual iterative method best for solving our system, our goal is to store the matrix in (2.19) so as to conserve space and make matrix-vector products as fast as possible.

We begin by analyzing the matrix. As we show below, aside from the two blocks of the form $\mathbf{UVU}^\dagger$, the matrix is sparse with all non-zero elements concentrated along 11 diagonals. This part of the matrix is therefore stored using the compressed diagonal mode. The remaining two blocks are composed of the dense matrix $\mathbf{U}$ and the diagonal matrix $\mathbf{V}$ (see Section 2.1). Since dense matrix-matrix products are costly ($O(M^3)$, where $M = n_x n_y$), we perform the matrix-vector products involving these blocks, $\mathbf{UVU}^\dagger\mathbf{C}$, one matrix-vector product at a time: $\mathbf{U}(\mathbf{V}(\mathbf{U}^\dagger\mathbf{C}))$. We thus store $\mathbf{U}$ in an ordinary two-dimensional array and $\mathbf{V}$ in one-dimensional array.

The rest of the matrix (aside from the identity matrix blocks) describes the onsite energies and nearest neighbor interactions in the tight-binding Hamiltonian. It is therefore not surprising that, if we order the supercell basis correctly, the matrix will have its non-zero elements concentrated along a few diagonals. We can arrange this by ordering the basis elements as in Figure 3.1. (Recall that the basis consists of $n_x n_y n_z$ orbitals, one for each site in the supercell representation of the device—see Section 2.1.) Since the supercell model enforces periodic boundary conditions by connecting sites on opposite edges of the supercell, there are two categories of orbitals to consider in determining the sparsity pattern of the matrix: those on the edges of the supercell and those in the interior. The analysis of orbital connectivity in Table 3.1 shows that there are only eleven different values of $i - j$ for which orbital $|i\rangle$ is connected with orbital $|j\rangle$ to produce a non-zero element in the matrix. Since one of them is the major diagonal $i - j = 0$ (on which the 1's in the unit matrix blocks lie), the part of the matrix excluding the blocks of the form $\mathbf{UVU}^\dagger$ has all non-zero elements concentrated on 11 diagonals. We thus use

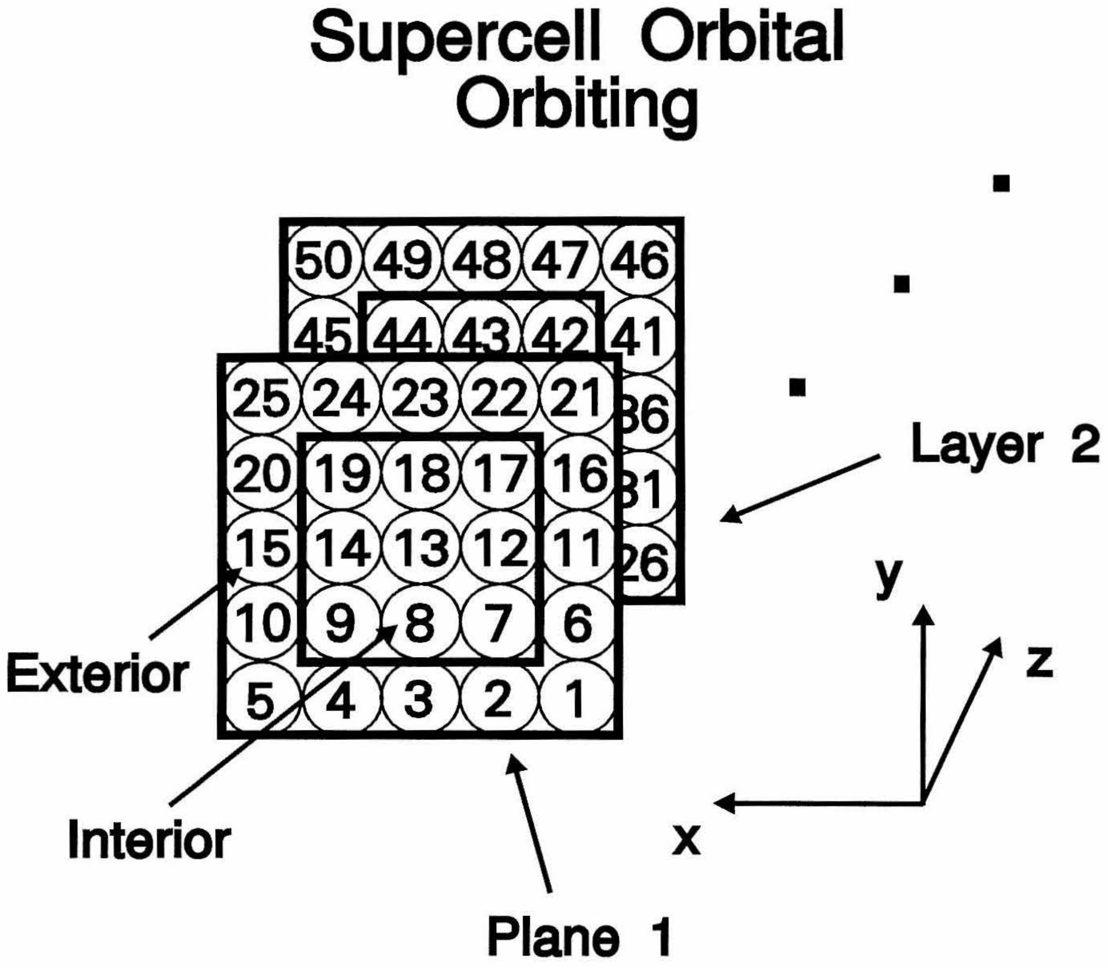

Figure 3.1: Ordering of supercell basis elements for a 5 × 5 supercell so that the matrix in (2.19) takes on a simple form with all non-zero elements concentrated along 11 diagonals. Each basis element corresponds to a row in the matrix, and each non-zero matrix element (outside the blocks representing the boundary conditions) arises from either a hopping matrix element between nearest neighbors or an onsite energy. Due to the periodic boundary conditions in the supercell method, it is important to recognize two different regions—the interior and the exterior—when determining the sparsity pattern of the matrix. See Table 3.1 for a list of the 11 non-zero diagonals, characterized by constant $i - j$.

Supercell Region (Location of i)	Type of Bond	$i-j$
Interior	$\rightarrow$	+1
	$\leftarrow$	-1
	$\downarrow$	$+M$
	$\uparrow$	-M
	$\otimes$	-M^2
	$\odot$	$+M^2$
	onsite	0
Exterior	$\rightarrow$	-$(M-1)$
	$\leftarrow$	$+(M-1)$
	$\downarrow$	-(M^2-M)
	$\uparrow$	$+(M^2-M)$

Table 3.1: The above table can be used to understand the sparsity structure of the matrix in (2.19). In the nearest neighbor rectangular lattice supercell model, each orbital is connected to its six nearest neighbor orbitals. Any two such orbitals, labeled by i and j, will have a value of $i-j$ found in the rightmost column above. A supercell of size $M \times M$ is assumed. The "Type of Bond" column indicates which orientation of nearest neighbor orbitals gives rise to each value of $i-j$ (assume i is at the base of the arrow and j is at the tip). The $x-$direction is to the left, the $y-$direction is up, and the $z-$direction is into the paper. (See Figure 3.1 for the labeling of orbitals in the supercell representation.)

the compressed diagonal storage mode for this part.

3.4 Benchmarks

We have thus far discussed several methods of dealing with the issue of storage and execution time in solving large sparse linear systems of equations. Of the methods discussed, we have implemented four to solve (2.19): LU decomposition, the conjugate gradient method, the quasi-minimum residual method with preconditioning and the quasi-minimum residual method without preconditioning. In Table 3.2 we summarize the memory requirements and execution times for the various methods. The test case for the benchmarks was a simulation of a double barrier structure with a 15 layer thick well, 5 layer thick barriers and 5 layer thick electrodes. A 5×5 supercell was used, and 100 transmission coefficients were calculated over the energy range from the electrode edge to the barrier edge.

As expected on the basis of the above discussion, the LU decomposition method required the most memory, as the LU factorization produced enough fill-ins to render the matrix dense for our problem. In addition, the method took somewhat longer to find solutions than the iterative methods. The conjugate gradient method, using storage by rows, afforded a great reduction in memory requirements and a modest reduction in execution time. The quasi-minimum residual method reduced the number of iterations compared to the conjugate gradient method, yielding a great improvement in execution time, and the compressed diagonal storage mode offered more storage efficiency. The quasi-minimum residual method with left preconditioning by the diagonal of $\mathbf{A}$ in (2.19) reduced the execution time slightly by reducing the number of iterations required for convergence. For most problems, however, we found that preconditioning gave only marginal performance improvement. The quasi-minimum residual method without preconditioning has been used for the simulations in this thesis.

The test case for comparing the various numerical methods considered in solving

Method	Storage Required (Mb)	Execution Time (mm:ss)
LU Decomposition Storage by Indices	12	7:31
Conjugate Gradient Storage by Rows	0.3	6:40
QMR w/ Preconditioning Compressed Diagonal	0.2	1:55
QMR wo/ Preconditioning Compressed Diagonal	0.2	2:08

Table 3.2: Comparison of the storage requirements and execution times for various methods of solving the sparse linear system arising in supercell calculations of quantum transport. The test case involved simulating a double barrier resonant tunneling structure with a 15 monolayer $GaAs$ well, 5 monolayer $AlAs$ barriers and 5 monolayer $GaAs$ electrodes represented by a 5×5 supercell. A series of 100 transmission coefficients for the device was calculated at evenly spaced points in the energy range from the $GaAs$ electrode edge to the $AlAs$ barrier edge. (Material parameters are the same as those used in Chapter 5, and $d_x = d_y = d_z = 0.2825nm.$) Note that the LU decomposition method proved highly inefficient in terms of storage due to the generation of a large number of fill-ins. The quasiminimum residual (QMR) method using the compressed diagonal matrix storage mode proved most efficient both in terms of storage and execution speed.

Supercell Size	Storage Required (Mb)	Execution Time (mm:ss)
5×5	0.2	2:08
10×10	0.9	11:08
15×15	2.9	32:15
20×20	7.5	86:06

Table 3.3: Storage requirements and execution times using the quasi-minimum residual method with the compressed diagonal storage mode. The test case involved simulating the same double barrier resonant tunneling structure as in the previous table. Various supercell sizes were used to represent the structure. A series of 100 transmission coefficients for the device was calculated at evenly spaced points in the energy range from the $GaAs$ electrode edge to the $AlAs$ barrier edge. Note how the storage requirements and execution times scale with the supercell size.

(2.19) was purposefully small since the LU decomposition method simply could not accommodate much larger problems on account of storage limitations. The quasi-minimum residual iterative method, however, can accommodate much larger problems, and in Table 3.3 we give an idea of how the memory requirements and execution times using this method scale with problem size. The same device as in Table 3.2 is simulated using different supercell sizes. For small problems, the dense blocks of (2.19) are negligible, and the non-zero diagonals contribute most to the memory requirements, so the memory scales as $O(n_x n_y n_z)$. For larger problems, the dense blocks contribute most, so the memory scales as $O(n_x^2 n_y^2)$. Likewise, in terms of execution time, for small problems, the compressed diagonal matrix-vector products dominate, so the execution time scales roughly as $O(n_x n_y n_z)$. For larger problems, the dense matrix-vector products dominate, so execution time scales more like $O(n_x^2 n_y^2)$. Deviations results from the fact that, as the problem size increases, the number of iterations to solve (2.19) also increases. (Krylov-space methods find solutions within N iterations, where N is the dimension of $\mathbf{A}$, without the effects of roundoff error—see Section 3.2.3.)

3.5 Concurrent Considerations

3.5.1 Overview

Thus far all our numerical pursuits have involved sequential algorithms running on single processor machines. As the cost of producing relatively high-powered workstations has fallen sharply, and the cost of manufacturing state-of-the-art supercomputers has remained fairly high, interest in parallel computing as an answer to the ever-increasing demands for higher performance has grown. Unlike typical single processor machines, parallel computers consist of a collection of processors or nodes linked for communication with one another in one of a number of topologies. Thus far parallel computers have demonstrated the promise of achieving cost-effective computation, albeit at a sacrifice in algorithmic and coding simplicity. In this section we address some of the issues pertinent to concurrent computation and how it could be applied to our problem.

3.5.2 Amdahl's Law

The suitability of a parallel environment to a particular algorithm depends on the degree to which the algorithm is parallelizeable and upon how much of the algorithm is inherently sequential. Amdahl's Law[7] gives an indication of the potential benefit of parallelizing an algorithm with this consideration in mind. Suppose a certain algorithm requires a fraction $1 - s$ of the execution time for sequential calculation. (The parameter s, in the range $[0, 1]$, is meant to reflect the parallelizeability of the algorithm.) Suppose, furthermore, that the algorithm is to be run on a parallel machine with N processors. Amdahl's Law states that the ratio of the execution time on a single processor to that on N processors is

$$\frac{T_{sequential}}{T_{concurrent}} = \frac{N}{N - (N-1)s}. \tag{3.19}$$

Thus, even as $N \rightarrow \infty$, the maximum speedup is a factor of $\frac{1}{1-s}$; so even if an algorithm is 2/3 parallelizeable ($s = 2/3$), one can obtain no more than a factor of 3 in execution speedup from running on a collection of nodes compared to running on a single node. When s approaches 1, however, parallel computers offer a substantial speedup. Figure 3.2 shows the speedup factor for an algorithm on an $N-$node machine as a function of the parallelizeable fraction s.

3.5.3 Topologies

Parallel computers come in a variety of topologies, or ways in which the nodes are connected for inter-node communications. Among the most popular today are the mesh topology, in which the nodes form a cubic array with nearest neighbor connections, and the hypercube topology, in which there are 2^n processors, one at each of the vertices of an $n-$cube with connections along the edges. Other topologies include the bus architecture and the omega network[8], for example. Needless to say, topology can affect efficiency of inter-node communications, and parallel algorithm design should keep topology in mind.

3.5.4 Load Balancing

In order to make the most efficient use of a parallel computer, the computation must be divided effectively among the nodes. This introduces the concept

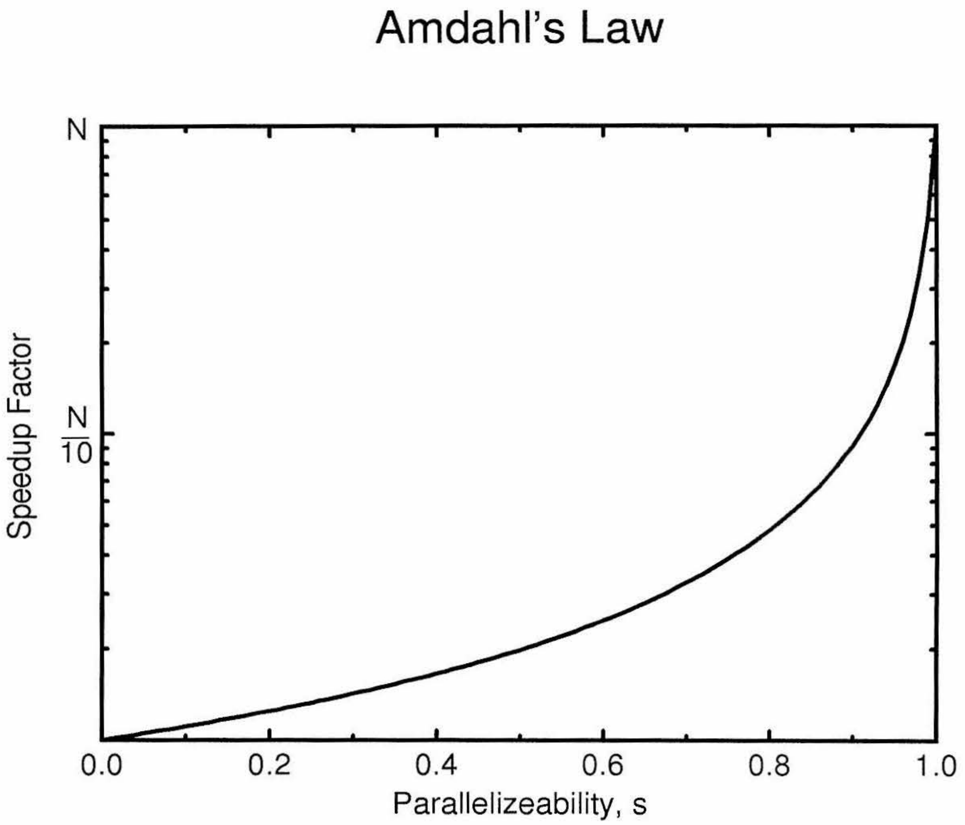

Figure 3.2: A plot of the speedup factor, $T_{sequential}/T_{concurrent}$, as a function of the fraction s of the execution time of an algorithm that can be distributed over N processors. Note the sharp rise in speedup as s approaches 1. Unless an algorithm is highly parallelizeable, much of the computational power of a concurrent computer remains untapped.

of load balancing, or dividing the computation up among the individual processors. Broadly speaking there are two approaches to load balancing: static and dynamic. Static load balancing partitions the work load only at the beginning of a computation and is well suited to problems such as finite element simulations on a fixed grid. Dynamic load balancing works throughout the execution of the algorithm to maintain balance as each node's workload changes. A calculation simulating the evolution of galaxies and clusters, where the spatial density of particles is changing, would be a good candidate for dynamic load balancing. In most cases, load balancing is crucial to the performance of a parallel algorithm and remains an area of intense research.

3.5.5 Implementations

Thus far we have discussed parallel computation from an abstract point of view. For coding and running programs on actual parallel machines, a software implementation is needed to provide basic functionality such as inter-node communication, data sharing, file access, and performance analysis. One example of such an implementation is the Express^{TM}[9] kernel and parallel toolkit, which provides extended capability to write parallel programs in FORTRAN and C. The package provides a library of routines to effect inter-node communications, processor synchronization, input/output services, and processor allocation and control. In this vein, Express^{TM} can also be used to parallelize sequential programs, easily distributing a simple loop over many processors, for example. More complicated algorithms present a greater coding challenge, and parallel programming is becoming a field of vigorous research.

3.5.6 Application to Supercell Calculations

Before concluding this section, we briefly discuss how concurrent computing could be applied to our problem. Three-dimensional supercell simulations present a formidable numerical challenge, both on the front of storage and on the front of execution time. Parallel computers offer help on both fronts. To speed the execution of simulations requiring small amounts of memory, we can simply treat a parallel machine as an array of independent processors. For example, to calculate a transmission coefficient curve of 100 different transmission coefficients, each at a different incident electron energy, we could run 100 copies of our sequential code, each on a separate processor (as long as the simulation size does not exceed the memory of a single processor). This case, known as the "embarrassingly parallel" case, would be relatively straightforward to implement on a parallel machine such as the Intel Delta, with nodes typically offering around 10 Mb of storage. This option is unavailable to larger simulations, however, as the storage would have to be spread across many processors. This case raises some thorny issues in algorithm design such as load balancing, for example. Needless to say, the solution of our problem on a parallel machine in this case would be a challenging and fascinating research problem.

Bibliography

[1] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipes* (Cambridge University Press, New York, 1988).

[2] The routine, Sparse1.3 by K. S. Kundert and A. Sangiovanni-Vincentelli, was acquired by e-mail request from netlib@ornl.gov.

[3] E. F. Van de Velde, *Concurrent Scientific Computing*, to be published.

[4] C. Lanczos, J. Res. Natl. Bur. Stand. **45**, 255 (1950).

[5] Y. Saad and M. H. Schultz, SIAM J. Sci. Stat. Comput. **7**, 856 (1986).

[6] R. W. Freund and N. M. Nachtigal, Technical Report 90.51, RIACS, NASA Ames Research Center, Dec. 1989.

[7] J. M. Levesque and J. W. Williamson, *A Guidebook to Fortran on Supercomputers* (Academic Press, Inc., San Diego, 1989).

[8] G. C. Fox, M. A. Johnson, G. A. Lyzenga, S. W. Otto, J. K. Salmon, D. W. Walker, *Solving Problems on Concurrent Processors* (Prentice Hall, Englewood Cliffs, New Jersey, 1988).

[9] *Express Fortran User's Guide, Version 3.0* Parasoft Corporation (1990).

Chapter 4

Neutral Impurities in Tunneling Structures

4.1 Introduction

4.1.1 Background

Our first application of the supercell model is to neutral impurities in tunneling structures. As discussed in Chapter 1, these nanoscale devices are strongly influenced by process imperfections and defects. In particular, neutral impurities can substantially alter the transmission properties of single and double barrier structures. We shall see that they can give rise to resonances whose position and strength are sensitive to the impurity location. In fact, an isolated impurity can produce negative differential resistance in a single barrier. A high concentration of impurities can yield a complex resonance structure that fluctuates with impurity configuration. In this chapter, we undertake a systematic study of these effects. We first present a brief background on impurities.

Impurities in bulk solids have been extensively investigated. A survey of impurities and other point defects in bulk materials can be found in a review by Pantelides[1]. The electronic structure and electronic levels of neutral impurities have been studied using a number of approaches. Cluster methods, such as the defect molecule model[2] and the atomic cluster model[3, 4], have been used to calculate energy levels of neutral impurities. The Extended Hückel theory method, developed by Walter and Birman[5], has been widely used to calculate electronic states. Self-consistent Green function methods[6] have also been employed. A recent optical study of neutral impurity levels can be found in an article by Monemar[7]. Additional topics such as elastic[8, 9] and inelastic[10] scattering from neutral impurities have also been considered.

Only recently have the effects of impurities on transmission in tunneling structures received attention. Resonant tunneling assisted by an energy level associated with a defect has been observed[11]. The authors use a single scattering center calculation and find that negative differential resistance can occur in a single barrier with isolated defects. Double barrier structures with a dilute concentration of impurities in the well have also been considered[12]. An average of the current density over impurity configurations was taken, and it was found that impurities produce a broadening in the well resonance and a reduction in its maximum.

In this thesis we examine three-dimensional quantum transport in single barrier and double barrier resonant tunneling structures with specific, three-dimensional impurity configurations. This allows us to address issues pertaining to interactions between impurities—the electron wave function is calculated taking all impurities in the device simultaneously into account. In addition we can study fluctuations in transmission properties resulting from different impurity configurations. Even for the case of a single isolated impurity, we find important differences between simulations in one, two and three dimensions.

4.1.2 Outline of Chapter

We first examine an isolated impurity in a single barrier. We study resonance shape and position as a function of material parameters and the location of the impurity within the barrier. We then consider level splitting and the effects on transmission in the case of two closely spaced impurities. Level splitting as a function of the distance between two impurities and their orientation within the lattice is examined. The level splitting is manifested in the transmission differently for different orientations of the impurity separation direction relative to the incident plane wave. We next study three-dimensional distributions of impurities in single and double barrier tunneling structures leading to a discussion of fluctuations. We present a current-voltage calculation for a single barrier with an isolated impurity, and we summarize some of the experimental evidence for tunneling via localized states in Section 4.3.

4.2 Simulation and Results

4.2.1 Isolated Impurity

We first consider an isolated impurity in a single barrier tunneling structure grown along the $z-$direction. We take the electrodes to have a band edge of $E_e = -1eV$ and an effective mass of $m_e = 0.0673m_0$, and the barrier to be $l_b = 9$ monolayers thick and to have a band edge of $E_b = 0eV$ and an effective mass of $m_b = 0.1m_0$. The impurity is placed in the middle layer of the barrier, and the discretization lengths used are $d_x = d_y = d_z = 0.2825nm \equiv a$. We represent the impurity by a single site with an onsite energy ΔU below that of the barrier. The hopping matrix element to the impurity site is the same as that in the rest of the barrier, $t \equiv -\hbar^2/2m_b a^2$. We can take the dimensionless quantity $\Delta U/t$ as a measure of impurity strength, so that $\Delta U/t < 0$ for an attractive impurity, and $\Delta U/t > 0$ for a repulsive impurity.

We plot, in Figure 4.1, the transmission versus incident electron energy for this structure using a few different values of $\Delta U/t$. For attractive impurities, if $|\Delta U|/t$ is large enough, there will be an impurity level between E_e and E_b, giving rise to

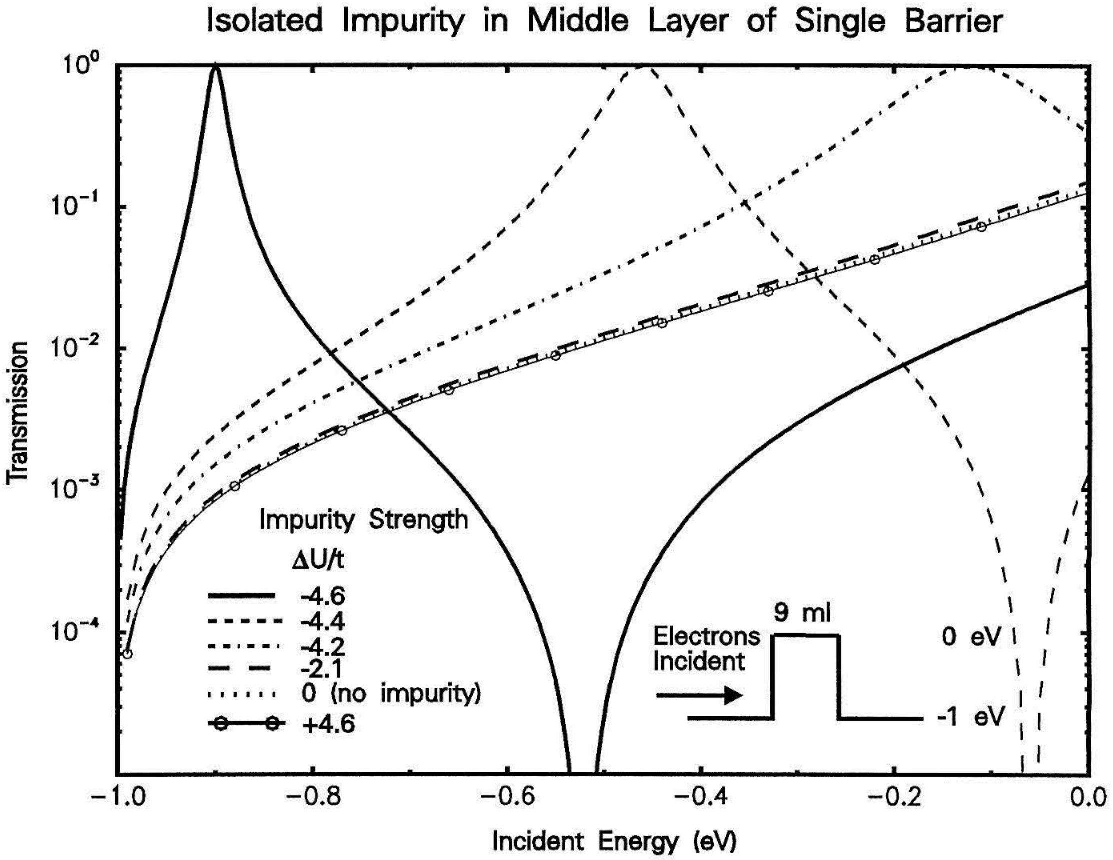

Figure 4.1: Isolated impurity in the middle layer of a single barrier tunneling structure. L_b = 9 monolayers, 13 × 13 supercell, $a = 2.825\AA$. $E_b = 0eV$, $m_b = 0.1m_0$, $E_e = -1eV$, $m_e = 0.0673m_0$. Various values of $\Delta U/t$ are used. Electrons are incident along the $z-$direction (i.e. with zero in-plane momentum).

a transmission resonance. At the resonance energy the impurity provides a locally favorable current path, as we can see in Figure 4.2, where the probability current density is plotted in the barrier plane containing the impurity. If $|\Delta U|/t$ is not large enough, there will be no impurity level in this range. Nonetheless, the impurity can still affect tunneling, as exhibited by the long-dashed curve. Repulsive impurities have less effect on the transmission, as seen from the curve marked with circles. The higher onsite energy of the repulsive impurities contributes in an averaged sense to an overall slightly higher barrier, thereby reducing transmission. In short we see that an isolated impurity (especially a strongly attractive impurity) can have a significant impact on tunneling.

In addition to a resonance, the transmission coefficient curve for $\Delta U/t = -4.6$ appears to have a zero near $E = -0.52eV$. This is due to interference caused by repetition of the supercells (and hence impurity sites) in the growth planes. Representing a well isolated impurity would require much larger supercells and a prohibitively large amount of computer memory. Nonetheless, the features we are interested in, namely the resonances, change little for supercells larger than about 13×13, so this size will suffice in most of our calculations.

We have thus far taken full advantage of our model by simulating an impurity in three dimensions. By constructing appropriate supercells, we can also simulate impurities in one and two dimensions. For example, to simulate an impurity in two dimensions, we would use a $1 \times n$ supercell; in one dimension, the supercells would consist of a single site. We use this versatility to show some important differences between tunneling calculations in one, two and three dimensions. Figure 4.3 shows the dependence of resonance width and normalized resonance energy, E/t, on the impurity strength, $\Delta U/t$, for a single barrier with an isolated attractive impurity in the middle layer. For a given impurity strength, simulations in different dimensions predict different resonance positions and widths. The resonance moves to higher energies as the dimension of the calculation is increased, due to the increasing number of directions in which the impurity bound state is confined.

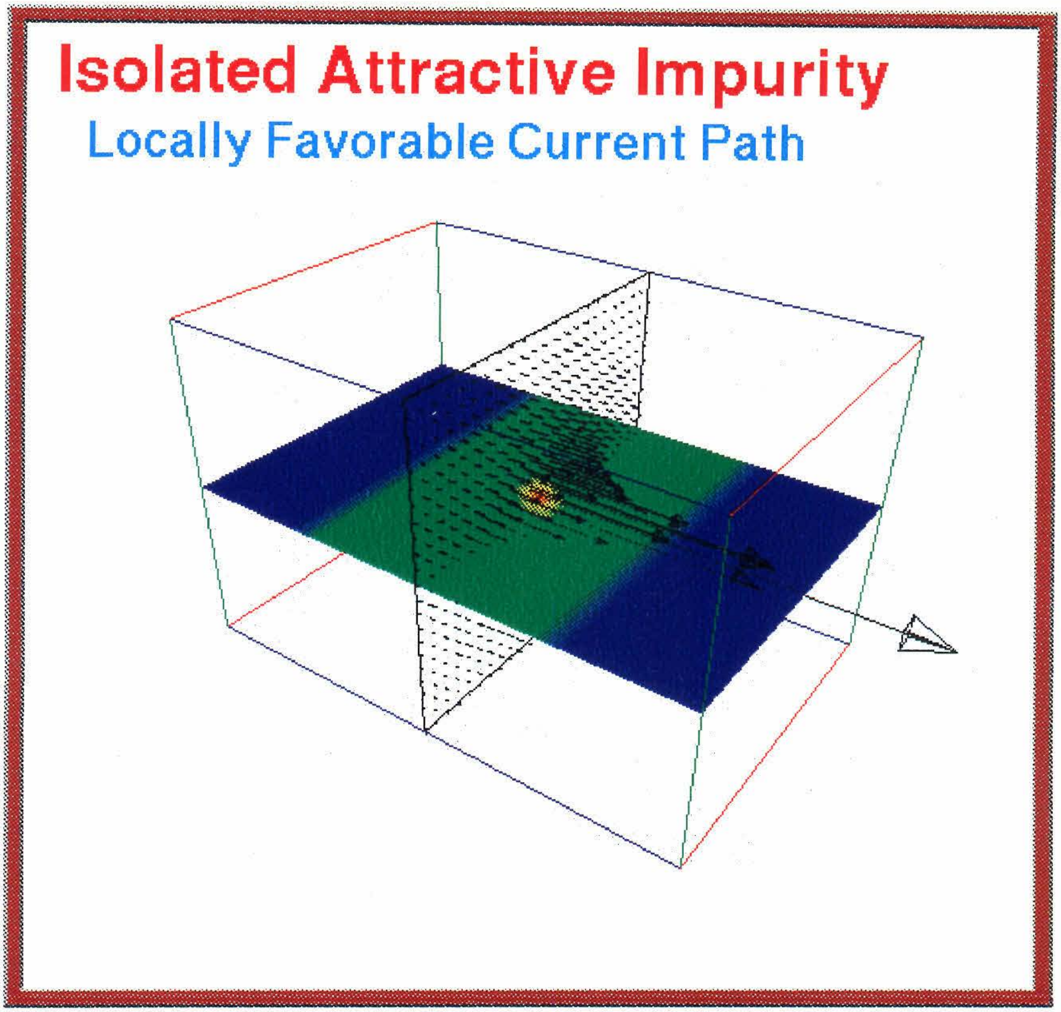

Figure 4.2: At the resonance energy, an impurity provides a locally favorable current path. Here, the probability current density is plotted in the plane containing the impurity. The device is that of Figure 4.1 with $\Delta U/t = -4.6$. The blue regions represent the electrodes, the green region represents the barrier, and the site in the middle is the impurity.

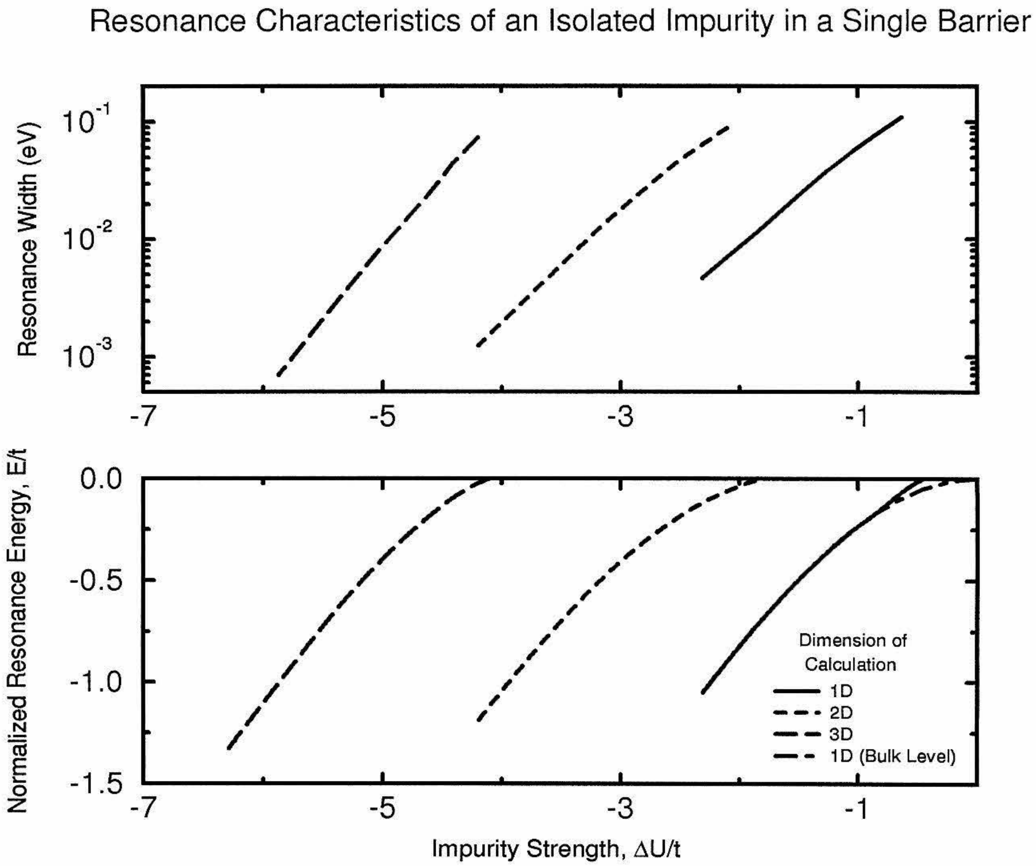

Figure 4.3: Resonance widths and normalized resonance energies, E/t, for an isolated impurity in the middle barrier layer of a single barrier as calculated in one, two and three dimensions. The dot-dashed curve is the analytical result for the resonance position of a single impurity in a bulk 1D sample. $L_b = 9$ monolayers, 1×1, 1×10 and 10×10 supercells are used in 1D, 2D and 3D respectively, $a = 2.825\AA$. $E_b = 0eV$, $m_b = 0.1m_0$, $m_e = 0.0673m_0$. Various electrode band edges $E_e \leq -1eV$ were chosen so that E_e was below the resonance level. Electrons are incident along the $z-$direction.

In one-dimensional simulations, it is confined only along the $z-$direction, whereas in three-dimensional simulations, it is confined in the lateral directions as well. When the resonance level rises, confinement along the $z-$direction grows weaker, due to the finite barrier thickness. Thus as the dimensionality increases and the resonance level rises, the resonance width increases, as shown in the top panel of Figure 4.3.

The finite thickness of the barrier can also affect the resonance position. For a strongly attractive impurity, the resonance position in a single barrier agrees with the level of the impurity in a bulk sample of barrier type material. In one dimension the bulk level is[13]

$$E/t = 2 - \sqrt{4 + \Delta U^2/t^2}, \tag{4.1}$$

and at high values of $|\Delta U|/t$, this agrees with the resonance position of an impurity in a single barrier. For weaker impurities, however, the single barrier resonances are at energy levels different from those of impurities in bulk samples (see Figure 4.3). Although a bound level always exists in bulk in one and two dimensions for $\Delta U/t < 0$[19], no such level exists for weak impurities in a single barrier of finite thickness. The finite extent along the $z-$direction does not support a bound state for very weakly attractive impurities.

These results on resonance position and width can be used to predict how neutral impurities might affect transmission in a single barrier. Whenever the impurity level lies above E_e, a transmission resonance can be expected. In this regard, we stress the important differences in the predictions of the one–, two– and three dimensional calculations. We also stress the importance of the finite barrier thickness in determining the resonance widths and positions, especially for weak impurities. Finally, as we saw earlier, even when there is no bound level between E_e and E_b, a neutral impurity can still affect transmission in this energy range.

4.2.2 Resonance Shape

In addition to impurity strength, the impurity location also impacts transmission. As an impurity is moved along the $z-$direction in a single barrier, the transmission resonance it produces changes shape and position. In Figure 4.4 we plot resonance position, resonance width and the maximum transmission coefficient as a function of impurity location in a 22 monolayer thick barrier. We find that the resonance moves to slightly higher energy as the impurity approaches the center of the barrier due to increasing confinement—the impurity site is surrounded by thicker walls. We find that the resonance width decreases as the impurity is moved toward the center of the barrier, another sign of increasing isolation from the electrodes. The maximum transmission increases to unity as the impurity approaches the middle layer of the barrier. It is clear that the maximum transmission increases faster than the resonance width decreases, so the transmission resonance grows stronger as the impurity is moved toward the middle layer of the barrier. From this we might expect that attractive impurities near the center of a barrier would play a larger role in the transmission than those near the edges. Indeed, both the resonance strength and position depend on the location of the impurity within the barrier.

4.2.3 Two Impurities

Having studied a single impurity, we now turn to the case of two attractive impurities. The interaction of two closely spaced impurities gives rise to a level splitting. The lower energy level corresponds to a state which is symmetric along the direction of separation of the impurities, and the higher energy level corresponds to an antisymmetric state. Each of these levels can result in a transmission resonance, depending upon the direction of the incident plane wave relative to the direction of separation of the two impurities. Whenever the direction of the incident plane wave has a component along the direction of separation of the two impurities, resonant tunneling can occur via both the symmetric and antisymmetric levels. When

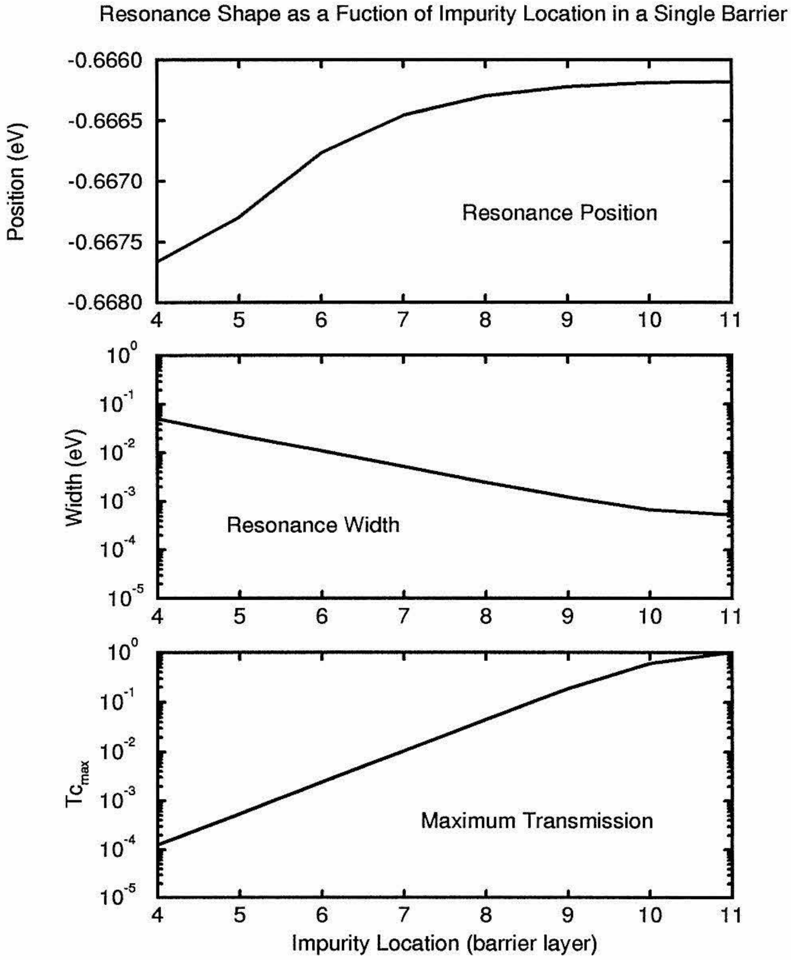

Figure 4.4: Resonance shape as a function of impurity location along the growth direction. The horizontal axis is labeled with the number of the barrier layer in which the impurity is located. The plot begins at layer 4, since the resonance shape is difficult to measure when the impurity is closer to the electrode. $L_b = 22$ monolayers, 13×13 supercell, $a = 2.825\AA$. $E_b = 0eV$, $m_b = 0.1m_0$, $E_e = -1eV$, $m_e = 0.0673m_0$, $\Delta U/t = -4.5$. Electrons are incident along the $z-$direction. Resonance width is the full width at half maximum.

the two directions are orthogonal, however, resonant tunneling can occur only via the symmetric level.

To illustrate this we examine the transmission through a single barrier with two impurities separated by a distance equal to five lattice constants. We plot the transmission coefficient versus energy for different relative orientations of the impurity separation direction and the incident plane wave direction. In Figure 4.5, the direction of the incident plane wave is fixed along the $z-$direction, and the impurity separation vector makes angle θ with this direction. The midpoint between the two impurities lies in the middle of the barrier. We note that for $\theta = 90^o$, i.e., the incident and separation directions are orthogonal, resonant tunneling occurs only via the symmetric level. As θ decreases, and the component of the incident plane wave direction along the separation direction increases, the resonance associated with the antisymmetric level increases in strength. The resonance widths of both the symmetric and antisymmetric resonances increase as θ decreases since the impurities are moved closer to the electrode-barrier interfaces. In Figure 4.6, we keep the impurities fixed at a separation of five lattice spacings along the y-direction in the middle plane of the barrier, and we vary the incident plane wave direction. As the in-plane momentum along the separation direction, q_y, is increased (holding $q_x = 0$), the antisymmetric resonance again grows stronger. The small variations in resonance position stem from the in-plane momentum energy shift as well as from finite supercell size effects. Thus the relative orientation of the incident direction and the impurity separation direction can play a significant role in the transmission properties of a tunneling structure.

Having examined orientation dependence, we next study level splitting as a function of impurity separation. We find that the level splitting decreases as the impurities are moved further apart. This is due to decreasing interaction between the impurities. Figure 4.7 shows the wave function in a single barrier with two closely spaced impurities. The system exhibits two bound levels, and the magnitude of the wave function along the line of the impurities is plotted at the symmetric

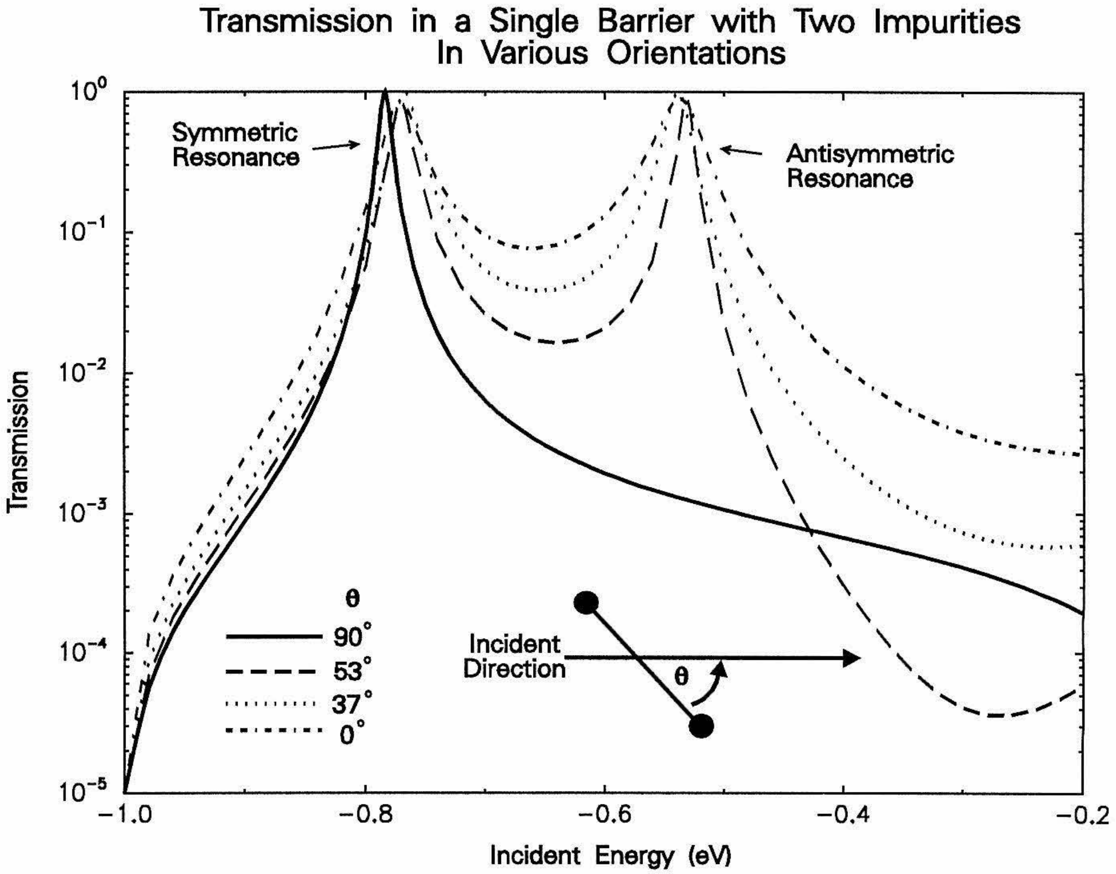

Figure 4.5: Transmission versus incident energy for a single barrier with two impurities separated by five lattice spacings. The plane wave is incident along the $z-$direction with which the impurity separation direction makes angle θ. The midpoint between the impurities lies in the middle of the barrier. $L_b = 13$ monolayers, 13×13 supercell, $a = 2.825 \mathring{A}$. $E_b = 0eV$, $m_b = 0.1m_0$, $E_e = -1eV$, $m_e = 0.0673m_0$, $\Delta U/t = -4.5$.

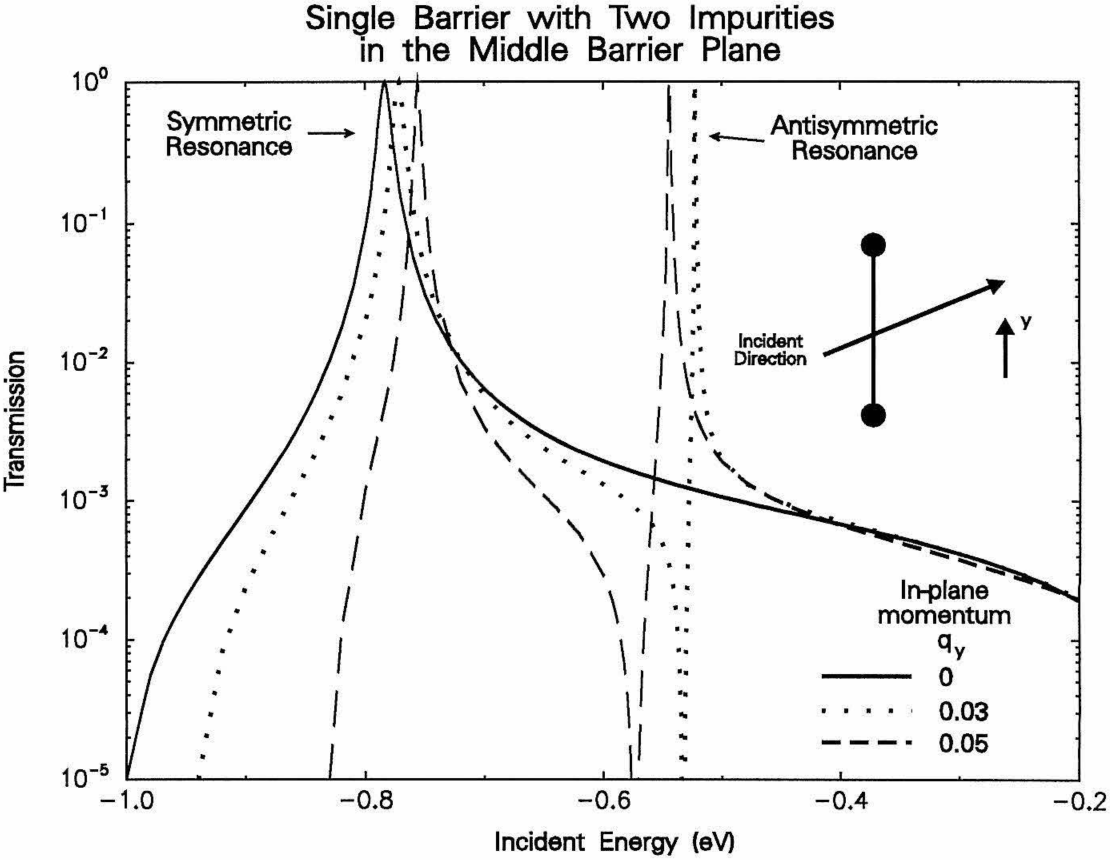

Figure 4.6: Transmission versus incident energy for a single barrier with two impurities separated by five lattice spacings along the y-direction in the middle barrier layer. The incident in-plane momentum along the separation direction, q_y, (measured in units of $\frac{\pi}{a}$) is varied ($q_x = 0$). $L_b = 13$ monolayers, 13×13 supercell, $a = 2.825\mathring{A}$. $E_b = 0eV$, $m_b = 0.1m_0$, $E_e = -1eV$, $m_e = 0.0673m_0$, $\Delta U/t = -4.5$.

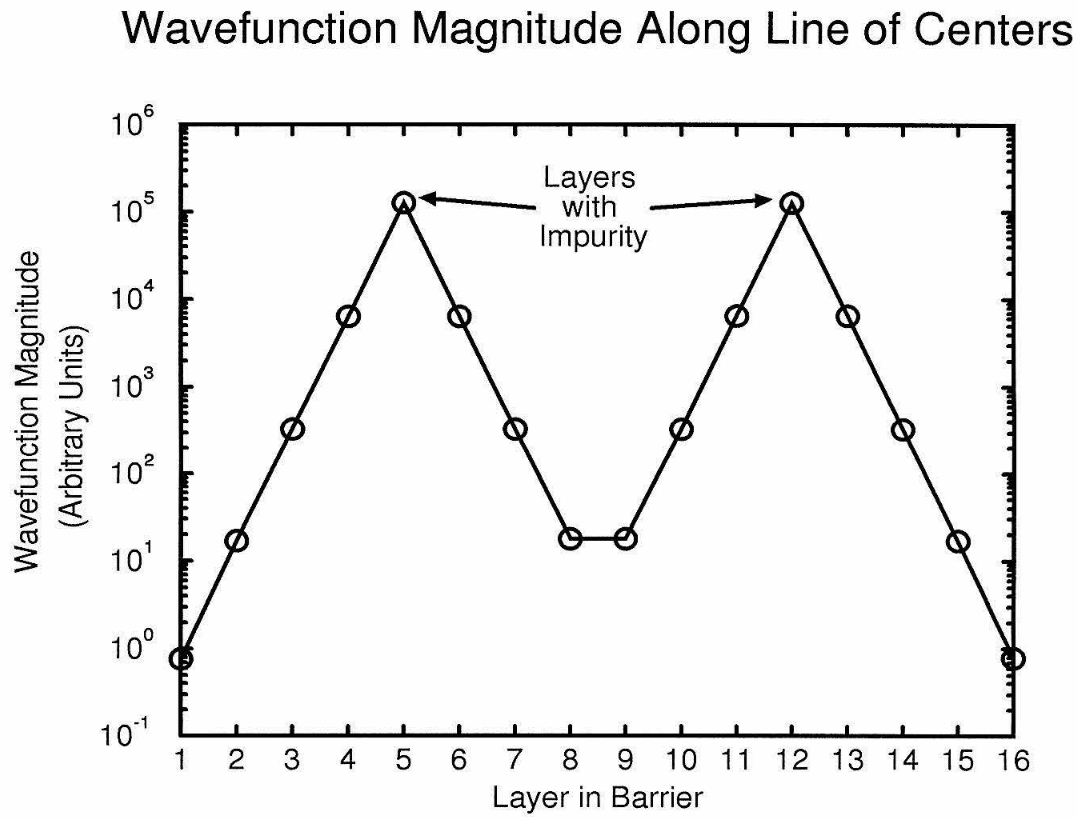

Figure 4.7: Symmetric state wave function magnitude along the line joining two impurities separated by 7 lattice spacings along the $z-$direction in a single barrier. (A line is drawn to guide the eye.) $L_b = 16$ monolayers, 10×10 supercell, $a = 2.825\mathring{A}$. $E_b = 20eV$, $m_b = 0.477m_0$, $E_e = 0eV$, $m_e = 0.0673m_0$, $\Delta U/t = -20$. Electrons are incident along the $z-$direction.

state level. This clearly demonstrates the exponential decrease in the wave function magnitude with distance from the impurities, which causes the level splitting to decrease as the impurities are moved further apart.

We now study the level splitting as a function of impurity separation. We simulate transmission through a single barrier with two closely spaced impurities in one, two and three dimensions and measure resonance level splitting as a function of impurity separation distance. In two and three dimensions, we are able to measure the splitting for different orientations of the impurity separation direction relative to the underlying square and cubic lattices. We find that, in all cases, the level splitting drops off exponentially with increasing distance between impurities along a given direction (see Figure 4.8). The splitting is not the same in all directions, however, for a given separation distance. In the bottom panel of Figure 4.8, we plot splitting versus separation for impurities separated along a cubic lattice axis (on-axis) as well as along the [011] direction. Not only is the splitting different in the two directions for a given separation distance, but it also drops off at a different exponential rate. When we calculate the splitting versus separation for impurities placed on-axis, the result is different when calculated in different dimensions unless the impurities are very strongly attractive. The top panel shows that the splitting is different in one and two dimensions for $\Delta U/t = -2.4$. In the bottom panel, $\Delta U/t = -50$, and the splitting is the same in one, two and three dimensions. The reason that the splitting for very strongly attractive impurities separated along a cubic lattice axis is the same in one, two and three dimensions is that the overlap of wave functions localized at the impurity sites is nearly the same in the three dimensions.

For the range of parameters we have chosen in Figure 4.8, we see that when impurities are separated by more than three lattice constants, the splitting should be negligible. When spaced closely, however, inter-impurity interactions can play a major role in determining resonance positions. In this case, the orientation of the impurity pair in the lattice should be taken into account. Thus there are

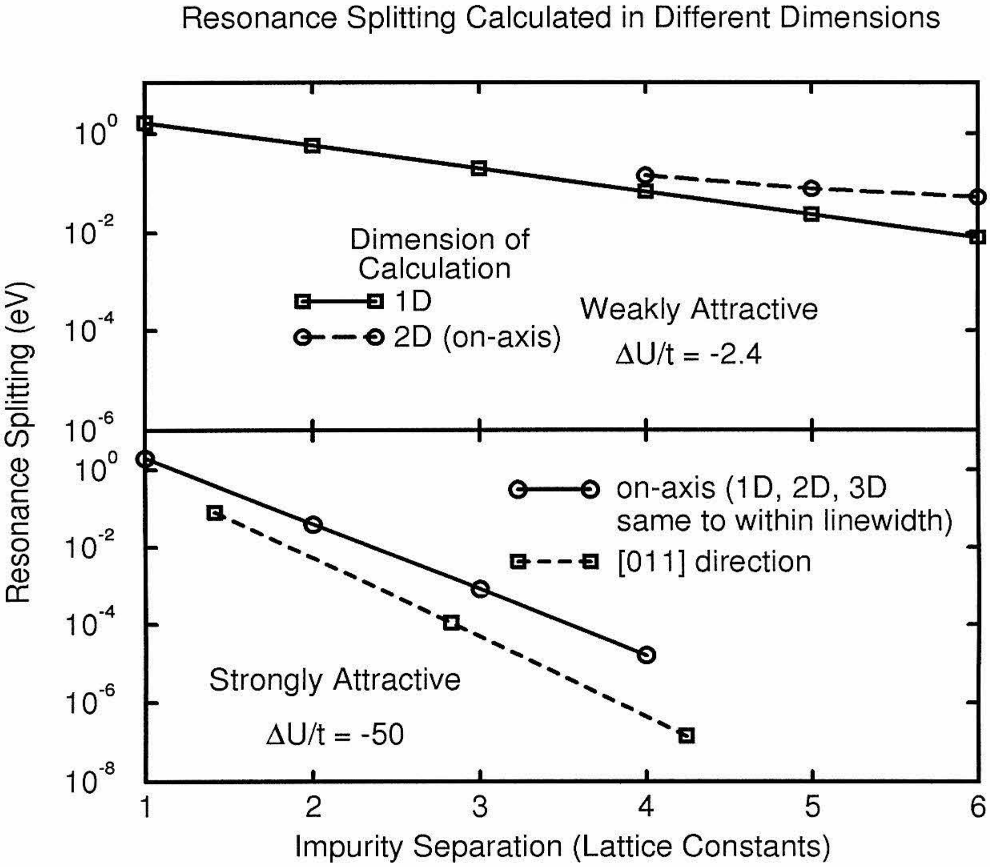

Figure 4.8: Resonance splitting as a function of impurity separation. For weakly attractive impurities separated along an axis (top panel) the splitting is different in one and two dimensions; for strongly attractive impurities (bottom panel) it is the same in one, two and three dimensions. Splitting also depends on the orientation of the impurity pair within the lattice. $L_b = 26$ monolayers, 1×1, 1×15 and 15×15 supercells used for 1D, 2D and 3D, respectively, $a = 2.825 \mathring{A}$. $m_b = 0.477 m_0$, $m_e = 0.0673 m_0$. In the top panel, $E_b = 0 eV$, $E_e = -2.4 eV$. In the bottom panel, $E_b = 50 eV$, $E_e = 0 eV$. Electrons are incident along the $z-$direction.

several factors that determine the resonance structure to which an impurity distribution gives rise. For isolated impurities, the location within the barrier plays the dominant role in determining the resonance positions and strengths. For pairs of closely spaced impurities, inter-impurity distance and orientation play the leading role. Based on these results, summarized in Figures 4.4, 4.5, and 4.8, we might expect that the shape of the transmission coefficient curve should fluctuate widely with configuration in a single barrier with a high concentration of impurities, where both impurity location and interactions are important.

4.2.4 Multiple Impurities

We examine a single barrier with a random distribution of attractive impurities. We calculate transmission for two different configurations of four impurities placed randomly among the sites of the 9 layers of 20×20 supercells representing the barrier. Figure 4.9 contains the results. Comparing with the transmission coefficients for an impurity-free single barrier, we see that the impurities give rise to several resonances of varying strengths and positions. Note also that the shape of the transmission coefficient curve is indeed very different for the two configurations.

Impurities in double barrier structures also affect transmission, as illustrated in Figure 4.10. We consider first the case of impurities in the well and then the case of impurities in the barriers. The top panel of Figure 4.10 shows the transmission coefficient curves for different concentrations of attractive impurities in the well. The lower onsite energy of these attractive impurities contributes in an averaged sense to a lower effective well band edge. As the impurity concentration is increased, this effective band edge moves down, and the $n = 1$ well resonance shifts down. In addition, the impurities in the well can give rise to new resonances, as shown by the solid curve.

The bottom panel of Figure 4.10 shows the transmission in a double barrier structure with attractive impurities in the barriers. Just as in the case with impu-

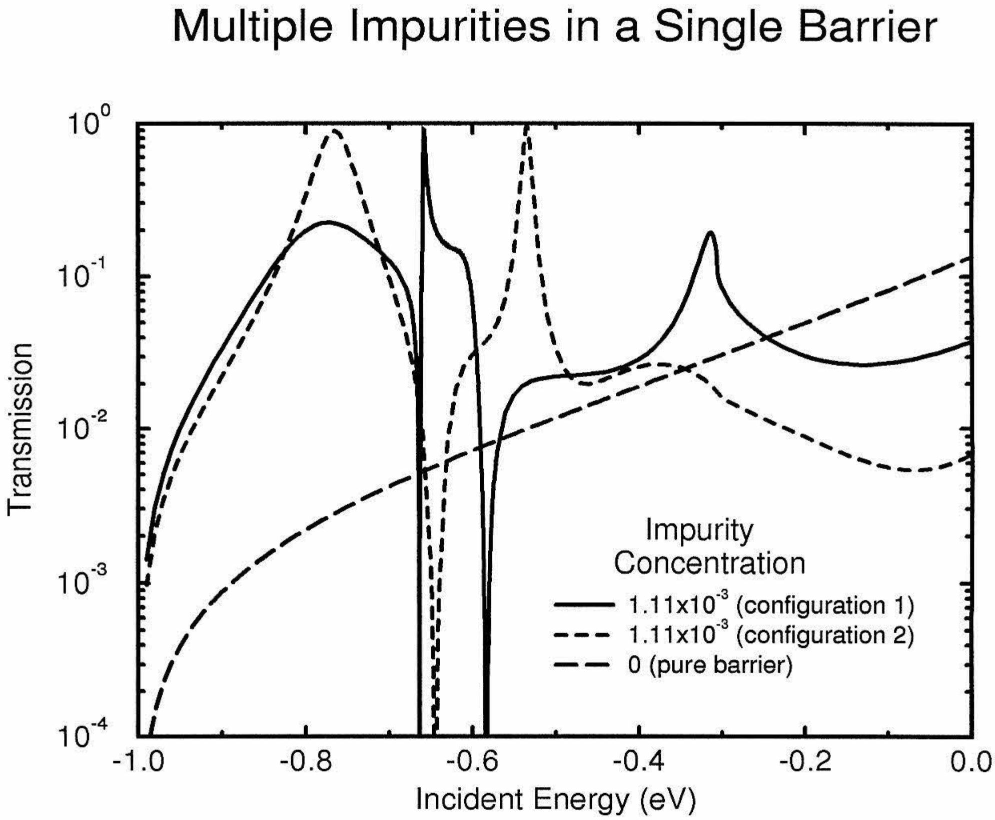

Figure 4.9: Single barrier with multiple impurities. $L_b = 9$ monolayers, 20×20 supercell, $a = 2.825\AA$. $E_b = 0eV$, $m_b = 0.1m_0$, $E_e = -1eV$, $m_e = 0.0673m_0$, $\Delta U/t = -4.5$. Impurity concentration is 1.11×10^{-3} (four impurities were distributed at random among the $9 \times 20 \times 20$ sites of the barrier). Transmission coefficients are shown for two different configurations. Electrons incident along the $z-$direction.

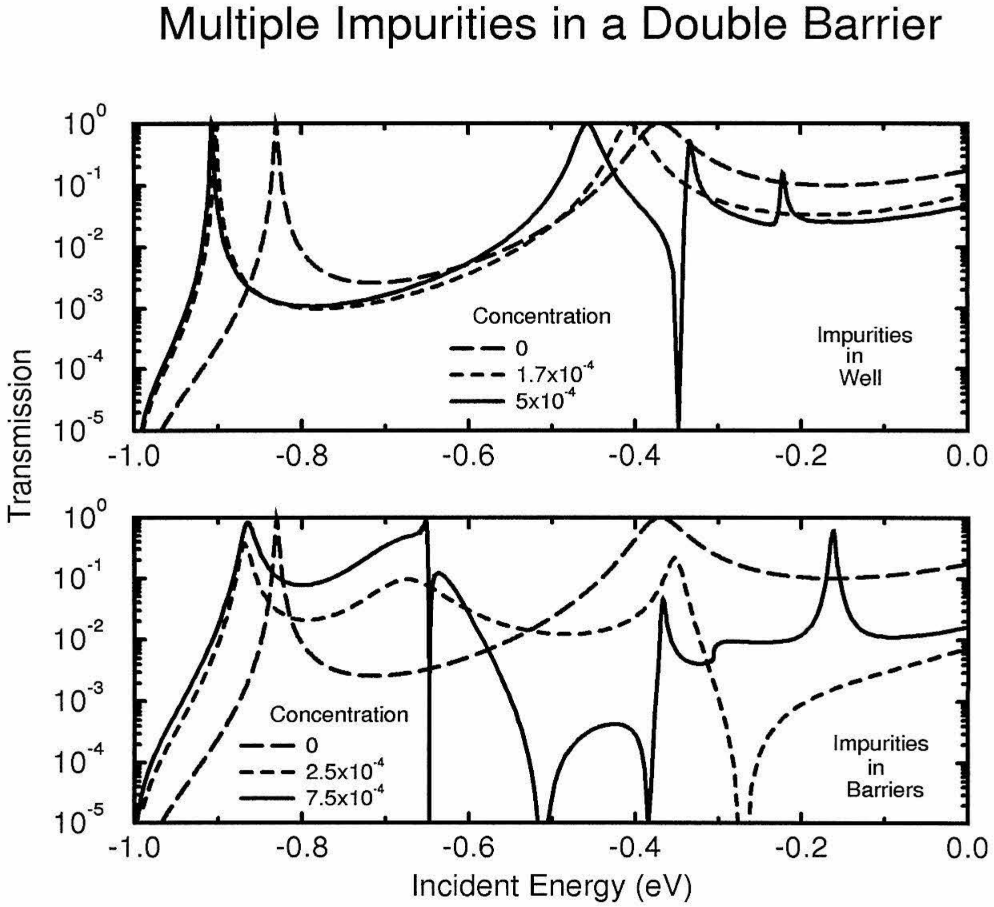

Figure 4.10: Double barrier with multiple impurities. $L_b = 5$ monolayers, $L_w = 15$ monolayers, 20×20 supercell, $a = 2.825 \mathring{A}$. $E_b = 0eV$, $m_b = 0.1m_0$, $E_e = -1eV$, $m_e = 0.0673m_0$, $\Delta U/t = -4.5$. For the case of impurities in the well, ΔU still refers to the difference between the impurity onsite energy and that of the barrier. The top panel shows the case of impurities in the well, and the bottom panel shows the case of impurities in the barriers. Electrons incident along the growth direction.

rities in the well, attractive impurities in the barrier can lower the effective barrier edge. This leads to a lowering and broadening of the $n = 1$ well resonance as seen in the figure. Again we notice new resonances of various strengths and positions.

In both single and double barrier structures, we have seen that impurities can give rise to resonances. The supercell size in the above calculations, 20×20, implies a cross-sectional area approximately $5.7nm$ on an edge. To simulate transport through a larger region of a device, we would need to perform configuration averaging over a large number of different impurity distributions. With a high impurity concentration, the wide variation in resonance structure for different local configurations as shown in Figure 4.9 would no longer allow impurities to produce distinct resonances when probed over a large area. Impurities would, however, still contribute collectively to the transmission by shifting and broadening well resonances in a double barrier or by increasing overall transmission in a single barrier, for example.

4.2.5 Current-Voltage Calculation

Thus far we have examined the effects of impurities on the transmission coefficients of tunneling devices. We have seen that impurities can shift and broaden resonances. Just as importantly, however, when probing devices over a small area, such as with scanning probe microscopy, impurities can give rise to new resonances. These resonances have important consequences for current-voltage characteristics in that they could give rise to negative differential resistance. Experimental evidence of negative differential resistance due to a locally favorable current path created by a donor in the well of double barrier has been presented[15].

In order to calculate the current at a particular bias, we need to integrate the transmission coefficient over the Fermi distribution in the emitter and over the in-plane momenta, $\mathbf{k}_{||}$, as described in Section 2.2.4. Fortunately, the integration over the direction of $\mathbf{k}_{||}$ can be performed analytically, since the transmission coefficient is almost perfectly isotropic, as we might expect from the symmetry of a device with an isolated, point-like impurity. To confirm this, we have plotted the transmission coefficient versus energy near resonance for plane waves incident with three different $\mathbf{k}_{||}$, all of the same magnitude, in the top panel of Figure 4.11. The integration over $|\mathbf{k}_{||}|$ must be performed analytically, however, since the transmission coefficient depends on this quantity in a non-trivial manner. In the bottom panel of Figure 4.11, we plot the transmission coefficients near resonance as a function of E_z, the energy corresponding to k_z in the emitter, for different $|\mathbf{k}_{||}|$. The resonances have slightly different widths and substantially different energy positions. We thus use (2.37) to calculate the current at $0K$ in our device.

We give here the results of our supercell calculation of the $0K$ current-voltage characteristic of a single barrier with an attractive impurity in the middle layer. We use the same material parameters as in section 4.2.1. The barrier is nine monolayers thick, and we take the Fermi level in the electrodes to be 0.05 eV above the band edge. In Figure 4.12 we plot the current density versus applied bias for this device. We see that the isolated impurity gives rise to substantial peak current and negative differential resistance as a result of resonant tunneling via the impurity level.

4.3 Comparison with Experiment

According to the calculations in this chapter and elsewhere[11], an impurity can give rise to a tunneling resonance by providing a locally favorable current path. Indeed, several experiments have reported "anomalous" transport features in various structures, and resonant tunneling has been proposed as the explanation. Before concluding this chapter, we mention a few of these experiments.

The first experiments we describe involve observation of anomalous negative differential resistance in the current-voltage characteristics of a device. Dellow et al.[15] have observed peaks in the current-voltage characteristics of a gated

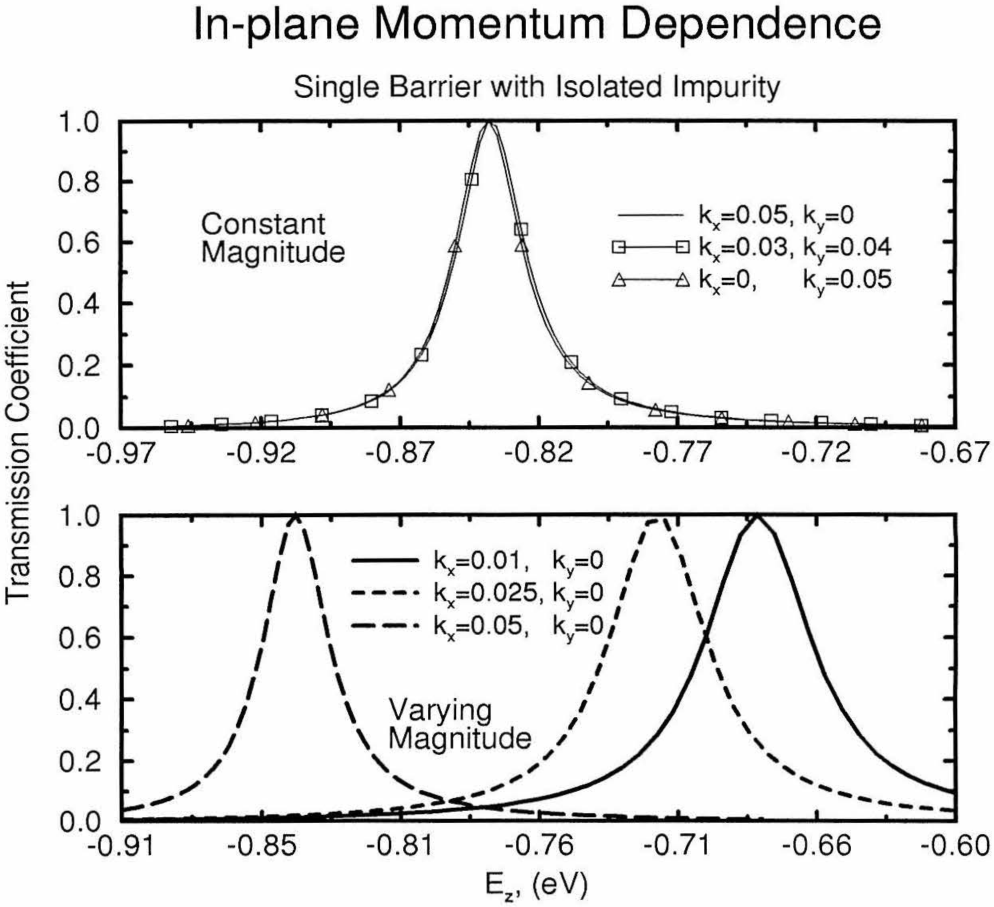

Figure 4.11: Dependence of transmission coefficient on in-plane momentum, $\mathbf{k}_{||}$. The device simulated is a single barrier with an isolated impurity in the middle barrier layer. $L_b = 9$ monolayers, 13×13 supercell, $a = 2.825\mathring{A}$. $E_b = 0eV$, $m_b = 0.1m_0$, $E_e = -1eV$, $m_e = 0.0673m_0$, $\Delta U/t = -4.5$. In the top panel the magnitude of $\mathbf{k}_{||}$ is held constant, and the direction is varied. In the bottom panel, the magnitude is varied along the $x-$direction. E_z is the electron energy corresponding to k_z in the emitter.

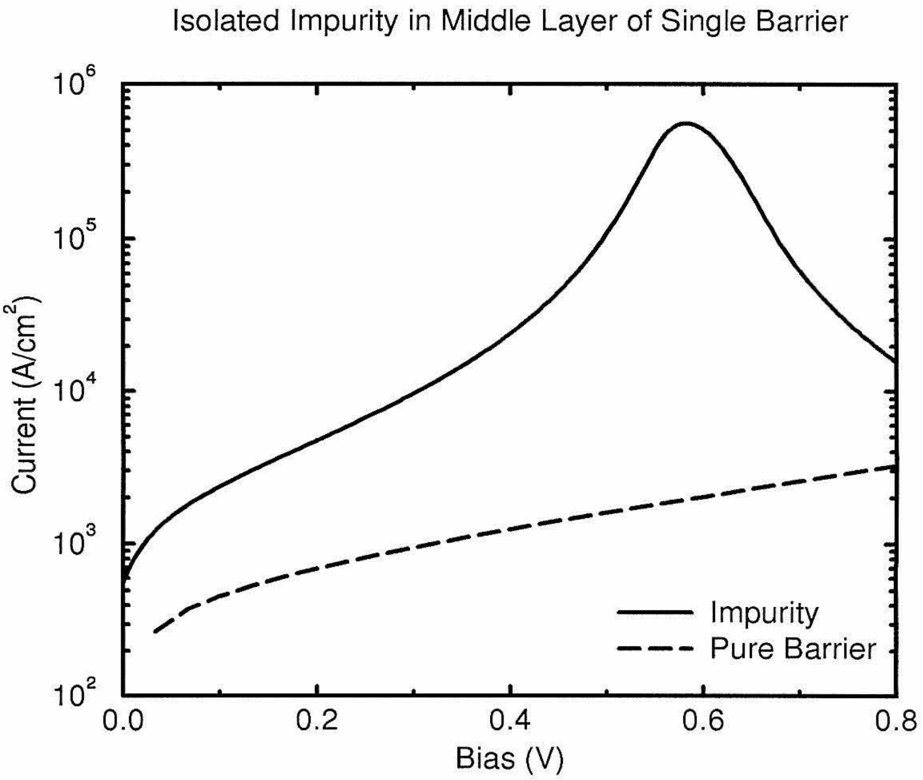

Figure 4.12: Negative differential resistance in a single barrier with an isolated impurity in the middle barrier layer. $L_b = 9$ monolayers, 13×13 supercell, $a = 2.825\mathring{A}$. $E_b = 0eV$, $m_b = 0.1m_0$, $E_e = -1eV$, $m_e = 0.0673m_0$, $\Delta U/t = -4.5$. Electrode Fermi level is $0.05eV$ above the electrode band edge, calculated at $0K$. Also shown for comparison is the current-voltage characteristic of a pure single barrier with the same parameters.

$GaAs/AlGaAs$ double barrier resonant tunneling diode well below the calculated resonant threshold. The gate bias was varied to control the cross-sectional area of the device, but the peaks remained in roughly the same position, ruling out lateral confinement and Coulomb blockade as explanations. Instead, a donor impurity in the quantum well is proposed as the origin of the peaks. In a quite different set of experiments, Tabe et al.[16] have found negative differential resistance in the tunneling spectroscopy of a $1.5nm$ thick SiO_2 film on degenerate Si when examining certain sites. The sites appeared as depressed areas when examined with a scanning tunneling microscope. Again, the negative differential resistance is ascribed to resonant tunneling through localized states in the oxide.

Other experiments have evidently shown tunneling through localized levels in different ways. Capasso et al.[17] have observed a series of narrow peaks in the low-temperature photocurrent-voltage characteristics of multiple-quantum-well $p-i-n$ junctions. The positions of the peaks were not the same from sample to sample, and the average peak height decreased with increasing barrier thickness, strongly suggestive of resonant tunneling through the barriers. In addition, steps in the capacitance-voltage curve were measured at the same positions as the peaks in the photocurrent. The explanation was that electrons dynamically stored in the wells leak out by tunneling through localized states in the barriers. Finally, Koch et al.[18] have measured large peaks in the tunneling conductivity through the oxide of a metal-oxide-silicon field-effect transistor (MOSFET) fabricated with Na^+ ions in the gate oxide. It was demonstrated that the tunneling current was spatially localized, and conduction through a filament or microshort in the oxide was ruled out. Thus the explanation was resonant tunneling through localized states.

4.4 Summary

We have explored several ways in which neutral impurities can play an important role in quantum transport in tunneling devices. We have found that an isolated impurity can give rise to a transmission resonance. The impurity provides a locally favorable current path at the resonant energy. We have investigated the variation of resonance shape and position with the location of an impurity in a single barrier and found that the resonance moves to higher energy, and that the resonance strength grows as the impurity is moved toward the center of the barrier. We have studied the interaction of two closely spaced impurities and found that the manifestation of level splitting in the transmission depends on the relation between the incident electron direction and the impurity separation direction. We have also seen how the level splitting is different when calculated in different dimensionalities, unless the impurities are strongly attractive. An analysis of single and double barriers with multiple impurities reveals that strongly attractive impurities can have a substantial impact on transmission. Depending on impurity concentration and the area over which a structure is probed, the impurities can shift and broaden resonances in a double barrier and increase overall transmission in a single barrier or give rise to new resonances. The influence of impurities thus depends on many factors including material parameters, location, distribution and concentration. In many situations, three-dimensional simulation is essential to understanding the physical phenomena for which the impurities are responsible.

Bibliography

[1] S. T. Pantelides, Rev. Mod. Phys. **50**, 797 (1970).

[2] C. A. Coulson and M. J. Kearsely, Proc. Royal Soc. **A241**, 433 (1957).

[3] R. P. Messmer and G. D. Watkins, Phys. Rev. B **7**, 2568 (1973).

[4] L. U. Dong and L. U. Fen, Chin. Phys. **1**, 472 (1981).

[5] D. G. Thomas, *II-VI Semiconducting Compounds* (Benjamin, New York, 1967).

[6] M. Scheffler, J. Bernholc, N. O. Lipari, and S. T. Pantelides, Phys. Rev. B **29**, 3269 (1984).

[7] B. Monemar, H. P. Gislason, and W. M. Chen, Phys. Rev. B **33**, 4424 (1986).

[8] K. C. Wong, J. Callaway, N. Y. Du, and R. A. LaViolette, Phys. Rev. B **43**, 1576 (1991).

[9] T. C. McGill and R. Baron, Phys. Rev. B **11**, 5208 (1975).

[10] K. C. Wong, N. Y. Du, J. Callaway, and R. A. LaViolette, Phys. Rev. B **41**, 12666 (1991).

[11] D. Stievenard, X. Letartre, and M. Lannoo, Appl. Phys. Lett. **61**, 1582 (1992).

[12] V. I. Sugakov and S. A. Yatskevich, Sov. Phys. Solid State **33**, 302 (1991).

[13] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics* (Addison-Wesley, Menlo Park, 1965).

[14] E. N. Economou, *Green's Functions in Quantum Physics* (Springer-Verlag, New York, 1967).

[15] M. W. Dellow, P. H. Beton, C. J. G. M. Langerak, T. J. Foster, P. C. Main, L. Eaves, M. Henini, S. P. Beaumont, and C. D. W. Wilkinson, Phys. Rev. Lett. **68**, 1754 (1992).

[16] M. Tabe and M. Tanimoto, Appl. Phys. Lett. **58**, 2105 (1991).

[17] F Capasso, K. Mohammed, and A. Y. Cho, Phys. Rev. Lett. **57**, 2303 (1986).

[18] R. H. Koch and A. Hartstein, Phys. Rev. Lett. **54**, 1848 (1985).

Chapter 5

Interface Roughness and Alloy Disorder

5.1 Introduction

5.1.1 Background

Our final application of the supercell model is to interface roughness and alloy disorder in double barrier resonant tunneling structures. Interface roughness is of interest since, at present, there is no way to avoid monolayer fluctuations in epitaxially grown structures, resulting in roughness at heterointerfaces[1, 2]. This roughness can cause scattering, altering transmission properties and degrading device performance characteristics such as the peak-to-valley current ratio. We shall see that interface roughness can play a substantial role in transmission when the scale of the roughness is on the order of the electron deBroglie wavelength. Likewise, alloy disorder is bound to exist in a ternary alloy region such as $Al_xGa_{1-x}As$, and this can have a significant impact on transmission when clusters on the scale of the deBroglie wavelength are present. In this chapter we study the impact of interface roughness and alloy disorder on the transmission properties of double barrier resonant tunneling structures.

Both alloy disorder and interface roughness have been studied theoretically and experimentally. Traditionally, calculated peak-to-valley current ratios in double barrier resonant tunneling diodes have been much higher than experimental values. Part of the discrepancy has been attributed to interface roughness, which can increase valley current via scattering. Chevoir and Vinter[3] have examined scattering assisted tunneling in double barrier resonant tunneling diodes via Fermi's golden rule. They find that interface roughness, alloy disorder and optical and acoustic phonons contribute to the valley current. Current-voltage characteristics in a $GaAs/AlGaAs$ double barrier structure with interface roughness have also been calculated within the coherent potential approximation[4]. It was found that scattering doesn't change peak current much, but it can raise valley current several orders of magnitude in a structure with thick barriers. In this calculation, the peak-to-valley current ratio grows quickly with barrier thickness and then saturates at around $7 - 10nm$, in good agreement with recent experiment[5, 6]. Apell[7] has made the important observation that the length scale of variations in the interface roughness relative to the electron deBroglie wavelength is important: when the roughness varies on a scale smaller than the deBroglie wavelength, the effect is minimal; when the two length scales are similar, the effect can be large. On a related topic, photoluminescence spectra in double barrier structures indicate that such large-scale roughness can give rise to hole localization in the quantum well[8].

In this thesis, we examine three-dimensional quantum transport in double barrier structures with interface roughness and alloy disorder. Our supercell model gives a precise, three-dimensional microscopic description of interface roughness and alloy disorder, allowing us to address issues such as ordering and clustering without configuration averaging. In addition we can calculate electron wave functions, explicitly showing how interface roughness and alloy disorder induce localization, altering transmission properties.

5.1.2 Outline of Chapter

We first calculate transmission coefficient curves for a series of double barrier structures with interface roughness characterized by different island sizes. We observe two effects: in-plane momentum ($\mathbf{k}_{||}$) scattering, which produces a broad bump just above the $n = 1$ resonance, and wave function localization, which broadens and downshifts the $n = 1$ resonance. We next calculate transmission coefficient curves for double barrier structures with alloy barriers. Different degrees of disorder and clustering in the barriers have different impacts on transmission. As the cluster size increases, the barriers grow less confining, broadening resonances and shifting them to lower energy. In addition, localized states arise, leading to new transmission resonances.

5.2 Simulation and Results

5.2.1 Interface Roughness

We begin by simulating double barrier structures with interface roughness. The structures we consider consist of a $L_w = 10$ monolayer $GaAs$ well and $L_b = 4$ monolayer $AlAs$ barriers. The electrodes are made of $GaAs$. Between the emitter and the left barrier and between the well and the right barrier, we insert a monolayer of interface roughness, composed of 50% $AlAs$ sites and 50% $GaAs$ sites, generated by a simulated annealing algorithm[9], in order to create interface islands. The size, λ, of the islands in such a layer is specified as twice the distance at which the autocorrelation for the site type function radially averaged and averaged over the supercell sites for that layer vanishes: $\langle S(\mathbf{r})S(\mathbf{r}+\lambda)\rangle = 0$, where $S(\mathbf{r}) = +1$ if site $\mathbf{r}$ is $AlAs$, and $S(\mathbf{r}) = -1$ if it is $GaAs$ (see Figure 5.1). The material parameters are $E_{GaAs} = 0eV$, $m_{GaAs} = 0.0673m_0$, $E_{AlAs} = 1.05eV$, $m_{AlAs} = 0.1248m_0$.

In Figure 5.2, we plot transmission coefficient curves near the $n = 1$ resonance for a series of double barrier structures with different island sizes. The incident

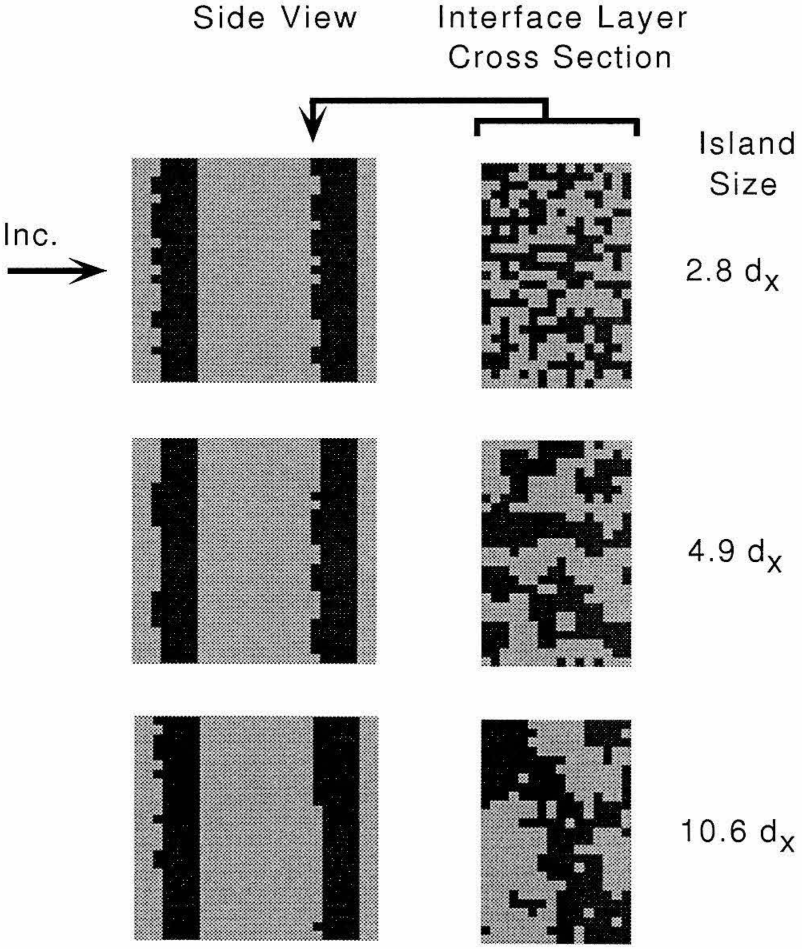

Figure 5.1: A double barrier structure with interface roughness consisting of two layers of 50% $AlAs$ and 50% $GaAs$ sites arranged using a simulated annealing algorithm. The light areas represent $GaAs$, and the dark areas represent $AlAs$. Different degrees of clustering are considered, leading to different average interface island sizes. In this example, a 16 × 25 supercell is used.

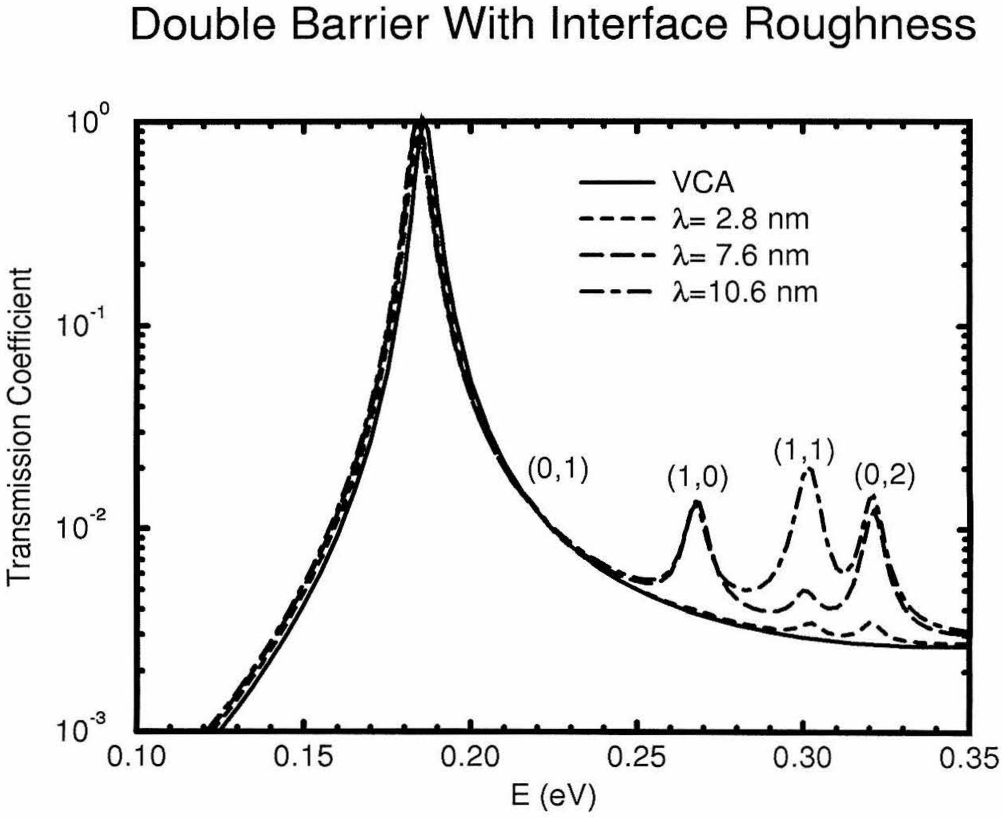

Figure 5.2: *GaAs*/*AlGaAs* Double barrier structure with a layer of interface roughness between the emitter and the left barrier and between the well and the right barrier. Each rough layer consists of 50% *AlAs* and 50% *GaAs*, arranged via a simulated annealing algorithm. $L_W = 12$, $L_b = 4$, $d_x = d_y = 1nm$, $d_z = 0.2825nm$, 16×25 supercell. Plane waves are incident along the $z-$direction ($\mathbf{k}_{||} = 0$). Calculations for rough layers with island sizes of $\lambda = 2.8$, 7.6 and 10.6nm are presented, along with a virtual crystal approximation calculation.

plane wave is chosen to have $\mathbf{k}_{||} = 0$. We also include, for reference, a transmission coefficient curve calculated in the virtual crystal approximation, where the layers of interface roughness were replaced with a fictitious material, whose effective mass and band edge are the averages of those of the constituents, $AlAs$ and $GaAs$. We notice satellite peaks above the $n = 1$ resonance in all cases except for the virtual crystal calculation. These peaks increase in strength with increasing island size, but they remain in the same position even though the interface roughness configurations vary considerably for the different cases shown.

The satellite peaks are due to in-plane momentum scattering in the supercell model. As stated in Section 2.1 a plane wave incident with in-plane momentum $\mathbf{k}_{||}^{inc}$ can scatter only to states with $\mathbf{k}_{||} = \mathbf{k}_{||}^{inc} + \mathbf{q}_{||}^{l,m}$, where $\mathbf{q}_{||}^{l,m} = (\frac{2\pi l}{N_x d_x}, \frac{2\pi m}{N_y d_y})$, $l = 1, \cdots, N_x$, $m = 1, \cdots, N_y$. In such a state, the band edge profile of the double barrier is just that of the double barrier at $\mathbf{k}_{||} = 0$ shifted up by $\Delta E_{lm} = E_{GaAs}(\mathbf{q}_{||}^{l,m}) - E_{GaAs}(\mathbf{0}) = 2t_{GaAs}(\cos\frac{2\pi l}{N_x} + \cos\frac{2\pi m}{N_y})$ (to within minor corrections due to the fact that $AlAs$ has a higher effective mass than $GaAs$). Thus when the total energy of the incident plane wave is $E_1 + \Delta E_{lm}$ (where E_1 is the $\mathbf{k}_{||} = 0$, $n = 1$ resonance energy), there should be a resonance related to the $n = 1$ resonance. The $\{\Delta E_{lm}\}$ thus determine the positions of the satellite peaks. The first few satellite peaks are shown in Figure 5.2, labeled by (l, m). Since the $\{\mathbf{q}_{||}^{l,m}\}$ (and hence the $\{\Delta E_{lm}\}$) are determined by the supercell dimensions $N_x d_x \times N_y d_y$ only, the peak positions don't vary with different configurations of interface roughness. As the supercell size increases, the $\{\Delta E_{lm}\}$ become more closely spaced until the peaks are separated by less than their widths. For very large supercell size (as would be needed to represent a macroscopic sample), the satellite peaks would coalesce into a broad bump above the $n = 1$ resonance.

The (l, m) satellite peak strength is determined by the extent of scattering into $\mathbf{k}_{||} = \mathbf{q}_{||}^{l,m}$; the larger the island size, the larger the scattering. As the island size is increased beyond that in Figure 5.2, a new phenomenon appears. The $n = 1$ resonance is broadened and downshifted, as shown in Figure 5.3 where we plot trans-

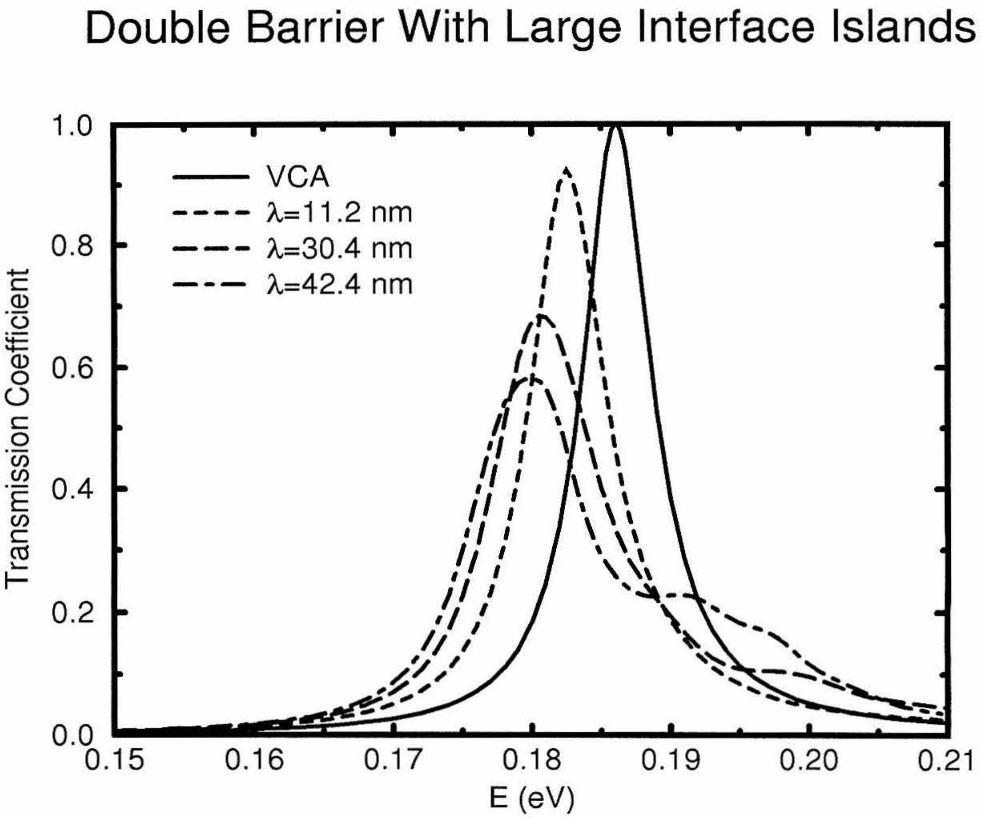

Figure 5.3: A *GaAs*/*AlGaAs* Double barrier structure with a layer of roughness between the emitter and the left barrier and between the well and the right barrier. Each rough layer consists of 50% *AlAs* and 50% *GaAs*, arranged via a simulated annealing algorithm. $L_W = 12$, $L_b = 4$, $d_x = d_y = 4nm$, $d_z = 0.2825nm$, 16×25 supercell. Plane waves are incident along the $z-$direction ($\mathbf{k}_{\|} = 0$). Calculations for rough layers with island sizes of $\lambda = 11.2$, 30.4 and 42.4nm are presented, along with a virtual crystal approximation calculation.

mission coefficient curves near the $n = 1$ resonance of double barrier structures with larger interface islands. This can be explained by wave function localization. As the interface island size increases beyond the electron deBroglie wavelength, the electrons begin to sense two regions: one with a wide well ($L_W = 13$) and one with a narrow well ($L_W = 12$). (In the case of microroughness, the electrons sense an average well width, in between that of the narrow and the wide wells.) The wide well region supports a lower resonance level than the structures with microroughness—hence the shift of the transmission peak to lower energy. In addition the wide well region is isolated from the collector by a thinner barrier (see Figure 5.1), accounting for some of the broadening. To give some intuition for localization in the case of large interface islands, we show, in Figure 5.4, a probability density isosurface for the electron wave function at the $n = 1$ resonance for the structure with $\lambda = 42.2nm$. It is evident that resonant transmission takes place mainly via the wide-well regions. The broad bump above the $n = 1$ peak in the transmission in Figure 5.3 is a result of $\mathbf{k}_{||}$ scattering into wide-well modes modulated by non-zero in-plane momentum. In addition, the transmission maximum is reduced on account of this scattering.

Thus we have seen two effects of interface roughness in a double barrier: $\mathbf{k}_{||}$ scattering can contribute to a broad bump above the $n = 1$ resonance, and, for large island sizes, tunneling restricted to the wide well can downshift and broaden the $n = 1$ resonance.

5.2.2 Alloy Disorder

We next examine alloy disorder in the barriers of a double barrier structure. We simulate structures with a $L_W = 12$ monolayer $GaAs$ well and $L_b = 10$ monolayer $Al_{0.5}Ga_{0.5}As$ barriers. The electrodes are again composed of $GaAs$. We plot transmission coefficient curves for such a structure in Figure 5.5, where the alloy barriers are composed of an uncorrelated random distribution of $AlAs$ and $GaAs$

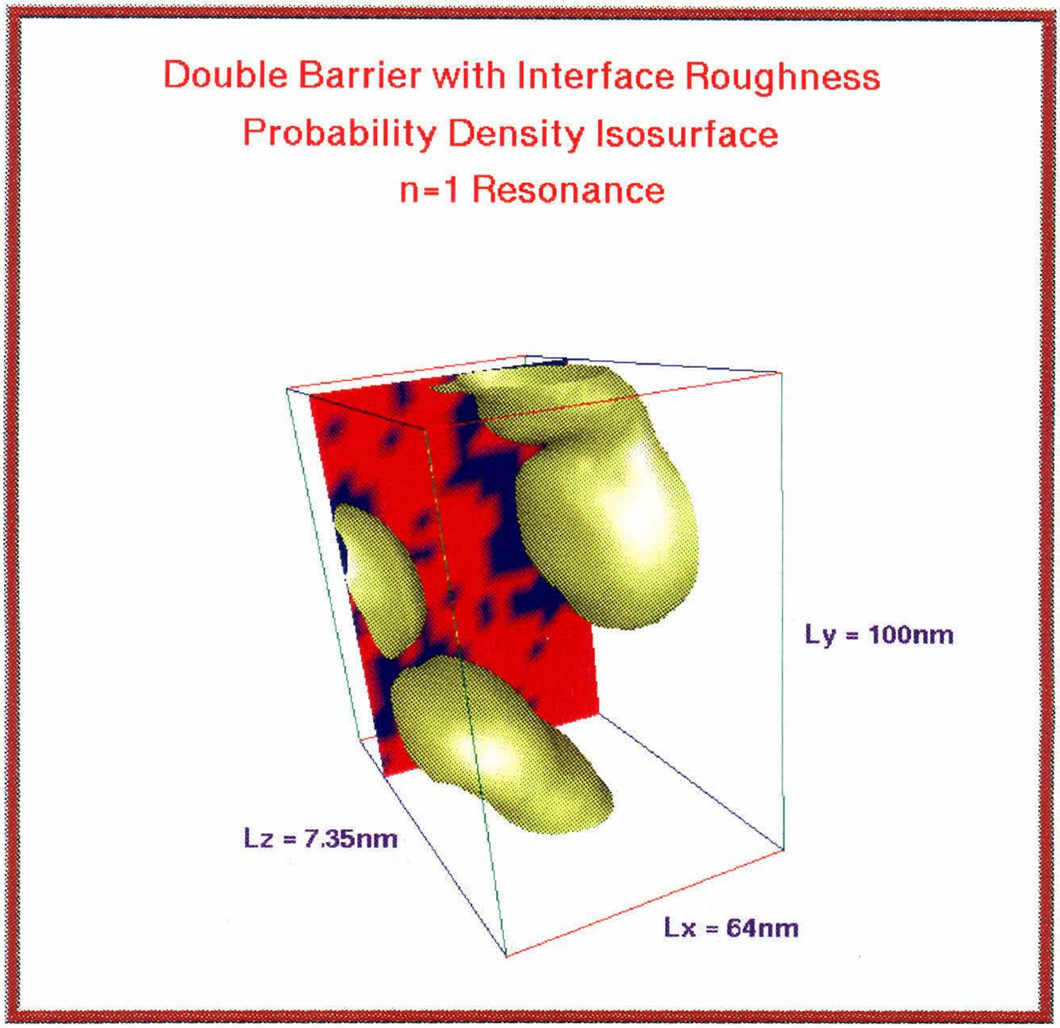
Double Barrier with Interface Roughness
Probability Density Isosurface
n=1 Resonance
Ly = 100nm
Lz = 7.35nm
Lx = 64nm

Figure 5.4: Probability density isosurface at the $n = 1$ resonance for the structure labeled $\lambda = 42.4nm$ in Figure 5.3. Also shown is a cross-section of the double barrier structure, showing the rough interface between the well and the right electrode. The red areas represent $AlAs$, and the blue areas represent $GaAs$.

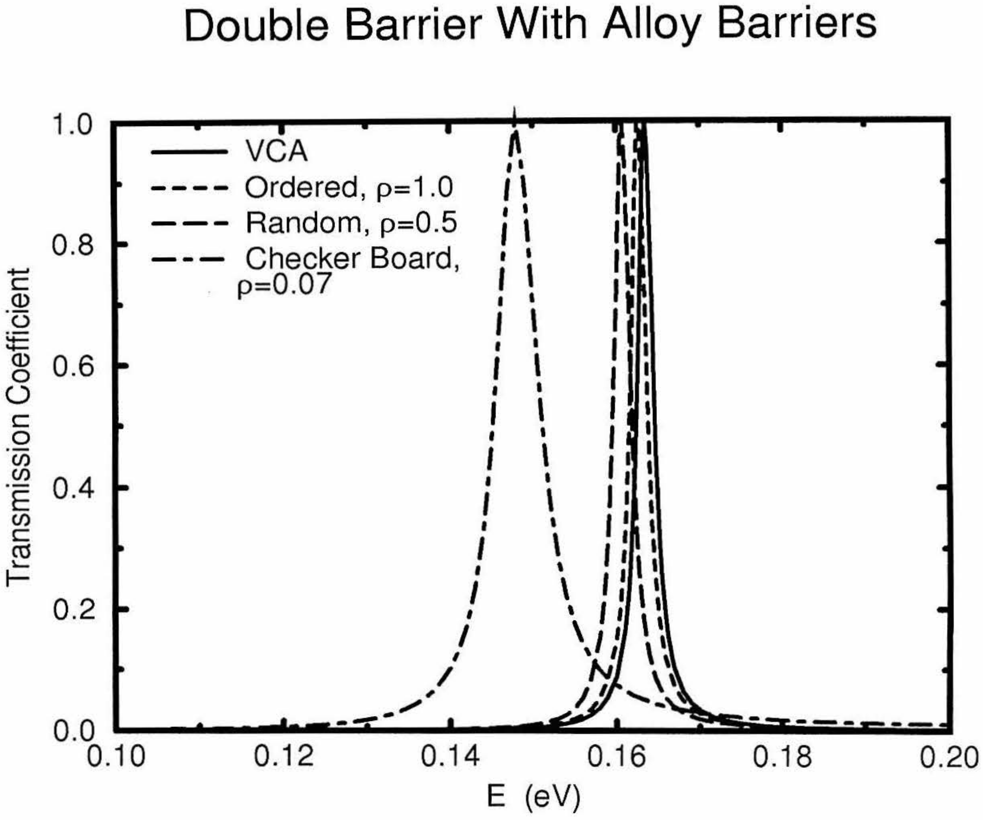

Figure 5.5: *GaAs/AlGaAs* double barrier structures with $Al_{0.5}Ga_{0.5}As$ alloy barriers. Transmission coefficient curves for different degrees of disorder, characterized by the short-range order parameter ρ are considered: one where each *AlAs* site is completely surrounded by *GaAs* sites (Ordered, $\rho = 1$), one where the *AlAs* and *GaAs* sites are distributed randomly (Random, $\rho = 0.5$), and one consisting of square regions of *AlAs* and *GaAs*, $0.2825nm$ on an edge, arranged in a checkerboard pattern (each barrier layer is identical; Checker Board, $\rho = 0.07$). Also shown for reference is a virtual crystal approximation calculation. $L_W = 12$, $L_b = 10$, $d_x = d_y = d_z = 0.2825nm$, 20×20 supercell. Plane waves are incident along the $z-$direction ($\mathbf{k}_{||} = 0$).

sites. For comparison, we include the results of a virtual crystal approximation calculation, where the alloy barriers are replaced by the same fictitious material as for the rough interfaces in the preceding section. We see that the virtual crystal approximation serves well, even though it neglects the rapid potential variations in the alloy regions. This is in line with our findings concerning interface roughness, where the roughness varied rapidly on the scale of the electron deBroglie wavelength, having little effect.

We can also use the supercell model to investigate alloy disorder. We plot, in Figure 5.5, transmission coefficient curves for a structure with alloy barriers where each $AlAs$ site is surrounded by $GaAs$ sites (Ordered), and for a structure with barriers composed of identical checkerboard layers, with square patches of $AlAs$ and $GaAs$ $0.2825nm$ on an edge (Checker Board). These structures may be characterized by the short-range order parameter ρ, defined as the ratio of the number of bonds connecting different types of sites (i.e., $GaAs - AlAs$) and the total number of bonds. For the ordered, random and checkerboard barriers, $\rho = 1.0, 0.5$ and 0.07, respectively. From the figure, it is clear that the $n = 1$ resonance is downshifted and broadened more and more with decreasing ρ.

Particularly striking are the shift and broadening for the checkerboard structure. This suggests that clustering may be important when the cluster size approaches the electron deBroglie wavelength. To examine clustering, we plot, in Figure 5.6, transmission coefficient curves for double barriers with alloy barriers with varying degrees of clustering. The barriers are generated with the same simulated annealing algorithm as the interface islands, layer by layer, and cluster size is characterized by the average island size in the layers. The larger the cluster, the more the $n = 1$ resonance is shifted down and broadened. Thus the effective barrier is less confining for larger clusters. For island sizes greater than about $5 - 6nm$, substantial new structure develops in the transmission coefficient curves. New peaks arise on account of localized states in the $GaAs$ clusters in the barriers. The peaks vary substantially in position depending on the cluster sizes and shapes

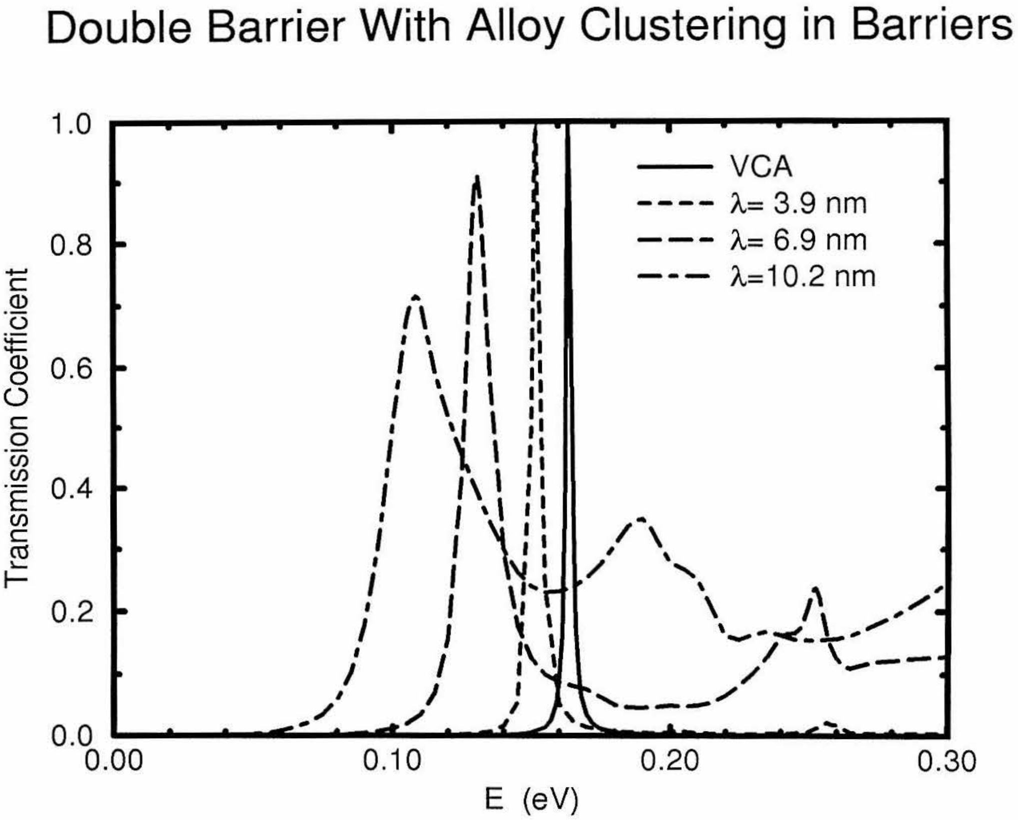

Figure 5.6: *GaAs/AlGaAs* double barrier structures with $Al_{0.5}Ga_{0.5}As$ alloy barriers generated by a simulated annealing algorithm. Transmission coefficient curves for different degrees of clustering, characterized by planar island sizes of $\lambda = 3.9$, 6.9 and $10.2nm$ are calculated. Also shown for reference is a virtual crystal approximation calculation. $L_W = 12$, $L_b = 10$, $d_x = d_y = 1nm$, $d_z = 0.2825nm$, 20×20 supercell. Plane waves are incident along the $z-$direction ($\mathbf{k}_{\|} = 0$).

for different alloy configurations.

5.3 Summary

We have investigated the effects of interface roughness and alloy disorder on the transmission properties of double barrier resonant tunneling structures. We found that interface roughness can cause both in-plane momentum scattering, which results in additional resonance structure above the $n = 1$ peak, and wave function localization, which downshifts and broadens the $n = 1$ resonance for island sizes on the order of the electron deBroglie wavelength. We have seen that alloy disorder can also downshift and broaden the $n = 1$ resonance, and that clustering becomes important when the cluster size is on the order of the electron deBroglie wavelength: wave function localization in clusters in the barriers can produce new transmission peaks.

Bibliography

[1] J. Christen, M. Grundmann, and D. Bimberg, J. Vac. Sci. Technol. B **9**, 2358 (1991).

[2] M. H. Bode and A. Ourmazd, J. Vac. Sci. Technol. B **10**, 1787 (1992).

[3] F. Chevoir and B. Vinter, Phys. Rev. B **47**, 7260 (1993).

[4] P. Johansson, Phys. Rev. B **46**, 12865 (1992).

[5] P. Gueret, C. Rossel, W. Schlup, and H. P. Meier, J. Appl. Phys. **66**, 4312 (1989).

[6] J. S. Wu, C. P. Lee, C. Y. Chang, K. H. Chang, D. G. Liu, and D. C. Liou, Sol. State Elec. **35**, 723 (1992).

[7] S. P. Apell, Phys. Scr. **43**, 630 (1991).

[8] T. H. Wang, X. B. Mei, C. Jiang, Y. Huang, J. M. Zhou, and G. Z. Yang, Appl. Phys. Lett. **62**, 1149 (1993).

[9] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

Chapter 6

Fluctuations in a Quantum Dot Waveguide

6.1 Introduction

6.1.1 Background

Our second application is to interface roughness and impurities in an electron waveguide. As we saw in Chapter 4, different configurations of defects lead to different effects on transmission. We shall see that different configurations of interface roughness in a quantum dot lead to fluctuations in transmission. An attractive impurity near the center of the dot can reduce these fluctuations, but the presence of more than a single impurity can lead to complex, impurity configuration dependent resonance structure, especially at high concentrations. If quantum devices are to become commercially viable as components in mass-produced circuits, statistical variations in imperfections from device to device and fluctuations must be considered. In this Chapter we discuss transmission fluctuations due to variations in interface (micro-)roughness and impurities in a quantum dot electron waveguide. We first present some background on quantum dot and quantum wire waveguides.

Quantum dots and quantum wire-shaped electron waveguides have been produced by a variety of techniques. One of the most prevalent involves epitaxial growth and lateral etching[1, 2]. A single barrier is grown using molecular beam epitaxy and then etched via x-ray lithography to produce lateral confinement on the scale of about 0.1μm. Another technique involves selective growth on a patterned substrate[3, 4, 5, 6]. Taking advantage of different growth rates along different crystal lattice directions, these techniques have produced wires with lateral confinement on the scale of about $50nm$. Fabrication of p-type Si quantum wires in n-type substrates has been achieved by selective implantation of focused ion beams of Ga[7]. Yet other methods use metal gate electrostatic confinement or strain gradients to produce lateral confinement[8].

On account of their small dimensions compared to semiconductor crystal lattice constants, most quantum wires exhibit structural variation. Interface roughness at the boundaries of the wires over the scale of a few monolayers is currently unavoidable. In addition, compositional variation, particularly due to impurities, is difficult to eliminate. As a consequence, the effects of these variations on device performance have drawn considerable attention.

Theoretical studies of quantum wires have revealed that interface roughness can alter the transmission spectra. A small width increase in one place in a quantum wire has been shown to produce dips in the well-known step-like conductance structure[9]. It has also been shown that cross-sectional area variations along a wire lead to a smearing of the peak-like structure of the average density of states plotted as a function of carrier energy[10].

Impurities in quantum wires have been studied both experimentally and theoretically. An isolated conductance peak observed below the turn-on of the first transverse mode in a narrow constriction has been attributed to resonant tunneling via a single impurity[11]. Degradation in the quantized conductance steps of a dual electron waveguide has been seen when the conductance channel is electrostatically steered into a scatterer[12]. Theoretical studies of an impurity in a narrow channel have revealed the ways in which scattering alters the transmission properties[13, 14, 15]. In these papers dips, peaks and shifts in the conductance and transmission coefficient curve features as a function of impurity location and strength have been calculated. Calculations involving a T-shaped quantum wire junction have shown that a repulsive impurity can both enhance and suppress transmission[16]. Impurities near the aperture of a waveguide have been shown to destroy quantized conductance[17], and ionized donors have been shown to affect the quantized conductance of point contacts in a way that reflects the detailed configuration of the impurities[18].

Other investigations of imperfections have been carried out, mostly by way of specific examples. If quantum devices are to become commercially viable as components of mass-produced circuits, however, statistical variations in imperfections from device to device and fluctuations must be considered. In this thesis we take advantage of the capability of our three-dimensional, supercell model of quantum transport to represent variation both along and perpendicular to the growth direction, allowing us to study novel geometries such as quantum wires and dots with structural and compositional variations.

6.1.2 Outline of Chapter

In Section 6.2.1 we give some examples of supercell calculations of the effects of imperfections such as interface roughness, impurities, and structural variations in a quantum dot. We then examine, in Section 6.2.2, fluctuations in the transmission resonance position, width and maximum due to different interface roughness configurations of the same statistical description in a quantum dot waveguide. In Section 6.2.3 we study the influence of a neutral impurity as a function of strength and location in a quantum dot with interface roughness. We find that an attractive impurity placed near the center of the waveguide can reduce transmission coefficient fluctuations from sample to sample. A high concentration of impurities in the dot, however, leads to a complex resonance structure that varies with impurity configuration. We summarize and conclude in Section 6.3.

6.2 Simulation and Results

6.2.1 Device Imperfections

We begin with an overview of some device imperfections in a quantum dot wave–guide treated with the supercell model. As in Chapter 4, the device electrodes are separated along the $z-$direction. A quantum dot with interface roughness and an impurity in the center, for example, is represented as in Figure 6.1. In Figure 6.2 we show the transmission coefficient for an ideal quantum dot and for dots with various imperfections. The ideal quantum dot is a $3.5nm \times 3.5nm \times 4.5nm$ cavity surrounded by confining walls with smooth interfaces and sandwiched between two electrodes along the $z-$direction. The center of the dot is taken as $x = y = z = 0$. The confining walls are made of barrier material, characterized by a band edge of $E_b = 1.05eV$ and an effective mass of $m_b = 0.1248m_0$, the cavity is composed of well material with a band edge of $E_w = 0eV$ and an effective mass of $m_w = 0.0673m_0$, and the electrodes have a band edge of $E_e = -1eV$ and an effective mass of $m_e = 0.1m_0$. A 13×13 supercell is used with discretization lengths $d_x = d_y = d_z \equiv a = 0.5nm$. The transmission coefficient curve is plotted for plane waves incident along the $z-$direction with no momentum in the $x-$ or $y-$directions. We see that, with these parameters, the first two transmission resonances occur at about $0.47eV$ and $0.85eV$.

Also plotted in Figure 6.2 is the transmission coefficient curve for a dot with an attractive neutral impurity in the center. The impurity is represented by a single site whose onsite energy is ΔU below that of the surrounding sites, and the hopping matrix element to the site, t, is the same as that in the surrounding well-type material. ΔU is thus positive for an attractive impurity and negative for a repulsive impurity. We shall use the dimensionless quantity $\Delta U/t$ as a measure

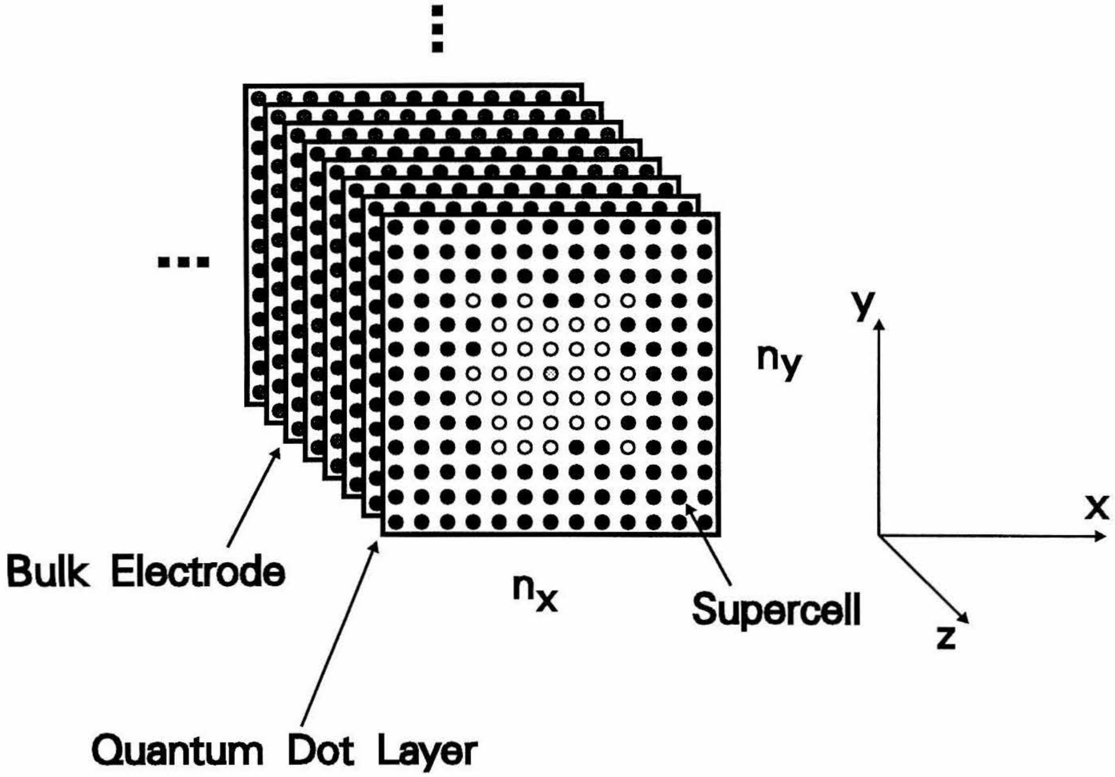

Figure 6.1: Supercell representation of a quantum dot electron waveguide with rough walls and an impurity in the cavity. The darkly shaded circles represent electrode material, the solid circles represent barrier material, the open circles represent the well material in the cavity and the lightly shaded circle represents an impurity.

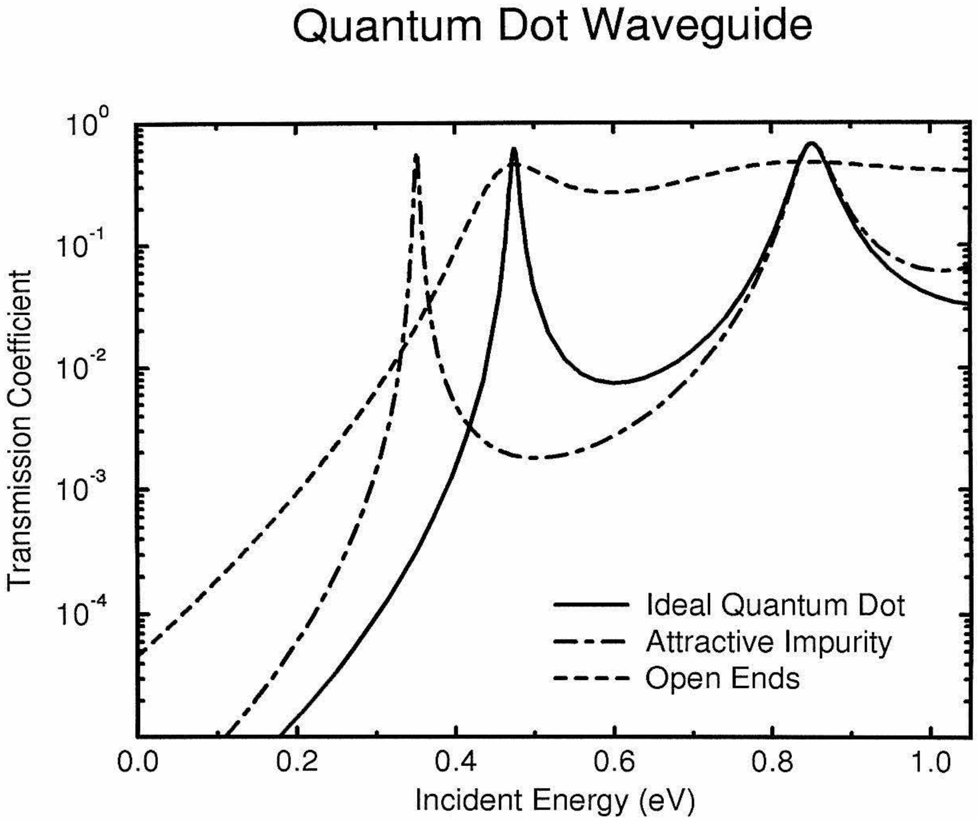

Figure 6.2: Transmission coefficient curves for quantum dots with device imperfections treated in the supercell model. The dots consist of a $3.5nm \times 3.5nm \times 4.5nm$ cavity surrounded by smooth confining walls, except in the case of the dashed curve, where the cavity is extended to the electrodes along the $z-$direction. The impurity is represented by a single site with $\Delta U/t \approx -3.1$. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

of impurity strength. For the curve shown in Figure 6.2, an attractive impurity with $\Delta U/t \approx -3.1$ was used ($t < 0$ by convention—see Eq. (2.2)). We see that the impurity lowers and sharpens the $n = 1$ resonance, but the $n = 2$ resonance remains nearly the same. This behavior will be explained in Section 6.2.3. Finally, in Figure 6.2, we show the transmission coefficient curve for a quantum dot where the cavity has been extended to the electrodes along the $z-$direction (Open Ends). Here the transmission resonances are broadened on account of reduced confinement of the resonant state.

In Figure 6.3 we show the transmission coefficient curve for a dot with interface roughness at the boundary between the cavity and the confining walls. The roughness consists of a $0.5nm$ shell which is a mixture of roughly 50% well material and 50% barrier material. The shell is constructed one site at a time, each site having a probability 0.5 of being well-type and 0.5 of being barrier-type, without correlation. Also plotted in the figure for reference are transmission coefficient curves for two ideal dots with smooth walls, whose dimensions represent the range of dimensions of the dot with rough walls. We see that the $n = 1$ transmission resonance for the dot with rough walls falls in between the $n = 1$ resonances of the two ideal dots.

It is thus evident that device imperfections on an atomic scale can alter transmission characteristics. Just as important, however, are the fluctuations from device to device due to variations in the imperfections. A set of quantum dots with the same statistical characterization of interface roughness, but different roughness configurations, could produce different transmission coefficient curves, leading to fluctuations from structure to structure. In order to calculate these fluctuations, we take advantage of the capability of our supercell model to simulate structures with three-dimensional variation for a variety of configurations. In the next section, we examine the impact of interface roughness variations on the transmission characteristics of a quantum dot.

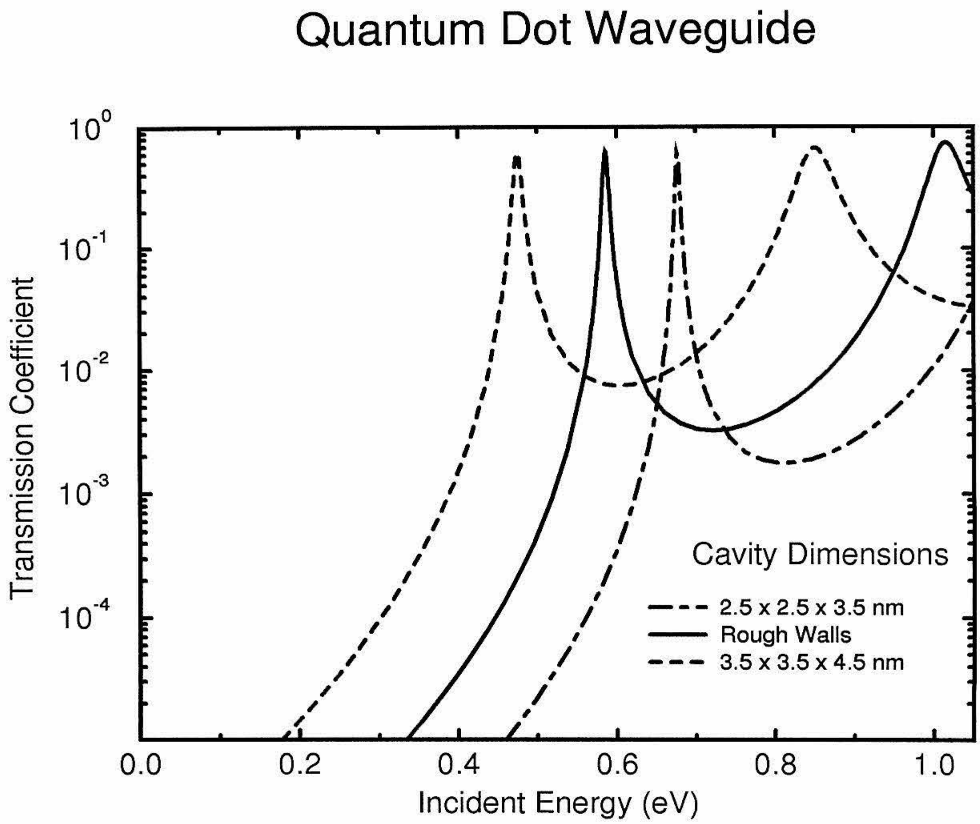

Figure 6.3: Transmission coefficient curves for a quantum dot with a $2.5nm \times 2.5nm \times 3.5nm$ cavity surrounded by a $0.5nm$ shell of interface roughness consisting of 50% well-type material and 50% barrier-type material randomly distributed without correlation. Also shown are transmission coefficient curves for two smooth-walled dots $2.5nm \times 2.5nm \times 3.5nm$ and $3.5nm \times 3.5nm \times 4.5nm$. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

6.2.2 Interface Roughness Fluctuations

To see the effects of interface roughness variation, we calculated transmission coefficient curves for ten dots with different roughness configurations. The description of the roughness is the same as in Section 6.2.1, and the results are plotted in the top panel of Figure 6.4, along with transmission coefficient curves for the same reference structures as in Figure 6.3. We see immediately that the resonance position varies over a range comparable to the resonance width. It is more difficult to see fluctuations in the resonance width and maximum transmission, so we plot these versus sample number in the bottom panel. The values are normalized so that the average width for the ten samples is 1, as is the average transmission maximum. We see that there is about a $10\% - 20\%$ variation in the resonance width and roughly a 5% variation in the maximum transmission. Also plotted for scale in Figure 6.4 are the resonance widths and maximum transmission coefficients for the two reference structures. Both the width and maxima for the ten samples show substantial fluctuation on this scale.

These fluctuations can be attributed to two sources of variation in the interface roughness surrounding the quantum dot: stoichiometric variation and variation in the configuration. Stoichiometric variation arises from the method used to generate the rough interfaces in Figure 6.4: each site in the shell of roughness is chosen with a probability of 0.5 of being well material and a probability of 0.5 of being barrier material. This means that the total number of barrier sites in the shell can vary, producing different effective levels of confinement. We can separate this variation from that of the configuration by constraining the stoichiometry in the shell. We plot, in the top panel of Figure 6.5, transmission coefficient curves for a set of ten rough-walled dots with constrained stoichiometry so that the total number of barrier sites in the shell is 134 (out of 266 total sites). Each dot is thus surrounded by the same amount of barrier material, but with a different configuration of roughness. In the bottom panel the width and maximum transmission

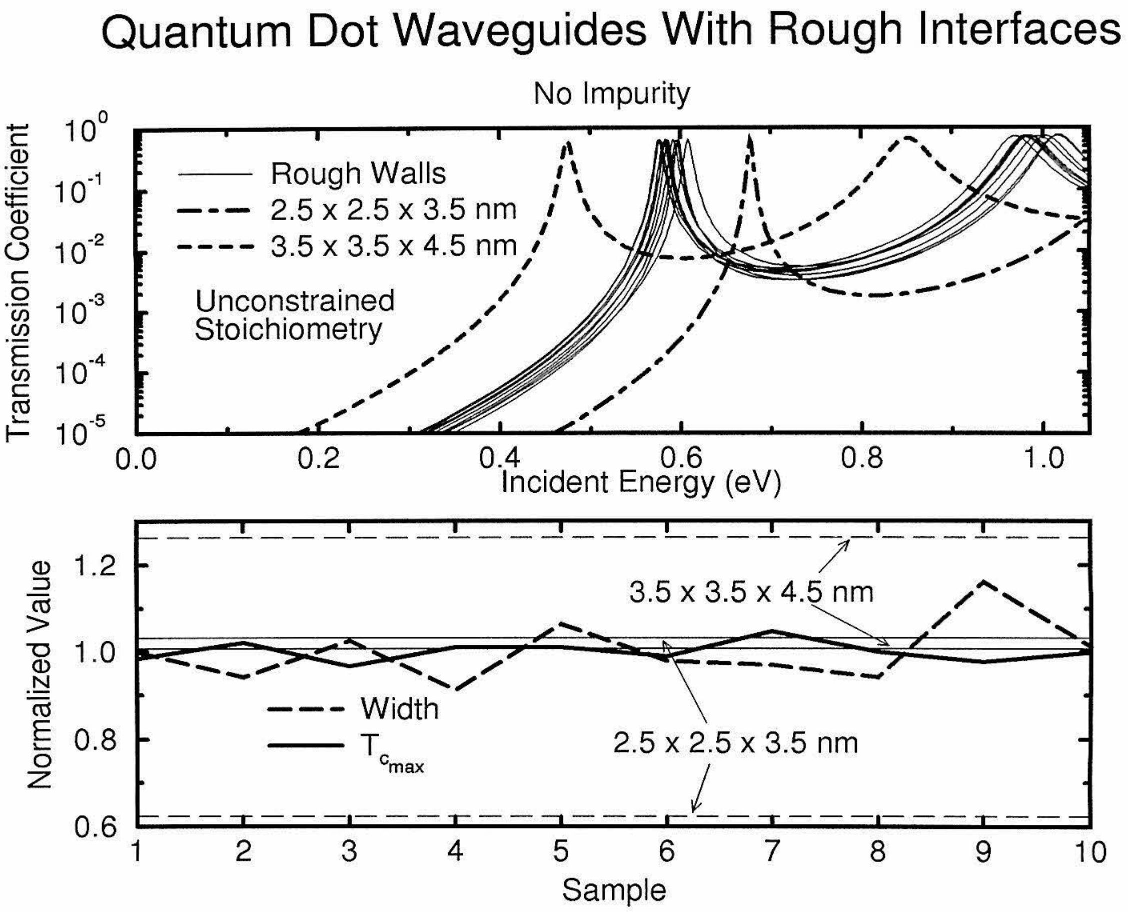

Figure 6.4: Top panel: transmission coefficient curves for quantum dots with ten different rough-walled configurations. The two reference curves from Figure 6.3 are also shown. Bottom panel: fluctuations in the resonance width and maximum transmission for quantum dots with ten different rough-walled configurations. Fluctuating values are normalized so that their mean is 1. Values for the two reference structures from Figure 6.3 are also plotted. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

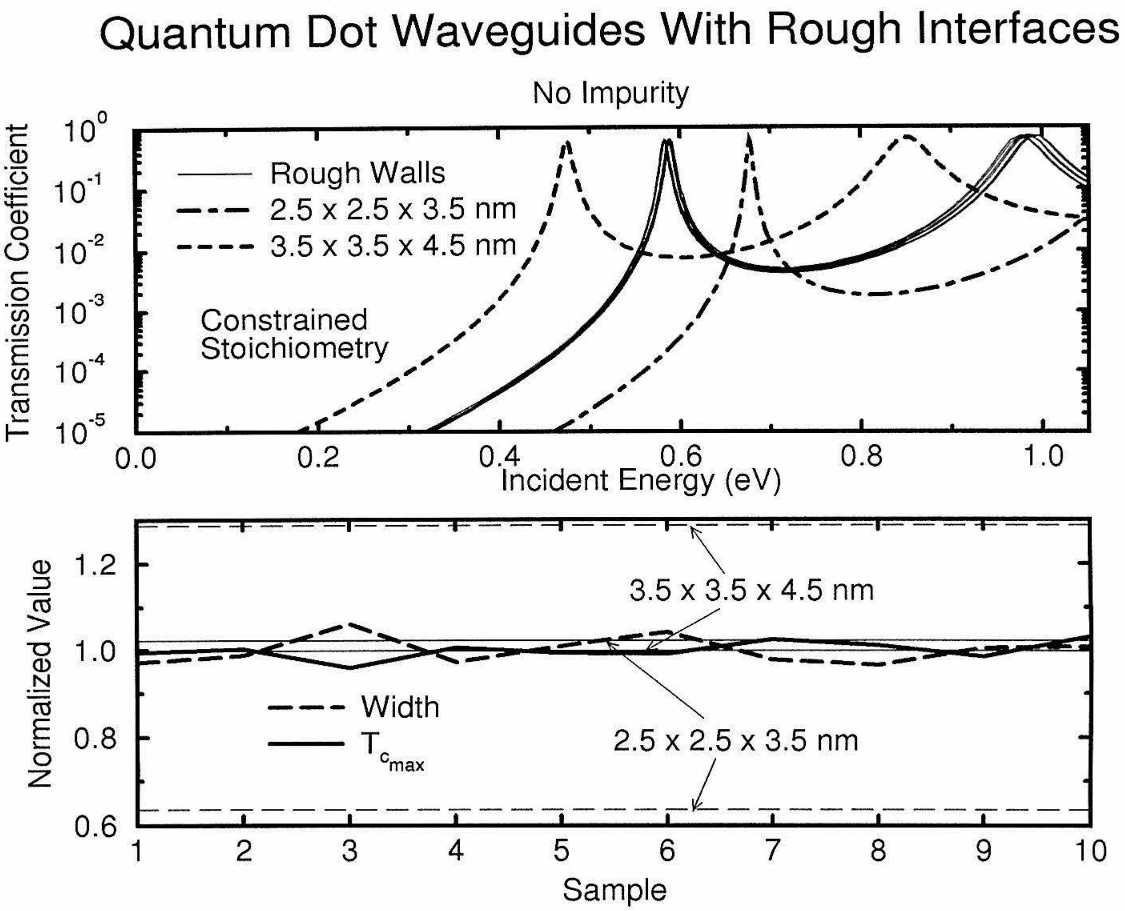

Figure 6.5: Top panel: transmission coefficient curves for quantum dots with ten different stoichiometrically constrained rough-walled configurations. The two reference curves from Figure 6.3 are also shown. Bottom panel: fluctuations in the resonance width and maximum transmission for quantum dots with ten different stoichiometrically constrained rough-walled configurations. Fluctuating values are normalized so that their mean is 1. Values for the two reference structures from Figure 6.3 are also plotted. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

coefficients are again plotted, as in Figure 6.4, along with those of the references structures. The transmission properties still fluctuate, though not as much as with the unconstrained stoichiometry.

These fluctuations can be understood through an analysis of the electron wave function at the resonance. We first calculate the total electron probability density in the quantum dot structure, including all sites in the supercells containing barrier material. We then calculate the total electron probability density in the $0.5nm$ shell of interface roughness and express this as a percentage of the total in the dot. At the $n = 1$ resonance in a dot with interface roughness, about 27.2% of the total electron probability density lies in the shell containing the roughness. Thus electrons sample the roughness substantially, and variations in the roughness configuration can be expected to have a significant impact. This suggests that, if the resonance mode could be altered so as to draw the resonant wave function away from the roughness, fluctuations might be reduced. How might this be accomplished? Figure 6.2 suggests an answer: an attractive impurity could lower the transmission resonance, drawing the wave function in toward the impurity site. In the next section we analyze impurities in a dot with rough walls in order to determine what impurity strength should be used and where the impurity should be located to achieve this.

6.2.3 Neutral Impurities

We begin with an analysis of impurity strength. We have calculated a series of transmission coefficient curves for a dot with interface roughness and an impurity in the center. The rough-walled dot is that of sample 1 in the previous section, and the impurity strength is varied from $\Delta U/t = -4.9$ (strongly attractive) to 2.2 (repulsive). The position, width and maximum transmission coefficient of the $n = 1$ resonance are plotted in Figure 6.6. We note that repulsive impurities have little effect on the transmission characteristics of the dot, and attractive impurities

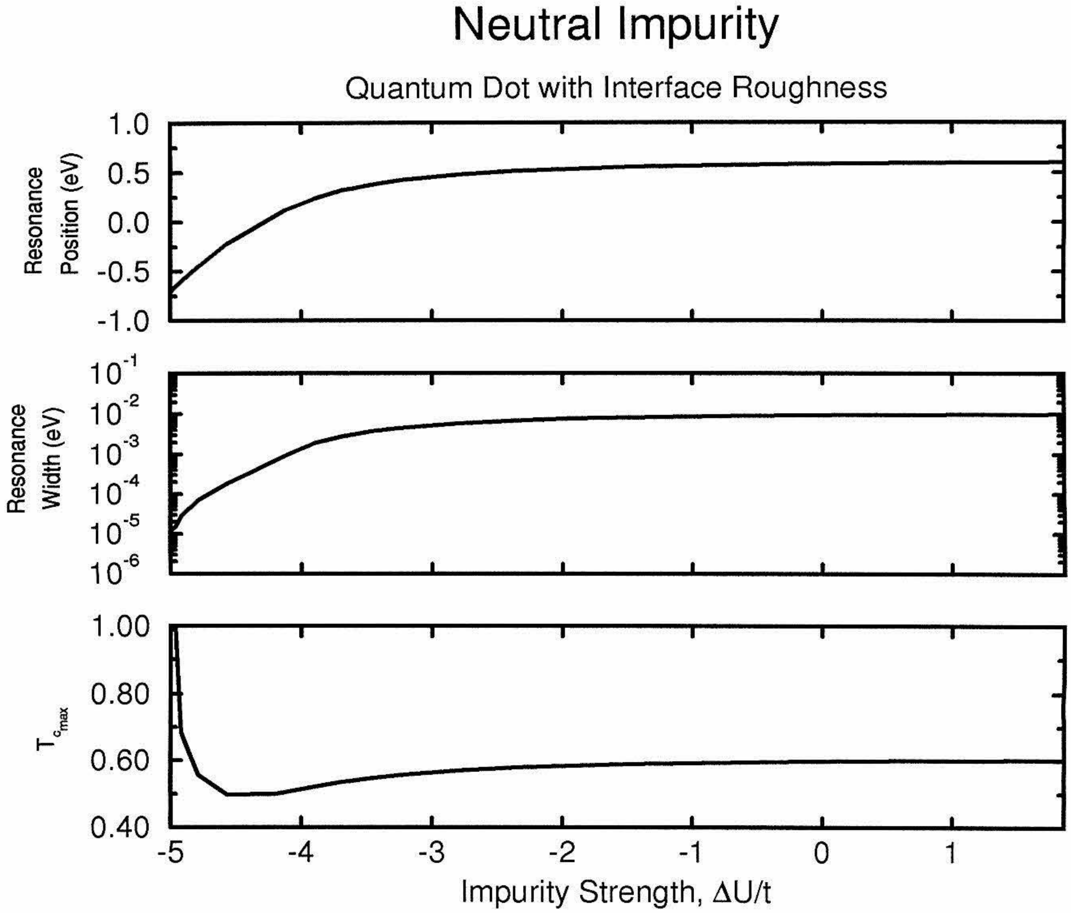

Figure 6.6: Characteristics of the $n = 1$ transmission resonance as a function of impurity strength, $\Delta U/t$, for a rough-walled dot with a neutral impurity in the center. The rough-walled dot is the same as that in Figure 6.3. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

have little effect above $\Delta U/t \approx -4$. In fact, we can divide the plots into two regimes, one where the $n = 1$ resonance has more of the character of the cavity mode of the dot (above -4), and one where it has more of the character of an impurity resonance (below -4). This division makes sense on two counts: first, a bound state for an attractive impurity does not exist above $\Delta U/t \approx -4$[19], so the impurity has a weaker influence in this regime; second, when we analyze the resonant wave function, it is similar to that for a dot without an impurity above $\Delta U/t \approx -4$ and similar to that for the quasi-bound state of an attractive impurity in a bulk region[20] below -4. The crossover between these two regimes is particularly striking in the bottom panel of Figure 6.6. Here the transmission maximum first decreases as the impurity attractive strength increases, owing to degradation of the cavity mode and then increases owing to the increasing strength of the impurity resonance. In fact, transmission reaches a minimum for an impurity strength around -4.5. In the top two panels, we see that the resonance moves toward lower energy and sharpens as the impurity attractive strength is increased below $\Delta U/t = -4$ on account of the increasing localization of the impurity bound state. Thus we see that choosing $\Delta U/t < -4$ should have the greatest effect in terms of reducing fluctuations due to interface roughness in the cavity.

We next examine impurity location. We analyze the two impurity strength regimes separately, as they give rise to different relationships between impurity location and resonance character. Weakly attractive impurities $\Delta U/t > -4$ can be analyzed as perturbing the cavity modes, whereas strongly attractive impurities $\Delta U/t < -4$ in a dot behave more like impurities in a single barrier structure[20].

In Figure 6.2 a weakly attractive impurity ($\Delta U/t = -3.1$) lowers the $n = 1$ transmission resonance by providing a slightly lower effective cavity band–edge. The $n = 2$ resonance, however, is changed little. This can be understood in terms of perturbation of the cavity modes by the impurity. The $n = 1$ cavity mode has an antinode in the center, at the location of the impurity. Thus this mode samples the impurity more than the $n = 2$ mode, which has a node in the center (see Figure 6.7).

Thus an impurity in the center affects the $n = 1$ mode more than the $n = 2$ mode. This type of analysis can be used to explain the data in Figure 6.8. Here we plot the $n = 1$ transmission resonance position, width and maximum for different values of the impurity location along the $z-$direction, keeping $x = y = 0$. The $n = 1$ resonance is lowered more when the impurity is in the center of the dot than when it is near the ends, as the $n = 1$ cavity mode is stronger in the center. The resonance narrows, and the maximum transmission increases as the impurity perturbs the $n = 1$ mode more toward the center, increasing symmetry and isolation from the electrodes. Likewise, in Figure 6.9, where we plot transmission coefficient curves for impurities at different $y-$locations and $x = z = 0$, the $n = 1$ transmission resonance is most strongly affected when $y = 0$, where the $n = 1$ cavity mode maximum occurs. Thus a weakly attractive impurity has the greatest effect on the $n = 1$ resonance of a dot when placed in the center.

A strongly attractive impurity in a dot, on the other hand, gives rise to an $n = 1$ resonance mode typical of an impurity localized state and can be analyzed as an isolated impurity in a single barrier structure[20]. In Figure 6.10, the $n = 1$ transmission resonance position, width and maximum are plotted for a strongly attractive impurity ($\Delta U/t = -4.9$) at different locations along the $z-$direction at $x = y = 0$. Here, the $n = 1$ resonance is lowered less for an impurity in the center than for an impurity near the ends of the dot. This behavior, opposite to that for the weakly attractive impurity in Figure 6.8, is a reflection of the increased confinement of the $n = 1$ (impurity-like) level when the impurity is in the center of the dot. Due to the increasing symmetry and isolation from the electrodes, the resonance is narrowest, and the transmission maximum is greatest for an impurity in the center of the dot. The lateral location dependence of the $n = 1$ resonance position for a strongly attractive impurity, shown in Figure 6.11, is the opposite of that for weakly attractive impurities—as the impurity is removed from the center, the confinement of the impurity level decreases, slightly lowering the $n = 1$ resonance.

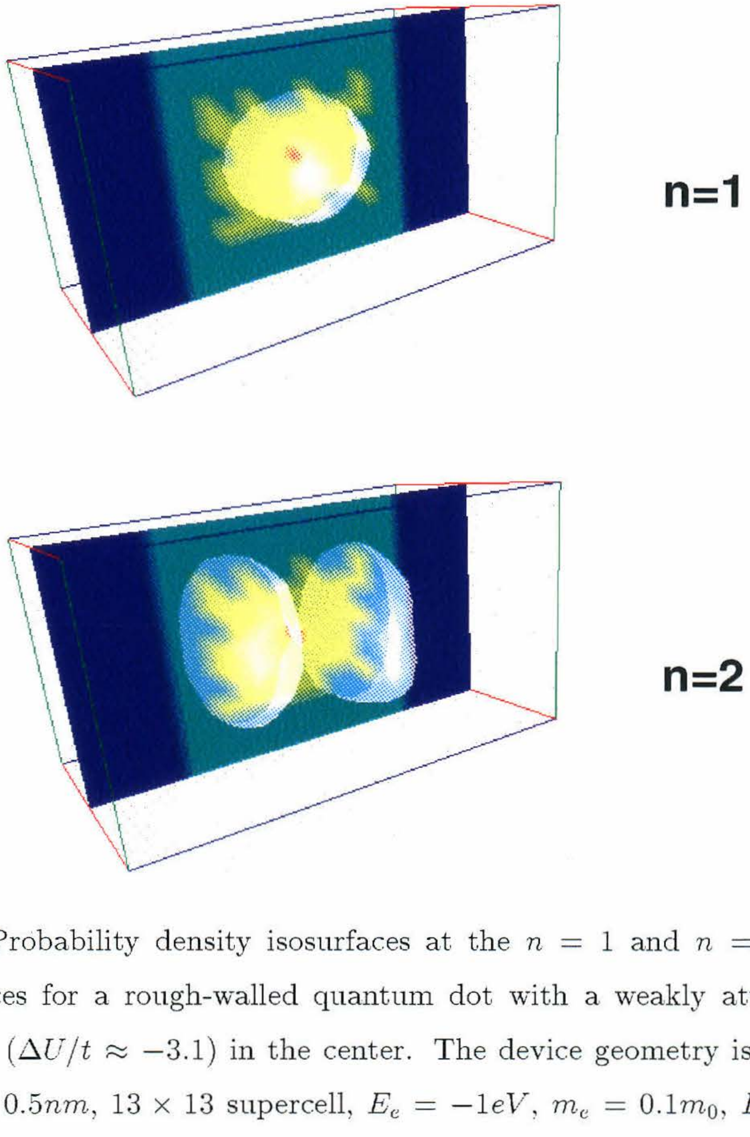

Figure 6.7: Probability density isosurfaces at the $n = 1$ and $n = 2$ transmission resonances for a rough-walled quantum dot with a weakly attractive neutral impurity ($\Delta U/t \approx -3.1$) in the center. The device geometry is that of Figure 6.6. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

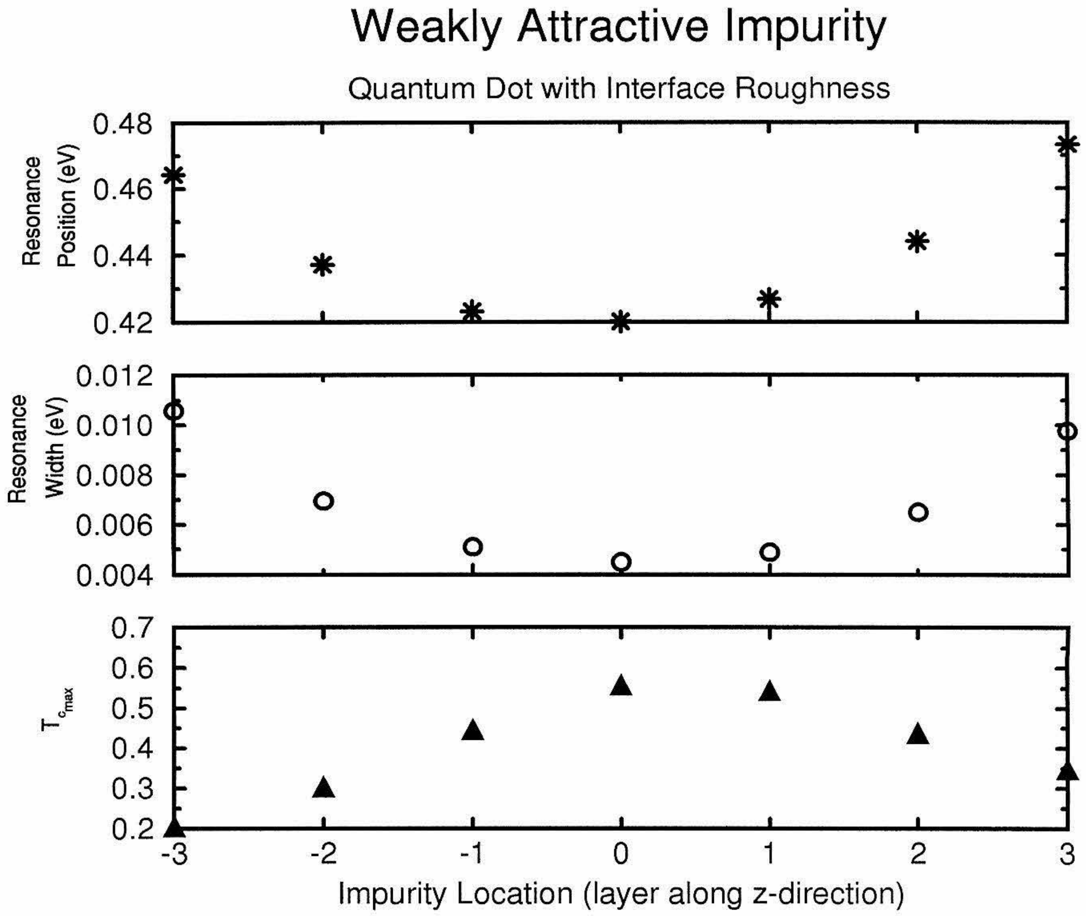

Figure 6.8: Characteristics of the $n = 1$ transmission resonance for a rough-walled dot with a weakly attractive impurity ($\Delta U/t \approx -3.1$) in different $z-$locations at $x = y = 0$. The rough-walled dot is the same as that in Figure 6.3. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

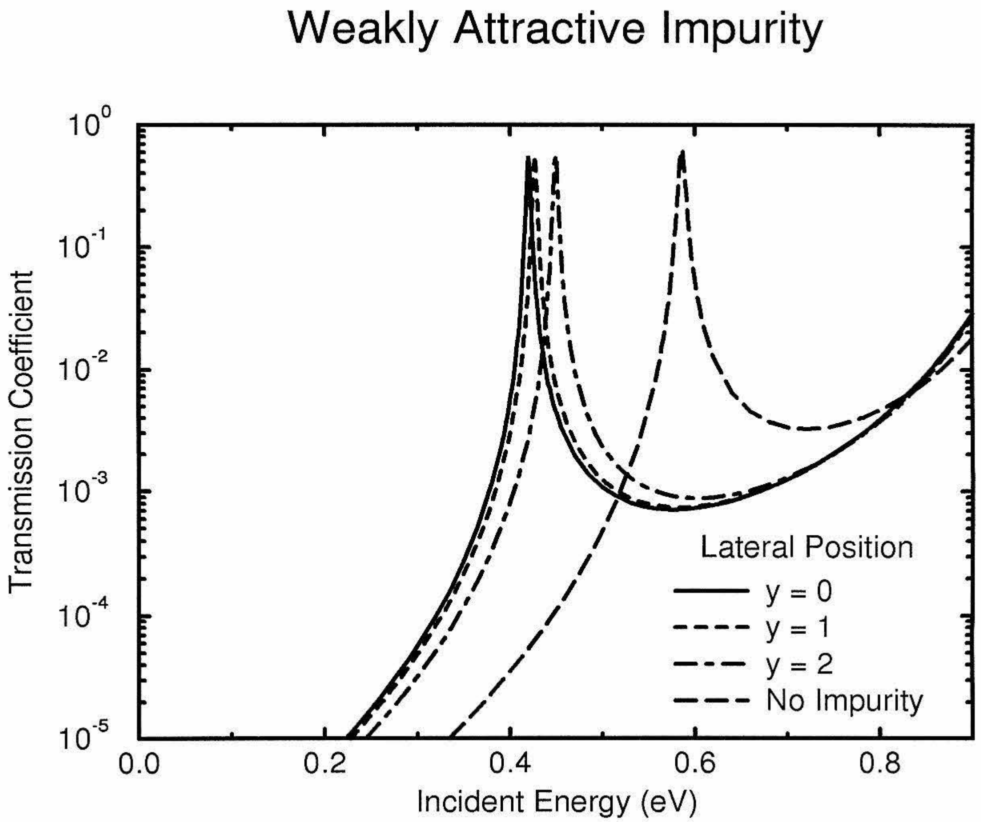

Figure 6.9: Transmission coefficient curves for a rough-walled dot with a weakly attractive impurity ($\Delta U/t \approx -3.1$) in different lateral locations at $x = z = 0$. The rough-walled dot is the same as that in Figure 6.3. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

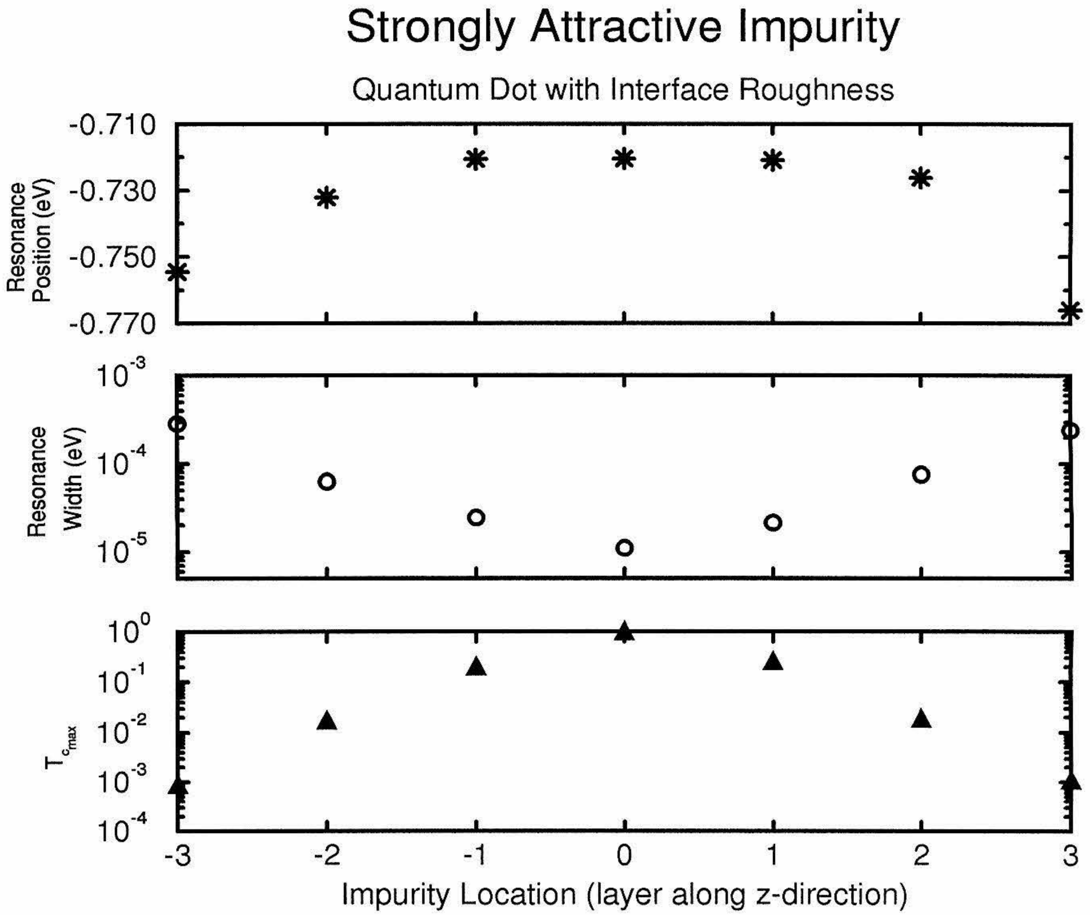

Figure 6.10: Characteristics of the $n = 1$ transmission resonance for a rough-walled dot with a strongly attractive impurity ($\Delta U/t \approx -4.9$) in different $z-$locations at $x = y = 0$. The rough-walled dot is the same as that in Figure 6.3. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

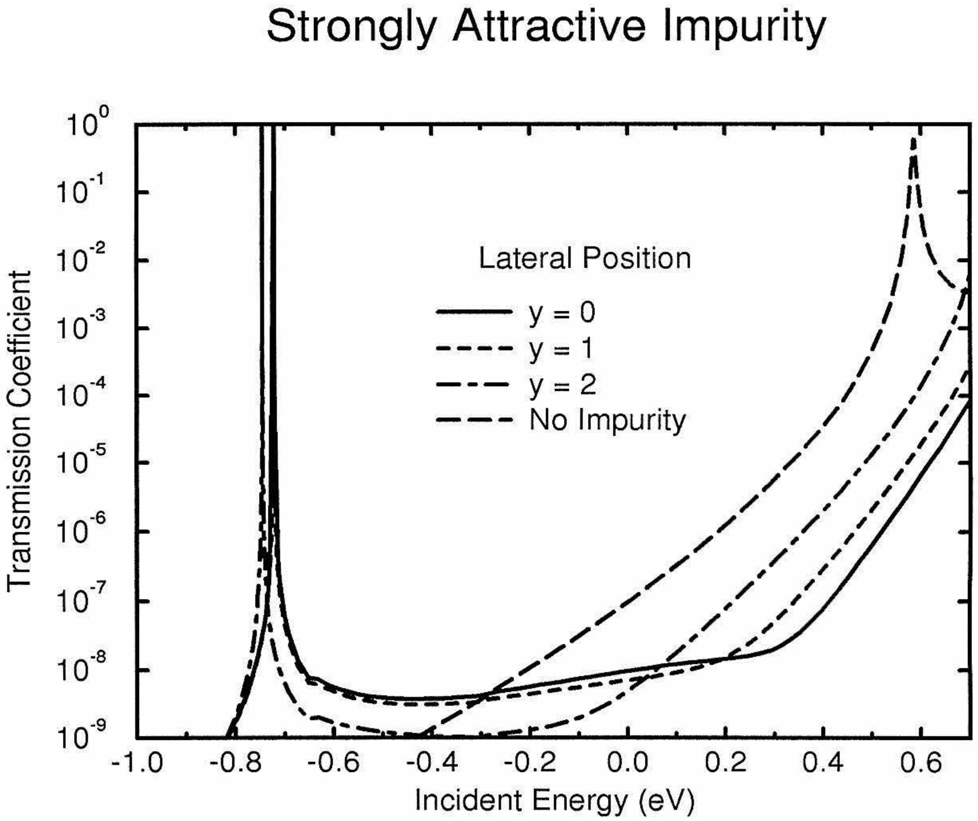

Figure 6.11: Transmission coefficient curves for a rough-walled dot with a strongly attractive impurity ($\Delta U/t \approx -4.9$) in different lateral locations at $x = z = 0$. The rough-walled dot is the same as that in Figure 6.3. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

An important observation in the case of strongly attractive impurities is that the $n = 1$ resonance position is nearly constant as long as the impurity is within a lattice constant or two of the center of the dot. The variation of resonance position over this range is less than that in the fluctuations of Figure 6.4. This suggests there may be some hope of reducing fluctuations in resonance position due to interface roughness if a strongly attractive impurity can be placed near the center of a quantum dot. Indeed, only 1.4% of the electron probability density associated with the $n = 1$ mode of a dot with an impurity at $x = y = z = 0$ with $\Delta U/t = -4.9$ lies in the shell of interface roughness. Thus the $n = 1$ mode of a dot with an impurity should sample the interface roughness less than without the impurity, leading to less fluctuation.

To analyze fluctuations in a dot with an impurity, we plot, in Figure 6.12, transmission coefficient curves for the same set of ten dots as in Figure 6.4, but with an impurity of strength $\Delta U/t = -4.9$ at $x = y = z = 0$. A glance at the figure reveals that the $n = 1$ resonance fluctuates over a much narrower energy range, as predicted. Here the standard deviation of the $n = 1$ resonance position for the ten samples is $0.0007eV$ compared with $0.008eV$ without the impurity. The resonance width and maximum transmission also fluctuate less, as shown in Figure 6.12. Here the widths and maximum transmission coefficients of the $n = 1$ resonances of the ten samples are plotted, normalized so that the average value is 1. Also shown for reference are the widths and maximum transmission coefficients of the $n = 1$ resonances of two ideal dots with an impurity of strength $\Delta U/t = -4.9$ at the center. Relative to the separation between ideal values, the width fluctuates less, and the maximum transmission coefficients are within 0.02 of unity. In the presence of the impurity, the standard deviation of the widths for the ten samples is 6.3%, and that for the maximum transmission coefficients is 0.5%; without the impurity, the standard deviation is 6.8% for the widths and 2.2% for the maximum transmission coefficients. (Standard deviations are given as a percentage of the average value for the ten samples.)

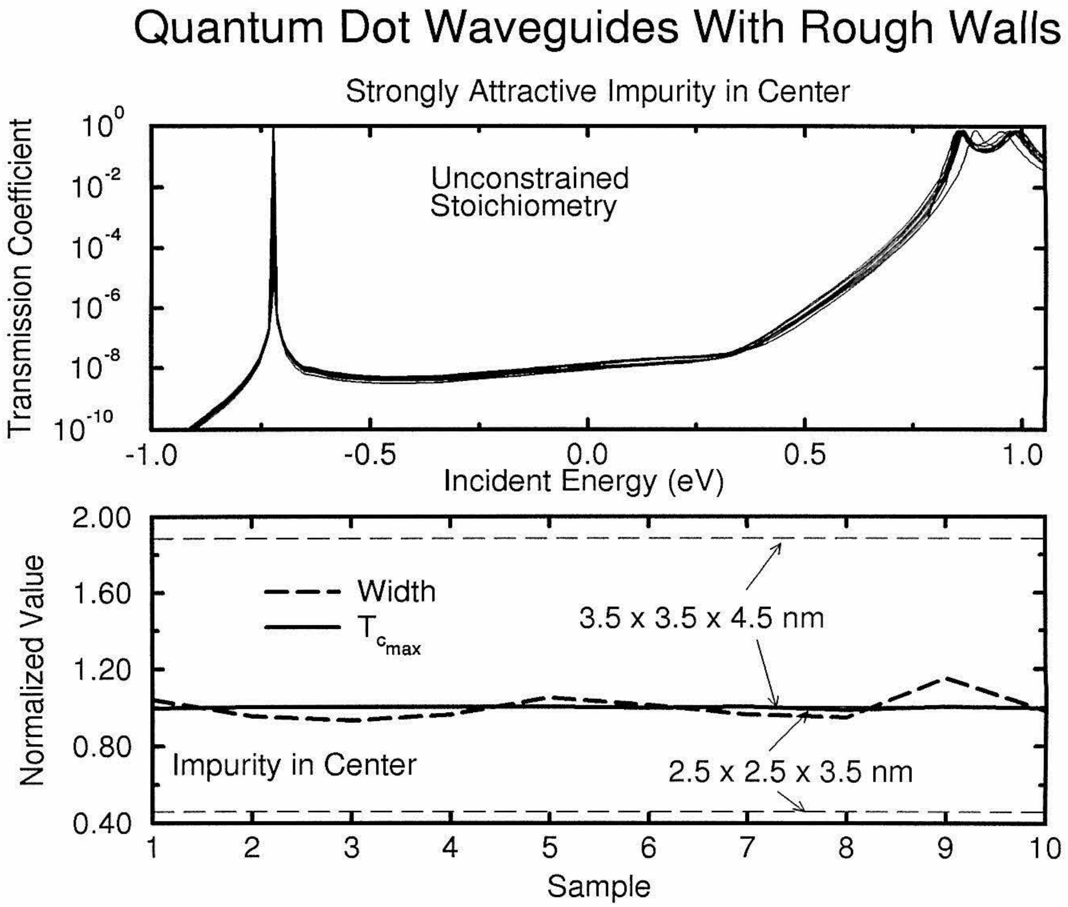

Figure 6.12: Top panel: transmission coefficient curves for quantum dots with ten different rough-walled configurations, each with a strongly attractive ($\Delta U/t \approx -4.9$) impurity in the center. Bottom panel: fluctuations in the resonance with and maximum transmission for quantum dots with ten different rough-walled configurations, each with a strongly attractive ($\Delta U/t \approx -4.9$) impurity in the center. Fluctuating values are normalized so that their mean is 1. Values for the resonance width and maximum transmission coefficient of $3.5nm \times 3.5nm \times 4.5nm$ and $2.5nm \times 2.5nm \times 3.5nm$ smooth-walled dots with an impurity in the center are also shown for reference. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

Although a strongly attractive impurity has maximal effect, even a moderately attractive impurity can reduce fluctuations due to interface roughness. Figure 6.13 shows the results for an impurity with $\Delta U/t = -3.97$, where about 9.1% of the probability density at the $n = 1$ resonance lies in the shell. The standard deviation for the widths is 6.9% and that for the maximum transmission coefficients is 2.3%.

Thus an attractive impurity near the center of a quantum dot can reduce fluctuations due to variations in interface roughness. In a set of quantum dots with a single impurity very close to the center, the transmission characteristics are more uniform than without an impurity. If the impurity location is not controlled precisely, however, or if multiple impurities are present, fluctuations will still pose a problem.

In fact, different impurity configurations at the same concentration can lead to completely different transmission spectra. To demonstrate this, we plot, in Figure 6.14, transmission coefficient curves for the rough-walled dot of sample 1 in Figure 6.4 with two different configurations of impurities in the cavity. Each configuration consists of 11 impurity sites placed at random among the 175 sites in the quantum dot. Also plotted in the figure is the transmission coefficient curve for the rough-walled dot without impurities. We see that the high concentration of impurities produces a complex resonance structure, whose peak positions, widths and maxima depend on the configuration. (The apparent 0's in the transmission are evidently supercell artifacts.)

6.3 Summary

We have examined the effects of atomic scale imperfections on the transmission properties of a quantum dot electron waveguide. We have seen that sample to sample variations in interface roughness in a waveguide can lead to fluctuations in the $n = 1$ transmission resonance position, width and maximum. We have also studied the effects of neutral impurities in quantum dots as a function of

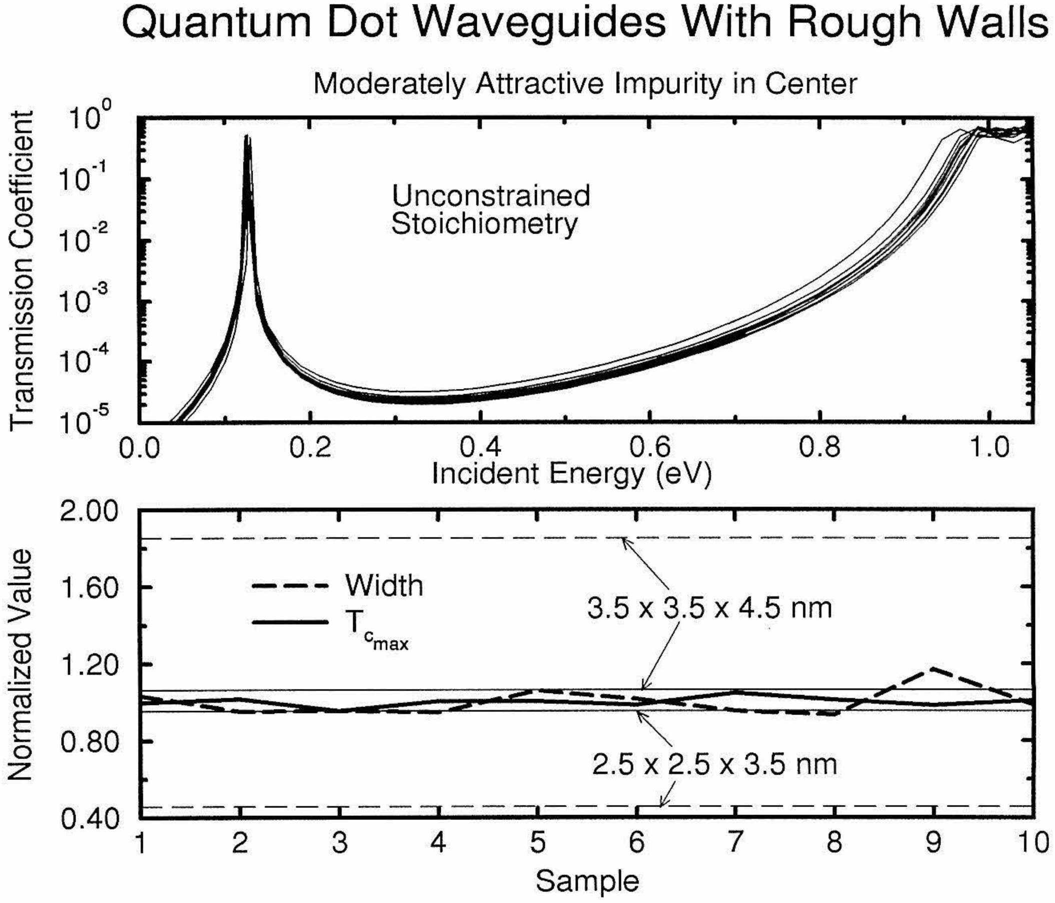

Figure 6.13: Top panel: transmission coefficient curves for quantum dots with ten different rough-walled configurations, each with a strongly attractive ($\Delta U/t \approx -4.9$) impurity in the center. Bottom panel: fluctuations in the resonance width and maximum transmission for quantum dots with ten different rough-walled configurations, each with a moderately attractive ($\Delta U/t \approx -3.97$) impurity in the center. Fluctuating values are normalized so that their mean is 1. Values for the resonance width and maximum transmission coefficient of $3.5nm \times 3.5nm \times 4.5nm$ and $2.5nm \times 2.5nm \times 3.5nm$ smooth-walled dots with an impurity in the center are also shown for reference. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

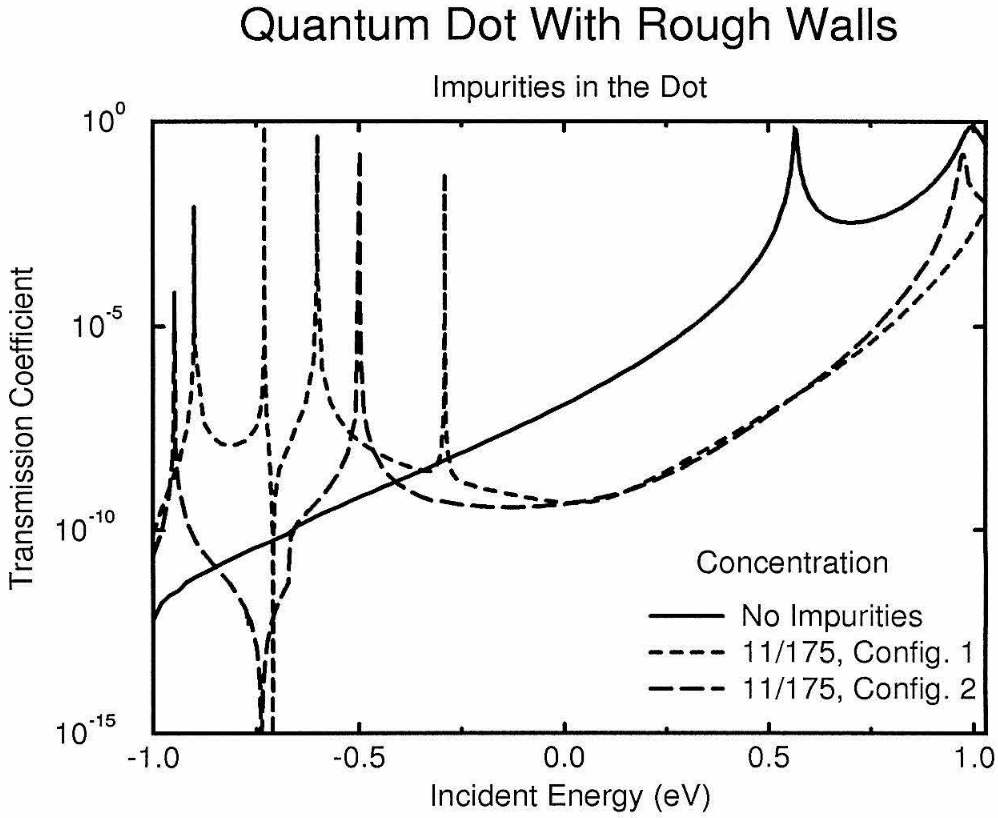

Figure 6.14: Transmission coefficient curves for a rough-walled quantum dot with a concentration of $0.063/a^3$ strongly attractive ($\Delta U/t \approx -4.9$) impurities in the cavity (11 impurity sites were chosen at random out of the 175 sites within the cavity). Also shown is the transmission coefficient curve for the rough-walled dot of Figure 6.3. $a = 0.5nm$, 13×13 supercell, $E_e = -1eV$, $m_e = 0.1m_0$, $E_b = 1.05eV$, $m_b = 0.1248m_0$, $E_w = 0eV$, $m_w = 0.0673m_0$. Plane waves are incident along the $z-$direction.

impurity strength and location and seen that an attractive impurity near the center of the dot draws the wave function at the $n = 1$ resonance in away from the interface roughness, reducing fluctuations. Nonetheless, the presence of more than a single impurity in the dot can lead to complex, impurity configuration dependent resonance structure, especially at high concentrations. Fluctuations thus pose a problem, both due to interface roughness and due to impurities. If quantum structures such as the quantum dot electron waveguide are to be used in devices produced on a large scale, the issue of fluctuations must be tackled. On the basis of the concept demonstration of fluctuation reduction by an isolated impurity near the center of a dot, there is hope that problems with atomic scale variation could be overcome. Research in this area should continue to prove challenging and rewarding.

Bibliography

[1] W. Chu, C. C. Eugster, and A. Moel, J. Vac. Sci. Technol. B **10**, 2966 (1992).

[2] K. Yoh, K. Kyomi, and A. Nishida, Jpn. J. Appl. Phys. Pt. 1 **31**, 4515 (1992).

[3] Y. Qian, J. M. Zhang, and J. Y. Xu, Superlatt. Microstruct. **13**, 241 (1993).

[4] K. Inoue, K. Kimura, and K. Maehashi, J. Cryst. Growth **127**, 1041 (1993).

[5] X. Q. Shen, M. Tanaka, and T. Nishinaga, J. Cryst. Growth **127**, 932 (1993).

[6] S. Tsukamoto, Y. Nagamune, and M. Nishioka, Appl. Phys. Lett. **62**, 49 (1993).

[7] H. Iwano, S. Zaima, Y. Koide, and Y. Yasuda, J. Vac. Sci. Technol. B **11**, 61 (1993).

[8] K. Kash, B. P. Van der Gaag, D. D. Mahoney, A. S. Gozdz, L. T. Florez, J. P. Harbison, and M. D. Sturge, Phys. Rev. Lett. **67**, 1326 (1991); *Nanostructure Physics and Fabrication*, edited by M. A. Reed and W. P. Kirk (Academic, San Diego, 1989).

[9] T. Itoh, S. Nobuyuki, and A. Yoshii, Phys. Rev. B **45**, 14,131 (1992).

[10] V. V. Mitin, Superlatt. Microstruct. **8**, 413 (1990).

[11] P. L. Mceuen, B. W. Alphenaar, and R. G. Wheeler, Surf. Sci. **229**, 312, (1990).

[12] C. C. Eugster, J. A. del Alamo, M. R. Melloch, and M. J. Rooks, Phys. Rev. B **46**, 10146 (1992).

[13] E. Tekman and S. Ciraci, Phys. Rev. B **42**, 9098 (1990).

[14] P. F. Bagwell, Phys. Rev. B **41**, 10354 (1990); A. Kumar and P. F. Bagwell, Phys. Rev. B **43**, 9012 (1991); P. F. Bagwell, T. P. Orlando, and A. Kumar, in *Resonant Tunneling in Semiconductors*, edited by L. L. Chang et al. (Plenum Press, New York, 1991) p. 417.

[15] C. S. Chu and R. S. Sorbello, Phys. Rev. B **40**, 5941 (1989).

[16] Y. Takagaki and D. K. Ferry, Phys. Rev. B **45**, 6715 (1992).

[17] D. van der Marel and E. G. Haanappel, Phys. Rev. B **39**, 7811 (1989); E. G. Haanappel and D. van der Marel, Phys. Rev. B **39**, 5484 (1989).

[18] J. A. Nixon, J. H. Davies, and H. U. Baranger, Phys. Rev. B **43**, 12638 (1991).

[19] E. N. Economou, *Green's Functions in Quantum Physics* (Springer-Verlag, New York, 1967).

[20] S. K. Kirby, D. Z.-Y. Ting, and T. C. McGill, to be published.