MODELING QUANTUM TRANSPORT IN NANOSCALE TRANSISTORS

A Thesis
Submitted to the Faculty of Purdue University
by Ramesh Venugopal

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

August 2003
To my beloved parents for their profound faith in education.
ACKNOWLEDGMENTS

I am deeply indebted to my thesis advisor Prof. Mark Lundstrom, for his help and advice through the course of my PhD study at Purdue University. Prof. Lundstrom is an excellent thesis advisor. Apart from his academic contributions, without which this thesis would not have been made possible, he spent a lot of his time training me to be a professional. He is a remarkable individual and I have truly benefited from the experience of working with him. I am also deeply indebted to Prof. Supriyo Datta for his valuable suggestions and guidance during the course of my research. My thanks to Dr. Ephraim Fischbach and Dr. David Janes for serving on my thesis committee.

Several individuals have contributed to this thesis at Purdue University. I am especially grateful to my colleague Dr. Zhibin Ren, who worked very closely with me for two years. I have had many fruitful discussions and learnt a lot from him during that time. I would also like to express my gratitude to Dr. Magnus Paulsson, Dr. Prashant Damle, Dr. Avik Ghosh, Junghoon Rhew, Anisur Rahman, Sayed Hasan, Jing Wang and Jing Guo for useful comments and discussions. Dr. Sebastien Goasguen and Dr. Uday Savagaonkar provided useful tips with computational issues for which I am grateful. I would also like to acknowledge Dr. Dejan Jovanovic for mentoring me during the course of my two summer internships at Motorola. Finally, I owe a lot to my brother Dr. Rajesh Venugopal who happened to be a fellow PhD student at Purdue University. It was his motivation and support that made me pursue graduate studies after a brief stint in the industry. There are many others at Purdue who have influenced me over the years and I would like to thank them all for their contributions, however indirect, to this thesis.

This work was supported by the Semiconductor Research Corporation, and through an Intel PhD Fellowship award.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xvi</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Overview</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Device Scaling</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Modeling non-equilibrium transport</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1 The Self-consistent simulation scheme</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Outline of the Thesis</td>
<td>8</td>
</tr>
<tr>
<td>2 Simulating Quantum Transport in Nanoscale MOSFETs: Ballistic Hole</td>
<td>13</td>
</tr>
<tr>
<td>Transport, Subband Engineering and Boundary Conditions</td>
<td></td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Theory</td>
<td>14</td>
</tr>
<tr>
<td>2.3 Results</td>
<td>20</td>
</tr>
<tr>
<td>2.4 Discussion</td>
<td>28</td>
</tr>
<tr>
<td>2.5 Summary</td>
<td>32</td>
</tr>
<tr>
<td>3 Simulating Quantum Transport in Nanoscale Transistors: Real versus</td>
<td>35</td>
</tr>
<tr>
<td>Mode-Space Approaches</td>
<td></td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>35</td>
</tr>
<tr>
<td>3.2 Theory</td>
<td>37</td>
</tr>
<tr>
<td>3.2.1 Real-Space solution</td>
<td>37</td>
</tr>
<tr>
<td>3.2.2 Mode-space solution</td>
<td>42</td>
</tr>
<tr>
<td>3.3 Results</td>
<td>46</td>
</tr>
<tr>
<td>3.4 Discussion</td>
<td>53</td>
</tr>
<tr>
<td>3.5 Summary</td>
<td>57</td>
</tr>
<tr>
<td>4 A Quantum Mechanical Treatment of Scattering in Nanoscale Transistors</td>
<td>59</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>4.2</td>
<td>A detailed scattering model</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Theory</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Results</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Discussion</td>
</tr>
<tr>
<td>4.3</td>
<td>A simple treatment of scattering using Büttiker probes</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Theory</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Results</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Discussion</td>
</tr>
<tr>
<td>4.4</td>
<td>Summary</td>
</tr>
<tr>
<td>5</td>
<td>A Quantum Mechanical Analysis of Channel Access Geometry and Series Resistance in nanoscale transistors</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>5.2</td>
<td>The modeling scheme</td>
</tr>
<tr>
<td>5.3</td>
<td>Results</td>
</tr>
<tr>
<td>5.3.1</td>
<td>The WN Device</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Effect of differing channel access geometry</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussion</td>
</tr>
<tr>
<td>5.5</td>
<td>Summary</td>
</tr>
<tr>
<td>6</td>
<td>Future Work</td>
</tr>
<tr>
<td></td>
<td>LIST OF REFERENCES</td>
</tr>
<tr>
<td></td>
<td>APPENDIX</td>
</tr>
<tr>
<td></td>
<td>Appendix A: The Self-energy calculation for the leads</td>
</tr>
<tr>
<td></td>
<td>Appendix B: From wavefunctions to Green’s functions</td>
</tr>
<tr>
<td></td>
<td>Appendix C: The relation between the probe self-energy and the low field mobility</td>
</tr>
<tr>
<td></td>
<td>VITA</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Self-consistent solution scheme</td>
<td>9</td>
</tr>
<tr>
<td>2.1</td>
<td>An ultra-thin body DG MOSFET structure with S/D doping of $10^{20}$ cm$^{-3}$ and an intrinsic channel</td>
<td>15</td>
</tr>
<tr>
<td>2.2</td>
<td>A pictorial representation of the Boltzmann solution for a generic hole subband. Filled circles represent states that are filled by the source Fermi level ($\mu_S$) and empty circles represent states filled by the drain Fermi level ($\mu_D$).</td>
<td>17</td>
</tr>
<tr>
<td>2.3</td>
<td>(a) The conduction band and the profile of the first subband in the 1.5 nm body n-MOSFET (“he” represents a heavy electron in the confinement direction). (b) The valence band and the profile of the first subband in the 1.5 nm body p-MOSFET (“hh” represents a heavy hole in the confinement direction). All energies are referred to the source Fermi level (0). Due to the high doping in the source/drain regions the electron subband is negative and the hole subband positive wrt. the reference. Note that the confinement energy of the hole is much higher than the electron. The bias state is $</td>
<td>V_{GS}</td>
</tr>
<tr>
<td>2.4</td>
<td>The $I_{DS}$ vs. $V_{GS}$ (log scale) and the $I_{DS}$ vs. $V_{DS}$ (linear scale) characteristics of the 1.5 nm body n (dotted line) and p (solid line) MOSFETs for a channel orientation of $\langle 100 \rangle$. Both devices exhibit near ideal subthreshold characteristics. The on-current of the p-MOSFET is $\sim 40%$ lower than that of the n-MOSFET.</td>
<td>23</td>
</tr>
<tr>
<td>2.5</td>
<td>2D charge density and injection velocity, extracted at the charge control point vs. $V_{GS}$ for the 1.5 nm body, n (dotted line) and p (solid line) MOSFETs. At the same gate voltage, the hole density is $\sim 10%$ greater than the electron density, but the hole injection velocity along $\langle 100 \rangle$ is $\sim 50%$ lower than the electron injection velocity due to the warped nature of the heavy hole subband.</td>
<td>24</td>
</tr>
<tr>
<td>2.6</td>
<td>2D charge density and injection velocity, extracted at the charge control point vs. $V_{DS}$ for the 1.5 nm body, n (dotted line) and p (solid line) MOSFETs. DIBL increases the charge density by $\sim 15%$. Injection velocities are higher than their equilibrium values due to degeneracy (Section 2.4).</td>
<td>25</td>
</tr>
</tbody>
</table>
2.7 The $I_{DS}$ vs. $V_{GS}$ (log scale) and the $I_{DS}$ vs. $V_{DS}$ (linear scale) characteristics of the 5 nm body n (dotted line) and p (solid line) MOSFETs for a channel orientation of $\langle 100 \rangle$. Both devices exhibit near ideal subthreshold characteristics. The on-current of the p-MOSFET is $\sim 25\%$ lower than that of the n-MOSFET.

2.8 2D charge density and injection velocity, extracted at the charge control point vs. $V_{GS}$ for the 5 nm body, n (dotted line) and p (solid line) MOSFETs. The crossover in the 2D charge density is due to an increased inversion layer capacitance for the electrons as a result higher subband occupancy. Note that higher subband occupancy considerably degrades the injection velocity for electrons, whereas hole velocity is relatively unaffected when compared against Fig. 2.5.

2.9 (a) Flux distribution in $\vec{k}$-space for the heavy hole band in the 5 nm body p-MOSFET along the $\langle 100 \rangle$ (left) and $\langle 110 \rangle$ (right) directions. (b) Flux distribution in $\vec{k}$-space for the light hole band in the 5 nm body p-MOSFET along the $\langle 100 \rangle$ (left) and $\langle 110 \rangle$ (right) directions. Note that in the on-state only one half of $\vec{k}$-space is occupied as drain injection is suppressed. Also, note that the $k$-axes are rotated by $45^\circ$ between the left and right pictures.

2.10 (a) The average valence band (averaged along $z$) profile within the $p^{++}$-$p^+$$n^{-}$-$p^+$-$p^{++}$, 1.5 nm body transistor in equilibrium ($V_{GS} = -0.6$ V) and in the on-state is plotted on the left. On the right we compare the average valence band obtained using floating boundary conditions (solid line) to that obtained using fixed boundary conditions (dotted line) in the on-state. (b) The charge density profiles obtained by simulating the $p^+$$n^{-}$-$p^+$ device using floating boundary conditions (solid line) are compared to those obtained using fixed boundary conditions (dotted line) in the on-state. Note that irrespective of the boundary condition used, macroscopic charge neutrality is always achieved.

3.1 (a) An ultra-thin body DG MOSFET structure with S/D doping of $10^{20}$ cm$^{-3}$ and an intrinsic channel (channel thickness = 1.5 nm). (b) The squeezed DG MOSFET structure used to investigate the effect of mode coupling (channel thickness = 0.75 nm).

3.2 (a) The solution to a 1D effective mass equation (step 1 of the mode-space solution) along slice $Z-Z$ (refer Fig. 3.1a) in the on-state ($V_{GS} = V_{DS} = 0.6$ V). (b) The local density of states from the real-space solution along slice $Z-Z$. 
3.3 The classical (inset) and quantum 2D density of states (DOS) along slice $Z - Z$ (refer Fig. 3.1) in the on-state. The position of the slice along the channel determines the subband energies (dotted lines).

3.4 Computed local density of states (LDOS) and the longitudinal subband profile (dotted line) along slice $X - X$ in the on-state. Broadening in the LDOS is due to quantum mechanical coupling to the S/D and oscillations are due to quantum mechanical reflections. The conduction band (solid line) is far below the subband due to confinement.

3.5 $I_{DS}$ vs. $V_{GS}$ and $I_{DS}$ vs. $V_{DS}$ characteristics for the model device (refer Fig. 3.1a) from real (line) and mode-space (circles) solution schemes. Close agreement ($\sim 1\%$) between the two solutions validates the approximations inherent in the decoupled mode solution.

3.6 The energy distribution of the on-current from the real (solid line) and mode-space (circles) solution schemes. The top of the first subband (dotted line) is also indicated to separate the thermionic and tunneling current components.

3.7 $I_{DS}$ vs. $V_{GS}$ and $I_{DS}$ vs. $V_{DS}$ characteristics for the model device (refer Fig. 3.1a) from a quantum (line) and Boltzmann (circles) treatment of the modes. The off-current from the quantum solution is higher than the Boltzmann solution due to source-to-channel tunneling. The on-current is lower as a result of self-consistency.

3.8 The 2D electron density and first subband profile, along the channel from a quantum (solid line) and classical (dashed line) treatment of the modes, in the on-state. Charge at the top of the source-barrier is primarily prescribed by gate electrostatics. Crossover between the quantum and classical charge profiles, is due to quantum reflections and tunneling.

3.9 $I_{DS}$ vs. $V_{GS}$ and $I_{DS}$ vs. $V_{DS}$ characteristics for the device with the squeezed channel (refer Fig. 3.1b) from mode-space (dashed line) and real-space (solid line) solution schemes. The mode-space solution does not account for mode coupling. Therefore, when the vertical potential profile is perturbed strongly, the decoupled mode-space solution exhibits inaccuracies.
4.1 (a) The longitudinal \( (E_l) \), transverse \( (E_{k,j}) \) and total \( (E) \) energy components are indicated. Note that elastic scattering (dotted arrows) mixes states with the same total energy. Inelastic scattering due to phonons (dash dotted arrows) mixes states with total energies separated by \( \hbar \omega \), where \( \hbar \omega \) is the energy of the phonon. In the ballistic case \( E_l \) and \( E_{k,j} \) can be treated independently and contributions from different transverse energies can be summed analytically at each \( E_l \). This is not the case once scattering is turned on. (b) Inelastic scattering due to a single phonon with energy \( \hbar \omega \), is pictorially represented. Phonon emission and absorption affects the in and out scattering functions \( (\Sigma^\text{in} \text{ and } \Sigma^\text{out}) \) for a state with a total energy \( E \).

4.2 The flowchart of the algorithm used to treat both, elastic and inelastic scattering in a 2D electron gas.

4.3 The self-consistently calculated (ballistic) subband profile in the on-state \( (V_{DD} = 0.4 \text{V}) \) for a double-gate MOSFET with a silicon film thickness of 1.5 nm. The S/D regions are degenerately doped \( (1 \times 10^{20} \text{ cm}^{-3}) \) and the channel length (demarcated by the dashed vertical lines) is 10 nm. The number of elastic scatterers is progressively from the drain to the top of the source to channel barrier as indicated by the arrow in order to generate Fig. 4.7.

4.4 (a) Simulated current spectra at the source and drain ends vs. the total electron energy \( E_l \) for the subband profile shown in Fig. 4.3 from both ballistic (solid lines) and elastic scattering (dashed lines) simulations. (b) Simulated current spectra at the source and drain end vs. the longitudinal electron energy \( E_{l,i} \) for the subband profile shown in Fig. 4.3 from both ballistic (solid lines) and elastic scattering (dashed lines) simulations.

4.5 (a) Simulated current spectra at the source and drain ends vs. the total electron energy \( E \), for the subband profile shown in Fig. 4.3 from both ballistic (solid lines) and inelastic scattering (dashed lines) simulations. Inelastic scattering is assumed to be caused by the emission and absorption of phonons with an energy of 63 meV. (b) Simulated current spectra at the source and drain ends vs. the longitudinal electron energy \( E_{l,i} \) for the subband profile shown in Fig. 4.3 from both ballistic (solid lines) and inelastic scattering (dashed lines) simulations.
4.6 Simulated current spectra at the source and drain ends vs. the longitudinal electron energy $E_l$, in the off-state from both ballistic (solid lines) and elastic scattering (dashed lines) simulations. The self-consistently calculated subband profile in the off-state is shown in the inset. Note that although scattering broadens the local density of states below the source to channel barrier (dotted vertical line), thus resulting in increased source to channel tunneling, the off-current from ballistic simulations is still higher. Ballistic simulations set an upper limit on both the off and the on current values.

4.7 (a) A pictorial representation of the essential physics of scattering (Also refer Fig. 4.3). The fraction of the scattered electrons that can surmount the source-to-channel barrier (Fig. 4.3) and make it back into the source, reduces as we move towards the drain. These carriers, whose total energy is $E$, are delineated by the cone. (b) The behavior of the on-current as a function of the scatterer number, which is progressively increased from the drain to the source (Fig. 4.3) as predicted by detailed, elastic scattering simulations. This plot indicates that scattering events occurring near the source end of the transistor degrade the on-current more severely as compared to scattering events occurring at the drain end, thus validating the essential physics of scattering.

4.8 The profile of a generic mode is illustrated along with the placement of Büttiker probes. Note that the Fermi level of the S/D contacts is fixed, while that of the probe is adjusted to preserve the number of carriers. The probe self-energy is related to the coupling strength.

4.9 (a) The current spectrum from the energy relaxed scattering model is compared against the ballistic limit in the on-state ($V_{GS} = V_{DS} = 0.4\, V$). Note that the drain current spectrum is relaxed in energy. (b) The current spectrum from the phase relaxed scattering model is compared against the ballistic limit, in the on-state. The source and drain spectra are identical in the presence of scattering, because this model relaxes channel directed momentum only.

4.10 (a) The local density of states in the on-state, from the ballistic (left) and energy relaxed scattering models (right). The first mode is also plotted (dotted line). Note that coherent oscillations in the LDOS are washed out when scattering is turned on. (b) The charge density spectrum from the ballistic (left) and energy relaxed scattering models (right), in the on-state. In the ballistic limit, the source and drain injected populations can be clearly identified. When scattering is turned on, these populations are mixed.
4.11 (a) $I_{DS}$ vs. $V_{GS}$ characteristics from the ballistic (solid line) and energy relaxed scattering models (dashed line) for $V_{DS} = 0.4\,V$. The off-current, from the scattering model is lower despite an increase in the tunneling current due to a broadening in the LDOS below the source-to-channel barrier. (b) The off-current vs. channel length from the ballistic (solid line) and energy relaxed scattering models (dashed line). Ballistic simulations are good enough to evaluate leakage and subthreshold characteristics.

4.12 (a) The self-consistent Fermi level of the Büttiker probes, from the energy relaxed scattering model, in the linear response region ($V_{DS} = 10\,mV$). (b) The extracted sheet resistivity in the on-state. Note the four components of the resistance; 1) quantum contact resistance, 2) S/D resistance 3) tip resistance and 4) channel resistance.

4.13 (a) $I_{DS}$ vs. $V_{DS}$ characteristics from the ballistic (solid line) and energy relaxed scattering models (dashed line) for $V_{GS} = 0.4\,V$. The on-current in the presence of scattering is $\sim 50\%$ of the ballistic limit. (b) The on-current vs. channel mobility is plotted to indicate that the ultimate performance of our device is primarily controlled by device parasitics.

4.14 (left) The current as a function of the scatterer number, which is progressively increased from the drain to the source is plotted from both, the energy relaxing (solid line with circles) and the phase relaxing (dashed line) scattering models. This plot indicates that the energy relaxed model captures the essential physics of scattering in transistors. (right) The current spectra vs. the longitudinal electron energy is plotted at the source and drain ends from a detailed treatment of phonon scattering (solid lines) and from the energy relaxing Büttiker probe model (dashed lines). The potential profile is fixed (Fig. 4.3) and the scattering strength is adjusted to obtain the same level of current flow from both models.

4.15 The profile of the first mode and the 2D electron density in the on-state with scattering turned on in the first half of the device (solid lines) and with scattering turned on in the second half of the device (dashed lines). Note that the potential drops in the source or the drain only when scattering is turned on. Also note that turning on scattering only in the second half of the device increases the 2D electron density in the channel.
5.1 (a) An ultra-thin body DG MOSFET structure with a S/D doping of $10^{20} \text{cm}^{-3}$ and an intrinsic channel (channel thickness = 1.5 nm). The S/D regions are 20 nm in length, the gate length is 10 nm and the S/D-channel junctions are abrupt. A slice of the device within which a 1D, $z$ directed effective mass equation is solved, is also indicated. (b) A thin body DG MOSFET with abruptly flared out S/D regions. The length of the flared regions is 10 nm, their thickness is 6.2 nm and the overall device length is 50 nm (same as that for device a). The region within the dashed box is the intrinsic device.

5.2 (a) The conduction and subband profiles along slices A-A (left) and B-B (right) for the WN device (Fig. 5.1b). Unprimed mode energies are illustrated using solid horizontal lines while primed mode energies are represented by dashed horizontal lines. The first unprimed mode within slice C-C is also superposed (right) and marked (triangles). (b) The charge distribution along slices A-A, B-B and C-C is presented. The net charge (dashed and dotted line) is obtained by summing charge contributions from the unprimed (solid line) and primed (dashed line) valleys.

5.3 (a) The energy resolved LDOS spectrum within the WN device (Fig. 5.1b) in the on-state, for the unprimed valleys. Note that all of the states within the narrow region is due to the unprimed valleys. (b) The energy resolved LDOS spectrum within the WN device from the primed valleys. States belonging to the primed valleys are reflected at the WN interface and decay exponentially within the narrow region. The first unprimed subband (dashed line) is also superposed on the plots. This subband is discontinuous at the WN interface due to an abrupt change in the width of the quantum well (from 6.2 nm to 1.5 nm).

5.4 (a) Simulated $I_{DS}$ vs. $V_{GS}$ characteristics for the UN (dashed line) and WN (solid line) devices at $V_{DS}=0.4$ V (includes the effect of scattering). Both devices exhibit the same off-current ($\sim7 \text{A/m}$) and subthreshold slopes. (b) Simulated $I_{DS}$ vs. $V_{DS}$ characteristics for the UN (dashed line) and WN (solid line) devices at $V_{GS}=0.4$ V. For the same off-current, the simulated on-current for the WN device is $\sim10\%$ lower than that of the UN device due to quantum mechanical reflections at the WN interface.
5.5 (a) The averaged conduction band profile (left) in the off-state (includes the effect of scattering) for the UN (dashed line) and WN (solid line) devices. Note the large built-in potential at the WN interface. The corresponding subband profiles are visualized on the right. Many subbands (dots) are occupied in the S/D regions of the WN device. These modes inject current into the single propagating mode within the narrow region due to scattering (modes line up as shown in Fig. 5.2b). The UN device exhibits single mode occupancy everywhere (dashed line). (b) The Büttiker probe potentials within the UN (triangles) and WN (solid line) devices indicates that all of the applied voltage drops within the unconducting channel in the off-state.

5.6 (a) The average conduction band profile (left) in the on-state (includes the effect of scattering) for the UN (dashed line) and WN (solid line) devices. Note that the source-to-channel barrier is depressed due to the large gate voltage and that differences in the on-current between the UN and WN devices reflects the ability of the large S/D regions within the WN device to inject and extract electrons from the narrow region. (b) The 2D charge density within the UN (dashed line) and WN (solid and dotted lines) devices in the on-state. The S/D regions within both, the UN and WN devices exhibit charge neutrality (2D charge equals the net 2D doping). Charge transfer from the narrow to the wide regions at the WN interface results in a depletion region within the intrinsic S/D of the WN device.

5.7 (a) The profile of the Büttiker probe potentials within the UN (dashed line) and WN (solid line) devices in the linear response regime ($V_{GS}=0.4\,V$ and $V_{DS}=0.05\,V$). Note that the applied voltage drops nonlinearly within the S/D regions of the WN device. Most of the voltage drops at the WN interface. (b) Derived sheet resistivity plots within both, the UN (dashed line) and WN (solid line) devices indicate the following resistances: 1) quantum contact resistance 2) S/D resistance 3) tip resistance and 4) channel resistance. In addition, the WN device also exhibits a large spreading resistance. This resistance which degrades the on-current of the WN device when compared to the UN device, saturates as the thickness of the S/D regions is increased beyond a certain value (inset).
5.8 (a) The current spectrum within the UN (dashed line) and WN (solid line) devices in the on-state with scattering. Inelastic scattering causes the source injected current spectrum to relax in energy. Within the UN device current at the source is injected into a single mode. Within the WN device the source current is injected into several unprimed (solid vertical lines) and primed (dashed vertical lines) modes. Scattering, at the WN interface enables these modes to inject current into a single mode within the narrow region. (b) The ballistically (non-selfconsistent) calculated current spectrum within the UN (dashed line) and WN (solid line) devices. In the absence of scattering, there is no energy relaxation of source injected current spectrum and S/D current spectra are symmetric. Coherent reflections, due to abrupt changes in device geometry greatly degrade current flow with the WN device when compared to the UN device.

5.9 Ballistically simulated (non-selfconsistent, potential profiles obtained from self-consistent scattering simulations are used) I-V characteristics for the UN (dashed line) and WN (solid line) devices indicate that the on-currents for the two device structures could differ by as much as \( \sim 250\% \). In contrast, once scattering is included, the I-V characteristics for the two devices differ by no more that \( \sim 10\% \).

5.10 The averaged conduction band profile within the WN device (on-state) at equilibrium (solid line) and on the application of a large drain voltage (dashed line, \( V_{DS} = 0.4 \text{ V} \)). It is clear that the conduction band edge is unchanged within the large S/D reservoirs as they maintain a near equilibrium distribution even when a large drain bias is applied. The conduction band within the intrinsic source (Fig 5.1b) floats to lower values to enforce charge neutrality (fig 5.6b).

5.11 (a) Simulated 2D charge and Büttiker probe potential profiles within a resistor with non-uniform contacts (inset) in the linear response regime. Note that the nature of the applied voltage drop is similar to the drop with the S/D regions of the WN device pictured in Fig. 5.7a. The quasi-Fermi potential drops steeply in the regions where the 2D charge density decreases to enforce current continuity. (b) Quantum mechanically simulated I-V characteristics (solid line) are compared against the I-V characteristics obtained using the domain decomposition algorithm (dashed line with triangles). The I-V characteristics of the intrinsic device is also illustrated (dashed line).
B.1 The sketch of a generic subband labeled $m$ from the source to the drain ($x$). There are $N_X$ discrete points and the length of the active device is $L$. Inside the S/D regions the potential is invariant wrt. $x$ and solutions to the 1D wave equation are plane waves.

C.1 The low bias conductance for a 40 nm resistor is plotted as a function of scattering strength, from analytical (solid line) expressions and self-consistent simulations (dashed line). The analytic conductance matches the simulated value to within $\sim 10\%$, indicating that our relation between the Büttiker probe strength and the mean free path (Eq. C.4) is physically correct.
ABSTRACT


As critical transistor dimensions scale below the 100 nm (nanoscale) regime, quantum mechanical effects begin to manifest themselves and affect important device performance metrics. Therefore, simulation tools which can be applied to design nanoscale transistors in the future, require new theory and modeling techniques that capture the physics of quantum transport accurately and efficiently. This thesis outlines a rigorous yet practical approach to model quantum transport in nanoscale silicon transistors based on the non-equilibrium Green’s function (NEGF) formalism. The objectives of this thesis are: 1) to implement the appropriate physics and methodology for nanoscale device modeling 2) to develop new TCAD (technology computer aided design) tools for quantum scale device simulation and 3) to examine and assess new features of carrier transport in futuristic nanoscale transistors. Technical issues are investigated by simulating transistors with a silicon-on-insulator (SOI) device geometry, which is widely accepted as the ideal device structure for scaling silicon technology to its fundamental limits. Simulation results presented in this thesis are used to highlight important aspects of carrier transport such as quantum coherence, confinement and tunneling, self-consistent band lineup, the nature of the voltage drop, the role of scattering and the behavior of the current versus voltage characteristics for transistors with differing geometries.
1. INTRODUCTION

1.1 Overview

For decades, progress in device scaling has followed an exponential curve, with the device density on a microprocessor doubling every three years. This has come to be known as Moore’s law [1]. Increase in the device density is made possible by shrinking the dimensions of each transistor. Shrinking transistor dimensions has resulted in increased transistor count and operating frequencies, thus enabling tremendous technological progress such as: microprocessors with greater than 100 million transistors and memory chips with greater than 1 Gb densities. The key dimension that is reduced from one transistor generation to the next is the gate length \( L_G \) (the minimum gate-length for present day devices is \( \sim 90 \) nm). Continued success in device scaling is necessary for maintaining the evolutionary technological improvements that have been the foundation for integrated circuit development and design this far. A group of leading companies publishes their projections for the next decade in an International Technology Roadmap for Semiconductors (ITRS-01) [2]. This roadmap projects a device gate-length of \( \sim 9 \) nm for the year 2016 [2]. Scaling devices to these dimensions is much more difficult and different when compared to current day scaling methodologies. This is because, metal-oxide-semiconductor (MOS) technology is approaching its limits at these dimensions. The most important issue to be addressed is how much further can aggressive device scaling be continued?. Although very difficult to answer, it is clear that new and revolutionary technologies will be needed to replace conventional MOS transistors as the driver behind electronic products in the future. Development of nanoscale transistors at the limit poses numerous practical and theoretical challenges that need to be surmounted if device scaling is to continue.
Device simulation requires new theory and modeling techniques that helps improve the understanding of device physics and design, for devices at the scaling limit.

Therefore the principle objectives of this thesis are: 1) to identify the appropriate physics and develop a methodology for device modeling which includes the correct physics, 2) to develop a new TCAD (technology computer aided design) tool for quantum level device simulation and 3) to examine and assess new features of carrier transport in nanoscale transistors. This thesis focusses on quantum effects and non-equilibrium, quasi-ballistic transport in extremely scaled transistors.

1.2 Device Scaling

There are two primary device structures that have being widely studied and used in CMOS (complementary MOS) technology. The first is the bulk structure, where a transistor is directly fabricated on the semiconductor substrate and the second, is called SOI (silicon-on-insulator), where a transistor is built on a thin silicon layer which is separated from the substrate by a layer of insulator. The bulk structure is relatively simple from a fabrication point of view and is still the standard structure in almost all CMOS based products.

Device scaling requires a balance between device functionality and device reliability. Both of these have to be maintained as one scales channel lengths to smaller sizes. To accomplish this, short channel effects (SCEs) have to be suppressed as much as possible. SCEs include threshold voltage \( V_{TH} \) variations versus channel length and drain induced barrier lowering (DIBL). Threshold voltage rolloff due to SCEs, results in a degraded subthreshold swing \( S \), which in turn renders it difficult to turn off a device, while DIBL results in a drain voltage dependent \( V_{TH} \), which complicates CMOS design at a circuit level. As critical transistor dimensions are scaled, reliability concerns become more pronounced. Unwanted leakage currents due to gate tunneling and junction tunneling rapidly increase, resulting in high off-state power dissipation.
Frank et al. have recently quantified a scale length ($\Lambda$) \[3\], that is a function of the maximum channel depletion width ($W_{dm}$), insulator thickness ($T_I$) and the ratio of the dielectric constants associated with the insulator and the silicon body ($\epsilon_{Si}/\epsilon_I$). To first order, their theory results in the following equation,

$$\Lambda = W_{dm} + \frac{\epsilon_{Si}}{\epsilon_I} T_I$$ \hspace{1cm} (1.1)

Depending on the complexity of the channel doping profile, this theory predicts that for a well tempered transistor, the minimum design $L_G$ lies between $\Lambda$ and $2\Lambda$. It is clear from Eq. 1.1 that a high body doping ($N_B$) results in reduced $W_{dm}$, and therefore a shorter scale length $\Lambda$ \[3\]. Of course, thinner $T_I$ or higher $\epsilon_I$ also helps device scaling. To achieve short scale lengths for bulk geometries, complicated super halo doping profiles have been proposed \[4\]. High gradient halos generated next to the source/drain (S/D) junctions are used to effectively screen the source end of the channel region from the influence of the drain induced electric fields. As the channel length varies around the nominal $L_G$, a shorter length causes the halo regions to merge, resulting in a higher $N_B$, which resists $V_{TH}$ rolloff. Using a ground plane in conjunction with these halo doping profiles, it has been shown that the bulk structure can be scaled down to $\sim$25 nm regime \[4\]. Below this length, scaling of bulk devices is limited by severe junction leakage. Junction leakage is large due to the high built-in electric fields (that are generated as a result of the halo implants) at the source (drain)-to-substrate p-n junctions.

Both, partially and fully depleted single-gate SOI MOSFET structures have been investigated as candidates for device scaling below $\sim$25 nm because they offer improved electrical isolation between the substrate and the active device region. However, since these MOSFETs (single-gate) have a thick buried oxide which cannot terminate the electric field lines from the drain end, they exhibit severe SCE \[5\] \[6\]. Recent studies indicate that ultra-thin body dual and tri-gate (DG, TG) SOI MOSFETs are better suited for ultimate scaling \[7\] \[8\] \[9\]. In these transistors, the presence of two or more gate electrodes significantly suppress the SCE observed in a
single-gate SOI geometry. Referring to Eq. 1.1, and approximating $W_{dm}$ by $T_{Si}$, it is clear that when $T_{Si}$ is scaled into the nanometer regime, so is the scale length. Also, since a high body doping is not needed in these transistors, band-to-band tunneling leakage is no longer a serious concern. The use of ultra-thin, lightly doped bodies results in a negligible $V_{TH}$ variation with body doping, increased channel mobility (due to reduced ionized impurity scattering and quantization effects [10] [11]) and a near ideal subthreshold swing (60 mV/dec). However, from a fabrication point of view, these structures may be difficult to build. Source/Drain (S/D) extension region resistances are an important concern when designing such structures because these resistances can be very high (due to the use of an ultra-thin body), thus limiting device performance. The use of undoped ultra-thin bodies, also reduces the possibility of adjusting $V_{TH}$ by varying the body doping. Gate stack engineering has to be performed to obtain an appropriate $V_{TH}$ either by employing new contact materials with desirable workfunctions, or maintaining an offset voltage between the different gate electrodes to mimic different workfunctions [12]. Quantum effects (subband splitting) can become significant as the confinement of carriers becomes stronger within ultra-thin bodies, which in turn translates into an increased sensitivity of $V_{TH}$ to the body thickness. This effect poses an additional challenge as fluctuations in $T_{Si}$ have to be strictly controlled. Despite the existence of these difficulties, the excellent scaling capability demonstrated by the ultra-thin body DG and TG structures cannot be underestimated. Another promising nanotechnology option is the carbon nanotube field-effect-tansistor (CNTFET), that has been demonstrated recently [13] [14]. This device (which can be conceptually viewed as a graphite sheet rolled into a tube along a specific direction), with planar or coaxial geometries seems to offer superior scaling properties when compared to silicon transistors due to its different physical properties which allow carriers to propagate ballistically (without scattering) over large distances. The unifying feature in all of these candidates for device scaling at the limit is that they are composed of large S/D reservoirs that inject and extract carriers into a narrow region (a part of which is gated) within which carriers are strongly
quantum confined. Therefore, this thesis is focused on developing a general simulation methodology that can be easily extended to model carrier transport in each of the aforementioned devices within a single quantum mechanical framework.

1.3 Modeling non-equilibrium transport

As MOSFETs scale to the nanometer regime, canonical carrier transport models are no longer capable of describing carrier transport accurately. These models are basically derived from the Boltzmann transport equation (BTE) by invoking several simplifying closure approximations [15]. They focus on scattering-dominant transport, which typically occurs in long channel devices. Nanoscale transistors, however, operate in the quasi-ballistic transport regime. Therefore simulation studies that rely on conventional models may inaccurately predict device performance and physics [16].

Quantum effects in nanoscale transistors can be grouped into two categories. The first set of effects arise due to carrier confinement along directions normal to the channel. Confinement quantizes the carrier energies thus resulting in 1) polydepletion (in the case of silicon transistors with poly gates) 2) a charge centroid shift away from the silicon/insulator interface and 3) a reduced density of states. These result in a degradation of the overall gate capacitance and also cause a dynamic shift in $V_{TH}$. The second set of effects arise due to a quantum mechanical coupling of the device to the S/D contacts and due to scattering within the device. These result in increased source-to-channel tunneling (which degrades the ability of the gate to turn off the device) and also cause a degradation of the on-current. A comprehensive model that captures all of these effects is clearly of interest.

1.3.1 The Self-consistent simulation scheme

The general scheme used to model carrier transport in nanoscale transistors is based on the non-equilibrium Green’s function formalism (NEGF). Since rigorous
derivations of this formalism are available in the literature [17] [18] [19] [20], this thesis focuses on applying this formalism to examine carrier transport in nanoscale MOSFETs without deriving the NEGF equations themselves. Within this simulation platform, the device is represented by a Hamiltonian \((H)\), which is coupled to two infinite reservoirs (S/D) (Fig. 1.1). The reservoirs are characterized by their respective Fermi levels \(\mu_S\) (source) and \(\mu_D\) (drain). Upon coupling the isolated device to the S/D reservoirs, the discrete levels within the device broaden into a continuous density of states (the extent of this broadening is a measure of the electron lifetime within the device, before it escapes into the S/D). Coupling of the active device to the S/D reservoirs is exactly accounted for by using appropriate self-energy matrices (\(\Sigma_S\) and \(\Sigma_D\) in Fig. 1.1). The self-energy method is also used to capture the effect of incoherent carrier transport within the device (denoted by \(\Sigma_{Scatt}\) in Fig. 1.1).

Coupling between the isolated device and the contacts results in a net charge transfer into or out of the device, thus resulting in a self-consistent potential (denoted by \(U_{SC}\) in Fig. 1.1). This potential is self-consistent because changes in \(U_{SC}\) modify the charge density \(\rho(\vec{r})\) inside the device, which in turn changes \(U_{SC}\) until both the charge and the potential attain consistent values. In equilibrium (when the S/D Fermi levels are shorted) the occupancy of the electronic states inside the device as well as the S/D reservoirs is dictated by a single Fermi-level \((\mu_S=\mu_D=E_F)\). However, if a non-zero voltage is applied between the source and drain terminals, the Fermi levels within these reservoirs separate by an amount \(V_{DS}\) (drain to source voltage). This difference in the Fermi-levels of the reservoirs causes one of the reservoirs to pump electrons into the device, while the other reservoir extracts electrons from the device resulting in a net flow of current. This is the non-equilibrium situation. At steady state, even in non-equilibrium, the potential \(U_{SC}\) and the charge density \(\rho\) attain self-consistent values.

The procedure to model the equilibrium and non-equilibrium situations is pictorially presented in Fig. 1.1. It is assumed that an initial guess for the electrostatic potential \((U_{SC})\) is available. Then, knowing \(H\), \(U_{SC}\), \(\Sigma_{S,D,Scatt}\) and \(\mu_{S,D}\), we can
compute the charge density, $\rho(\vec{r})$, using the well-known NEGF equations. Knowing the charge density, Poisson’s equation is solved for the new $U_{SC}$. If we wish to model electron-electron interactions (significant in very small devices), the solution to Poisson’s equation can be augmented using sophisticated techniques such as Hartree-Fock or Density functional theory. Once the new $U_{SC}$ is obtained, we recalculate $\rho$ using the NEGF equations. The two steps: $U_{SC}$ to $\rho$ and $\rho$ to $U_{SC}$ are repeated until self-consistency is achieved. It should be noted that when solving the NEGF equations numerically, it is necessary to discretize the Hamiltonian operator in a specific basis. The Hamiltonian that is used in all of this thesis is a single band effective mass Hamiltonian. This Hamiltonian, when discretized in a real-space basis yields a real-space charge density $\rho(\vec{r})$ (on solving the NEGF equations). However, in some cases (Chapter 5), it may be convenient to discretize the Hamiltonian in some other basis (not real-space). In such cases a more powerful point of view is that the local electron density is the same as the diagonal elements of the density matrix ($\rho$) expressed in a real-space basis. The calculation of the density matrix is the central problem in quantum transport theory and once computed, all the quantities of interest such as the local charge density and terminal currents can be derived from the density matrix [21] [20].

Based on the discussion above, the overall solution procedure for modeling carrier transport in a nanoscale transistors consists of the following steps (Fig. 1.1),

1. Given a particular type of device and contacts, we first make a choice of an appropriate basis (or representation) in which all the operators are discretized. Two types of basis functions, namely, the full real-space and the mode-space are used to discretize operators in this thesis. Both basis functions constitute a complete and orthonormal basis expansion set (Chapters 3, 4 and 5).

2. The next step is to write down a suitable Hamiltonian matrix for the device in the chosen basis. A single band effective mass approximation is used to obtain the Hamiltonian operator.
3. The contact self-energy functions $\Sigma_S$ and $\Sigma_D$, and the self-energy matrices that represent incoherent scattering events ($\Sigma_{\text{Scatt}}$) are then calculated. Details of these calculations are presented in chapters 3, 4 and 5 respectively.

4. To begin the self-consistent procedure, we start with a guess for the self-consistent Hartree potential, $U_{SC}$, in the chosen basis. This guess may be obtained from a classical simulator.

5. Using all the information gathered in the steps above, and the contact Fermi levels $\mu_S$ and $\mu_D$ the NEGF equations are solved to obtain the density matrix.

6. Once the density matrix has been calculated, Poisson’s equation is solved to obtain a new self-consistent potential $U_{SC}$.

7. Steps 5 and 6 are iterated until both, the potential and charge density converges to within an acceptable tolerance level.

8. The converged $\rho$ is used to calculate terminal currents.

1.4 Outline of the Thesis

This thesis in conjuction with [22] outlines a comprehensive simulation scheme that can be applied to design and evaluate the performance of futuristic nanoscale transistors. Although the modeling concepts presented in this work can be applied to study carrier transport in a variety of device structures, this thesis primarily uses a simple, ultra-thin, SOI, DG MOSFET as the example device for presenting the simulation approach and highlighting quantum effects. The thesis is divided as follows:

1. Chapter 2 presents a self-consistent solution of the Boltzmann and Poisson equations in the ballistic (no scattering) limit for both n and p-channel DG MOSFETs. It should be noted that in case of electrons, the nature of the $E - k$ relationship is such that a single band effective mass Hamiltonian can be
Fig. 1.1. Schematic of a device coupled to contacts. $H$ is the device Hamiltonian, and $U_{SC}$ is the self-consistent Hartree potential. The contact Fermi levels are $\mu_S$ and $\mu_D$. These are fixed by the applied voltage. The effect of coupling the device to the semi-infinite S/D contacts is captured through the self-energy matrices, $\Sigma_S$ and $\Sigma_D$. Incoherent processes are treated through $\Sigma_{Scatt}$. 
easily derived. However, in case of holes the form of the \( E - k \) relationship is highly nonlinear. Therefore a multiband effective mass Hamiltonian needs to be used for modeling hole transport. The computational burden associated with performing quantum simulations using such a Hamiltonian is large. Chapter 2 presents a simplification that can be applied to obtain a reasonable comparison between the ballistic performances of identical n and p-MOSFETs. The primary objectives of chapter 2 are to introduce the concept of mode-space, highlight the role self-consitency, discuss boundary conditions and examine performance related properties such as inversion layer charge, threshold voltage and thermal injection velocity. Chapter 2 discusses subband engineering issues that arise due to electron and hole quantization in DG MOSFETs and has been published in [23].

2. Chapter 3 generalizes the simple injection and reflection ideas outlined in chapter 2 through the use of the NEGF formalism. The NEGF formalism is applied to self-consistently simulate electron transport (ballistically) in n-channel DG MOSFETs using both real-space and decoupled mode-space discretizations of the active device Hamiltonian. This chapter highlights the efficiency and accuracy of the decoupled mode-space solution, discusses its inherent approximations and examines its realm of validity. Chapter 3 demonstrates that the decoupled mode-space solution is a computationally efficient method to treat electron transport in confined (one or two dimensional) device structures where the device geometry does not exhibit abrupt variations. This work has been published in [24].

3. Chapter 4 extends the ballistic NEGF solution scheme presented in the previous chapter to include the effects of scattering. Results from a detailed treatment of scattering (which is computationally demanding) within the NEGF framework is used to qualitatively motivate a simple treatment of scattering which is inspired by the concept of Büttiker probes. Büttiker probe based models
phenomenologically simulate the effect of scattering due to all possible mechanisms (e.g., ionize impurity, phonon, surface roughness etc.) as a perturbation to the active device Hamiltonian. Chapter 4 relates this perturbation strength to experimentally measured low field mobilities, compares momentum and energy-relaxing scattering models and highlights quantum effects associated with incoherent transport in nanoscale transistors. This work, which enables the application of the NEGF formalism to analyze and design nanoscale transistors has been published in [25].

4. Chapter 5 applies the NEGF formalism to examine the effects of differing device geometries on device performance, with and without scattering effects. This chapter uses a coupled version of the mode-space solution because devices with abruptly varying geometries are simulated and compared. Results presented in this chapter indicate that devices with largely differing geometries exhibit near identical characteristics once scattering is turned on, but exhibit largely differing characteristics once scattering is turned off. A simple procedure to analyze and design single, dual and tri-gate SOI MOSFETs is presented and verified by rigorous simulation results. Also, a parallel implementation of the self-consistent algorithm on a Linux cluster using a message passing interface is discussed. This work will be published in [26].

5. Chapter 6 summarizes all of the work presented in the previous chapters and lists potential directions for future work.
2. SIMULATING QUANTUM TRANSPORT IN
NANOSCALE MOSFETS: BALLISTIC HOLE
TRANSPORT, SUBBAND ENGINEERING AND
BOUNDARY CONDITIONS

2.1 Introduction

Recent industrial trends indicate that the silicon-on-insulator (SOI) geometry, with its inherently improved short channel immunity, could be the device structure that drives CMOS technology in the future [27] [7]. Also, CMOS scaling trends have progressed to a point where it is of interest to assess the ultimate performance limits of nanoscale transistors [2]. In this regard several papers examining the ballistic limits of n-channel SOI transistors have been published recently [28] [29] [30]. A comprehensive analysis of p-channel transistors has yet to be reported because of the complicated bandstructure for holes. Therefore, the objectives of this chapter are to describe a 2D simulator which is capable of modeling ballistic hole transport in thin body SOI p-MOSFETs including the effect of hole band nonparabolicity and to use this simulator to explore subband engineering for p-MOSFETs. Since boundary conditions require special care for ballistic simulations, we also discuss the boundary conditions used in this and subsequent chapters and also in our previous work [28].

As the channel length is reduced from one transistor generation to the next, it becomes necessary to reduce the SOI film thickness in order to maintain a tolerable degree of short channel effect (threshold voltage rolloff and DIBL) [3]. Reducing the SOI film thickness causes the carriers in the inversion layer to be distributed over fewer subbands. This in turn, substantially affects the electrical characteristics of short channel SOI MOSFETs. Recent studies have presented strategies to enhance
the performance of n-channel SOI MOSFETs by utilizing this subband structure [31]. However, such subband engineering cannot be performed for n-channel devices in isolation, and it is of interest to see how carrier quantization affects the performance of p-MOSFETs. Therefore we use our 2D simulator for holes to examine the effects of subband engineering in ultra-thin body p-MOSFETs. The importance of 2D electrostatic effects such as DIBL and off-equilibrium transport are emphasized as part of this study. Simulated device performance for p-MOSFETs is compared against the performance of identical n-MOSFETs in order to provide a clear perspective of subband engineering in thin body SOI transistors at the ballistic limit.

This chapter is organized as follows: 1) Sec. 2.2 outlines the modeling scheme 2) Sec. 2.3 presents simulation results for n and p channel MOSFETs with identical device geometries 3) Sec. 2.4 discusses the special boundary conditions that have been used to model ballistic transport and 4) Sec. 2.5 summarizes key findings.

2.2 Theory

In this section we present the self-consistent two-dimensional (2D) solution scheme that has been used to evaluate the hole distribution and source-to-drain current in thin body, fully depleted, ballistic (no scattering) p-MOSFETs. Quantum confinement and band nonparabolicity effects are included in the solution scheme. The simulated device structure is an idealized DG MOSFET with the Si/SiO₂ interface perpendicular to the ⟨100⟩ direction (Fig. 2.1). The y dimension is treated as infinite and the potential is assumed to be translationally invariant in the y direction. A grid spacing of a (0.4 nm) along the x direction and b (0.15 nm) along the z direction has been used in this simulation study.

Hole transport through this idealized structure is modeled by including the effect of the warped heavy and light hole bands through the following $E - \mathbf{k}$ relationship,

$$E(k) = \frac{-\hbar^2 k^2}{2m_0} \left[ A + \sqrt{ \left( B^2 + C^2 \frac{k_x^2 k_y^2 + k_y^2 k_z^2 + k_x^2 k_z^2}{k^4} \right) } \right]$$  (2.1)
where the negative sign represents the heavy hole and the positive sign the light hole bandstructure. A, B and C are constants with an associated uncertainty and determined from cyclotron resonance experiments. In our analysis the A, B and C coefficients are 4.22, 0.78 and 4.8 respectively [32]. It should be noted that in a bulk semiconductor, the light and heavy hole bands are degenerate around the \( \vec{k} = 0 \) point as indicated in Eq. 2.1. Also, in general, a nonlinear \( E - \vec{k} \) relationship for holes implies a non-zero, \( \vec{k} \) dependent coupling between the light and heavy hole bands. To include these coupling effects accurately, a \( 6 \times 6 \) Hamiltonian operator that is \( \vec{k} \) dependent needs to be discretized in two dimensions [33]. Such a solution is computationally expensive because of the size of the Hamiltonian and the associated \( \vec{k} \)-space grid. Therefore, to reduce the computational cost and at the same time capture the essential physics of ballistic hole transport including the effect of non-parabolicity, we use simplifying but physical approximations.

To explain the solution scheme we focus our attention on a thin \( z \) directed vertical slice at a fixed \( x \) location within the device as shown in Fig. 2.1. Carriers within this slice are strongly quantum confined along \( z \) and reside in a discrete set of subbands. The effect of strong quantum confinement is the removal of the degeneracy between the light and heavy hole bands around the \( \vec{k} = 0 \) point (\( k_z \) discrete) and an increased
separation between adjacent subbands away from the $\vec{k} = 0$ point. Therefore, in modeling hole transport in ultra-thin body transistors, we assume that the separation between adjacent subbands is large enough (strong confinement) to ignore band-to-band coupling. Also, by focusing on the valence band maximum ($k_x = k_y = 0$), individual subband energies and their corresponding wave functions (envelope) can be obtained by solving a one-dimensional (1D), $z$ directed, single band effective mass equation of the form,

$$\frac{\hbar^2}{2m^*_z} \frac{\partial^2 \Psi_i(x, z)}{\partial^2 z} - qV(x, z)\Psi_i(x, z) = E_i(x)\Psi_i(x, z)$$

(2.2)

In Eq. 2.2, $m^*_z$ is the effective mass along the confinement direction, $E_i(x)$ and $\Psi_i(x, z)$ are the eigen energies and wavefunctions for subband $i$, at slice $x$ and $V(x, z)$ is the 2D potential profile. The effective mass used in Eq. 2.2 is,

$$m^*_z = \frac{m_0}{A \mp B}$$

(2.3)

where the negative sign represents the heavy hole and the positive sign the light hole band respectively. Note that $\Psi_i$ in Eq. 2.2 is indeed the envelope function for holes along the gate confinement direction. In deriving Eq. 2.2, we restricted our focus to the valence band maximum ($k_x = k_y = 0$), thereby eliminating the complex band warping and coupling present in the valence band. Therefore, the reciprocal effective mass tensor becomes a diagonal matrix for the heavy and light hole bands. This approximation of a parabolic bandstructure for holes in the confinement direction, enables us decouple the channel and confinement directions thus greatly reducing the associated computational burden [34]. It has been shown to work well for double-gate MOSFETs [24]. Note that the warped $E(k_x, k_y)$ surface ($k_z = 0$ in Eq. 2.1) associated with each hole subband includes the essential characteristics of the corresponding $E(k_x, k_y)$ relationship obtained from a more general calculation based on the Luttinger Hamiltonian within the energy range of interest [33]. These approximations and the bandstructure used in this analysis have also been shown to reproduce measured C-V characteristics of p-MOS capacitors accurately in the literature [35].
A solution to Eq. 2.2 at every \( x \) location within the device yields a set of subband profiles. Since coupling between different subbands is ignored and the transport is ballistic, a 1D Boltzmann solution can be directly solved to obtain the charge density and current contribution from each subband. The Boltzmann solution technique for a generic subband profile is illustrated in Fig. 2.2. The subband is coupled to the source and drain reservoirs which are characterized by their Fermi levels \( \mu_S \) and \( \mu_D \) respectively. It should be noted that the subband energy is the effective potential energy of the hole and that holes flow up the potential hill unlike electrons. The subband profile shown in Fig. 2.2 can be spatially resolved into two regions: 1) points to the left of the subband minimum \( E_{i,\text{min}} \) and 2) points to the right of the subband minimum. In region 1, holes injected from the source reservoir with \( x \) directed energies greater than \( E_{i,\text{min}} \) are reflected off the source-to-channel barrier. Therefore, both the \( +k_x \) and the \( -k_x \) states representing \( x \) directed energies greater than \( E_{i,\text{min}} \) are in equilibrium with the source (filled circles in Fig. 2.2). Holes with \( x \) directed energies
less than $E_{i,\text{min}}$ are not reflected by the source-to-channel barrier in the Boltzmann picture. Since there is no barrier reflection and the transport is ballistic, the $+k_x$ states are in equilibrium with the source (filled circles in Fig. 2.2) and the $-k_x$ states are in equilibrium with the drain (empty circles in Fig. 2.2) for $x$ directed energies less than $E_{i,\text{min}}$. A similar analysis applies for holes in region 2.

The charge density within a device is primarily determined by electrostatics. Therefore, when evaluating the 2D hole density (given a subband profile), we assume that the $E(k_x, k_y)$ relationship for the heavy and the light hole subbands is parabolic and is characterized by a mass that yields the same 2D density of states (DOS) as the corresponding non-parabolic $E(k_x, k_y)$ relation ($k_z = 0$ in Eq. 2.1). This approximation is necessary to obtain analytical expressions for the charge density as outlined in the subsequent paragraph. The 2D DOS effective mass can be expressed as [36],

$$m_d^* = \frac{m_0}{2\pi} \int_0^{2\pi} \frac{d\theta}{[A + \sqrt{((B^2 + C^2 \cos^2 \theta \sin^2 \theta))}]}$$

where, $\theta$ is the polar angle in the $k_x - k_y$ plane. Using this mass and the incidence and reflection arguments illustrated in Fig 2.2, the 2D hole density in region 1 can be expressed as,

$$p_1(x) = p_{2D,i} \ln(1 + e^{-\bar{\mu}_S}) + \frac{1}{\sqrt{\pi}} \int_0^{\bar{E}_{i,\text{min}}} \frac{-d\bar{E}_x}{\sqrt{-\bar{E}_x}} \sqrt{\bar{\delta} - 1/2(\bar{E}_x - \bar{\mu}_S)}$$

$$+ \ln(1 + e^{-\bar{\mu}_D}) - \frac{1}{\sqrt{\pi}} \int_0^{\bar{E}_{i,\text{min}}} \frac{-d\bar{E}_x}{\sqrt{-\bar{E}_x}} \sqrt{\bar{\delta} - 1/2(\bar{E}_x - \bar{\mu}_D)}$$

Both Fermi-Dirac statistics and spin degeneracy have been included in Eq. 2.5. Note that all energies in Eq. 2.5 have been normalized by $k_B T$ and that the tilde implies that the energies are specified relative to the local subband energy $E_i(x)$. The 2D density factor for subband $i$ ($p_{2D,i}$) is $m_d^* k_B T / 2\pi \hbar^2$. A similar expression for the 2D hole density in region 2 can be obtained.

$$p_2(x) = p_{2D,i} \ln(1 + e^{-\bar{\mu}_S}) - \frac{1}{\sqrt{\pi}} \int_0^{\bar{E}_{i,\text{min}}} \frac{-d\bar{E}_x}{\sqrt{-\bar{E}_x}} \sqrt{\bar{\delta} - 1/2(\bar{E}_x - \bar{\mu}_S)}$$
\[
+ \ln(1 + e^{-\tilde{\mu}_D}) + \frac{1}{\sqrt{\pi}} \int_0^{\bar{E}_{i,\text{min}}} \frac{-d\tilde{E}_x}{\sqrt{-\tilde{E}_x}} \tilde{\delta}_{-1/2}(\tilde{E}_x - \tilde{\mu}_D) \] (2.6)

This 2D charge density is distributed along the confinement direction at each \(x\) location using the corresponding wave function \(|\Psi_i(z)|^2/b\) in order to obtain the 3D density profile. A summation over all the subbands yields the final charge density that is input to the Poisson solver for self-consistent solutions. Note that this treatment captures all of the quantum effects associated with hole confinement and can also treat the effect of hole wavefunction penetration into the insulator regions.

A nonlinear version of the Poisson equation which uses floating boundaries (Section 2.4) is solved in order to expedite convergence between the Poisson and transport iterations. Nonlinearity is forced into the Poisson equation by expressing the 3D hole density in terms of a corresponding quasi Fermi level at each node of the 2D grid using,

\[
p(x, z) = N_V \tilde{\delta}_{1/2}[\frac{-qV_{old} - E_F}{k_B T}]
\]

and

\[
E_p = -qV_{old} - k_B T \tilde{\delta}_{-1/2}[\frac{p}{N_V}]
\]

where, \(N_V\) is the effective DOS in the valence band, \(V_{old}\) the potential from the previous iteration and \(\tilde{\delta}_{-1/2}^{-1}\) the inverse Fermi integral of order \(-1/2\) \cite{37}. Details of the nonlinear technique to solve Poisson’s equation can be found in \cite{38} \cite{39} \cite{22}. The maximum difference in potential between successive iterations is used as a measure of convergence.

Once convergence is achieved, the terminal current can be evaluated at any point within the device as it is a conserved quantity. We evaluate the current at the bottom of the source-to-channel barrier (\(E_{i,\text{min}}\) point in Fig. 2.2) because this is a point where there is no reflection. Since this is a one time calculation, the terminal current, including the effect of band nonparabolicity is obtained by explicitly summing fluxes in \(\vec{k}\)-space. It should be noted that although the confinement direction is \(\langle 100 \rangle\), the channel could be oriented along either the \(\langle 100 \rangle\) or the \(\langle 110 \rangle\) direction. Therefore
the occupancy factor (Fermi function) associated with a specific $\vec{k}$-state needs to be correctly defined. If the channel is oriented along $\langle 100 \rangle$, all $+k_x$ states are filled using $\mu_S$ and the $-k_x$ states are filled using $\mu_D$. However, if the channel orientation is changed to $\langle 110 \rangle$, we perform a 45° rotation of the $(k_x, k_y)$ axis to obtain a modified 2D $E(k'_x, k'_y)$ relation. In this new orientation, all of the $+k'_x$ states are filled using $\mu_S$ and the $-k'_x$ states are filled using $\mu_D$ given the subband minimum. The expression for the current density from subband $i$ including spin degeneracy is,

$$I_i = \frac{q}{2\pi^2} \times$$

$$\int_{-\infty}^{\infty} \int_{0}^{\infty} \left[ \frac{\bar{\nu}(k_x, k_y)dk_xdk_y}{1 + \exp\left(\frac{\mu_S - (E(k_x, k_y) + E_{i,\text{min}})}{k_B T}\right)} - \frac{\bar{\nu}(k_x, k_y)dk_xdk_y}{1 + \exp\left(\frac{\mu_D - (E(k_x, k_y) + E_{i,\text{min}})}{k_B T}\right)} \right]$$

(2.8)

Note that the velocity in Eq. 2.8, includes the direction dependence associated with a specific channel orientation. Summing each subband contribution yields the total terminal current for a particular channel orientation. Analogous expressions for the charge density and current in ultra-thin body, fully depleted n-MOSFETs were presented in [40].

### 2.3 Results

In this section we compare the ballistic performance of two ultra-thin body n and p MOSFETs. The first device has a silicon body thickness of 1.5 nm while the second device has a body thickness of 5 nm. Both devices have the geometry shown in Fig. 2.1 with a source/drain doping (S/D) of $10^{20}\text{cm}^{-3}$ and an intrinsic channel. The junctions are abrupt and the oxide thickness is 1.5 nm. In order to minimize short channel effects, a channel length of 20 nm has been assumed in all of the simulations. A power supply voltage ($V_{DD}$) of 0.6 V is used, and, to facilitate performance comparison, the threshold voltage (current value = $1\mu\text{A}/\mu\text{m}$) of each device is calibrated to $0.25V_{DD}$ by adjusting the gate workfunction. Electron and
hole wavefunction penetration into the insulator regions is not considered, and the focus is primarily on the on-state ($V_{GS} = V_{DS} = V_{DD}$).

The ballistic current is a product of the charge density and the carrier velocity. The charge density within a device is primarily prescribed by 2D electrostatics and the velocity (for a given charge density) by the bandstructure. Therefore the effects of 2D electrostatics and bandstructure on device performance can be explained by examining the charge density and carrier velocity at a point within the device. We examine these quantities at the “charge control point.” To illustrate the charge control point we plot the 2D potential and subband profiles within the 1.5 nm body, n and p MOSFET structures in the on-state (Fig. 2.3). Although the potential profile is a 2D function ($x$ and $z$ dimensions), the subbands are a function of $x$ alone and represent the effective potential energy of the confined 2D carriers. Each subband exhibits a maximum (minimum for holes) as one moves from the source to the drain. This point is referred to as the charge control point because it is a point where there is no carrier reflection due to the potential profile. Note that due to the thin nature of the body (1.5 nm), both transistors exhibit single subband occupancy. Also note that since the effective mass ($0.98m_0$) of the electron in the confinement direction is higher than that of the corresponding heavy hole ($m_0/A - B$), the confinement energy of the heavy hole (distance between $E_V$ and the subband) is higher than that of the electron (distance between $E_C$ and the subband).

Figure 2.4 presents the ballistic current vs. voltage characteristics of the 1.5 nm body n and p MOSFETs. Both transistors exhibit a near ideal subthreshold swing of 61 mV/dec due to the thin body and the double gate geometry. Despite exhibiting identical subthreshold behavior, the on-current from the ballistic p-MOSFET is only 60% of that from the ballistic n-MOSFET for a channel orientation of $\langle 100 \rangle$. To explain this difference we plot the 2D charge density and the corresponding non-equilibrium injection velocity (derived from the current) at the charge control point in Fig. 2.5. It is clear from Fig. 2.5 that although the hole density is slightly higher than the corresponding electron density as a result of 2D electrostatics, the hole injection
Fig. 2.3. (a) The conduction band and the profile of the first subband in the 1.5 nm body n-MOSFET ("he" represents a heavy electron in the confinement direction).

(b) The valence band and the profile of the first subband in the 1.5 nm body p-MOSFET ("hh" represents a heavy hole in the confinement direction). All energies are referred to the source Fermi level(0). Due to the high doping in the source/drain regions the electron subband is negative and the hole subband positive wrt. the reference. Note that the confinement energy of the hole is much higher than the electron. The bias state is $|V_{GS}| = 0.45 \, V$ and $|V_{DS}| = 0.6 \, V$. 
velocity is considerably (~50%) lower than the corresponding electron velocity. This indicates that the heavy hole is indeed heavy along both the confinement and channel directions unlike the heavy electron which is heavy along the confinement but light along the channel (unprimed valley, 0.19\textit{m}_0). This analysis also indicates that in thin body SOI devices under conditions of single subband occupancy, the ballistic p-MOSFET will always largely under perform the n-MOSFET irrespective of channel length scaling. Therefore bandstructure engineering is essential in order to increase p-MOSFET performance.

An excellent analytical analysis of ballistic device performance in thin body DG MOSFETs assuming single subband occupancy and 1D electrostatics has been presented in the literature [41]. We examine the validity of this 1D treatment of the electrostatics by plotting the 2D charge density and velocity at the charge control point as a function of |\textit{V}_{DS}| for the 1.5 nm body n and p-MOSFETs (Fig. 2.6). In the absence of 2D effects, the charge at the top of the barrier is determined by the gate overdrive in the on-state (evaluated at |\textit{V}_{DS}| = 0, |\textit{V}_{GS}| = 0.6 V). On including
Fig. 2.5. 2D charge density and injection velocity, extracted at the charge control point vs. $V_{GS}$ for the 1.5 nm body, n (dotted line) and p (solid line) MOSFETs. At the same gate voltage, the hole density is $\sim 10\%$ greater than the electron density, but the hole injection velocity along $\langle 100 \rangle$ is $\sim 50\%$ lower than the electron injection velocity due to the warped nature of the heavy hole subband.

2D effects it is clear from Fig. 2.6 that the charge at the control point is increased as a result of DIBL in case of both n and p MOSFETs. For the p-MOSFET, this increase in charge density is $\sim 15\%$ despite using an ultra thin body and a relatively long channel. The charge density difference translates to an increased difference ($\sim 25\%$) in the ballistic current as a result of degeneracy (see Sec. 2.4). A similar trend is also observed in case of the n-MOSFET. Note that due to degeneracy effects, the saturation nonequilibrium injection velocities for both electrons ($\sim 1.9 \times 10^7$ cm/s) and holes ($\sim 1 \times 10^7$ cm/s) are much higher than their equilibrium injection values ($\sim 1.5 \times 10^7$ cm/s for electrons and $\sim 8 \times 10^6$ cm/s for holes).

On increasing the body thickness to 5 nm, the difference between the on-currents of the n and p MOSFETs reduces to $\sim 25\%$ (Fig. 2.7). To understand this reduced difference, we focus on the 2D charge density and injection velocity at the control point once again (Fig. 2.8). The injection velocity for the hole is $\sim 85\%$ of that for the electron (unlike the 1.5 nm body) and the 2D hole density is $\sim 10\%$ less than the corresponding electron density, thus resulting in the observed difference in the
Fig. 2.6. 2D charge density and injection velocity, extracted at the charge control point vs. $V_{DS}$ for the 1.5 nm body, n (dotted line) and p (solid line) MOSFETs. DIBL increases the charge density by $\sim$15%. Injection velocities are higher than their equilibrium values due to degeneracy (Section 2.4).

on-current. To explain this behavior of the charge density and injection velocity, we note that only $\sim$40% of the inversion electron charge resides in the first unprimed subband in case of the 5 nm n-MOSFET. Whereas, $\sim$80% of the hole charge continues to occupy the first heavy hole band in the p-MOSFET. Increased occupancy of the high energy subbands degrades the electron injection velocity because the average channel effective mass is increased [31]. For holes, since the light hole band is light in the confinement and channel directions, the hole injection velocity is improved slightly as the light hole population is increased, thus resulting in a reduced difference between the electron and hole injection velocities. The effective gate capacitance of the n-MOSFET is greater than the p-MOSFET as the effective 2D DOS for electrons is greater than that of holes (primed bands have a higher DOS compared to the light hole band). Therefore the inversion charge density at a high gate voltage is higher for electrons than holes (crossover point in Fig. 2.8 is the voltage at which high energy subbands start contributing to the electron charge).
Fig. 2.7. The $I_{DS}$ vs. $V_{GS}$ (log scale) and the $I_{DS}$ vs. $V_{DS}$ (linear scale) characteristics of the 5 nm body n (dotted line) and p (solid line) MOSFETs for a channel orientation of (100). Both devices exhibit near ideal subthreshold characteristics. The on-current of the p-MOSFET is $\sim$25% lower than that of the n-MOSFET.

Fig. 2.8. 2D charge density and injection velocity, extracted at the charge control point vs. $V_{GS}$ for the 5 nm body, n (dotted line) and p (solid line) MOSFETs. The crossover in the 2D charge density is due to an increased inversion layer capacitance for the electrons as a result higher subband occupancy. Note that higher subband occupancy considerably degrades the injection velocity for electrons, whereas hole velocity is relatively unaffected when compared against Fig. 2.5.
It should be noted that increasing the body thickness from 1.5 nm to 5 nm reduces the inversion layer capacitance for both the n and p-MOSFETs as the charge centroid is pushed away from the Si/SiO$_2$ interface. Also, for electrons we noted that the saturation injection velocity reduced with higher subband occupancy. The net effect of the effective gate capacitance and injection velocity degradation is a $\sim$30\% reduction in the n-MOSFET performance as we scale the body thickness from 1.5 nm to 5 nm. For the p-MOSFET we observed that increasing the body thickness actually increased the injection velocity slightly (for the same charge density). Therefore the degradation in performance as we scale the body thickness from 1.5 nm to 5 nm is primarily due to capacitance degradation alone and is $\sim$15\%. This indicates that p-MOSFETs are less sensitive to subband engineering arising from reducing the body thickness as compared to n-MOSFETs.

Finally, we examine the on-current performance of the 5 nm body p-MOSFET as a function of channel orientation. Although Eq. 2.1 indicates that holes in the heavy hole band are heavier along $\langle 110 \rangle$ as opposed to $\langle 100 \rangle$, we find that for the same 2D hole density, the current in the 5 nm body p-MOSFET is $\sim$20\% higher for a $\langle 110 \rangle$ channel as compared to a $\langle 100 \rangle$ channel. To explain this behavior, we plot the net hole flux distribution in $\vec{k}$-space along the $\langle 100 \rangle$ (left) and $\langle 110 \rangle$ (right) directions for both the heavy and light hole bands in Fig. 2.9 (Since the transistor is operating in the on-state only one-half of the $\vec{k}$-states are occupied). Figure 2.9a shows that most of the heavy holes are located near the $\langle 110 \rangle$ direction, and that these holes are responsible for the net flux irrespective of whether the channel is oriented along $\langle 110 \rangle$ or $\langle 100 \rangle$. Figure 2.9b shows that the light hole flux displays the opposite trend; most of the light holes are located near the $\langle 100 \rangle$ direction. Therefore, no simple treatment of the direction dependence of the hole current is possible and one needs to explicitly sum over all the flux components in $\vec{k}$-space to study the effect of channel orientation.

For bulk p-MOSFETs, it has been observed experimentally that the on-current for $\langle 100 \rangle$ oriented channels is $\sim$7\% higher than that of $\langle 110 \rangle$ oriented devices [42]. This behavior is the opposite of what we observed in our 5 nm body p-MOSFET. Note that
in the 5 nm body device, although \( \sim 80\% \) of the hole density resides in the first heavy hole band and only \( \sim 20\% \) in the first light hole band (note the maximum value on the \( \vec{k} \) axes in Fig. 2.9a and Fig. 2.9b), the light hole band contributes \( \sim 35\% \) of the total hole current along \( \langle 100 \rangle \). This clearly indicates that if the population in the light hole band is increased, the difference in the total current between the \( \langle 100 \rangle \) and \( \langle 110 \rangle \) directions will reduce and may even be reversed. Our simplified hole band structure ignores subband-to-subband coupling and scattering. These effects may have to be included in order to capture the physics of hole transport more accurately.

2.4 Discussion

In simulating ballistic transport in MOSFETs, great care must be exercised in treating the boundaries. We impose floating boundary conditions when solving Poisson's equation. This boundary condition is realized by setting,

\[
\hat{x} \cdot \nabla V = 0
\]  

(2.9)

at the source and drain ends of the simulation domain. Conventional transport models use fixed potential boundary conditions assuming equilibrium statistics and charge neutrality at the contacts. Under ballistic transport conditions, equilibrium statistics no longer apply [43], and our use of the floating boundary condition helps us restrict our focus to the intrinsic ballistic device without explicitly including the large scattering dominated contacts to which it is coupled. Note that the information about the drain bias (Fermi levels) is used in solving the transport equation for the hole density. Therefore, even if the potential is allowed to float, it cannot float to any arbitrary value. The potential floats relative to the source/drain Fermi levels.

To explain the floating boundary condition, we modify our idealized 1.5 nm body DG MOSFET structure by appending heavily doped \( 5 \times 10^{20}/\text{cm}^3 \) \( p^{++} \) regions to the left and right ends of the device. The averaged valence band profile for the modified device structure under equilibrium conditions \( (V_{GS} = -0.6V, V_{DS} = 0V) \) is shown in Fig. 2.10a (this profile is obtained by weighting the valence band energy with the
Fig. 2.9. (a) Flux distribution in $\vec{k}$-space for the heavy hole band in the 5 nm body p-MOSFET along the $\langle 100 \rangle$ (left) and $\langle 110 \rangle$ (right) directions. (b) Flux distribution in $\vec{k}$-space for the light hole band in the 5 nm body p-MOSFET along the $\langle 100 \rangle$ (left) and $\langle 110 \rangle$ (right) directions. Note that in the on-state only one half of $\vec{k}$-space is occupied as drain injection is suppressed. Also, note that the $k$-axes are rotated by 45° between the left and right pictures.
local charge density and integrating over the confinement direction). The presence of the heavily doped $p^{++}$ regions creates a large potential barrier for holes injected from the source and drain contacts. This barrier gives rise to strong reflections which maintain a near equilibrium distribution in the $p^{++}$ regions even when a large bias is applied to the drain. Therefore, a fixed potential boundary condition based on charge neutrality can be used when solving Poisson’s equation. The $p^{++}$ regions maintain a near equilibrium distribution even under bias.

On examining the nonequilibrium ($|V_{GS}| = |V_{DS}| = 0.6 \, V$) potential profile (Fig. 2.10a) of the modified device, it is clear that the averaged valence band is unchanged in the heavily doped $p^{++}$ regions but floats to a higher value in the $p^+$ source region of the intrinsic device. This observation can be explained by examining Fig. 2.2. Under equilibrium conditions, both the $+k_x$ and the $-k_x$ states in the $p^+$ source are filled by a single Fermi level resulting in zero net current. As the drain bias is increased to higher and higher values, the drain injected (negative) half of the hole distribution is suppressed in region 1 (empty circles in Fig. 2.2). Therefore, all of the charge in this region is comprised of source injected carriers. Note that with increasing $V_{GS}$, the source-to-channel barrier height decreases. This reduces the fraction of the source injected carriers that is reflected back and most of the charge in region 1 exists in the $+k_x$ states. Although nearly one half of the $\vec{k}$-space occupancy is suppressed, 2D electrostatics requires that charge neutrality has to be maintained (integrated doping equals the 2D hole density) at every point within region 1. To achieve charge neutrality, the hole density residing in the $+k_x$ states in region 1 nearly doubles between the equilibrium and nonequilibrium bias conditions. To accommodate this increased charge in the $+k_x$ half, the valence band has to float to a higher energy (Fig. 2.10a). Increased occupancy of the high energy $+k_x$ states is what we refer to as degeneracy. Carriers in the heavily doped $p^{++}$ region are predominantly backscattered by the built in barrier, so the potential in this region is unchanged with drain bias.

Now, if we remove the $p^{++}$ regions but use the floating boundary condition to terminate the $p^+$ regions, we observe the potential behavior plotted in Fig. 2.10a.
Fig. 2.10. (a) The average valence band (averaged along z) profile within the $p^{++}-p^+-n^-p^+-p^{++}$, 1.5 nm body transistor in equilibrium ($V_{GS} = -0.6 \, V$) and in the on-state is plotted on the left. On the right we compare the average valence band obtained using floating boundary conditions (solid line) to that obtained using fixed boundary conditions (dotted line) in the on-state. (b) The charge density profiles obtained by simulating the $p^+-n^-p^+$ device using floating boundary conditions (solid line) are compared to those obtained using fixed boundary conditions (dotted line) in the on-state. Note that irrespective of the boundary condition used, macroscopic charge neutrality is always achieved.
Figure 2.10a demonstrates that the averaged valence band under bias for the device with floating boundaries \((p^+ - n^- - p^+)^-\) is identical to that for device with fixed boundaries \((p^{++} - p^+ - n^- - p^+ - p^{++})^-\) within the region of interest \((p^+ - n^- - p^+)^-\). 2D charge density profiles also exhibit a similar match in the \(p^+ - n^- - p^+\) region (Fig. 2.10b) thus indicating that the floating boundary condition does capture the effect of coupling a ballistic device to a scattering contact.

### 2.5 Summary

This chapter described a self-consistent 2D simulation scheme for modeling ballistic transport in ultra-thin body, fully depleted, SOI p-MOSFETs including the effect of band nonparabolicity. While doing so, the use of floating boundary conditions was highlighted and validated. The simulator was used to examine the physics of subband engineering on hole transport. Our simulations indicate that under conditions of single subband occupancy, the ballistic n-MOSFET will significantly outperform the ballistic p-MOSFET. They also indicate that 2D effects are strong in these short channel devices and need to be accounted for to obtain an accurate ballistic limit. Subband engineering (thinning the SOI film) improves the performance of ballistic n-MOSFETs significantly, because both gate capacitance and injection velocity are increased. For ballistic p-MOSFETs, the inversion capacitance is increased, but there is a small degradation in the hole injection velocity. Hence the performance improvement is significantly smaller. The heavy hole band flux is higher along \(\langle 110 \rangle\) as compared to the \(\langle 100 \rangle\) direction for a given charge density. The reversed trend applies to the light hole subbands. Therefore, if most of the hole charge is in the heavy hole band, a channel orientation of \(\langle 110 \rangle\) yields a higher current as compared to a \(\langle 100 \rangle\) channel orientation. If the light hole population increases, this difference is reduced and channel orientation does not affect p-MOSFET device performance significantly.
Although we have used a simple treatment of the valence bands, the behavior of the hole current with respect to channel orientation presented in this paper is consistent with a more rigorous treatment which uses a $6 \times 6$ Luttinger Hamiltonian to describe the hole bandstructure [44]. The ideas presented in this chapter can be extended to treat ballistic transport including the effect of warped bandstructures in general. To do so, the analytical calculations for hole density and current will have to be replaced by numerical calculations. We believe that rigorous solutions are definitely desirable and need to be developed to understand the physics of hole transport in SOI transistors, but feel that simple approaches such as ours, provides valuable insight efficiently until rigorous solutions are rendered tractable. The results presented in this chapter point out the importance of a careful treatment of the valence bandstructure when simulating p-MOSFETs.
3. SIMULATING QUANTUM TRANSPORT IN NANO SCALE TRANSISTORS: REAL VERSUS MODE-SPACE APPROACHES

3.1 Introduction

As transistor dimensions are scaled, quantum effects, which affect the threshold voltage (confinement), gate capacitance (charge centroid shift), off-current (source barrier tunneling) and gate leakage begin to manifest themselves, and semiclassical methods that disregard these effects are inadequate in capturing the physics of ballistic transport. In chapter 2, we presented a simple solution to the ballistic, semiclassical Boltzmann transport equation based on injection and reflection arguments. While this solution captured quantum confinement effects and provided a reasonable estimate of the ballistic on-current, it did not treat the effects of source-to-channel tunneling, which increases the off-current (and may pose an ultimate scaling limit). Also, the intuitive solution scheme is not general and cannot be easily extended to treat the effects of scattering. Therefore, this chapter focuses on extending the simple injection and reflection arguments presented in chapter 2), within a more general simulation framework which is based on the non-equilibrium Green’s function formalism (NEGF) [45] [20]. The NEGF formalism will be applied to model electron transport in the rest of this thesis.

Quantum modeling approaches rely on a self-consistent solution of the Schrödinger and Poisson equations in order to obtain the charge distribution and current for a specific device geometry. A solution to the Schrödinger equation can be pursued at varying levels of complexity depending on the nature of the device under study and the desired degree of accuracy. For bulk devices, 2D solutions based on the Green’s function formalism have been demonstrated by two groups [46] [47]. These solutions,
which are based on a real-space discretization of the full 2D effective mass Hamiltonian, are computationally expensive. Such discretization is essential in order to treat quantum transport accurately in bulk MOSFETs because the confining effect of the gate is lost as carriers move from the source to the drain. However, this is not the case in thin body SOI MOSFETs where mobile charges are quantum confined all the way from the source to the drain (due to the thin body). For such geometries, the computational burden associated with the real-space solution can be greatly reduced (without compromising on accuracy) by expanding the 2D Hamiltonian in coupled or decoupled mode-space [48] [40] (characteristic modes of the Hamiltonian in the confinement direction) and by treating the first few occupied modes. This chapter benchmarks the simplified, decoupled mode-space solution against the rigorous 2D real-space solution by applying them to simulate electron transport in a thin body, fully depleted, DG n-MOSFET structure. Simulation results from the two approaches are compared and while doing so, the various approximations inherent in the decoupled mode-space solution, its realm of validity and the generality of the real-space solution are discussed. This chapter aims to describe the numerical methods that one can use to simulate quantum transport in different transistor geometries with specific emphasis on the decoupled mode-space solution scheme. Quantum boundary conditions are also derived, and the quantum mechanical features of the simulation results are highlighted.

The chapter is divided into the following sections: 1) Sec. 3.2 presents the real-space and mode-space solutions succinctly. The size of the problem associated with each method is highlighted. 2) Sec. 3.3, compares simulation results obtained by applying both techniques to model ballistic transport in a thin body, fully depleted, DG MOSFET. 3) Sec. 3.4, critically examines the validity for the mode-space solution. 4) Sec. 3.5 summarizes key findings.
3.2 Theory

The simulated device structure is shown in Fig.1a. A uniform rectangular grid with a grid spacing of \(a\), along the \(x\) direction and \(b\), along the \(z\) direction is used. The device is represented by a Hamiltonian matrix that is coupled to two infinite reservoirs, the source and drain. The source/drain (S/D) extension regions are terminated using open boundary conditions (no \(x\) dependence of the potential), and the Fermi level in these regions is specified by the applied voltage. The width of the device is assumed to be large, and the potential is assumed to be translationally invariant along the width (\(y\) dimension). A single band effective mass Hamiltonian is used to model carrier transport.

3.2.1 Real-Space solution

This section briefly explains the real-space simulation method with specific emphasis on the self-energy concept (which is used to quantum mechanically couple the active device to the infinite S/D contacts) and the size of the problem. In this simulation method, the effective mass Hamiltonian is discretized in 2D real-space (\(x, z\)), to solve for its retarded Green’s function. We begin by expanding the 3D Hamiltonian for the device in terms of \(\delta(x - x')\delta(z - z')\) and \(e^{ik_jy}/\sqrt{W}\), where the plane wavefunction, \(e^{ik_jy}/\sqrt{W}\), represents the device width (\(W\)). The quantum number, \(k_j\), corresponds to the eigenenergy, \(E_{k_j} = \hbar^2 k_j^2/2m_y^*\), where \(m_y^*\) is the electron effective mass in the \(y\) direction. The real-space delta functions, \(\delta(x - x')\) and \(\delta(z - z')\) with eigenvalues \(x'\) and \(z'\) respectively, combined with \(e^{ik_jy}/\sqrt{W}\), form a complete and
Fig. 3.1. (a) An ultra-thin body DG MOSFET structure with S/D doping of $10^{20}$ cm$^{-3}$ and an intrinsic channel (channel thickness = 1.5 nm). (b) The squeezed DG MOSFET structure used to investigate the effect of mode coupling (channel thickness = 0.75 nm).
orthogonal expansion function set. On expansion, the Hamiltonian in block diagonal form is

\[
H = \begin{bmatrix}
  h(x, z) + E_{k_1} I & 0 & \cdots & 0 \\
  0 & h(x, z) + E_{k_2} I & \cdots & 0 \\
  0 & \cdots & \ddots & 0 \\
  \cdots & \cdots & \cdots & h(x, z) + E_{k_j} I \\
  \cdots & \cdots & \cdots & 0 & \ddots
\end{bmatrix}
\] (3.1)

where I is the identity matrix and

\[
h(x, z) = \begin{bmatrix}
  \alpha_0 & \beta & 0 & \cdots & 0 \\
  \beta & \alpha_1 & \ddots & \cdots & 0 \\
  0 & \ddots & \ddots & \ddots & 0 \\
  \cdots & 0 & \alpha_{N_X} & \beta \\
  \cdots & \cdots & 0 & \beta & \alpha_{N_X+1}
\end{bmatrix}
\] (3.2)

is block tridiagonal. Note that Eqs. 3.1 and 3.2 represent infinite matrices. This is because the device is coupled to infinite leads which have not yet been replaced by appropriate boundary conditions. If we think of the device as being composed of vertical slices adjoining each other, then the \(\alpha\)'s represent coupling along the \(z\) direction within each slice while the \(\beta\)'s represent coupling between adjacent slices. The \(\alpha\)'s, with indices less than one, represent successive slices going into the source contact while those with indices greater than \(N_X\) represent successive slices into the drain. The \(\alpha\)'s and \(\beta\)'s are themselves block matrices and are,

\[
\alpha[x] = \begin{bmatrix}
  2t_x + 2t_z - qV_1(x) & -t_z & 0 & \cdots \\
  -t_z & 2t_x + 2t_z - qV_2(x) & \ddots & 0 \\
  0 & \ddots & \ddots & -t_z \\
  \cdots & 0 & -t_z & 2t_x + 2t_z - qV_{N_Z}(x)
\end{bmatrix}
\] (3.3)

\[
\beta = \begin{bmatrix}
  -t_x & 0 & \cdots \\
  0 & -t_x & \cdots \\
  \cdots & 0 & -t_x
\end{bmatrix}
\] (3.4)
The $V$'s (Eq. 3.3), represent the potential along a vertical slice at site $x$ and $t_x$ and $t_z$, the coupling energies between adjacent grid points in $x$ and $z$ respectively. These site coupling energies are given by,

$$t_x = \frac{\hbar^2}{2m^*_{x}a^2} \quad \text{and} \quad t_z = \frac{\hbar^2}{2m^*_{z}b^2} \tag{3.5}$$

It is clear from Eq. 3.1 that blocks representing different plane wave states (along the device width) are decoupled, as there is no scattering within the device. Also, $E_{kj}$ ranges between 0 and $+\infty$, accounting for all possible plane wave states. There is no restriction on the solution domain and it can be easily extended to include the insulator regions provided changes in the electron effective mass is correctly accounted for within the insulator and at the silicon/insulator interface when discretizing the effective mass Hamiltonian.

For each plane wave eigenenergy $E_{kj}$, we write the retarded Green’s function relevant to 2D transport as,

$$G(E) = [EI - [h(x, z) + E_{kj}I]]^{-1} = [E(k_x, k_z)I - h(x, z)]^{-1} \tag{3.6}$$

where the in-plane energy is defined as $E(k_x, k_z) \equiv E - E_{kj}$ (note that that $E(k_x, k_z)$ is just a notation to identify the in-plane energy and that the Hamiltonian in the $X-Z$ plane is in a real-space basis). We then account for the infinite leads by introducing an appropriate self-energy function [20]. Details of the self-energy ($\Sigma$) calculation are presented in appendix 6. The self-energy represents the effects on the finite device Hamiltonian due to the outgoing wavefunctions from an impulse excitation within the device [21]. It allows us to eliminate the huge S/D reservoirs and work solely within the device subspace whose dimensions are much smaller. The size of the self-energy matrix is $(N_X \times N_Z)^2$. On including the self-energies, the final form of the Green’s function matrix is

$$G[E(k_x, k_z)] = [E(k_x, k_z)I - h(x, z) - \Sigma_S - \Sigma_D]^{-1} \tag{3.7}$$

Note that the Green’s function in a real-space representation, has a size of $(N_X \times N_Z)^2$. 

Once the retarded Green’s function is evaluated, electron density and terminal current can be computed. We define a new quantity in terms of the lead self-energies.

\[ \Gamma \equiv i(\Sigma - \Sigma^\dagger) \] (3.8)

Physically the function \( \Gamma \) determines the electron exchange rates between the S/D reservoirs and the active device region [19]. But in general it can be viewed as the measure of interaction strength due to any perturbation source. It should be noted that, although the device itself may be in a non-equilibrium state, electrons are injected from the equilibrium S/D reservoirs (Fermi level is uniquely fixed for all carriers based on the applied voltage). The spectral density functions due to the S/D contacts can be obtained as [20]

\[ A_S = G\Gamma_S G^\dagger \quad \text{and} \quad A_D = G\Gamma_D G^\dagger \] (3.9)

where \( \Gamma_S \equiv i(\Sigma_S - \Sigma_S^\dagger) \), and \( \Gamma_D \equiv i(\Sigma_D - \Sigma_D^\dagger) \) (for clarity, we use \( \Gamma_S \) or \( \Gamma_D \) to denote matrices the same size as \( G \), with nonzero diagonal blocks \( \Sigma_S - \Sigma_S^\dagger \) or \( (\Sigma_D - \Sigma_D^\dagger) \)). The spectral functions are \( (N_X \times N_Z)^2 \) matrices. Although full in general, only the diagonal entries are of importance as they represent the state density at each lattice node. Therefore significant savings in computational cost are derived by solving for the diagonal blocks of the spectral functions as opposed to the entire matrix using a recursive algorithm [30]. The source-related spectral function is filled up according to the Fermi function in the source contact, while the drain-related spectral function is filled up according to the Fermi function in the drain contact. Both the source and drain spectral functions contribute to the 3D electron density, which, for each in-plane energy is [20],

\[
n [E(k_x, k_z)] = \frac{1}{2\pi ab} \int_0^{+\infty} D : \left[ f(\mu_S, E(k_x, k_z) + E_{k_j}) A_S + f(\mu_D, E(k_x, k_z) + E_{k_j}) A_D \right] dE_{k_j}
\] (3.10)
where \( f \) is the Fermi-Dirac statistics function, and \( D = (2/\pi \hbar)\sqrt{m^*_y/2E_{k_j}} \), represents the transverse mode state density (including spin degeneracy). Since the spectral functions (Eq. 3.9) depend on the in-plane energy alone, they can be moved out of the integral in Eq. 3.10 which reduces to [20],

\[
n[E(k_x, k_z)] = \frac{1}{ab} \sqrt{\frac{m^*_y k_B T}{2\pi^3 \hbar^2}} \times \left[ \mathfrak{F}_{-1/2} (\mu_S - E(k_x, k_z)) A_S + \mathfrak{F}_{-1/2} (\mu_D - E(k_x, k_z)) A_D \right]
\]

(3.11)

where the Fermi-Dirac integral, \( \mathfrak{F}_{-1/2} \), accounts for all transverse mode contributions (for an analytical approximation to \( \mathfrak{F}_{-1/2} \) see [37]). To obtain the total 3D electron density, we need to integrate Eq. 3.11 over \( E(k_x, k_z) \) and sum contributions from every conduction band valley. The 3D electron density is fed back to the Poisson equation solver for self-consistent solutions.

Once self-consistency is achieved, the terminal current can be expressed as a function of the transmission coefficient [19]. The transmission coefficient from the source contact to the drain contact is defined in terms of the Green’s function as [20],

\[
T_{SD} = \text{Trace}[\Gamma_S \Gamma_D G^\dagger]
\]

(3.12)

The transmitted current at each in-plane energy (including spin degeneracy) is,

\[
I[E(k_x, k_z)] = \frac{q}{\hbar^2} \sqrt{\frac{m^*_y k_B T}{2\pi^3}} \times \left[ \mathfrak{F}_{-1/2} (\mu_S - E(k_x, k_z)) - \mathfrak{F}_{-1/2} (\mu_D - E(k_x, k_z)) \right] T_{SD} [E(k_x, k_z)]
\]

(3.13)

The total current, like the 3D charge, is obtained by integrating over all \( E(k_x, k_z) \) and summing over all conduction band valleys.

### 3.2.2 Mode-space solution

In this simulation scheme, the Green’s function is solved for in a mode-space representation. These modes replace the \( \delta(z - z') \) dependence of the basis, when compared
to the real-space solution. This approach greatly reduces the size of the problem and provides sufficient accuracy when compared to full 2D spatial discretization. We begin by solving a 1D, \( z \) directed, effective mass equation for each vertical device slice along \( x \), to obtain a set of eigenenergies and eigenfunctions (modes) along the gate confinement direction. The equation that is solved is

\[
-\frac{\hbar^2}{2m^*_z} \frac{\partial^2}{\partial z^2} \Psi_i(x, z) - qV(x, z) \Psi_i(x, z) = E_i(x) \Psi_i(x, z)
\]  

(3.14)

where, \( m^*_z \) is the electron effective mass in the \( z \) direction, \( \Psi_i(x, z) \), the wavefunction, and \( E_i(x) \) the eigenenergy for subband \( i \) at slice \( x \) respectively. As with the real-space solution, the simulation domain in the confinement direction can be extended to include the insulator regions. Each vertical slice has a width, \( a \), and within each slice, all quantities are assumed to be constant in the \( x \) direction.

The 3D Hamiltonian for the device is expanded in terms of \( \delta(x - x')\Psi_i(x, z) \) and \( e^{ik_jy}/\sqrt{W} \). The new basis functions, \( \delta(x - x')\Psi_i(x, z) \) and \( e^{ik_jy}/\sqrt{W} \) also constitute a complete and orthogonal expansion functions set. The Hamiltonian in this representation is

\[
H = \begin{bmatrix}
    h[E_1(x) + E_{k_j}] & 0 & \cdots & \cdots & \cdots \\
    0 & h[E_2(x) + E_{k_j}] & 0 & \cdots & \cdots \\
    0 & \cdots & \ddots & \cdots & 0 \\
    0 & 0 & \cdots & h[E_i(x) + E_{k_j}] & 0 \\
    0 & 0 & 0 & 0 & \ddots
\end{bmatrix}
\]  

(3.15)

where,

\[
h_i = \begin{bmatrix}
    2t_x + E_i(1) + E_{k_j} & -t_x & 0 & \cdots \\
    -t_x & 2t_x + E_i(2) + E_{k_j} & \ddots & 0 \\
    \cdots & 0 & \ddots & -t_x \\
    \cdots & 0 & -t_x & 2t_x + E_i(N_X) + E_{k_j}
\end{bmatrix}
\]  

(3.16)

is the Hamiltonian for subband \( i \), with a planewave eigenenergy \( E_{k_j} \). The subband index \( (i) \) in Eq. 3.16 runs over all subbands (replaces \( z \) in real-space calculations). Numbers 1 to \( N_X \) in parenthesis replace the position \( x \) because of the discretization.
The block diagonal nature of the Hamiltonian in Eq. 3.15 indicates that in the ballistic limit, subband coupling is neglected in the mode-space solution scheme (see Sec. 3.4). In the case of bulk MOSFETs, inversion layer electrons become unconfined as we move towards the drain end of the device and strong electric fields near the drain couple different modes even in the ballistic limit. Therefore a mode-space solution, assuming decoupled modes is no longer applicable. Full real-space discretization provides the only accurate scheme to treat quantum ballistic transport in such devices.

Knowing the Hamiltonian for each subband, we write the retarded Green’s function relevant to 1D transport as \[ G(E) = \left[ EI - h \left[ E_i(x), E_{kj} \right] - \Sigma \right]^{-1} = \left[ E_l I - h \left[ E_i(x) \right] - \Sigma \right]^{-1} \] (3.17)

where, the longitudinal \((x)\) energy \(E_l \equiv E - E_{kj}\) (replaces \(E[k_x, k_z]\)) of the real-space solution). The self-energy for the leads (\(\Sigma\)) is a function of the longitudinal energy alone, and can derived based on the analysis in appendix 6 as

\[
\Sigma(E_l) = \begin{bmatrix}
-t_x e^{ik_lx} & 0 & \cdots \\
0 & \cdots & 0 \\
\cdots & 0 & -t_x e^{ik_Nx}
\end{bmatrix}
\] (3.18)

where, \(E = E_{kj} + E_i(n) + 2t_x(1 - \cos k_l a)\). The subband energy at the contact boundary (source or drain) is \(E_i(n)\) and the subscript \(l\) represents the longitudinal dependence of \(k\).

From a computational point of view, the size of the problem is measured by the size of the Hamiltonian. In a real-space representation the size of Hamiltonian is defined by the total nodal number in the 2D mesh, namely \((N_X \times N_Z)^2\); while in the mode-space representation, every subband is treated individually, and the size of the 1D Hamiltonian for each subband, is measured by the nodal number along the channel direction, namely \((N_X)^2\). In case of thin body, fully depleted SOI MOSFETs, strong body confinement causes the separation between modes to be large in energy. Therefore the Fermi level populates only a few modes even at high bias. Hence in practice, calculations, including the lowest few modes provide the desired accuracy.
This, coupled with the reduced size of the mode-space Hamiltonian, implies that the mode-space approach provides enormous savings in computational burden without loss in accuracy as compared to a real-space solution which implicitly treats all modes (including mode coupling).

The mode-space spectral density functions due to the S/D contacts are analogous to those in Eq. 3.9. They differ from the real-space solution in that their diagonal entries represent the local density of states at site $x_i$ for mode $i$. The 2D electron density for mode $i$ at a longitudinal energy $E_l$ is

$$n_i(E_l) = \frac{1}{a} \sqrt{\frac{m_e^* k_B T}{2\pi^3 \hbar^2}} \left[ \mathfrak{F}_{-1/2}(\mu_S - E_l)A_S + \mathfrak{F}_{-1/2}(\mu_D - E_l)A_D \right]$$  

The net 2D electron density for each mode is obtained by integrating Eq. 3.19 over $E_l$. This 3D electron density at each lattice node of our 2D real-space grid is obtained by multiplying $n_i$ with the corresponding distribution function $|\Psi_i(x, z)|^2/b$, and by summing over all subbands (index $i$) and conduction band valleys. Since the eigenvalue problem is solved exactly along the gate confinement direction, quantum effects associated with confinement and asymmetric gate design can be handled correctly within the mode-space modeling scheme. Once self-consistency is achieved, the terminal current is evaluated by summing contributions from each mode and conduction band valley. For a detailed description of the self-consistent solution scheme, see [40].

While presenting the self-consistent solution scheme in this section, the NEGF equations were simply stated without any rigorous derivation. Although a rigorous derivation of these equations is available in the literature ([17] [18] and the references therein), an intuitive understanding of these equations can be obtained by motivating their derivation from a wavefunction perspective. Therefore, appendix 6 translates key equations in a wavefunction description of the decoupled mode-space solution into the language of the NEGF formalism. Although appendix 6 indicates that the two methodologies are equivalent when modeling ballistic transport in ultra-thin body SOI MOSFETs, it should be noted that the NEGF formalism provides a general simulation platform that can be naturally extended to treat the effects of scattering
and to model carrier transport in novel devices such as carbon nanotube and molecular transistors.

### 3.3 Results

The simulated device structure (Fig. 3.1a) is an ultra thin body, fully depleted, symmetric, n-MOSFET with S/D regions doped at $10^{20} \text{cm}^{-3}$ and an intrinsic channel. The gate length is 10 nm, and there is no gate-to-S/D overlap. The junctions are abrupt, and the oxide thickness for both top and bottom gates is 1.5 nm. A body thickness of 1.5 nm, and a power supply ($V_{DD}$) of 0.6 V has been used in this simulation study. The gate work function (4.25 V) has been adjusted to yield a threshold voltage ($V_T$) of 0.15 V. Gate oxides are treated as infinite potential barriers for electrons in all of the simulations.

In order to highlight the quantum effects that one observes in nanoscale transistors and validate the simplified mode-space approach, we compare internal quantities from the real and mode-space solutions. It should be noted that the real-space solution, implicitly includes all modes and their coupling effects (if any). In Fig. 3.2a, we plot results from the solution to a 1D effective mass equation (modes), along slice $Z - Z$ (refer Fig. 3.1a) in the on-state ($V_{GS} = V_{DS} = 0.6V$). The local density of states (LDOS) spectrum vs. in-plane energy ($E[k_x, k_z]$), obtained from the real-space solution is also plotted alongside in Fig. 3.2b. Light areas in Fig. 3.2b represent regions of high state density while dark areas signify low state density. Spatially, the LDOS goes to zero at the silicon/oxide interface (infinite potential barriers at $z = 0$ and 1.5 nm) and exhibits single or multiple maxima points. If we compare this plot, to the result obtained from the first step of the mode-space solution (refer Fig. 3.1a and Fig. 3.2a), we find that the spatial behavior of the LDOS along $Z - Z$, is clearly captured by the mode-space solution. Note that, on superposing the mode energies onto the LDOS spectrum (dotted lines in Fig. 3.2b), we find that the maximum density of states occurs at in-plane energies that are higher that the corresponding subband
energy (non-classical behavior). This observation can be explained by examining Fig. 3.3, where we plot the classical and quantum 2D density of states (DOS) as a function of the in-plane energy for slice $Z - Z$ (note that the subband energy depends on the location of the slice in $x$).

In the classical case (inset), the 2D DOS ($x-z$ plane) is a convolution of discrete delta functions (subbands) with a 1D DOS, that has a $1/\sqrt{E(k_x)}$ dependence. This is because, in the classical case there is no information about the quantum mechanical coupling of the device to the S/D reservoirs. Since, $x$ is treated as a free dimension, the classical 2D DOS exhibits singularities around each subband energy. On the other hand, the real-space solution includes coupling information through the self-energy terms associated with the source and drain. These self-energies are composed of real and imaginary parts. The effect of the real part is to shift the maxima of the DOS in energy, while that of the imaginary part is to broaden the singularity in the classical DOS around each subband energy (dotted lines in Fig. 3.3) leading to a tail in the quantum DOS below each subband as shown in Fig. 3.3. This tail in the DOS is the cause of source-to-channel tunneling. Note that, all of the quantum effects in the channel direction, are a result of coupling the active device Hamiltonian to the S/D reservoirs through the self-energy terms.

In order to capture quantum effects along the channel within the mode-space solution, we couple each subband to the S/D reservoirs through specific self-energy terms calculated using Eq. 3.18. On introducing this coupling, we obtain the final form of the mode-space solution, whose LDOS is illustrated in Fig. 3.4. We plot the LDOS spectrum along slice $X - X$ (Fig. 3.1a). The conduction band (solid line) and the first subband profile (dotted line) along the channel, is superposed on this plot. With the inclusion of coupling information, this LDOS spectrum is identical to that obtained from the real-space solution for the energy range considered. The presence of a forbidden energy region between the conduction and first subband and the broadening of the LDOS around the first subband energy is clearly visible. It should be noted that the source-to-channel barrier is with respect to the subband profile as opposed
Fig. 3.2. (a) The solution to a 1D effective mass equation (step 1 of the mode-space solution) along slice $Z - Z$ (refer Fig. 3.1a) in the on-state ($V_{GS} = V_{DS} = 0.6V$). (b) The local density of states from the real-space solution along slice $Z - Z$. 
Fig. 3.3. The classical (inset) and quantum 2D density of states (DOS) along slice $Z-Z$ (refer Fig. 3.1) in the on-state. The position of the slice along the channel determines the subband energies (dotted lines).

to the conduction band edge and that tunneling occurs at energies much greater than the classical conduction band energy. States injected from the drain end of the device are reflected off a large barrier under high drain bias and interfere strongly. States injected from the source that have energies slightly above or less than the source barrier also interfere, resulting in the observed quantum oscillations in the LDOS. At energies much greater than the subband maxima, injected states are free and there is no visible quantum interference effect. The oscillations in the LDOS give rise to local oscillations in the 3D electron density, as charge density is a convolution of the S/D injected LDOS and the corresponding Fermi function (Eq. 3.10 and Eq. 3.19).

Figure 3.5, compares the $I_{DS}$ vs. $V_{DS}$ and $I_{DS}$ vs. $V_{GS}$ characteristics for our model device from real and mode-space solutions. It is clear from Fig. 3.5 that the two solution schemes are in close agreement with each other, thus indicating that the mode-quantum solution, which is computationally inexpensive (order of magnitude less in computational time), is an attractive simulation tool for modeling thin body, SOI n-MOSFETs in the ballistic limit. It should be noted that although the 3D charge exhibits local oscillations, the current vs. voltage characteristics from both
Fig. 3.4. Computed local density of states (LDOS) and the longitudinal subband profile (dotted line) along slice $X - X$ in the on-state. Broadening in the LDOS is due to quantum mechanical coupling to the S/D and oscillations are due to quantum mechanical reflections. The conduction band (solid line) is far below the subband due to confinement.

real and mode-space solutions are smooth. This is because current is a function of the transmission coefficient (Eq. 3.13), which depends on the overall potential profile from the source to the drain. Since local charge density oscillations are washed out when solving Poisson’s equation for the potential, the potential profile and hence current is a smooth function of the applied voltage.

The on-current spectrum vs. in-plane/longitudinal energy from real and mode space simulations is plotted in Fig. 3.6. The maximum of the first subband is also indicated (dotted line) to separate the thermionic and tunneling current components. The spectrum indicates that conduction in this thin body MOSFET is essentially through the first mode (only one peak). Also, exact agreement between the real and mode-space simulation results, highlights the validity of the approximations inherent in the decoupled mode-space model (Sec 3.4). Figure 3.6, indicates that tunneling carriers constitute $\sim 25\%$ of the simulated on-current. Therefore it is expected that a 1D Boltzmann treatment of a mode [29] [43], which does not include tunneling effects would under predict the on-current. To verify if this is the case, we compare
Fig. 3.5. $I_{DS}$ vs. $V_{GS}$ and $I_{DS}$ vs. $V_{DS}$ characteristics for the model device (refer Fig. 3.1a) from real (line) and mode-space (circles) solution schemes. Close agreement ($\sim 1\%$) between the two solutions validates the approximations inherent in the decoupled mode solution.

the self-consistent current-voltage characteristics, obtained from a quantum and a Boltzmann (classical) treatment of modes, in Fig. 3.7.

Figure 3.7 indicates that the simulated on-currents from the Boltzmann solution are remarkably close to, but slightly higher than those predicted by the quantum solution even though the tunneling component is missing in the classical solution. This close agreement is the result of self-consistency and can be understood by examining the subband profile and charge density along the channel from the two solution schemes (Fig. 3.8). Note that in the on-state, the 2D charge density at the subband maximum is prescribed by gate electrostatics irrespective of whether the charge is due to tunneling or thermionic emission. Therefore, the subband maximum in case of the Boltzmann solution is lowered to obtain roughly the same charge as in the quantum case. Also note that all of this charge is thermionic in nature in case of the Boltzmann model, whereas it has both tunneling and thermionic components in case of the quantum model. Since tunneling carriers have a lower velocity due to their lower longitudinal energy, the quantum model predicts a lower on-current compared to its classical analogue in the on-state.
Fig. 3.6. The energy distribution of the on-current from the real (solid line) and mode-space (circles) solution schemes. The top of the first subband (dotted line) is also indicated to separate the thermionic and tunneling current components.

Fig. 3.7. $I_{DS}$ vs. $V_{GS}$ and $I_{DS}$ vs. $V_{DS}$ characteristics for the model device (refer Fig. 3.1a) from a quantum (line) and Boltzmann (circles) treatment of the modes. The off-current from the quantum solution is higher than the Boltzmann solution due to source-to-channel tunneling. The on-current is lower as a result of self-consistency.
Fig. 3.8. The 2D electron density and first subband profile, along the channel from a quantum (solid line) and classical (dashed line) treatment of the modes, in the on-state. Charge at the top of the source-barrier is primarily prescribed by gate electrostatics. Crossover between the quantum and classical charge profiles, is due to quantum reflections and tunneling.

In the off-state, all of the current is due to tunneling. Therefore, the quantum model predicts a degraded subthreshold swing and higher off-current as compared to its classical counterpart. From a ballistic simulation viewpoint, it seems that in order to model the on-current accurately, a Boltzmann treatment of the mode along the channel direction is adequate. However, the main advantage of the mode-quantum solution is that the self-energy concept used to model the S/D contacts can, and has been extended to treat scattering [12]. Also, source-to-channel tunneling imposes a scaling limit on the channel length (for lengths below 10 nm). The Boltzmann solution cannot capture this tunneling limit. For a detailed comparison of the classical and quantum models in mode-space, when applied to thicker body transistors, refer to [28].

3.4 Discussion

The mode-space discretization of the Hamiltonian greatly reduces the size of the problem as compared to a full 2D spatial discretization ($N_x^2$ as opposed to $(N_x \times N_z)^2$). It is important to look at the conditions under which this approach provides
good simulation accuracy. In this section, we analytically expand the Schrödinger equation invoking the mode-space representation, and assess the approximations made in simplifying the Hamiltonian. This analysis explains why the decoupled mode-space approach provides the high degree of simulation accuracy in case of thin body DG and SG SOI MOSFETs.

We start with the 2D Schrödinger equation in the $x-z$ domain (the $y$ dimension is decoupled from the $x-z$ domain, and can therefore be treated separately)

$$-rac{\hbar^2}{2m^*_{x}} \frac{\partial^2}{\partial^2 x} \Phi(x,z) - \frac{\hbar^2}{2m^*_{z}} \frac{\partial^2}{\partial^2 z} \Phi(x,z) - qV(x,z)\Phi(x,z) = [E - E_k] \Phi(x,z) \quad (3.20)$$

Left multiplying the mode-space eigenvectors and performing the integration in real-space we obtain

$$\int [\delta^*(x - x')\Psi^*_i(x,z)] \left[ -\frac{\hbar^2}{2m^*_{x}} \frac{\partial^2}{\partial^2 x} \Phi(x,z) \right] dx \, dz + \int [\delta^*(x - x')\Psi^*_i(x,z)] \left[ -\frac{\hbar^2}{2m^*_{z}} \frac{\partial^2}{\partial^2 z} - qV(x,z) \right] \Phi(x,z) \, dx \, dz = [E - E_k] \int [\delta^*(x - x')\Psi^*_i(x,z)] \cdot \Phi(x,z) \, dx \, dz \quad (3.21)$$

The third term in Eq. 3.21 becomes,

$$[E - E_k] \int [\delta^*(x - x')\Psi^*_i(x,z)] \cdot \Phi(x,z) \, dx \, dz = [E - E_k] \tilde{\Phi}_i(x') \quad (3.22)$$

Note that $\tilde{\Phi}_i(x')$, is the expansion coefficient of $\Phi(x',z)$, with respect to the mode-space eigenvector $\Psi_i(x',z)$ as defined by

$$\Phi(x',z) = \sum_{i=1}^{\infty} \tilde{\Phi}_i(x') \Psi_i(x',z) \quad \text{and} \quad \int \Psi^*_i(x',z) \Psi_j(x',z) dz = \delta_{ij} \quad (3.23)$$

where, $\delta_{ij}$ is the usual Kronecker delta. We can rewrite the second term in Eq. 3.21 as,

$$\int \Psi^*_i(x',z) \left[ -\frac{\hbar^2}{2m^*_{z}} \frac{\partial^2}{\partial^2 z} - qV(x',z) \right] \Phi(x',z) \, dz = E_i(x') \tilde{\Phi}_i(x') \quad (3.24)$$

Finally the first term in Eq. 3.21 can be expressed as

$$\int \delta(x - x')\Psi^*_i(x,z) \left[ -\frac{\hbar^2}{2m^*_{x}} \frac{\partial^2}{\partial^2 x} \Phi(x,z) \right] \, dx \, dz$$
\[\begin{align*}
= & \int \delta(x - x') \left[ -\frac{\hbar^2}{2m^*_x} \frac{\partial^2}{\partial^2 x} \{\Psi^*_i(x, z)\Phi(x, z)\} \right] \, dx \, dz \\
- & \int \delta(x - x') \Phi(x, z) \left[ -\frac{\hbar^2}{2m^*_y} \frac{\partial^2}{\partial^2 x} \Psi^*_i(x, z) \right] \, dx \, dz \\
- & 2 \int \delta(x - x') \left[ -\frac{\hbar^2}{2m^*_x} \frac{\partial}{\partial x} \Psi^*_i(x, z) \frac{\partial}{\partial x} \Phi(x, z) \right] \, dx \, dz \tag{3.25}
\end{align*}\]

Note that the second term in Eq. 3.25 reduces to \(-\hbar^2/2m^*_x \partial^2/\partial^2 x' \tilde{\Phi}_i(x')\) after integration. If we assume that for all \(x\) (the shape of a mode does not change along the channel)

\[\frac{\partial \Psi^*_i(x, z)}{\partial x} = 0, \tag{3.26}\]

Equation (3.20) becomes,

\[-\frac{\hbar^2}{2m^*_x} \frac{\partial^2 \tilde{\Phi}_i(x')}{\partial^2 x'} + E_i(x') \tilde{\Phi}_i(x') = [E - E_{k_j}] \tilde{\Phi}_i(x'). \tag{3.27}\]

Equation 3.27 is the decoupled mode-space transformation of the 2D Hamiltonian invoking the assumption represented by Eq. 3.26. Equation 3.27 is indeed a 1D differential equation and greatly reduces the size of the original 2D problem. Note that Eq. 3.27 has two implications. The first is that subbands with different energies do not couple and the second is that some coupling information of a subband with itself is also lost.

Although the potential profile varies from the source to drain, if \(V(x, z)\) retains the same shape in the \(z\) direction, at different locations along the channel, the eigenfunctions are the same at each \(x\) location, even though the eigenvalues are different. As a result, Eq. 3.27 is satisfied. In the case of SOI MOSFETs with uniform thin bodies, there is little room for the potential to vary vertically. Therefore, Eq. 3.27 is a valid approximation and results obtained from the decoupled mode-space solution are in close agreement with real-space simulation results. This observed agreement between the real and decoupled mode-space solutions holds true in the case of thicker bodies (upto 5 nm) as long as the device has a uniform SOI geometry. For a wavefunction description of ballistic transport in decoupled mode-space see appendix 6.

If we perturb the uniformity of the potential profile by squeezing the channel region of our model device (Fig. 3.1b) and compare the real and mode-quantum
Fig. 3.9. $I_{DS}$ vs. $V_{GS}$ and $I_{DS}$ vs. $V_{DS}$ characteristics for the device with the squeezed channel (refer Fig. 3.1b) from mode-space (dashed line) and real-space (solid line) solution schemes. The mode-space solution does not account for mode coupling. Therefore, when the vertical potential profile is perturbed strongly, the decoupled mode-space solution exhibits inaccuracies.

solutions for a fixed potential profile, we see that the $I_{DS}$ vs. $V_{DS}$ characteristics of the real-space approach no longer agrees with that of the mode-space approach as indicated in Fig. 3.9. We know that in our model device transport is through the first subband. Therefore the mismatch between the real and mode-quantum solutions is because the mode-space solution does not completely capture the effect of a subband coupling with itself although a part of this information is built into the mode-space Hamiltonian as seen from its tridiagonal nature (Eq. 3.2). The differences in current at high $V_{DS}$ could be as high as $\sim 15\%$. Note that the current from the real-space solution is lower. It has been pointed out that including the flared out portions of the S/D contacts would have a similar effect and this reduction in current can be thought of as arising due to a quantum spreading resistance that is not captured by the mode-space solution [49]. In the case of bulk transistors, channel depletion widths can vary considerably from the source to drain, resulting in significant changes in the vertical confinement potential profiles. Therefore the mode-space approach becomes inappropriate for bulk device simulations or simulation of devices composed
of heterostructures. A coupled mode-space representation, which includes all of the terms in Eq. 3.25, would be appropriate for such structures [50].

3.5 Summary

We presented two approaches based on the NEGF formalism (real and decoupled mode-space), each with a different degree of complexity, that can be used to simulate 2D MOSFET structures under non-equilibrium conditions. These approaches were compared and contrasted by using them to simulate an ultra small DG n-MOSFET (with a body thickness of 1.5 nm). In doing so, quantum effects that are observed in nanoscale transistors, the treatment of open boundaries and the importance of self-consistency were highlighted. We showed that the real-space solution, which is the most general, is computationally expensive due to the 2D nature of the Hamiltonian; while the decoupled mode-space solution, which is specifically applicable to thin body, fully depleted, SOI device geometries is inexpensive (yet accurate) for two reasons: 1) A 1D Hamiltonian is used in the mode-space solution and 2) only few modes need to be considered as modes with high energies are not occupied by electrons and do not contribute to transport. For the model 1.5 nm body DG MOSFET, the simulation time per bias point on a one processor SUN workstation (300 MHz) was 40 secs in case of the mode-space solution as opposed to 1.5 hrs for the real-space solution. Finally, the validity of the mode-space solution and its regime of applicability were discussed by simulating a device with a squeezed channel region. These simulations indicated that including the flared out regions of the S/D contacts or simulating heterostructure devices at an effective mass level, would require a coupled mode-space simulation scheme.
4. A QUANTUM MECHANICAL TREATMENT OF SCATTERING IN NANOSCALE TRANSISTORS

4.1 Introduction

In all of the previous chapters, we discussed carrier transport in the ballistic limit. Real devices operate below this limit due to scattering. Therefore, this chapter focuses on extending the ballistic NEGF simulation scheme presented in the previous chapter, to include the effects of scattering in nanoscale transistors. We focus on the SOI device geometry and extend the decoupled mode-space solution (Sec. 3.2.2) to include the effects of scattering on electron transport. The self-energy concept (Eq. 3.18), previously used to capture coupling between the active device region and the semi-infinite leads is modified to simulate scattering as a perturbation to the device Hamiltonian.

Scattering within a device can be due to several reasons. Microscopically, electrons are confined within a very narrow channel and the channel is sandwiched between insulators in SOI devices. These insulator surfaces are never perfectly smooth, and the semiconductor lattice is never defect free. Therefore electrons in the channel experience surface roughness scattering. Both, channel carrier densities and impurity concentrations are typically very large. Therefore there is significant electron-electron and electron-impurity scattering within the device. Moreover, devices typically operate at relatively high temperatures (greater than 300 K) resulting in strong phonon-electron interactions. All these mechanisms need to be considered for an accurate treatment of scattering. Currently, solutions to the Boltzmann equation using Monte Carlo methods offer the best platform for modeling the effects of scattering in detail. Although comprehensive in their treatment of scattering, these simulation platforms do not treat quantum effects rigorously (they use the density gradient or effective
potential treatments [51] [52]). As critical device dimensions are shrunk quantum effects begin to manifest themselves strongly. These effects need careful consideration because they affect critical device performance metrics.

Within the NEGF formalism, a detailed treatment of the various scattering processes (e.g. surface roughness, phonon, impurity etc.) is possible. The first part of this chapter explains the procedure to extend the ballistic, decoupled mode-space technique to treat both, elastic as well inelastic scattering mechanisms at a detailed level within the NEGF simulation platform. However, due to the large computational cost involved in a detailed simulation, we use the insights obtained from detailed simulations to motivate an approximate treatment of scattering using the concept of Büttiker probes, in the second part of this chapter. Although introduced phenomenologically by Büttiker [53] [19], a comparison with the detailed NEGF scattering treatment provides insights and allows us some control over the approximation. Büttiker probes are used to simulate the effect of scattering due to all possible mechanisms and including the effect of degeneracy. Apart from its computational viability, the main advantages of the phenomenological scattering model are: 1) the use of a single parameter to treat scattering 2) the existence of an analytic relation between this parameter an equivalent low-field mobility (appendix 6). From a design perspective, it is important to be able to calibrate the parameters used in our quantum mechanical model to mimic an equivalent low field mobility because the low field mobility can be measured experimentally. Such a calibration enables us to relate our results to those obtained from conventional simulation tools, thus serving as a benchmark which can be used to validate simpler models. It also helps fit our parameters to measured device data in a meaningful fashion, thus enabling the application of device simulation to device design studies.

This chapter describes the numerical methods used to treat the effects of scattering transport with specific emphasis on the phenomenological, Büttiker probe based model. Section 4.2 presents the theory and simulation results for a detailed treatment
of scattering within the NEGF formalism and Sec. 4.3 presents the theory and simulation results for a phenomenological treatment of scattering using Büttiker probes.

4.2 A detailed scattering model

In the previous chapter we described the procedure to solve the NEGF equations in decoupled mode-space, self-consistently with the 2D Poisson equation in order to model ballistic transport in nanoscale SOI transistors. The overall procedure to solve these equations to capture the effects of scattering is similar. We briefly review the steps common to both ballistic and scattering transport models for the SOI transistor geometry (Fig. 3.1a).

1. A 2D solution to the Schrödinger equation is obtained by solving two 1D problems, one in the direction normal to the channel (z), which yields the electron distribution in the confinement direction and the subband profiles, and the other, along the channel direction based on the subband profiles, which yields the electron concentration in the transmission direction.

2. The 2D electron density for each subband is distributed over the silicon body (normal to the channel) according to the subband wavefunctions.

3. The 2D Poisson’s equation is solved using this electron density to obtain a new potential profile. This potential profile is used to solve the Schrödinger once again and the calculation cycle repeated until convergence is achieved.

4.2.1 Theory

In SOI transistor structures considered in this thesis (Fig. 3.1a), the silicon film is thin and wide. Therefore, electrons propagating from the source to the drain can be modeled as a 2D gas. This implies that in order to treat the effects of scattering in detail, two coordinates in energy need to be used. The first coordinate is the electron energy associated with the transport dimension (x in Fig. 3.1a), which will
be referred to as the longitudinal energy \(E_l\) and the second coordinate is the electron energy associated with the width dimension \(y\) in Fig. 3.1a), which will be referred to as the transverse energy \(E_{k_j}\). These components, along with the total electron energy \(E = E_l + E_{k_j}\), are illustrated in Fig. 4.1a. In the ballistic case (chapter 3) both, the longitudinal and the transverse energy of each electron was preserved as it propagates from the source to the drain (or vice versa). Therefore, it was possible to decouple these energy components and account for contributions from the transverse modes analytically, through the use of Fermi Dirac integrals (as in Eqs. 3.11 and 3.13). However, once scattering is turned on, the longitudinal and transverse energy components are no longer decoupled. Isotropic elastic scattering mixes electrons in states with the same total energy, while isotropic inelastic (phonon) scattering mixes electrons in states with different total energies as illustrated in Fig. 4.1a. Therefore, contributions from transverse modes cannot be simply accounted for as in the ballistic case, thus greatly increasing the computational burden associated with a detailed treatment of scattering.

Our treatment of scattering is similar to that used by Lake and Datta in their work on simulating resonant-tunneling diodes [21]. We begin our calculations by assuming that,

1. The scattering rates \(\tau^{\text{in}}, \tau^{\text{out}}\) or the alternately, the scattering functions \(\Sigma^{\text{in}}, \Sigma^{\text{out}}\) are dependent only on the total energy of the electrons \(E = E_l + E_{k_j}\). Note once again that \(E_{k_j}\) is the transverse electron energy.

2. The scattering rates are functions of the longitudinal coordinate \(x\) but not the transverse coordinate \(y\), implying that plane waves are the transverse eigenfunctions of the Hamiltonian even in the presence of scattering.
The calculation begins with the device Hamiltonian for subband $i$ at a transverse energy of $E_{k_j}$ in decoupled mode-space. In discrete form this Hamiltonian is,

$$H_i[h_i] = \begin{bmatrix}
2t_x + E_i(1) + E_{k_j} & -t_x & 0 & \cdots \\
-t_x & 2t_x + E_i(2) + E_{k_j} & \ddots & 0 \\
\vdots & \ddots & \ddots & -t_x \\
\vdots & 0 & -t_x & -2t_x - E_i(N_x) + E_{k_j}
\end{bmatrix}$$

(4.1)

where $h_i$ is the discretized version of the left hand side of Eq. 3.27 on a finite difference grid. Note that this equation is identical to Eq. 3.16 and that the matrix in Eq. 4.1 has a size of $N_x^2$, where $N_x$ is the number of grid points along $x$ (Fig. 3.1a). The quantity we are interested in is the retarded Green’s function for subband $i$, which can be expressed as,

$$G_i(E_l, E_{k_j}) = [E_l - H_i - \Sigma_{S/D}(E_l) - \Sigma_{Scatt}(E)]^{-1}$$

$$= [E_i I - h_i - \Sigma_{S/D}(E_l) - \Sigma_{Scatt}(E)]^{-1}$$

(4.2)

where, $E_l = E - E_{k_j}$, $\Sigma_{S/D}$, is the self-energy matrix which describes the coupling between the active device and the semi-infinite S/D leads (Eq. 3.18) and $\Sigma_{Scatt}$ is the self-energy matrix which accounts for the effects of both elastic and inelastic scattering processes within the device.

In this work we assume that scattering is a local interaction. Therefore, the self-energy matrix, $\Sigma_{Scatt}(E)$ (in Eq. 4.2), is diagonal for both elastic as well as inelastic scattering processes. The elements of the self-energy matrix for scattering can be obtained by defining a broadening function as,

$$\Gamma_{Scatt}[E] = j \left[ \Sigma_{Scatt}(E) - \Sigma_{Scatt}^\dagger(E) \right]$$

(4.3)

where $j = \sqrt{-1}$. This function was first introduced in chapter 3 (Eq. 3.9), to describe the electron exchange rate ($\Gamma_{S/D}/\hbar$) between the the active device and the S/D reservoirs in the ballistic limit. Here, $\Gamma_{Scatt}/\hbar$, describes the rate at which electrons
in or outscatter into a state with a total energy $E$ due to elastic and (or) inelastic scattering processes. Since $\Gamma_{\text{Scatt}}(E)$ is a measure of the net scattering rate, it can be expressed as a sum of in and out scattering functions [19].

$$\Gamma_{\text{Scatt}}[E] = \Sigma_{\text{in}}^{\text{Scatt}}[E] + \Sigma_{\text{out}}^{\text{Scatt}}[E]$$

(4.4)

Invoking the self-consistent Born approximation, the in and out scattering functions for isotropic elastic scattering and isotropic inelastic scattering (due to a single optical phonon with an energy of $\hbar\omega$) can be expressed as [19] [45],

$$\Sigma_{\text{in}}^{\text{elastic}}[E] = D_{\text{elastic}} \cdot G_n[E]$$

$$\Sigma_{\text{out}}^{\text{elastic}}[E] = D_{\text{elastic}} \cdot G_p[E]$$

(4.5)

and

$$\Sigma_{\text{in}}^{\text{inelastic}}[E] = D_{\text{inelastic}} \cdot [(N\omega + 1)G_n(E + \hbar\omega) + N\omega G_n(E - \hbar\omega)]$$

$$\Sigma_{\text{out}}^{\text{inelastic}}[E] = D_{\text{inelastic}} \cdot [(N\omega + 1)G_p(E - \hbar\omega) + N\omega G_p(E + \hbar\omega)]$$

(4.6)

where, $D$, is a constant representing the interaction strength felt by electrons due to elastic and inelastic scattering processes, $G_n(E)$ is the electron correlation function, $G_p(E)$ is the hole correlation function and $N\omega$, is the number of phonons of energy $\hbar\omega$ in equilibrium with the semiconductor lattice (at a specific temperature). Note that when expressed in a real-space basis, the diagonal entries of $G_n(E)/2\pi$ and $G_p(E)/2\pi$ in Eqs. 4.5 and 4.6 represent the net electron and hole density spectra vs. total energy at each grid point on the finite difference grid (along $x$ in Fig. 3.1a). The matrices, $G_n(E)$ and $G_p(E)$ have the same size as the Hamiltonian for mode $i$ ($N_X^2$).

Also note that Eq. 4.6 assumes that the phonon bath is in equilibrium and that the phonon number at a specific temperature is given by the Bose-Einstein statistic.

$$N\omega = \frac{1}{e^{\frac{\hbar\omega}{k_B T}} - 1}$$

(4.7)

where $T$ is the lattice temperature. Equation 4.6 can be intuitively understood by examining the cartoon in Fig. 4.1b (left). Electrons can be scattered into empty states
Fig. 4.1. (a) The longitudinal ($E_l$), transverse ($E_{kj}$) and total ($E$) energy components are indicated. Note that elastic scattering (dotted arrows) mixes states with the same total energy. Inelastic scattering due to phonons (dash dotted arrows) mixes states with total energies separated by $\hbar \omega$, where $\hbar \omega$, is the energy of the phonon. In the ballistic case $E_l$ and $E_{kj}$ can be treated independently and contributions from different transverse energies can be summed analytically at each $E_l$. This is not the case once scattering is turned on. (b) Inelastic scattering due to a single phonon with energy $\hbar \omega$, is pictorially represented. Phonon emission and absorption affects the in and out scattering functions ($\Sigma^{\text{in}}$ and $\Sigma^{\text{out}}$) for a state with a total energy $E$. 
at a total energy $E$ from filled states at an energy of $E + \hbar \omega$ by the emission of an optical phonon with energy $\hbar \omega$. The in scattering strength due to this mechanism is proportional to the interaction strength, the electron density at $E + \hbar \omega$ and the number of phonons plus one ($N_\omega + 1$). Electrons with a total energy of $E - \hbar \omega$ can also scatter into empty states at energy $E$ due to phonon absorption. The in scattering strength for this mechanism is again proportional to the interaction strength, the electron density at energy $E - \hbar \omega$ and the phonon number ($N_\omega$). The total in scattering strength at energy $E$ is therefore a sum of the two contributions. The net out scattering strength can be similarly understood from the cartoon in Fig. 4.1b (right).

Knowing the in and out scattering strengths (Eqs. 4.5 and 4.6) for elastic and inelastic scattering process, one can use Eqs. 4.3 and 4.4 to derive an expression for the imaginary part of the scatterer self-energy (to be used in Eq. 4.2).

$$\text{Imag}[\Sigma_{\text{Scatt}}(E)] = -\frac{1}{2}[\Sigma_{\text{in}}(E) + \Sigma_{\text{out}}(E)]$$ (4.8)

Since the self-energy matrix, $\Sigma_{\text{Scatt}}(E)$, is causal, the real part of the self-energy matrix can be derived from its imaginary part (Eq. 4.8) through a Hilbert transformation. However, it has been demonstrated that the real part of the self-energy matrix (which generates a potential shift) is minor compared to the Hartree potential obtained from Poisson’s equation [54]. Therefore, when modeling scattering transport we ignore the real part of $\Sigma_{\text{Scatt}}$ and retain only its imaginary part. Equation 4.8 is used to evaluate the retarded Green’s function for mode $i$ using Eq. 4.2. Knowing the Green’s function, we can calculate the electron and hole correlation functions for mode $i$ using Eqs. 4.5 and 4.6 [19].

$$G_{ni}^n[E_l, E_{kj}] = G_i[E_l, E_{kj}][\Sigma_{S/D,i}^{\text{in}}(E_l, E_{kj}) + \Sigma_{\text{Scatt}}^{\text{in}}(E)] G_{ni}^\dagger[E_l, E_{kj}]$$

$$G_{pi}^p[E_l, E_{kj}] = G_i[E_l, E_{kj}][\Sigma_{S/D,i}^{\text{out}}(E_l, E_{kj}) + \Sigma_{\text{Scatt}}^{\text{out}}(E)] G_{pi}^\dagger[E_l, E_{kj}]$$ (4.9)

Note that the in and out scattering functions in Eq. 4.9, have two components: 1) Due to both momentum and energy relaxing scattering process within the device.
(Fig. 4.1b), and 2) Due to a coupling between the active device and the semi-infinite S/D leads. Since the S/D contacts are treated as ideal reservoirs, equilibrium statistics can be imposed at the ends of the active device, resulting in the following expressions for the in and out scattering functions due to the S/D (for mode $i$) [19].

$$
\Sigma_{\text{in}}^{\text{S/D},i}[E_l, E_{k_j}] = \begin{bmatrix}
\Gamma_{S,i}(E_l)f(E - \mu_S) & 0 & 0 & \cdots \\
0 & \cdots & \cdots & 0 \\
0 & \cdots & \cdots & 0 \\
0 & \cdots & \cdots & \Gamma_{D,i}(E_l)f(E - \mu_D)
\end{bmatrix}
$$

$$
\Sigma_{\text{out}}^{\text{S/D},i}[E_l, E_{k_j}] = \begin{bmatrix}
\Gamma_{S,i}(E_l)[1 - f(E - \mu_S)] & 0 & 0 & \cdots \\
0 & \cdots & \cdots & 0 \\
0 & \cdots & \cdots & 0 \\
0 & \cdots & \cdots & \Gamma_{D,i}(E_l)[1 - f(E - \mu_D)]
\end{bmatrix}
$$

(4.10)

where, $\Gamma_{S,i}$ and $\Gamma_{D,i}$ are the broadening functions due to coupling between the active device and the S/D leads (Eq. 3.8) and the $f$’s are the S/D Fermi distribution functions. Note that only the first and last slices of the active device couple to the S/D leads. Therefore, the in and out scattering matrices in Eq. 4.10, have non-zero entries at indices 1 and $N_X$.

On substituting Eqs. 4.5, 4.6 and 4.10, into Eq. 4.9, we obtain the electron and hole correlation functions for mode $i$ at each longitudinal and transverse energy. In order to obtain the electron and hole density spectra vs. the total energy $E$, for mode $i$, Eq. 4.9 needs to be integrated along lines of constant total energy in the $E_l$-$E_{k_j}$ plane as shown in Fig. 4.1a. Such an integration results in,

$$
G_i^{\text{in}}[E] = \int_0^\infty \frac{2}{\pi \hbar} \sqrt{\frac{m^*_{y,i}}{2E_{k_j}}} G_i(E_l, E_{k_j}) \Sigma_{\text{in}}^{\text{S/D},i}(E_l, E_{k_j}) G_i^\dagger(E_l, E_{k_j}) dE_{k_j}
$$

$$
G_i^{\text{out}}[E] = \int_0^\infty \frac{2}{\pi \hbar} \sqrt{\frac{m^*_{y,i}}{2E_{k_j}}} G_i(E_l, E_{k_j}) \Sigma_{\text{out}}^{\text{S/D},i}(E_l, E_{k_j}) G_i^\dagger(E_l, E_{k_j}) dE_{k_j}
$$

(4.11)
where $m_{y,i}^*$ is the $y$ directed (Fig. 3.1a) effective mass for electrons in mode $i$. Note that singularities exist in Eq. 4.11. These singularities are due to the 1D DOS associated with the transverse modes along the device width ($y$) (Eq. 4.11 includes spin degeneracy). These singularities, which pose challenges to the self-consistent solution, are eliminated by transforming variables and by using a non-uniform energy grid. Although the upper limit on the integral in Eq. 4.11 is $+\infty$, in practice the integration limit is restricted to a finite value, namely $E_{k_j}(\text{max})$, which ensures that the longitudinal electron energy is far below the subband minimum when $E_{k_j}$ is greater than $E_{k_j}(\text{max})$.

The net electron and hole correlation functions at a specific total energy $E$, are obtained by summing Eq. 4.11 over all conduction band valleys and subbands. Diagonal entries of these correlation functions are used in Eqs. 4.5 and 4.6. Equations 4.2, 4.5, 4.6, 4.8 and 4.11 constitute a self-consistent loop. The calculation begins with an initial guess for the in and out scattering functions for both, elastic and inelastic scattering processes (Eqs. 4.5 and 4.6). Knowing these functions, the retarded Green’s function is evaluated at each point in the $E_l-E_{k_y}$ plane (Fig. 4.1a) using Eqs. 4.8 and 4.2. These Green’s functions are used to evaluate the electron and hole correlation functions for mode $i$, using Eqs. 4.9 and 4.11. The correlation functions obtained by summing Eq. 4.11 over all the subbands and valleys are used to evaluate a new guess for the in and out scattering functions in Eqs 4.5 and 4.6. The entire calculation cycle is repeated until self-consistency is achieved (the scatter self-energy converges). Since the scatterer self-energy matrices used in Eq. 4.2, include contributions from all valleys and subbands, our treatment of scattering accurately captures both inter subband and inter valley scattering effects. Diagonal elements of the converged electron correlation functions for each mode (Eq. 4.9) yield the 2D electron density at each point in the $x$ (Fig. 3.1a) direction ($n_i = G_{i}^n/2\pi a$). This 2D charge density is distributed in the confinement direction ($z$) according to the corresponding mode eigenfunctions to obtain the 3D electron density at each node of the $x-z$ grid. This 3D charge is summed over all the valleys and subbands and input to Pois-
son’s equation to obtain a refined 2D electrostatic potential profile. This refined 2D electrostatic potential profile is used to reset the decoupled mode-space Hamiltonian (Eq. 4.1) and the entire charge density calculation is repeated until the electrostatic potential converges to within set tolerance limits. Thus, the solution scheme to treat scattering in detail includes two self-consistent loops (a self-consistent loop to model electron transport including scattering effects and an outer Poisson’s loop) as shown Fig. 4.2, which illustrates the flow of the algorithm.

Once self-consistency is achieved (ie: the self-energy matrix for elastic and inelastic scattering and the electrostatic potential profile has converged), the S/D current spectra for mode $i$ vs. the total electron energy $E$, is calculated using,

$$I_S[E] = \frac{q}{h} \int_{0}^{\infty} \frac{2}{\pi \hbar} \sqrt{\frac{m^{*}_{y,i}}{2E_{k_j}}} \times$$

$$\text{Trace} \left[ \Sigma^{\text{in}}_{S,i}(E_l, E_{k_j}) G^P_{i}(E_l, E_{k_j}) - \Sigma^{\text{out}}_{S,i}(E_l, E_{k_j}) G^H_{i}(E, E_{k_j}) \right] dE_{k_j}$$

$$I_D[E] = \frac{q}{h} \int_{0}^{\infty} \frac{2}{\pi \hbar} \sqrt{\frac{m^{*}_{y,i}}{2E_{k_j}}} \times$$

$$\text{Trace} \left[ \Sigma^{\text{in}}_{D,i}(E_l, E_{k_j}) G^P_{i}(E_l, E_{k_j}) - \Sigma^{\text{out}}_{D,i}(E_l, E_{k_j}) G^H_{i}(E, E_{k_j}) \right] dE_{k_j}$$

where the integration over the transverse modes ($E_{k_j}$) is once again performed along lines of constant total energy $E$, in the $E_l - E_{k_j}$ plane (Fig. 4.1a). These spectra are visualized and discussed in Sec. 4.2.2. The net source to drain (drain to source) current is obtained by integrating Eq. 4.13 over the total energy $E$ and by summing contributions from all valleys and subbands.

Before we conclude this section, we note that in order to obtain a well converged self-energy matrix for elastic and inelastic scattering processes, and a well converged electrostatic potential profile, the integrals in Eq. 4.11 have to be performed on a fine energy grid. Typical values of the grid spacing used in our calculations lie between $\sim 0.5$-$0.8$ meV along both, the longitudinal and the transverse energy axes (Fig. 4.1a). For devices studied in this thesis, the typical number of grid points in the $x$ dimension is $\sim 120$ and the power supply voltage is $\sim 0.4$ V. Even under the assumption of single
Fig. 4.2. The flowchart of the algorithm used to treat both, elastic and inelastic scattering in a 2D electron gas.
mode occupancy, a single non-selfconsistent (wrt. both the self-energy loop and the Poisson loop) iteration cycle requires the inverse of \( \sim 1000 \times 1000 \) \((E_l - E_k)\) grid), \( \sim 120 \times 120 \) matrices. Therefore, simulating the effects of scattering in detail poses a huge computational challenge. All of the detailed scattering calculations presented in this chapter were performed on a Linux cluster using Matlab-6.1. Parallelization of the energy grid was achieved by using an MPI (Message Passing Interface) toolbox for Matlab, developed at the University of Granada in Spain [55].

4.2.2 Results

The simulated device structure (Fig. 3.1a) is an ultra-thin body, fully depleted, symmetric, dual gate n-MOSFET with the S/D regions doped at \(10^{20} \text{cm}^{-3}\) and an intrinsic channel. The gate length is 10 nm and there is no gate-to-S/D overlap. The S/D extensions are 10 nm and the junctions are abrupt. In order to highlight quantum effects and the effect of scattering transport, we choose a thin silicon body (1.5 nm), which exhibits single mode occupancy. The oxide thickness (1 nm) and power supply voltage \((V_{DD} = 0.4 \text{ V})\) are set based on the ITRS [2]. We adjust the gate workfunction for both the top and bottom gates to obtain a ballistic off-current of \(10 \mu A/\mu m\), consistent with the ITRS requirement. Gate oxides are treated as infinite potential barriers for electrons in all of our simulations. For a detailed treatment of scattering, the input parameter is the perturbation strength \(D\), in Eqs. 4.5 and 4.6. Since values of this parameter are difficult to find in the literature (especially for 2D carriers), we choose a reasonable perturbation strength for both, elastic and inelastic scattering processes, which ensures that the scattering time \((\Gamma_{\text{Scatt}}/\hbar, \text{Eq. 4.3})\) is comparable to the semiclassical device transit time at carrier energies around the source Fermi level (0 eV). Also, since our primary objective is to derive qualitative insight from a rigorous treatment of scattering, results presented in this section are from non-selfconsistent simulations (wrt. Poisson’s equation) that use fixed electrostatic potential profiles. These profiles were obtained from self-consistent, ballistic simulations.
Fig. 4.3. The self-consistently calculated (ballistic) subband profile in the on-state ($V_{DD} = 0.4V$) for a double-gate MOSFET with a silicon film thickness of 1.5 nm. The S/D regions are degenerately doped ($1 \times 10^{20} \text{cm}^{-3}$) and the channel length (demarcated by the dashed vertical lines) is 10 nm. The number of elastic scatterers is progressively from the drain to the top of the source to channel barrier as indicated by the arrow in order to generate Fig. 4.7

In order to obtain qualitative insight and demonstrate the rigorous scattering model, we compare internal quantities with and without scattering within our model device. Figure 4.4a illustrates the current spectrum vs. total electron energy (Eq. 4.13) from both, ballistic as well as elastic scattering simulations (the subband profile for these calculations is shown in Fig. 4.3). In the ballistic limit, electrons enter the device from the source and leave through the drain. As they do so, both, the longitudinal as well as the transverse energy components are preserved for each electron (Fig. 4.1a). Therefore, the ballistic source and drain current spectra are symmetric when plotted vs. the total energy ($E$). In Fig. 4.4a. Also, since the number of electrons is conserved throughout the device, the net source current (integrated over the total energy coordinate) has the same magnitude as the net drain current (while reading this plot, it should be noted that source injected current is positive for electrons entering the device, while the drain collected current is negative for electrons leaving the device). If we compare this ballistic spectrum to the one obtained from scattering simula-
tions (with elastic scattering turned on), we observe that although the source and drain current spectra continue to remain symmetric, the magnitude of the current is reduced. This behavior can be explained by noting that although elastic scattering processes preserve the total electron energy \( E \) (Fig. 4.1a), they cause a degradation of the channel directed energy (momentum) \( E_l \), which in turn reduces the on-current when compared to the ballistic limit. This effect is illustrated in Fig. 4.4b, where the current spectra at the source and drain ends are plotted vs. the channel directed energy, \( E_l \). Figure 4.4b clearly indicates that electrons leave the device with reduced longitudinal energies (when compared to the source injected spectrum) as a result of elastic scattering within the active device region. Note once again that although the source and drain current spectra vs. \( E_l \) are not symmetric, scattering preserves the number of electrons within the device. Therefore the integrated source injected and drain collected currents still sum to zero as in the ballistic case.

Next, we present results which illustrate the effect of electron-phonon interactions on the on-state of a transistor (the subband profile used in these calculations is plotted in Fig. 4.3). We assume that inelastic scattering is caused by optical phonons with an energy of 63 meV. Note that since the optical phonon energy is high, the number of optical phonons (Eq. 4.7) present within the device is very small (\( \sim 0.1 \)) even at room temperature (300 K). Therefore, in and out scattering events due to phonon emission (the scattering strengths for phonon emission are proportional to \( N_\omega + 1 \) as shown in Eq. 4.6) are favored when compared to in and out scattering events due to phonon absorption (the scattering strengths for phonon absorption are proportional to \( N_\omega \)). This is clearly demonstrated in Fig. 4.5a, where we plot the current spectrum vs. the total electron energy (Eq. 4.13) from both, ballistic and phonon scattering simulations. The drain current spectrum vs. the total electron energy is relaxed when compared to the source current spectrum once inelastic scattering is turned on, because of phonon emission events. Since the scattering strength used in these calculations is not large (the on-current with scattering is \( \sim 65\% \) of the ballistic value), Fig. 4.5a indicates that many of the source injected electrons con-
Fig. 4.4. (a) Simulated current spectra at the source and drain ends vs. the total electron energy $E$, for the subband profile shown in Fig. 4.3 from both ballistic (solid lines) and elastic scattering (dashed lines) simulations. (b) Simulated current spectra at the source and drain end vs. the longitudinal electron energy $E_L$, for the subband profile shown in Fig. 4.3 from both ballistic (solid lines) and elastic scattering (dashed lines) simulations.
continue to retain their high energy values even at the drain. In fact, the minima of the drain current spectrum is displaced from the maxima of the source injected spectrum along the total energy axis by approximately 63 meV, thus implying that most of the source injected electrons have undergone a single phonon emission event. Inelastic scattering also strongly relaxes the directed momentum of the source injected electron stream as shown in Figure 4.5b, where we plot the source and drain current spectra vs. the longitudinal electron energy, $E_l$. As in the case of elastic scattering, electrons leave the device with reduced channel directed energies and the current spectrum at the drain is uniformly smeared out over the entire $E_l$ axis. At low values of $E_l$, the drain current spectrum is cutoff by the subband edge, and at high values of $E_l$ it is cutoff by the source Fermi function as shown in Fig. 4.5b. In subsequent sections we will demonstrate that all of these aforementioned features in the current spectra are captured by our simple phenomenological treatment of scattering, which is based on the Büttiker probe concept.

Figure 4.6, plots the current spectrum vs. the longitudinal electron energy from both, ballistic as well as elastic scattering simulations in the off-state (the subband profile used in this calculation is shown in the inset of Fig. 4.6). Note that scattering introduces a broadening in the LDOS (Eq. 4.2) even at longitudinal energies that are below the source-to-channel barrier. Therefore, the source injected current spectrum from the scattering simulation exhibits an increased tunneling leakage at longitudinal energies below the source-to-channel barrier when compared to the ballistic current spectrum. However, it should be noted that since the scatterer self-energies for both, elastic and inelastic scattering processes are self-consistently (Eqs. 4.11, 4.5, 4.6, 4.3 and 4.2) calculated from the electron and hole correlation functions, the broadening in the LDOS below the source-to-channel barrier is not uncontrolled. The self-consistent loop guarantees that the scatterer self-energy smoothly decays to zero as the LDOS reduces to zero in the classically forbidden regions below the source-to-channel barrier (Fig. 4.6). Therefore, the ballistically calculated off-current ($10\mu A/\mu m$) is always higher than the off-current obtained from scattering simulations ($6.7\mu A/\mu m$). Bal-
Fig. 4.5. (a) Simulated current spectra at the source and drain ends vs. the total electron energy $E$, for the subband profile shown in Fig. 4.3 from both ballistic (solid lines) and inelastic scattering (dashed lines) simulations. Inelastic scattering is assumed to be caused by the emission and absorption of phonons with an energy of 63 meV. (b) Simulated current spectra at the source and drain ends vs. the longitudinal electron energy $E_L$, for the subband profile shown in Fig. 4.3 from both ballistic (solid lines) and inelastic scattering (dashed lines) simulations.
Fig. 4.6. Simulated current spectra at the source and drain ends vs. the longitudinal electron energy $E_l$, in the off-state from both ballistic (solid lines) and elastic scattering (dashed lines) simulations. The self-consistently calculated subband profile in the off-state is shown in the inset. Note that although scattering broadens the local density of states below the source to channel barrier (dotted vertical line), thus resulting in increased source to channel tunneling, the off-current from ballistic simulations is still higher. Ballistic simulations set an upper limit on both the off and the on current values.

Ballistic simulations set an upper limit on both, the off and the on current in nanoscale transistors. We demonstrate that this behavior of the off-current is also captured by our simple phenomenological scattering model.

4.2.3 Discussion

In this section, we use our detailed scattering model to verify a simple analytical theory that describes the essential physics of scattering in nanoscale transistors [56]. This theory will be used to ascertain the validity of the phenomenological Büttiker probe based scattering models that will be presented in the following section. To understand the essential physics of scattering, we consider the on-state of a MOSFET (Fig. 4.3) with electrons thermally injected from the source, undergoing scattering in the channel and being collected by the drain. According to [56], although scattering
can occur anywhere inside the device, only those scattering events that occur in the low field region near the source have the largest effect on the on-current. To understand why the importance of backscattering reduces from the source to the drain, we consider a single electron injected from the source into the channel with a total energy, $E$ (note that $E_l$ is its longitudinal energy). Figure 4.7a, tracks this electron in 2D momentum space (initial state is the solid arrow in Fig 4.7a). Now assume that this electron undergoes a single elastic, isotropic scattering event in the channel (final state is the dotted arrow in Fig 4.7a). For this electron to make it ballistically back to the source, its longitudinal energy should be greater than $\epsilon_{\text{max}}$ (Fig. 4.3). It is clear from Fig. 4.7a, that only a small cone of electrons have enough longitudinal energy to backscatter into the source and that this cone reduces as we move towards the drain ($\epsilon_{\text{max}}$ increases relative to the local subband edge, and the dotted circle approaches the solid circle in Fig. 4.7a). Therefore, scattering near the source affects the current more strongly than near the drain.

Figure 4.7b shows the effect of scatterer placement on the device performance from the detailed, elastic scattering model. This figure was generated by fixing the potential profile (Fig. 4.3) and increasing the number of scatterers progressively from the drain towards the source. When the scatterer number is zero, the detailed scattering model yields the ballistic on-current, but as the scatterer number is increased, the simulated on-current slowly reduces initially and then drops sharply thereafter. Note that our detailed treatment of elastic scattering causes the longitudinal energy of the source injected electrons to relax, as discussed in Sec. 4.2.2. Therefore, electrons scattering near the drain lose much of their longitudinal energies, thus making it difficult for a backscattered electron to reflect back to the source. Hence, scattering near the drain affects the on-current mildly. However, as the number of scatterers increases towards the source (black dots in Fig. 4.3), electrons scattering very near the source-to-channel barrier do backscatter back into the source as they have not dissipated enough longitudinal energy, thus reducing the on-current significantly. It should be noted that our detailed scattering model considered the effect of elastic scattering alone, and
Fig. 4.7. (a) A pictorial representation of the essential physics of scattering (Also refer Fig. 4.3). The fraction of the scattered electrons that can surmount the source-to-channel barrier (Fig. 4.3) and make it back into the source, reduces as we move towards the drain. These carriers, whose total energy is $E$, are delineated by the cone. (b) The behavior of the on-current as a function of the scatterer number, which is progressively increased from the drain to the source (Fig. 4.3) as predicted by detailed, elastic scattering simulations. This plot indicates that scattering events occurring near the source end of the transistor degrade the on-current more severely as compared to scattering events occurring at the drain end, thus validating the essential physics of scattering.
that the addition of any other scattering mechanism serves to relax the longitudinal energy even more. This further reduces the probability (non-self-consistently) of an electron backscattering all the way to the source from the drain end of the device. The behavior of the on-current vs. the number of scatterers, obtained from rigorous scattering simulations clearly verifies the essential physics presented in [56].

4.3 A simple treatment of scattering using Büttiker probes

In the previous section we extended the ballistic NEGF simulation scheme presented in chapter 3, to include the effects of elastic and inelastic scattering processes at a detailed level. Although insightful, a detailed treatment of scattering within the NEGF formalism is computationally demanding and cannot be applied to analyze transistor design issues within a reasonable period of time. Also, the scattering strengths used to characterize phase and energy relaxing scattering processes cannot be easily mapped onto measured low-field mobility data. Therefore, a comparison to semiclassical simulation results or to experiments is difficult when scattering is simulated at a detailed level. A simpler treatment of scattering, which is computationally efficient and captures quantum effects that are important in SOI transistor structures is clearly of interest. This section presents a phenomenological treatment of scattering based on a simple approximation inspired by Büttiker, often used in mesoscopic physics. The scattering model presented in this section, is a one parameter fitting model and the parameter used to model the strength of scattering can be easily related to a low-field mobility as shown in appendix 6. This section begins with a description of the procedure to implement the simple scattering model within the NEGF formalism. Next, this model is applied to simulate the full range current vs. voltage characteristics of a nanoscale MOSFET self-consistently. While doing so interesting quantum mechanical features of the solution scheme are highlighted. This section concludes with a discussion of the behavior of this phenomenological model with regards to the essential physics of scattering (Sec. 4.2.3).
4.3.1 Theory

The simulated device structure is shown in Fig. 3.1a. In the ballistic limit, there are only two reservoirs connected to a device, namely, the source and drain contacts. These contacts inject carriers into and extract carriers from the active device while conserving the current (net current at the source contact equals the net current at the drain contact). In the presence of scattering, Büttiker probes can be used to model dissipative transport phenomenologically within the transistor [53] [57]. These Büttiker probes perturb the Hamiltonian of the device in a manner similar to the source and drain contacts and can also be viewed as reservoirs coupled to the device. However, the fundamental difference between the real S/D reservoirs and those represented by Büttiker probes is that the probes can only change the electron energy/momentum and not the electron number within the device. This implies that one can view a Büttiker probe as extracting electrons from the device, perturbing the energy/momentum of those electrons and reinjecting an equal number back into the system with a different energy/momentum distribution. The coupling energy between the device and the probes can be adjusted to vary the scattering strength smoothly from zero (ballistic transport) to a high value (diffusive transport) as illustrated in appendix 6.

Note that when treating scattering in detail, the in and out scattering functions that described both, elastic and inelastic processes (Eqs. 4.5 and 4.6) were not related to the corresponding broadening functions (Eq. 4.3) by simple Fermi functions. They had to be self-consistently computed from the electron and hole correlation functions (Eq. 4.11). However, within the Büttiker probe approximation, we assume that individual scattering centers couple to the device in a manner identical to the S/D reservoirs. Therefore, the broadening functions for the probes are related to the in and out scattering functions by simple Fermi functions. This assumption of a Fermi Dirac distribution within a probe (scattering center), leads to a significant reduction in the computational burden associated with treating the effects of scattering in the NEGF
solution scheme. Unlike the real S/D contacts, the scattering contacts do not have well defined Fermi levels, their individual Fermi potentials are adjusted to achieve carrier conservation within the device (scattering shuffles electrons in momentum space, but preserves the number of carriers). Carrier conservation at each scattering center (zero probe current) guarantees current continuity through the transistor even in the presence of scattering.

To implement the phenomenological scattering model, we start with the 1D Hamiltonian in the transmission direction ($x$) for mode $i$ (Eq. 4.1). Next, we attach semi-infinite, 1D, Büttiker probes to each mode $i$, as shown in Fig. 4.8. Two unknowns, namely the bandwidth and the band center define the density of states within a probe. These parameters cannot be chosen arbitrarily since they affect the Büttiker probe approximation. However, a comparison with the detailed NEGF scattering formalism ensures that the approximation gives physically reasonable results when the density of states within the probes coincides with that of the device. This is discussed further in Sec 4.3.2. Each probe is characterized by a coupling energy $U_p$ and by a Fermi-level $\mu_p$ (Fig. 4.8). The quantity that we are interested in is the retarded Green’s function for mode $i$ within the active device region. The Green’s function, at a specific longitudinal electron energy ($E_l = E - E_{k_j}$) is,

$$G_i[E_l] = [E_lI - h_i - \Sigma_i(E_l)]^{-1}$$

(4.13)

where, $\Sigma_i(E_l)$, is the self-energy matrix which accounts for all of the effects associated with coupling a finite device to infinite S/D reservoirs and the effect of scattering within the device. It is,

$$\Sigma_i[E_l] = 
\begin{bmatrix}
-t_x e^{j k_x, 1 a} & 0 & 0 & \cdots \\
0 & -\frac{|U_2|^2}{t_x} e^{j k_x, 2 a} & 0 & \cdots \\
\vdots & \vdots & \vdots & \ddots \\
\vdots & \vdots & -\frac{|U_{N X - 1}|^2}{t_x} e^{j k_x, N X - 1 a} & \ddots \\
\vdots & \vdots & \vdots & \ddots & -t_x e^{j k_x, N X a}
\end{bmatrix}$$

(4.14)
Fig. 4.8. The profile of a generic mode is illustrated along with the placement of Büttiker probes. Note that the Fermi level of the S/D contacts is fixed, while that of the probe is adjusted to preserve the number of carriers. The probe self-energy is related to the coupling strength.

On including the self-energy matrix (Eq. 4.14), the size of the discrete Green’s function matrix for mode $i$ is $N^2_X$. Since Büttiker probes represent isolated scattering centers, the self-energy matrix is diagonal and has non-zero entries only at those points where a probe has been introduced. Note that the self-energy for a probe is proportional to $U_p$ (Eq. 4.14). Just as $t_x$ (Eq. 3.5) represents the coupling strength between adjacent points within the device, $U_p$ represents the coupling strength between points in the device and points in the probe. If this energy is large, it implies that an electron in the device can easily scatter into the probe. When $U_p$ tends to zero, there is no coupling between the device and the probes and electrons ballistically traverse the active device region. An analytic relation between the the probe coupling energy and the mean free path is presented in appendix 6. It should be noted that although we use 1D probes, the nature of the probes can be easily modified by adjusting the probe self-energy. This may provide a useful technique for mimicking scattering processes in other systems.
Once the retarded Green’s function is evaluated, electron density and current due to injection from the S/D contacts and all of the Büttiker probes can be easily computed. By defining the broadening functions for the S/D reservoirs and the probes as, \( \Gamma_i = j(\Sigma_i - \Sigma_i^\dagger) \), the state spectral functions due to injection from the S/D and all probes for mode \( i \) can be expressed as,

\[
A_n^i[E_l] = G_i[E_l] \Gamma_n^i[E_l] G_i^\dagger[E_l] = |G_{i}^{mn}|^2 \Gamma_i^n
\]

where \( n \) runs over all the contacts (including the S/D). Note that \( A_n^i \) is a matrix with the same size as \( G_i \), and that its diagonal entries constitute the LDOS due to injection from probe \( n \). Since \( G_{i}^{mn} \) (with a running index \( m \)) is the \( n \)th column of \( G_i \), one does not need to calculate the entire \( G \) (computationally expensive) in order to obtain the spectral function. Only those columns corresponding to S/D contacts or Büttiker probe positions need to be calculated. Thus it is clear that in the ballistic case, we need only the first and last columns of \( G \), while with scattering turned on everywhere, the entire \( G \) needs to be evaluated. Transmission between any two reservoirs labeled \( m \) and \( n \) is,

\[
T_{i}^{mn}[E_l] = \text{Trace}[\Gamma_i^m G_i \Gamma_i^n G_i^\dagger] = \Gamma_i^m |G_{i}^{mn}|^2 \Gamma_i^n
\]

Knowing the LDOS (Eq. 4.15), the electron correlation function at node \( m \), for mode \( i \), including the effect of all scattering centers and the S/D is,

\[
G_{i}^{n}[E_l, m] = 2 \sum_n \int_0^\infty \sqrt{\frac{m_{y_i}^*}{2\hbar^2 E_{k_j}}} (A_i^n)^{mn}[E_l] f \left[ \mu_n, E_l + E_{k_j} \right] dE_{k_j}
\]

where \( n \) is the reservoir index that runs over all the probes and the S/D, \( f \) is the Fermi-Dirac function for reservoir \( n \) and \( \sqrt{\frac{m_{y_i}^*}{2\hbar^2 E_{k_j}}} \) is the transverse mode state density for subband \( i \) (including spin degeneracy). Since the spectral function, \( (A_i^n)^{mn} \) depends on the longitudinal energy alone (Eq. 4.15), it can be moved out of the integral in Eq. 4.17 which reduces to,

\[
G_{i}^{n}[E_l, m] = \frac{1}{\hbar} \sqrt{\frac{2m_{y_i}^* k_B T}{\pi}} \sum_n (A_i^n)^{mn}[E_l] \tilde{F}_{-1/2}[\mu_n - E_l]
\]
where the argument of the $\mathcal{F}_{-1/2}$ function has been normalized by $k_B T$. Note that the assumption of a Fermi function for each probe, enables an analytical integration of Eq. 4.17 along the $E_{k_j}$ axis in Fig. 4.1a, thus reducing the computational burden when compared to Eq. 4.11.

In a similar fashion, the net current at reservoir $m$ including contributions from all reservoirs ($n$), modes (labeled by $i$) and valleys is,

$$I^m[E_l] = \frac{q}{\hbar^2} \sum_i \sqrt{\frac{2m^*_i k_B T}{\pi^3}} \sum_n T_i^{mn} \left[ \mathcal{F}_{-1/2}(\mu_m - E_l) - \mathcal{F}_{-1/2}(\mu_n - E_l) \right]$$  (4.19)

Note that the index $n$ includes the S/D contacts as well and that the transmission between nodes $m$ and $n$ is as specified by Eq. 4.16. Also note that while the Fermi level of the S/D contacts is fixed by the applied voltage, the Fermi level of the Büttiker probes has to be determined from current continuity. Current continuity requires that the net current at each probe equals zero. This implies that

$$I^m = \int_{-\infty}^{\infty} I^m[E_l] dE_l = 0 \quad \text{or} \quad I^m[E_l] = 0 \quad \text{for each } E_l$$  (4.20)

at each probe (probe indices are denoted by $m$) [58]. Equation 4.20 imposes a set of constraints on the Fermi potentials of the probes. These constraining equations are solved for the probe Fermi levels.

Note that Eq. 4.20 provides two options to ensure that the net current at each scattering center equals zero. Both of these options are examined as they represent different phenomenological treatments of scattering. In the first model, we assume that the net current at each scatterer, when integrated over all $E_l$ equals zero i.e. $\int_{-\infty}^{\infty} I^m[E_l] dE_l = 0$. This requirement implies that electrons from all modes are fully thermalized at each probe according to the corresponding probe Fermi potential and temperature (the local distribution is Fermi-Dirac). It also implies that the probe Fermi potentials are just position and not energy dependent. We will refer to this treatment of scattering as the energy relaxed probe model because the longitudinal energy of the electrons within the device is relaxed due to scattering. In the second model we assume that the current at each scatterer equals zero for each $E_l$. Therefore,
Eq. 4.20 is trivially satisfied. However, in this model, the Fermi level of the Büttiker probes are both position and longitudinal energy dependent because we obtain a set of chemical potentials for the probes at each $E_l$ (Note that although we refer to the chemical potential of the probe as a Fermi potential for mathematical convenience, the distribution locally is not Fermi-Dirac). We will refer to this model as the phase breaking scattering model because the channel directed energy of the electrons is not relaxed although the channel directed momentum is relaxed. Irrespective of our choice of the scattering model, it should be noted that carrier populations in different modes are mixed as a result of scattering because the net current at a probe includes a sum over all modes. Therefore both models capture the effect of intervalley scattering.

In case of the energy relaxed model, the position dependent probe potentials can be adjusted iteratively using Newton’s method because the constraining equations are nonlinear (due to the integration over $E_l$). The Jacobian matrix for the Newton iteration scheme is numerically evaluated as (Eqs. 4.16 and 4.19),

$$J_{mm} = \frac{\partial I_m}{\partial \mu_m} = \frac{q}{\hbar^2} \int_{-\infty}^{\infty} \sum_i \frac{2m^*_{y,i}k_BT}{\pi^3} \left\{ \frac{\partial \tilde{\gamma}_{-1/2}(\mu_m - E_l)}{\partial \mu_m} \sum_n T_{mn} \right\} dE_l$$

$$J_{mn} = \frac{\partial I_m}{\partial \mu_n} = -\frac{q}{\hbar^2} \int_{-\infty}^{\infty} \sum_i \frac{2m^*_{y,i}k_BT}{\pi^3} \left\{ T_{mn} \frac{\partial \tilde{\gamma}_{-1/2}(\mu_n - E_l)}{\partial \mu_n} \right\} dE_l \quad (4.21)$$

and the corrections to the probe potentials during the solution searching iterations as,

$$\Delta \mu_{\text{probes}} = -J^{-1}I_{\text{probes}} \quad (4.22)$$

Note that the S/D Fermi levels are specified by the applied voltage, therefore the size of the Jacobian in Eq. 4.21 is $(N_X - 2)^2$ although the summation index $n$, in Eq. 4.21, includes the S/D contacts. For the phase-breaking model, we directly solve for the occupancy function $\tilde{\gamma}_{-1/2}(\mu_m - E_l)$, at each longitudinal energy using a linear solution scheme of the form $A\tilde{\gamma}_{-1/2} = B$ because there is no energy integral to be dealt with. Here, $A$ and $B$, which are $E_l$ dependent are,

$$A^{mn} = \frac{q}{\hbar^2} \sum_i \frac{2m^*_{y,i}k_BT}{\pi^3} \sum_{n \neq m} T_{im}$$
\[ A^{mn} = -\frac{q}{\hbar^2} \sum_i \sqrt{\frac{2m^*_{y,i} k_B T_i}{\pi^3}} T^{mn}_i \]
\[ B^m = \left[ T^{m1} \tilde{s}_{-1/2}(\mu_s - E_l) + T^{mN} \tilde{s}_{-1/2}(\mu_d - E_l) \right] \quad (4.23) \]

In Eq 4.23, the summation index \( n \) runs over all the reservoirs including the S/D contacts for the diagonal terms and over the probes alone for the off-diagonal terms of \( A \), while index \( m \) runs over the probes alone (\( \mu_s \) is the source Fermi level and \( \mu_d \) the drain Fermi level). We mainly focus on the energy relaxing probe model in this chapter because this model seems to capture the essential physics of scattering more accurately within a nanoscale transistor when compared to the phase breaking model (Sec. 4.3.3).

Once the probe distribution functions have been evaluated using either Eqs. 4.21 and 4.22, or Eq. 4.23, the net 2D charge density \( n_i \) for mode \( i \) can be calculated by integrating Eq. 4.17 over \( E_l \). This 2D charge density is distributed in the confinement direction \( z \) according to the corresponding mode eigenfunctions to obtain the 3D electron density at each node of the \( x-z \) grid. The 3D charge is summed over all the valleys and subbands and input to Poisson’s equation to obtain a refined 2D electrostatic potential profile for self-consistent simulations.

### 4.3.2 Results

The simulated device structure (Fig. 3.1a) is identical to the structure used to demonstrate the detailed scattering model. However, since the Büttiker probe based scattering model is computationally efficient, we present self-consisistent simulations and current vs. voltage characteristics in this section. For scattering simulations, the input parameter is a position dependent low field mobility (appendix 6, provides an analytical relation between the mobility and scattering strength). Measured low field mobilities of \( \sim 200 \text{ cm}^2/\text{V-s} \) (at a charge density of \( \sim 10^{13}/\text{cm}^{-2} \)) have been reported in the literature, for silicon film thicknesses of \( \sim 5 \text{ nm} \) [59] [60] [61]. As the silicon film thickness is reduced, there are two competing effects which affect the
mobility. Increased surface roughness scattering reduces the low field mobility, while confining all of the electrons to the unprimed bands raises the low field mobility [31]. Therefore, as an approximation, we assume that these effects cancel, and that a low field mobility of 200 cm$^2$/V-s is a reasonable value to use in the channel region (our channel charge is also $\sim 10^{13}$/cm$^{-2}$ in the on-state). For the S/D extensions which are highly doped, we use a doping dependent mobility model that yields a value of 55 cm$^2$/V-s at a donor doping concentration of $\sim 10^{20}$cm$^{-3}$.

In order to highlight quantum effects and present a general picture of how the Büttiker probe models work, we compare internal quantities with and without scattering once again. Figure 4.9 shows the self-consistent current spectrum vs. the longitudinal electron energy (Eq. 4.19) in the on-state ($V_{GS} = V_{DS} = 0.4V$) from both, the energy-relaxing (Fig. 4.9a), and the phase-breaking scattering models (Fig. 4.9b). The ballistic current is superposed on each of the figures for comparison. If we compare this ballistic spectra to the one obtained from the energy relaxed Büttiker probe model, we observe that the source injected current is reduced in magnitude, and the drain collected current no longer mirrors the source current distribution. This is because this model strongly relaxes the channel directed energy of the electrons injected from the source. These electrons leave the device with lower longitudinal energies because they lose energy due to scattering. The phase breaking model, (Fig. 4.9b) on the other hand, preserves the longitudinal energy of electrons, thus resulting in a symmetric spectrum (similar to the ballistic case). It should be noted that although the current spectra are symmetric, the current magnitude is reduced when compared to the ballistic case, because back scattered electrons lose their directed momentum due to scattering. We will mainly focus on results obtained using the energy relaxing model for the rest of this section. An in depth comparison of the energy and phase relaxing scattering models against a detailed treatment of scattering is presented Sec 4.3.3.

Figure 4.10 compares the LDOS (Eq. 4.15) and the 2D charge density spectra (Eq. 4.17) vs. the longitudinal electron energy, with and without scattering ($V_{GS} =$
Fig. 4.9. (a) The current spectrum from the energy relaxed scattering model is compared against the ballistic limit in the on-state ($V_{GS} = V_{DS} = 0.4 \, \text{V}$). Note that the drain current spectrum is relaxed in energy. (b) The current spectrum from the phase relaxed scattering model is compared against the ballistic limit, in the on-state. The source and drain spectra are identical in the presence of scattering, because this model relaxes channel directed momentum only.
$V_{DS} = 0.4V$). The first mode is superposed on each of the plots in Fig. 4.10 (dotted white line) to indicate the effective potential energy of electrons. In the ballistic case, there is no phase relaxation within the device. Therefore, states injected from the drain end of the device undergo reflections and interfere strongly to the right of the source-to-channel barrier. This interference results in coherent oscillations in the LDOS as seen in Fig. 4.10a (left). When scattering is turned on throughout the device, phase information of the electrons within the device is randomized and the energy levels are broadened when compared to the ballistic case. Therefore, all of the interference effects are washed out as seen in Fig. 4.10a (right). Note that both, the ballistic LDOS and the LDOS with scattering, exhibit non-zero values below the source-to-channel barrier resulting in source-to-channel tunneling. In the ballistic case, the charge density spectrum (the square root of the charge is plotted to resolve low charge densities at the drain end) can be resolved into two components; one due to source injection and the other due to drain injection. The source injected charge propagates from the source to the drain without any energy relaxation, resulting in a ballistic peak in the charge spectrum at the drain end, as seen in Fig 4.10b (left). The drain injected charge on the other hand, is completely reflected by the source-to-channel barrier in the on-state. When scattering is turned on, the longitudinal energy of the electrons is relaxed (in case of the energy relaxing model) and the source and drain populations can no longer be distinguished. Also, any coherencies in the charge spectrum are also washed out as seen in Fig. 4.10b (right).

Next, we look at the full range $I_{DS}$ vs. $V_{GS}$ characteristics for the model device with and without scattering in Fig. 4.11a. Current in the off-state ($V_{GS} = 0, V_{DS} = 0.4V$), is comprised of two components; source-to-channel tunneling, for longitudinal energies below the source-to-channel barrier, and diffusion, for energies above the barrier. Scattering, in general, broadens the LDOS as discussed earlier. Therefore, the tunneling current component is increased in the off-state as a result of scattering, when compared to the ballistic limit. However, the detailed NEGF scattering model indicates that the scattering rate should be proportional to the LDOS at each energy
Fig. 4.10. (a) The local density of states in the on-state, from the ballistic (left) and energy relaxed scattering models (right). The first mode is also plotted (dotted line). Note that coherent oscillations in the LDOS are washed out when scattering is turned on. (b) The charge density spectrum from the ballistic (left) and energy relaxed scattering models (right), in the on-state. In the ballistic limit, the source and drain injected populations can be clearly identified. When scattering is turned on, these populations are mixed.
(Eqs. 4.5, 4.6, 4.3 and 4.2). Therefore, by choosing the energy band of the Büttiker probes to coincide with the local energy band within the device (Eq. 4.14), we ensure that the broadening in the LDOS due to scattering and thus the increase in the tunneling current is not uncontrolled. This is particularly important when modeling low power devices where the off-current is dominated by tunneling.

The diffusion current component reduces when scattering is turned on, because the degenerate thermal injection velocity is reduced when compared to the ballistic limit [56]. Therefore, the cumulative effect of reduced diffusion and increased tunneling in the presence of scattering, is that the ballistic off-current is always higher than the off-current in the presence of scattering (in agreement with Fig. 4.6). This is clearly seen from Fig. 4.11b, where we plot the off-current versus channel length for channel lengths down to 5 nm. The ballistic off-current sets an upper limit on the leakage current and provides a fairly accurate picture of the subthreshold behavior as transistors are scaled to smaller dimensions.

The Fermi level of each scatterer is adjusted to conserve current in the presence of scattering (Eqs. 4.21 and 4.22). This quantity, which is analogous to the quasi Fermi level computed in semiclassical models, has a clear physical interpretation. Unlike the quasi Fermi level (which is derived from the charge instead of the current), the Fermi level of the Büttiker probes is an actual representation of how the potential drops from the source to the drain. Figure 4.12a, plots this quantity in the linear region ($V_{DS} = 10\text{mV}$) of operation. In the off-state, the channel resistance is high. Therefore all of the applied voltage drops in the channel region of the device. As the gate voltage increases, the channel conductivity increases, and the voltage dropped in the channel is reduced. This leads to a flattened Fermi potential profile in the channel and large voltage drops in the S/D regions. In the ballistic limit, there is no mechanism for internal voltage drop and all of the $V_{DS}$ is dropped across the contact/device interface, resulting in a finite ballistic current. Note that the source (drain) Fermi potential represents the Fermi energy of source (drain) injected carriers. In the off-state there is very little current flow within the device and near equilibrium
Fig. 4.11. (a) $I_{DS}$ vs. $V_{GS}$ characteristics from the ballistic (solid line) and energy relaxed scattering models (dashed line) for $V_{DS} = 0.4\,V$. The off-current, from the scattering model is lower despite an increase in the tunneling current due to a broadening in the LDOS below the source-to-channel barrier. (b) The off-current vs. channel length from the ballistic (solid line) and energy relaxed scattering models (dashed line). Ballistic simulations are good enough to evaluate leakage and subthreshold characteristics.
conditions prevail at both, the source/device and the drain/device interfaces. However, as the device is turned on, the distribution at the source/device interface (and the drain/device interface) is strongly off-equilibrium in order to maintain a large current flow. This leads to the observed discontinuity in the Fermi potential at the contact/device interface as seen in Fig. 4.12a.

In the linear region, it is possible to derive the sheet resistivity versus position along the channel, from the probe Fermi levels using [4],

$$\rho_{sh} = \frac{\partial \mu}{\partial x} \frac{I_{DS}}{W} \quad (4.24)$$

The derived resistivity, at low drain and high gate voltage ($V_{DS} = 10\text{mV}, V_{GS} = 0.4V$) is plotted in Fig. 4.12b. The sheet resistivity can be divided into the following regions: 1) quantum contact resistance, 2) S/D extension resistance, 3) tip resistance and 4) channel resistance. Note that the gate modulates the channel resistance and a fraction of the tip resistance only. Therefore, in our device (with a channel mobility of 200 cm$^2$/V-s), the on-current is primarily limited by the S/D parasitic resistances (S/D mobility is just 55 cm$^2$/V-s). Figure 4.13a, plots the $I_{DS}$ vs. $V_{DS}$ characteristics in the on-state ($V_{GS} = V_{DD} = 0.4V$). For a ballistic off-current of 10 $\mu A$/um, the simulated on-current in the presence of scattering, is only 50% of the ballistic limit due to S/D and tip parasitics. To reinforce this point, we plot the on-current as a function of channel mobility in Fig. 4.13b. The channel mobility is progressively increased from zero to very high values. It is clear from Fig. 4.13b, that the on-current saturates at $\sim 55\%$ of the ballistic limit, and does not increase with increasing channel mobility because of the parasitic resistances. It is expected that transistors with a double-gate geometry will yield twice the on-current when compared against those with a single gate (SG) geometry for the same level of off-current. This expectation will definitely be met in the ballistic limit. However, if series resistance limits device performance, the performance of a DG MOSFET is degraded to a greater extent than a transistor with a SG geometry. Therefore, the performance benefit expected from DG transistors may not be met. Our model enables us compare DG and SG MOSFETs from the ballistic to the diffusive limit and serves as a valuable design
Fig. 4.12. (a) The self-consistent Fermi level of the Büttiker probes, from the energy relaxed scattering model, in the linear response region ($V_{DS} = 10 \text{ mV}$). (b) The extracted sheet resistivity in the on-state. Note the four components of the resistance; 1) quantum contact resistance, 2) S/D resistance 3) tip resistance and 4) channel resistance.
4.3.3 Discussion

In this section we assess the behavior of the energy and phase relaxing scattering models in order to ascertain their applicability to modeling electron transport within a MOSFET. To understand the physics of scattering that is captured by each model, we compare their behavior to the essential physics of scattering which was discussed in Sec. 4.2.3. We choose to compare the energy and phase relaxing scattering models non-selfconsistently, in order to avoid the complicated behavior associated with self-consistent electrostatics [54]. Self-consistency is important, and we discuss it briefly at the end of this discussion section.

Figure 4.14a (left) shows the effect of scatterer placement on the device performance from both, the energy and phase relaxing models. This figure was generated by fixing the potential profile (Fig. 4.3) and increasing the number of scatterers progressively from the drain towards the source. When the scatterer number is zero, both models yield the ballistic limit. But as the scatterer number is increased, the energy relaxing scattering model shows almost no change in current initially and a linear drop thereafter. Since this model causes the longitudinal electron energies to relax, electrons scattering near the drain find it difficult to make it back to the source (Fig. 4.7a), and hence the on-current is initially unaffected. However, as the number of scatterers increases towards the source, electrons scattering very near the source-to-channel barrier do backscatter back into the source as they have not dissipated enough longitudinal energy, thus reducing the current. The phase breaking model shows distinctly different results. The transmitted current decreases with increasing number of scatterers irrespective of their location. This is because, in this model, electrons can reverse direction without losing their channel directed energy. Therefore, scattering anywhere in the channel can reflect electrons back into the
Fig. 4.13. (a) $I_{DS}$ vs. $V_{DS}$ characteristics from the ballistic (solid line) and energy relaxed scattering models (dashed line) for $V_{GS} = 0.4\, V$. The on-current in the presence of scattering is $\sim 50\%$ of the ballistic limit. (b) The on-current vs. channel mobility is plotted to indicate that the ultimate performance of our device is primarily controlled by device parasitics.
source. Therefore, an assessment of the two models clearly indicates that the energy relaxing scattering model captures the essential physics of scattering. Figure 4.14a (right), plots the current spectra vs. the longitudinal electron energy from both, a detailed treatment of phonon scattering (Fig. 4.5) and the energy relaxing Büttiker probe model at the source and drain ends of the device, in the on-state (Fig. 4.3). These spectra were generated by fixing the potential profile and varying the scattering strength to obtain the same level of current flow within the device (to within ~10%)

Figure 4.14 (right), indicates that the energy relaxing, Büttiker probe model overestimates the extent of the longitudinal energy relaxation due to scattering. However, the close match in the qualitative features of the current spectra clearly indicates this computationally efficient model is an attractive technique for simulating dissipative transport in nanoscale transistors.

Fig. 4.14. (left) The current as a function of the scatterer number, which is progressively increased from the drain to the source is plotted from both, the energy relaxing (solid line with circles) and the phase relaxing (dashed line) scattering models. This plot indicates that the energy relaxed model captures the essential physics of scattering in transistors. (right) The current spectra vs. the longitudinal electron energy is plotted at the source and drain ends from a detailed treatment of phonon scattering (solid lines) and from the energy relaxing Büttiker probe model (dashed lines). The potential profile is fixed (Fig. 4.3) and the scattering strength is adjusted to obtain the same level of current flow from both models.
Having selected the Büttiker probe based, energy relaxing scattering model, we use it to examine the importance of self-consistent electrostatics on the essential physics of scattering. We divide the device into two halves and consider two cases; In the first case, scattering (mobility of 100 cm$^2$/V-s) is turned on in the first half of the device while the second half is ballistic, and in the second case scattering is turned on in the second half of the device while the first half is ballistic. Our self-consistent simulation results using the energy relaxing model indicate that irrespective of the region where scattering is turned on, the on-current is significantly degraded (∼43% in the first case and ∼33% in the second case) when compared to the ballistic limit. To understand this behavior, we plot the self-consistent subband profile and 2D charge density for both cases in Fig. 4.15. When scattering is introduced only in the first half of the device, the on-current is strongly degraded when compared to the ballistic limit due to backscattering of electrons at and around the top of the source-to-channel barrier (Fig. 4.15). Scattered electrons can easily re-enter the source region as they still preserve most of their channel directed energy and hence reduce the net current. When scattering is turned on in the second half of the device alone, the reduction in on-current is due to a complicated interplay between self-consistency and the effect of scattering [54]. The ballistic stream of source injected electrons entering the second half of the device undergoes scattering. Once reflected, if the mean free path of these electrons is comparable to the channel length (mean free path is ∼10 nm for a mobility of 100 cm$^2$/V-s) a fraction of these scattered electrons makes it back to the source. This is the first mechanism that reduces the on-current. The second mechanism is because the electrons reflected in the second half of the device result in an increased 2D electron density in the channel as shown in Fig. 4.15 (dotted line). Self-consistency, causes the subband potential to float to higher energies in the channel and also broadens the potential profile from the source to the drain. An increased source-to-channel barrier combined with a broadened potential profile (which increases the probability of a backscattered electron to re-enter the source as seen from Fig. 4.7a) causes the on-current to decrease further when scattering is
Fig. 4.15. The profile of the first mode and the 2D electron density in the on-state with scattering turned on in the first half of the device (solid lines) and with scattering turned on in the second half of the device (dashed lines). Note that the potential drops in the source or the drain only when scattering is turned on. Also note that turning on scattering only in the second half of the device increases the 2D electron density in the channel.

turned on the second half of the device. This self-consistent behavior of the on-current as a function of scatterer placement clearly indicates that scattering is important not only at the source but throughout the channel in nanoscale transistors. This behavior of the on-current, when the channel length is comparable to or shorter than the mean free path has also been observed when scattering is treated rigorously using the Green’s function formalism by [54]. The energy relaxed, Büttiker probe based scattering model thus captures all the essential physics of scattering within a MOSFET, including self-consistent effects in a simple and elegant fashion.

4.4 Summary

This chapter presented two techniques to quantum mechanically treat the effects of scattering transport in nanoscale, SOI transistors. At the detailed level we implemented a rigorous treatment of scattering within the NEGF simulation platform. Results obtained by applying the rigorous scattering model to simulate a model de-
vice, were used as a benchmark to validate a simple, computationally efficient, phenomenological treatment of scattering. We then applied this simple model to treat the effects of dissipative transport in an ultra-thin body (1.5 nm), DG, n-MOSFET. In doing so, quantum effects that are observed in nanoscale transistors, the role of scattering, effect of parasitics and the importance of self-consistent electrostatics were highlighted and discussed.

Our phenomenological scattering model is a one parameter model, and the parameter we use can be analytically related to a low field mobility. Such a relation, enables the use of our model to explore novel physics and device design issues in nanoscale transistors because it can be calibrated to experimentally measured mobility data. This scattering model is also very useful because the detailed NEGF formalism can be used to motivate changes to the nature of the Büttiker probes to mimic the effects of scattering in different material systems such as carbon nanotubes and molecules by appropriately changing the probe self-energy. Our simple model provides an excellent tradeoff between increased computational cost and the physics of scattering that needs to be captured in devices of the future.
5. A QUANTUM MECHANICAL ANALYSIS OF CHANNEL ACCESS GEOMETRY AND SERIES RESISTANCE IN NANOSCALE TRANSISTORS

5.1 Introduction

The push towards achieving small transistors with the requisite device characteristics has resulted in the exploration of novel device structures that have been recently reported in the literature [62] [27]. The channel length for some of these devices is of the order of 10 nm or lower. As critical transistor dimensions continue to shrink, the role of channel access geometry must be understood. Therefore, the primary objective of this chapter is to apply all of the concepts previously developed in this thesis to understand and highlight the effects of differing access geometries on device performance.

Several device designs with single, double or tri-gate geometries are being considered as candidates for scaling silicon devices to the limit [63] [10] [11]. Despite the nature of the gate, all of these geometries share a common characteristic- they are composed of a thin silicon body sandwiched between insulators and coupled to large source/drain (S/D) reservoirs. One such device, a double-gate (DG), silicon-on-insulator (SOI) transistor, is illustrated in Fig. 5.1b. Within the intrinsic device (the region within the dashed box in Fig. 5.1b), electrons are strongly quantum confined and reside in a discrete set of subbands. However, as these electrons propagate from the intrinsic device into the S/D regions (which are typically wide in order to reduce the resistance), the subband structure is relaxed, and there is significant mixing of electrons in different modes. Therefore, modeling such structures requires the capability to seamlessly couple these wide and narrow regions, while capturing all of the quantum effects within the narrow region accurately. The effects of scattering also
need to be accounted for in both the wide and the narrow regions, because scattering which is present in real devices, permits electrons from several modes in the wide region to couple into a few modes within the narrow region. We simulate the effects of scattering phenomenologically using the concept of Büttiker probes, as discussed in the previous chapter. Also, the quantum mechanical parameters in our scattering model are calibrated to mimic different low-field mobilities in different regions of the transistor. Such a mapping is useful because low-field mobilities continue to be relevant even at the short length scales considered in this study [64] [65]. The overall simulation technique solves the NEGF equations in coupled mode-space along with Poisson’s equation self-consistently.

Electron transport from wide to narrow regions has been extensively examined in the physics literature through the study of quantum point contacts [66] [67]. These studies, (which were nonself-consistent) compared the low bias conductance behavior (at temperatures close to 0 K) of abrupt wide-to-narrow devices against devices where the channel access geometry was gradually flared out (also referred to as adiabatic). They concluded that for specific thicknesses of the narrow region, the low bias conductance is unaffected by the channel access geometry. When studying the low bias conductance properties around the Fermi surface (at low temperatures), it should be noted that if the Fermi level is well above the subband edge within the narrow region (achieved by changing the thickness), the large S/D reservoirs act as “reflectionless contacts” [19]. The term reflectionless implies that electrons at the Fermi surface can transmit from the narrow to the wide regions with negligible probability of reflection. In such cases, devices with differing channel access geometries could exhibit the same low bias conductance around the Fermi energy. However, electrons close to the subband edge (large wavelengths) may experience large reflections.

In nanoscale transistors the current spectrum is distributed over a wide range of energies (and not just around the Fermi energy) due to the high voltages that are applied. Also, the alignment of the Fermi energy relative to the subband edge is determined by self-consistent electrostatics. Therefore the large S/D reservoirs cannot
Fig. 5.1. (a) An ultra-thin body DG MOSFET structure with a S/D doping of 10^{20} \text{cm}^{-3} and an intrinsic channel (channel thickness = 1.5 nm). The S/D regions are 20 nm in length, the gate length is 10 nm and the S/D-channel junctions are abrupt. A slice of the device within which a 1D, \textit{z} directed effective mass equation is solved, is also indicated. (b) A thin body DG MOSFET with abruptly flared out S/D regions. The length of the flared regions is 10 nm, their thickness is 6.2 nm and the overall device length is 50 nm (same as that for device a). The region within the dashed box is the intrinsic device.
be assumed as reflectionless contacts, thus implying that their geometry may affect device characteristics. Moreover, a 10-20% difference in the series resistance due to differing channel access geometries is an important effect that needs to be emphasized when evaluating different device designs. Also, the effect of scattering, which can be safely neglected in QPC’s operating at low temperatures cannot be ignored in transistors within which temperatures exceed 300 K. We will highlight the importance of scattering by comparing results from both ballistic (no scattering) as well as scattering simulations for nanoscale, silicon transistors with differing channel access geometries. Due to the aforementioned reasons, conclusions from the mesoscopic physics literature need to be carefully examined for nanoscale silicon transistors operating at room temperature.

Device design and optimization through the use of simulations requires the ability to quickly perform a large number of simulations within a reasonable amount of time. Even for the idealized device structure with the dimensions illustrated in Fig. 5.1b, the simulation time could be very large. Computational burden is greatly increased because a large number of subbands need to be included when simulating the flared out S/D reservoirs quantum mechanically. Therefore, an approximate treatment of the S/D reservoirs which are usually much larger than those in Fig. 5.1b is clearly useful. We examine one such approximation wherein the device in Fig. 5.1b is decomposed into an intrinsic component (region within the dashed box), and a parasitic component (the S/D reservoirs). The intrinsic device is explicitly simulated, and the extrinsic characteristics which include the effect of the S/D reservoirs are estimated from the intrinsic current vs. voltage (I-V) characteristics through the addition of appropriate parasitic resistances. We present a recipe which can be applied to examine design issues in nanoscale transistors and highlight the appropriate boundary conditions that need to be used when simulating the intrinsic device.

The chapter is divided into the following sections: Section 5.2 presents the solution scheme. Section 5.3 presents simulation results which explain the role of self-
consistency, scattering and channel access geometry on device performance. Section 5.4 discusses device design issues and Section 5.5 summarizes key findings.

5.2 The modeling scheme

The simulated device structures are shown in Fig. 5.1. The solution scheme begins with a real-space discretization of the active device Hamiltonian. In discrete form the block tridiagonal, 2D device Hamiltonian ($x-z$ plane) is given by Eq. 3.2.

If we think of the device as being composed of vertical slices (Fig. 3.1a) adjoining each other, the $\alpha$'s in Eq. 3.2 represent the discretized, one-dimensional (1D) effective mass Hamiltonian along the $z$ direction within each slice (Eq. 3.3) and the $\beta$'s (Eq. 3.4) represent coupling between adjacent slices. The $\alpha$'s, with indices less than one, represent successive slices going into the source contact while those with indices greater than $N_X$ represent successive slices into the drain. There are $N_X$ slices within the active device region. The overall size of the discretized active device Hamiltonian in a real-space basis for each conduction band valley is $(N_X \times N_Z)^2$, where $N_Z$ is the number of grid points along $z$, within each slice of the device.

The choice of a real-space basis, although natural, leads to a computational problem. If we use a grid spacing of $a=0.25$ nm and $b=0.1$ nm to discretize the Hamiltonian for the device pictured in Fig. 5.1b, then $N_X=200$ and $N_Z=62$. Therefore the size of the active device Hamiltonian is $\sim 12000 \times 12000$. The computational burden of performing a large number of matrix operations on such a Hamiltonian is prohibitively expensive. Therefore, we transform the basis from real to mode-space. Chapter 3 presented the decoupled mode-space approach, which can be applied to accurately simulate electron transport in device structures with uniform geometries (Fig. 5.1a). However, when the device geometry exhibits abrupt changes (Fig. 5.1b), the decoupled mode-space approximation (Eqs. 3.26 and 3.27) is no longer valid and all of the terms appearing in a modal expansion of the real-space Hamiltonian (Eq. 3.21) need to be retained. Mathematically, this operation is a basis transformation where the
new basis is constructed by solving for the eigenfunctions of $\alpha$ (Eq. 3.3). As with the decoupled mode-space solution scheme (chapter 3), a significant reduction in computational burden is obtained because only a few modes need to be included even within the large S/D regions of the device in Fig. 5.1b. In order to illustrate the basis transformation (which will be referred as the coupled mode-space transformation) we consider a simple example where $N_Z=3$, $N_X=3$ and assume that the lowest two modes are all that need to be included. For this example the basis transformation matrix (mode-space basis in terms of the real-space basis) looks like,

$$U = \begin{bmatrix}
1 & 2 & 3 & 1 & 2 & 3 \\
\times & 0 & 0 & \times & 0 & 0 \\
\times & 0 & 0 & \times & 0 & 0 \\
\times & 0 & 0 & \times & 0 & 0 \\
0 & \times & 0 & 0 & \times & 0 \\
0 & \times & 0 & 0 & \times & 0 \\
0 & \times & 0 & 0 & \times & 0 \\
0 & 0 & \times & 0 & 0 & \times \\
0 & 0 & \times & 0 & 0 & \times \\
0 & 0 & \times & 0 & 0 & \times \\
\end{bmatrix} \quad (5.1)$$

In Eq. 5.1, the crosses mark non-zero entries corresponding to the mode eigenfunctions which are expressed in a real-space representation. The size of the basis transformation matrix is $(N_X \times N_Z) \times (N_X \times \text{number of modes})$. Note that each column of $U$ expresses the modes in real-space and that the columns are ordered into blocks of size $N_X$ (in this case 3) as labeled in Eq. 5.1. Different columns of $U$ are orthonormal to each other, thus indicating that $U$ is a unitary transformation matrix. The device Hamiltonian in the new basis is

$$h_{\text{mode}} = U^\dagger h[x, z]U \quad (5.2)$$
The size of active device Hamiltonian in mode-space is \((N_X \times \text{number of modes})^2\). For the device structure in Fig. 5.1b, we use the lowest five modes to effect the basis transformation, thus resulting in a Hamiltonian matrix which is 1000×1000 in size (the original Hamiltonian in real-space had a size of 12000×12000 for each conduction band valley). When expressed in coupled mode-space, the Hamiltonian has clear physical meaning. Diagonal blocks (size \(N_X^2\)) represent the decoupled mode-space Hamiltonian for each mode (Eq. 3.16), while off-diagonal blocks represent coupling between modes. For device structures such as those in Fig. 5.1a (in which the confining potential along \(z\) retains its shape as one moved from the source to the drain), we demonstrated that the off-diagonal blocks within the coupled mode-space Hamiltonian can be ignored and individual modes can be treated as decoupled (3.4). However, when the confining potential changes abruptly as electrons flow from a wide to narrow region (Fig. 5.1b), the off-diagonal blocks of the Hamiltonian are significantly greater than zero. These non-zero elements in the mode-space Hamiltonian indicate that different modes couple as a result of changes in the device geometry, even in the absence of real scattering (geometric scattering).

Once the mode-space Hamiltonian is computed, we can evaluate the Green’s function in coupled mode-space at each injection energy (denoted by \(E_L\) and referred to as the longitudinal energy).

\[
G_{\text{mode}} = \left[ E_L I - h_{\text{mode}} - \Sigma_{\text{S/D}}(E_L) - \Sigma_{\text{Scatt}}(E_L) \right]^{-1}
\]

(5.3)

In Eq. 5.3, \(\Sigma_{\text{S/D}}\), represents the self-energy matrix which accounts for the coupling between the active device and the source-drain (S/D) contacts and \(\Sigma_{\text{Scatt}}\), is the self-energy matrix which models the effect of scattering within the device. The procedure to set up the self-energies in coupled mode-space is discussed in the subsequent paragraph. Once the Green’s function has been evaluated, all the quantities of interest such as the local charge density and the terminal currents can be derived by extending the procedures listed in chapters 3 4. These quantities which are computed in coupled mode-space, can then be mapped onto the real-space basis by applying the inverse of the transformation used in Eq. 5.2.
Deep inside the S/D contacts the electrostatic potential is assumed to be invariant (along $x$). Therefore, diagonal blocks of the Hamiltonian ($\alpha$’s in Eq. 3.2) repeat themselves within the source ($\alpha$’s with indices less than 1) and the drain ($\alpha$’s with indices greater than $N_X$). Since the mode-space basis is composed of the eigenvectors of $\alpha$, changing the basis from real to mode-space diagonalizes $\alpha$ within the S/D regions (appendix 6).

$$\alpha_{\text{mode}} = \begin{pmatrix}
\epsilon_1 + 2tx & 0 & \ldots & \ldots & \ldots \\
0 & \epsilon_2 + 2tx & 0 & \ldots & \ldots \\
\ldots & 0 & \epsilon_3 + 2tx & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \epsilon_Nx + 2tx
\end{pmatrix} \quad (5.4)$$

The $\epsilon$’s in Eq. 5.4 represent the cutoff energies of the various modes within the contacts. The coupling matrix $\beta$, in Eq. 3.4, remains unchanged as it is proportional to the identity matrix (this matrix couples adjacent grid point along the $x$ dimension within the same mode). Therefore, we can visualize the contacts as $N_Z$ separate 1D conductors in parallel. In this representation the self-energy matrix can be expressed by extending the results in Eq. 3.18. For the simple example that we considered to illustrate the basis transformation (Eq. 5.1), the self-energy matrix is,

$$\Sigma_{\text{S/D}}(E_L) = \begin{pmatrix}
-t_xe^{ik_1a} & 0 & \ldots & \ldots & \ldots \\
0 & 0 & \ldots & \ldots & \ldots \\
\ldots & 0 & -t_xe^{ik_2a} & 0 & \ldots \\
\ldots & \ldots & 0 & -t_xe^{ik_2a} & 0 \\
\ldots & \ldots & \ldots & \ldots & -t_xe^{ik_2a}
\end{pmatrix} \quad (5.5)$$

where $i=\sqrt{-1}$. The superscript on the $k$’s in Eq. 5.5 refer to the mode index (the wave vector is specified with respect to the cutoff energy of the mode) and the subscript to
the site index along $x$. Since the first and the last slices of the active device are the only regions which couple to the S/D contacts, the self-energy matrix representing the contacts has non-zero entries only at sites 1 and 3 (or alternatively 1 and $N_X$ for the general case). It should be noted that the contact self-energy matrix has the same size as the coupled mode-space Hamiltonian.

Scattering within our modeling framework is treated phenomenologically using Büttiker probes [57] [53]. Within this phenomenological framework scattering processes can be viewed as another set of contacts which couple to the device just like the source and drain. This implies that the self-energy describing scattering has the same form as Eq. 5.5. The probe self-energy in terms of the coupling strength for the example used to illustrate Eqs. 5.1 and 5.5 is,

$$
\Sigma_{\text{Scatt}}(E_L) =
$$

$$
\begin{bmatrix}
1 & 2 & 3 \\
\frac{|P_1|^2}{t_x} e^{ik\frac{a}{l}} & 0 & \ldots \\
\ldots & \frac{|P_2|^2}{t_x} e^{ik\frac{a}{l}} & 0 \\
\ldots & 0 & \frac{|P_3|^2}{t_x} e^{ik\frac{a}{l}} \\
\ldots & \ldots & 0 \\
\ldots & \ldots & \ldots
\end{bmatrix}
$$

where the coupling energy $|P|^2/t_x$, is adjusted to mimic the same low field mobility for electrons in each mode. The subscripts and superscripts on the probe strength in Eq. 5.6 have the same interpretation as in Eq. 5.5. The Fermi level of the probes at the ends of the device ($\mu_1$ and $\mu_{N_X}$ in Fig. 5.1b) is fixed by the applied bias, while that of the probes in the interior of the device ($\mu_2$ up to $\mu_{N_X-1}$ in Fig. 5.1b) is adjusted to ensure current continuity (net source injected current equals the net drain collected current). Since the probe Fermi levels are adjusted by forcing the net current at each probe to equal zero ($\int I_{\text{probe}}(E_L) dE_L = 0$), the scattering model
presented in this work causes a relaxation of the channel directed (or longitudinal) electron energies, and results in strong inter valley scattering (scattering between subbands from different valleys). Note that the Fermi potential of all the probes within a single vertical slice is shorted as shown in Fig. 5.1b. Such a treatment is justified because there is no net current flow along the $z$ dimension. The probe Fermi levels (derived from current continuity) are a direct measure of how applied voltage drops from the source to the drain (along $x$). These Fermi levels are visualized and discussed in Sec. 5.3.

The calculation cycle begins with a guess for the two-dimensional (2D) potential $V(x, z)$. Knowing the potential profile, the real-space Hamiltonian (Eq. 3.2) is transformed to coupled mode-space using Eqs. 5.1 and 5.2. Büttiker probes are introduced (Eq. 5.6) and a Newton method is used to adjust the probe Fermi levels to ensure current continuity. These Fermi levels are then used to compute the density matrix in coupled mode-space for each conduction band valley. These density matrices are then transformed (inverse of the transformation in Eq. 5.2) to real-space and in that representation their diagonal entries yield the three-dimensional (3D) charge density at each node of the $(x, z)$ grid. The 3D charge (summed over all valleys and spins) is used to solve Poisson’s equation to obtain a new potential profile, and the entire calculation cycle is repeated until a self-consistent solution is achieved.

In order to accurately sample the injection from the S/D leads, a very fine longitudinal energy grid with a spacing of $\Delta E_L = 0.5 \text{mV}$ or less needs to be used. For typical values of power supply voltages listed in the International Technology Roadmap for Semiconductors (ITRS) [2], the energy grid is composed of $\sim 1500$ points. Manipulating matrices of size $1000 \times 1000$, $\sim 4500$ times ($3 \times 1500$ as there are three conduction band valleys) is a computationally challenging task (because all the columns of the Green’s function in Eq. 5.3 need to be computed). Thus, even when solved in coupled mode-space, the computational time per bias point for the device structure in Fig. 5.1b is $\sim 7$ hours on 45, 1.2GHz processors of a Linux Cluster. All calculations in this work were performed on a Linux cluster using Matlab-6.1. Parallelization of
the energy grid was achieved using a Message Passing Toolbox (MPI) interface for Matlab, developed at the University of Granada in Spain [55].

5.3 Results

The simulated device structures are illustrated in Fig. 5.1. Both devices are ultra thin body, fully depleted, symmetric, n-MOSFETs with S/D regions doped at $10^{20}\text{cm}^{-3}$ and an intrinsic channel. Their gate lengths are 10 nm, and there is no gate-to-S/D overlap. The junctions are abrupt, and the oxide thickness for both top and bottom gates is 1 nm. A body thickness of 1.5 nm, and a power supply ($V_{DD}$) of 0.4 V has been used in this simulation study. The oxide regions are treated as infinite potential barriers for electrons. The effect of channel access geometry on device performance is examined by comparing the characteristics of a device with abruptly flared out S/D regions (Fig. 5.1b), against those of a device with uniformly narrow S/D regions (Fig. 5.1a) while fixing the overall device length (50 nm). A S/D mobility of 55 cm$^2$/V-s and a channel mobility of 200 cm$^2$/V-s is used to capture the effects of scattering. Henceforth, we shall refer to the device with the flared out geometry as the wide-to-narrow (WN) device and the device without the flared out S/D as the uniformly-narrow (UN) device.

5.3.1 The WN Device

In order to highlight quantum effects, especially within the WN device (Fig 5.1b), we plot the self-consistently calculated subband energies (obtained by solving for the eigenvalues of $\alpha$ in Eq. 3.3) and charge densities along the slices labeled A-A, B-B and C-C in Fig. 5.2 (for the on-state, $V_{GS}=V_{DS}=0.4$ V). For the WN device, the thickness of the wide S/D is 6.2 nm and these regions are terminated using hard-wall (infinite potential) boundary conditions along the z dimension. Therefore, charge in these regions is quantized. However, unlike the narrow region which being extremely thin, exhibits single subband occupancy, charge within the wide S/D is distributed
over several subbands. Away from the WN constriction (section A-A), the shape of the electrostatic potential (along $z$) and hence the modes (eigenfunctions of $\alpha$ in Eq. 3.3) do not change appreciably along the $x$ dimension. Therefore, different modes within section A-A can be treated as decoupled because the off-diagonal terms in the mode-space Hamiltonian (Eq. 5.2) are close to zero (Sec. 3.4). Along slice A-A, the electron population from each mode can be viewed as being distributed according to the individual mode eigenfunctions. In slice A-A, all of the charge resides in the lowest five unprimed (valleys with a heavy effective mass along $z$, and two fold degenerate) and the lowest two primed subbands (valleys with a light effective mass along $z$, and two fold degenerate). Since the mode eigenfunctions for the low energy unprimed modes are evanescent at the device center (note that these energies are below the conduction band maximum in Fig. 5.2a), the spatial charge distribution for the unprimed valleys is peaked away from the centerline of the device. On the other hand, the charge distribution for the primed valleys is parabolic, with a peak at the center of the wide regions (all of the primed mode energies are above the conduction band maximum as shown in Fig. 5.2a). The net charge (summed over all the bands) as expected, exhibits a uniform distribution except at the ends, where it smoothly drops to zero conforming to the hard-wall boundary conditions.

The internal picture along slice B-B (one node to the left of the constriction) is much different from that along A-A (Fig. 5.1b). The shape of the modes (wrt. $z$) change rapidly along the $x$ dimension as electrons propagate from a wide (6.2 nm) to a narrow (1.5 nm) quantum well (Sec. 3.4). Therefore, along slice B-B, individual mode energies and mode eigenfunctions do not provide an accurate description of the effective potential energy and spatial distribution of electrons. Non-zero values of the off-diagonal elements within the mode-space Hamiltonian (Eq. 5.2) results in intra mode scattering (scattering between modes belonging to the same conduction band valley) even in the ballistic limit. In fact the charge density distribution (along $z$) is peaked at the device center for both the primed and unprimed valleys (Fig. 5.2b, center). Electrons from the unprimed valleys respond to the channel directed electric
field with a light effective mass \(0.19 m_0\) and hence have large wavelengths (inversely proportional to the effective mass), while those from the primed valleys respond with a heavy effective mass (on average) and have short wavelengths. Therefore, the horizontal distance (along \(x\), away from the oxide regions) over which the electron density increases (from zero) to the values along section A-A is greater for the unprimed valleys (as illustrated in Fig 5.2b, center). Since the electron waves have propagating solutions only within the narrow constriction, individual valley and the net charge distributions along B-B exhibit sharply increasing trends between the top and bottom silicon-oxide interfaces (0 and \(1.5\) nm). At the constriction (section C-C), there is only one propagating mode (from the unprimed valleys) within the energy range over which electrons are injected into the narrow region. All the other modes are evanescent and decay exponentially (note that the electron density from the primed valleys is much below that from the unprimed valley along section C-C). The electron distribution within section C-C is peaked at the device center and conforms to the hard-wall boundary conditions used to terminate the Hamiltonian at the oxide-silicon interface.

The internal pictures presented so far clearly indicate that the solution scheme captures all of the quantum effects within the WN geometry accurately. In order to discuss the role of scattering, we plot the subband energies along section B-B in Fig 5.2a (right). The subband energy of the lowest unprimed mode at the constriction (C-C) is also superposed on this plot (marked line). Note that although these energies represent a rough estimate of the potential energy for electrons (due to the mode coupling issues discussed earlier), it is clear that scattering (especially intervalley) is an important aspect of transport within the WN structure. Scattering enhances the ability of the evanescent modes to transfer charge into the propagating modes at the constriction. If the effect of scattering were not included, there would be no interaction between the primed and unprimed modes at sections B-B and C-C despite their being energetically very close. Although these internal plots correspond to the on-state \((V_{GS} = V_{DS} = V_{DD})\) of operation, the overall picture within the WN device
Fig. 5.2. (a) The conduction and subband profiles along slices A-A (left) and B-B (right) for the WN device (Fig. 5.1b). Unprimed mode energies are illustrated using solid horizontal lines while primed mode energies are represented by dashed horizontal lines. The first unprimed mode within slice C-C is also superposed (right) and marked (triangles). (b) The charge distribution along slices A-A, B-B and C-C is presented. The net charge (dashed and dotted line) is obtained by summing charge contributions from the unprimed (solid line) and primed (dashed line) valleys.
is similar for all states of biasing. The internal picture throughout the UN device resembles that along section C-C (the UN device exhibits single mode occupancy everywhere).

Interesting features of the solution scheme are illustrated in Fig. 5.3 where the energy (channel directed energy) resolved local density of states (LDOS) within the WN device is visualized in the on-state (with scattering) for both the unprimed and primed valleys. White areas in the figure represent a large density of states, while dark regions indicate a low LDOS. The profile of the first unprimed subband is also superposed on the plots in order to provide a rough estimate of the minimum potential energy for electrons in different regions of the device (note that the subband profile is discontinuous at the WN interface, because of an abrupt change in the width of the quantum well from 6.2 nm to 1.5 nm). The well demarcated subband structure in the LDOS spectrum away from the WN interface is gradually lost due to mode coupling as one approaches the WN interface. Reflections at the WN interface result in a reduced LDOS at the constriction, which then sharply increases into the large S/D reservoirs (also evident in the 2D charge density plotted in Fig. 5.6b). Figure 5.3a clearly indicates that only the unprimed valleys contribute to the LDOS within the narrow region. Modes belonging to the primed valleys are completely reflected and decay exponentially within the narrow region. These modes indirectly inject current into the narrow region through inelastic scattering processes. Inelastic scattering also destroys coherent oscillations in the LDOS and the LDOS decays to zero within the forbidden regions below the first unprimed subband energy.

5.3.2 Effect of differing channel access geometry

Figure 5.4 presents the self-consistently simulated $I_{DS}$ vs. $V_{GS}$ and $I_{DS}$ vs. $V_{DS}$ characteristics for both the UN and WN device geometries, including the effects of scattering. For the same gate work function both geometries exhibit identical off-currents ($\sim 7$ A/m), subthreshold slopes (subthreshold swing is $\sim 80$ mV/dec) and
Fig. 5.3. (a) The energy resolved LDOS spectrum within the WN device (Fig. 5.1b) in the on-state, for the unprimed valleys. Note that all of the states within the narrow region is due to the unprimed valleys. (b) The energy resolved LDOS spectrum within the WN device from the primed valleys. States belonging to the primed valleys are reflected at the WN interface and decay exponentially within the narrow region. The first unprimed subband (dashed line) is also superposed on the plots. This subband is discontinuous at the WN interface due to an abrupt change in the width of the quantum well (from 6.2 nm to 1.5 nm).
threshold voltages (Fig. 5.4a). Also, for the same off-current, the simulated on-current for the device with a WN geometry is only $\sim 10\%$ lower than that of the device with a UN geometry. Note that this difference in the on-current represents the worst case scenario. In reality, the geometry of the S/D regions would not be hyper-abrupt, but would flare out gradually in order to reduce the parasitic gate-to-S/D capacitances. For such geometries it is possible that the differences in the on-current between UN and WN devices could be even lower than 10%. The output characteristics for both device structures are weakly saturated because of the low power supply voltage used in this study and due to series resistance effects. Simulated I-V characteristics for the WN device can be understood by viewing the large S/D regions as ideal reservoirs. These regions which maintain a near equilibrium distribution (due to multiple subband occupancy and scattering), inject and extract electrons from the narrow region while maintaining macroscopic charge neutrality. Therefore, current flow depends on the ease with which electrons can enter (exit) and propagate through the narrow region. We compare the transport properties of the UN and WN devices by focusing on the internal picture once again.

The conduction band profiles (averaged along $z$) in the off-state ($V_{DS}=0.4\,\text{V}, V_{GS}=0\,\text{V}$) for the WN and UN devices are compared in Fig. 5.5a (left). Note that although the conduction band profiles for the UN and WN devices match within the narrow region, there is a large built-in potential at the interface between the wide and narrow regions of the WN device (Fig. 5.5a, left). To explain the origin of this built-in potential we note that both the wide and the narrow S/D regions are degenerately doped ($10^{20}\,\text{cm}^{-3}$). For the same degenerate bulk doping value, the equilibrium Fermi level within a neutral narrow, n-type semiconductor is much higher than the conduction band edge of a comparably doped, neutral wide semiconductor, because only one subband is occupied in the narrow semiconductor while several are occupied in the wide semiconductor. Therefore, when the wide and narrow portions are intimately connected, electrons spill from the narrow into the wide region (regions of high to low Fermi energy), and a built-in potential develops to equalize the Fermi levels.
Fig. 5.4. (a) Simulated $I_{DS}$ vs. $V_{GS}$ characteristics for the UN (dashed line) and WN (solid line) devices at $V_{DS}=0.4$ V (includes the effect of scattering). Both devices exhibit the same off-current ($\sim 7$ A/m) and subthreshold slopes. (b) Simulated $I_{DS}$ vs. $V_{DS}$ characteristics for the UN (dashed line) and WN (solid line) devices at $V_{GS}=0.4$ V. For the same off-current, the simulated on-current for the WN device is $\sim 10\%$ lower than that of the UN device due to quantum mechanical reflections at the WN interface.
Fig. 5.5. (a) The averaged conduction band profile (left) in the off-state (includes the effect of scattering) for the UN (dashed line) and WN (solid line) devices. Note the large built-in potential at the WN interface. The corresponding subband profiles are visualized on the right. Many subbands (dots) are occupied in the S/D regions of the WN device. These modes inject current into the single propagating mode within the narrow region due to scattering (modes line up as shown in Fig. 5.2b). The UN device exhibits single mode occupancy everywhere (dashed line). (b) The Büttiker probe potentials within the UN (triangles) and WN (solid line) devices indicates that all of the applied voltage drops within the unconducting channel in the off-state.
The corresponding subband profiles (eigenvalues of $\alpha$ in Eq. 3.3) are visualized in Fig. 5.5a (right). It should be remembered that within the WN device, these subbands have clear physical meaning away from the WN constrictions. They have been visualized in order to qualitatively understand the off-state behavior of the I-V characteristics. Note once again that these subband energies are discontinuous at the interface between the wide and the narrow regions due to an abrupt change in the width of the quantum well (6.2 nm to 1.5 nm). Since the first subband is all that matters within the narrow region, high energy modes are not plotted within this region in Fig. 5.2a (right). Note that scattering (both geometric and real) between different modes at the WN interface ensures that electrons get injected into the single mode propagating within the narrow region (this mode lines up between the third unprimed and the first primed subbands as shown in Figs. 5.2a and 5.5a). However, electrons injected into this mode from the source are reflected off the large source-to-channel barrier (within the narrow region) in the off-state. Such a barrier is also present within the UN device (where the subband has a clear meaning everywhere) as seen from Fig. 5.2a (right). Since the height of this barrier (governed by the gate work function) is the same for both the UN and WN devices, it is clear that their off-currents should be identical. In Sec. 5.2, we mentioned that the Fermi level of each Büttiker probe (which represents an isolated scattering center) is adjusted to ensure current continuity. Therefore, these Fermi potentials can be treated as a measure of how the applied source-to-drain voltage drops within the device. In the off-state, the channel is unconducting and we expect all of the applied voltage to drop in the channel region irrespective of the device geometry. This is clearly evident from Fig. 5.5b, where the scatterer Fermi levels are plotted along the channel for both, the UN and WN geometries.

The self-consistently calculated conduction and subband profiles in the on-state (with scattering) are plotted in Fig. 5.6a. In the on-state, the potential barrier for electrons is suppressed (compared to Fig 5.5a) due to the large voltage applied to the gate electrodes. The on-current of the WN device depends on the ability of the
large S/D reservoirs to inject (extract) electrons from the narrow region and the transport properties of the narrow region itself. As mentioned earlier inter subband scattering aids this injection and extraction process. For the UN device, the S/D reservoirs are natural extensions of the device itself. However, this is not the case within the WN device as evidenced by the 2D charge density (integrated along \( z \)) plots in Fig. 5.6b. In our discussion of the off-state we mentioned that an intimate connection between comparably doped, neutral, wide and narrow n\(^+\) regions results in a transfer of charge from the narrow to the wide region. Such a charge transfer occurs at the wide-to-narrow source and drain interfaces within the WN device. This charge transfer results in a depletion region which extends into S/D regions of the intrinsic device (Fig. 5.6b). As the drain voltage is increased from zero to \( V_{DD} = 0.4 \, V \), the n\(^+\)(wide) – n\(^+\)(narrow) junction at the source becomes reverse biased thereby extending the depletion region further into the intrinsic source. On the other hand, the positive drain voltage forward biases the n\(^+\)(narrow) – n\(^+\)(wide) junction on the drain side thus shrinking the depletion region as shown in Fig. 5.6b. Note that transport properties of the UN device are identical to those of the intrinsic WN device because they have same S/D extension and channel mobilities and also exhibit single mode occupancy. Therefore, observed differences in the on-current between the UN and WN devices (for the same off-current) are due to the depletion effects coupled with quantum mechanical reflections at the WN interface.

The effect of self-consistency is also evident in Figure 5.6b. Self-consistent electrostatics causes macroscopic charge neutrality (net 2D charge equals the net 2D doping) to be achieved within both, the large S/D reservoirs as well as the narrow S/D extensions (away from the WN interface) of the WN device. Due to the abruptness of the S/D-to-channel junctions and the high doping within these narrow S/D extensions, the effect of fringing fields from the gate within the WN device has little or no effect on the on-state charge density. In fact, the 2D charge profile within the UN device matches that of the WN device almost identically in the narrow region as shown in Fig. 5.6b.
Fig. 5.6. (a) The average conduction band profile (left) in the on-state (includes the effect of scattering) for the UN (dashed line) and WN (solid line) devices. Note that the source-to-channel barrier is depressed due to the large gate voltage and that differences in the on-current between the UN and WN devices reflects the ability of the large S/D regions within the WN device to inject and extract electrons from the narrow region. (b) The 2D charge density within the UN (dashed line) and WN (solid and dotted lines) devices in the on-state. The S/D regions within both, the UN and WN devices exhibit charge neutrality (2D charge equals the net 2D doping). Charge transfer from the narrow to the wide regions at the WN interface results in a depletion region within the intrinsic S/D of the WN device.
In the on-state the channel is conducting, so a significant fraction of the applied source-to-drain voltage drops within the S/D reservoirs and extensions. As illustrated in Fig. 5.7a, where the Büttiker probe Fermi potentials are plotted in the on-state \((V_{GS}=0.4\, \text{V}, V_{DS}=50\, \text{mV})\), the applied voltage drops linearly within the S/D regions of the UN device. Within the WN device, however, almost all of the voltage drop in the heavily doped S/D regions occurs at the interface between the wide and the narrow regions. We can interpret the nature of this voltage drop by deriving a sheet resistivity \(\frac{\partial \mu}{\partial x} / I_{DS} W\) at low \(V_{DS}\), which is plotted in Fig 5.7b [4]. Both the UN and WN devices exhibit the following resistances: 1) quantum contact resistance 2) S/D resistance 3) tip resistance and 4) channel resistance. In addition, the WN device exhibits a large quantum-mechanical spreading resistance at the interface between the wide and the narrow regions. Note that within the S/D reservoirs of the WN device, the large number of propagating modes causes the sheet resistivity to be lower than that of the UN device (single mode conduction). Within the intrinsic portion of the WN device, however, the nature of the voltage drop as well as the sheet resistance looks identical to that of the UN device (same mobilities and single mode occupancy).

Next, we examine the effect of the wide-to-narrow aspect ratio. If we progressively increase the thickness of the S/D regions (beyond 1.5 nm, which corresponds to the UN device) and examine the on-current, we obtain the trend shown in the inset of Fig. 5.7b. As the thickness of the S/D regions is increased, the bulk S/D resistance decreases (number of propagating modes increases), while the spreading resistance increases (quantum mechanical reflections), thus resulting in an optimum value of the on-current (\(\sim 5\%\) higher than that of the UN device). A further increase in the thickness of the S/D regions causes the increasing spreading resistance to dominate, thus resulting in a decrease in the on-current. However, beyond a certain thickness value (wide-to-narrow aspect ratios exceeding 4), the spreading resistance and hence the on-current saturates at a value which is \(\sim 10\%\) less than that of the UN device. This behavior of the resistivity and the on-current in the presence of scattering clearly
Fig. 5.7. (a) The profile of the Büttiker probe potentials within the UN (dashed line) and WN (solid line) devices in the linear response regime ($V_{GS}=0.4$ V and $V_{DS}=0.05$ V). Note that the applied voltage drops nonlinearly within the S/D regions of the WN device. Most of the voltage drops at the WN interface. (b) Derived sheet resistivity plots within both, the UN (dashed line) and WN (solid line) devices indicate the following resistances: 1) quantum contact resistance 2) S/D resistance 3) tip resistance and 4) channel resistance. In addition, the WN device also exhibits a large spreading resistance. This resistance which degrades the on-current of the WN device when compared to the UN device, saturates as the thickness of the S/D regions is increased beyond a certain value (inset).
demonstrates that an understanding of the intrinsic device (Fig. 5.1b) is sufficient to explain the transport properties of devices with different S/D geometries.

Figure 5.8a plots the self-consistently calculated current spectrum at the source and drain contacts (slices 1 and $N_X$ in Fig. 5.1) as a function of the channel directed energy for both the UN and the WN devices in the on-state (while reading the plot it should be noted that the source injected current is positive for electrons entering the device, while the drain collected current is negative for electrons leaving the device). Since Büttiker probes simulate the effect of inelastic scattering, we observe that the drain collected current is relaxed in energy when compared to the source injected current for both, the UN and the WN devices. Within the UN device which exhibits single mode occupancy, the source injected current smoothly increases for energies above the subband edge and finally tails to zero as the current at the source end is cutoff by the source Fermi function. Within the WN device, current at the source is injected into several subbands (their energies are indicated by the vertical lines in Fig. 5.8a). Therefore the current spectrum for the WN device exhibits a stepped behavior, with each step occurring at the energy at which a specific mode starts propagating. There is a monotonic increase in the current spectrum (till it is cut off by the source Fermi function) because both the primed and the unprimed modes within the wide source reservoir can inject current into the single unprimed mode within the narrow region due to scattering.

If we compare this picture to the current spectrum obtained by ballistically simulating the WN device non-selfconsistently (where the 2D potential profile is fixed from the scattering simulations), we observe large qualitative differences in the current spectrum (Fig. 5.8b) for the UN and WN devices. Since scattering is turned off, there is no energy relaxation of the source injected current. Therefore, the source and drain current spectra are symmetric. Also, the source current spectrum for the WN device does not increase monotonically. Coherent quantum mechanical reflections at the WN interface generate strong oscillations in the current spectrum and greatly degrade the current flowing through the WN device when compared to the
Fig. 5.8. (a) The current spectrum within the UN (dashed line) and WN (solid line) devices in the on-state with scattering. Inelastic scattering causes the source injected current spectrum to relax in energy. Within the UN device current at the source is injected into a single mode. Within the WN device the source current is injected into several unprimed (solid vertical lines) and primed (dashed vertical lines) modes. Scattering, at the WN interface enables these modes to inject current into a single mode within the narrow region. (b) The ballistically (non-selfconsistent) calculated current spectrum within the UN (dashed line) and WN (solid line) devices. In the absence of scattering, there is no energy relaxation of source injected current spectrum and S/D current spectra are symmetric. Coherent reflections, due to abrupt changes in device geometry greatly degrade current flow with the WN device when compared to the UN device.
Fig. 5.9. Ballistically simulated (non-selfconsistent, potential profiles obtained from self-consistent scattering simulations are used) I-V characteristics for the UN (dashed line) and WN (solid line) devices indicate that the on-currents for the two device structures could differ by as much as $\sim 250\%$. In contrast, once scattering is included, the I-V characteristics for the two devices differ by no more that $\sim 10\%$.

UN device (for example, note the clear reduction in current at energies around the first primed mode when compared against Fig. 5.8a)). A ballistic simulation (non-selfconsistent) of the I-V characteristics for the UN and WN devices indicates that the computed current values could differ by as much as $\sim 250\%$ (as opposed to $\sim 10\%$, when scattering is turned on) due to coherent reflections caused by an abrupt change in the device geometry (Fig. 5.9). Such differences have been recently reported in the literature [68] [69]. These results clearly emphasize the importance of including the effects of scattering when performing a detailed design analysis of nanoscale transistors operating at room temperature.

5.4 Discussion

In this section we discuss an approach which can be applied to design and assess the performance of nanoscale silicon transistors efficiently. Device design and optimization requires the ability to perform a large number of simulations within a
reasonable time period. Even for the small device dimensions considered in this paper (Fig. 5.1b), the simulation time per bias point on a Linux cluster was very large (∼7 hrs). The computational burden was large because we explicitly included the large S/D reservoirs within our quantum simulation domain. In real devices these regions have an area which is much larger than 6.2 nm × 10 nm in order to reduce the silicide resistance (not considered in our study), and are virtually impossible to handle within a quantum mechanical framework. Therefore, we examine the possibility of decomposing the simulation domain into an intrinsic (region within the dashed box in Fig. 5.1b and parasitic components (large S/D reservoirs). Knowing the characteristics of the intrinsic device and the parasitic resistances, the actual device performance (extrinsic characteristics) can be obtained by solving the following equations.

\[
V_{GS} = V_{GS}^{\text{intrinsic}} - I_{DS}R_S \\
V_{DS} = V_{DS}^{\text{intrinsic}} - I_{DS} [R_S + R_D]
\] (5.7)

where \(V_{DS}\) and \(V_{GS}\) are the extrinsic voltages. Note that in order to simulate the intrinsic device only a few modes need to be included when expanding the device Hamiltonian in mode-space (Eq. 5.2). Also note that within the intrinsic device each mode can be treated as decoupled (Sec. 3.4). Therefore, the computational burden for quantum mechanically simulating the intrinsic device is extremely small. The parasitic resistances of the extrinsic S/Ds can be extracted using a classical simulator and used in Eq. 5.7.

Simulation of the intrinsic device requires an appropriate set of boundary conditions. Within the intrinsic device the transport equation (NEGF) is solved by imposing fixed boundary conditions on the Fermi potentials of the source and drain contacts (specified by \(V_{DS}\)). However, we impose floating boundary conditions when solving Poisson’s equation. This boundary condition is realized by setting,

\[
\hat{x} \cdot \nabla V = 0
\] (5.8)
at the source and drain ends of the simulation domain. Conventional transport models use fixed potential boundary conditions assuming equilibrium statistics and charge neutrality at the contacts. Under non-equilibrium conditions, equilibrium statistics no longer apply within the intrinsic device [43], and our use of the floating boundary condition helps us capture the effect of coupling the intrinsic device to the large scattering dominated contacts. Note that even if the potential is allowed to float, it cannot float to any arbitrary value. The potential floats relative to the source/drain Fermi levels in order to achieve charge neutrality (Fig. 5.6b). To explain the floating boundary condition, we plot the averaged conduction band profile for the device shown in Fig. 5.1b under equilibrium ($V_{GS} = 0.4\,V, V_{DS} = 0\,V$) and non-equilibrium ($V_{GS} = 0.4\,V, V_{DS} = 0.4\,V$) conditions (with scattering) in Fig. 5.10. The large S/D regions maintain a near equilibrium distribution even when a large bias is applied to the drain, because of scattering between the many modes which contribute to conduction in this region. Therefore, the electrostatic potential within this region is unaffected even when a large bias is applied to the drain as shown in Fig. 5.10. However, within the intrinsic device the conduction band floats to lower values as the drain bias is increased (Fig. 5.10). Under equilibrium conditions, both the $+k_x$ and the $-k_x$ states in the intrinsic source are filled by a single Fermi level resulting in zero net current. As the drain bias is increased to higher and higher values, the drain injected half of the electron distribution is suppressed within the intrinsic source. To achieve charge neutrality in this region (Fig. 5.6b), the conduction band floats to lower energies, an effect that is captured through the imposition of floating boundary conditions (Sec. 2.4).

In order to verify our domain decomposition algorithm we simulate a resistor (using NEGF) with ideal, non-uniform contacts in the linear response regime (inset of Figure 5.11a). The resistor has the dimensions of the wide S/D reservoirs for the device in Fig 5.1b. The left contact to the resistor is 6.2 nm deep, while the right contact is 1.5 nm deep. The dimension of the right contact is chosen to simulate the coupling of the source reservoir to the narrow intrinsic device. All the regions of the
Fig. 5.10. The averaged conduction band profile within the WN device (on-state) at equilibrium (solid line) and on the application of a large drain voltage (dashed line, $V_{DS}=0.4$ V). It is clear that the conduction band edge is unchanged within the large S/D reservoirs as they maintain a near equilibrium distribution even when a large drain bias is applied. The conduction band within the intrinsic source (Fig 5.1b) floats to lower values to enforce charge neutrality (fig 5.6b).

resistor which are not covered by contacts, are terminated using hard-wall boundary conditions. Simulated charge density and Büttiker probe potentials for this resistor are plotted in Fig. 5.11a. As expected, reflections at the right edge causes the charge density to drop at the right end of the resistor. The 2D charge density gradually increases to the charge neutral value over a few Debye lengths away from the right contact. Since the charge density is non-uniform, the applied voltage drops non-linearly. The quasi-Fermi potential (of the Büttiker probes) drops rapidly in regions where the charge density is reduced in order to conserve current. The nature of this voltage drop is similar to the actual profile of the Büttiker probe potentials illustrated in Fig 5.7a (for the WN device). Based on our calculations in the linear response region, the estimated parasitic source (drain) resistance was $\sim 57$ ohm$-\mu$m. This resistance value was used to estimate the I-V characteristics of the WN device shown in Fig 5.1b, using the intrinsic I-V and Eq. 5.7. A comparison (Fig. 5.11) between the I-V characteristics extracted using our domain decomposition algorithm
Fig. 5.11. (a) Simulated 2D charge and Büttiker probe potential profiles within a resistor with non-uniform contacts (inset) in the linear response regime. Note that the nature of the applied voltage drop is similar to the drop with the S/D regions of the WN device pictured in Fig. 5.7a. The quasi-Fermi potential drops steeply in the regions where the 2D charge density decreases to enforce current continuity. (b) Quantum mechanically simulated I-V characteristics (solid line) are compared against the I-V characteristics obtained using the domain decomposition algorithm (dashed line with triangles). The I-V characteristics of the intrinsic device is also illustrated (dashed line).
and the actual I-V (from a quantum simulation of the entire device) characteristics presented in Fig 5.4b (for the WN device) indicates that our domain decomposition scheme provides a fairly accurate and efficient technique for analyzing and designing nanoscale transistors [12] [70]. The S/D reservoirs and the intrinsic device can be treated independently.

5.5 Summary

We presented a computationally efficient method to quantum mechanically simulate electron transport within nanoscale transistors including the effect of scattering. This modeling scheme, which is based on an expansion of the effective mass Hamiltonian in coupled mode-space, was applied to nanoscale transistors with differing channel access geometries. Our simulation study indicated that devices with very different channel access geometries exhibit nearly identical current-voltage characteristics (no more than \(\sim 10\%\) differences in the worst case) once the effect of scattering was included. However, when simulated ballistically, devices with different access geometries yielded largely differing current-voltage characteristics due to coherent reflections (as much as \(\sim 250\%\) in the worst case).

We then proposed a domain decomposition algorithm which can be applied to efficiently assess and design nanoscale transistors at the end of the roadmap. This algorithm divided the device into intrinsic and parasitic components, simulated each component independently, and extracted the overall device characteristics through a simple interpolation scheme. A comparison of the extracted data against data obtained from rigorous quantum mechanical simulations of the entire device clearly showed that this domain decomposition algorithm is an accurate method for analyzing the transport properties of transistors at the end of the roadmap.
6. FUTURE WORK

This thesis addressed device physics, modeling and design issues for nanoscale, silicon transistors. The main accomplishments of this thesis are,

1. In chapter 2, we presented a modeling scheme for simulating ballistic hole transport in thin body, fully depleted, SOI p-MOSFETs including the effect of band nonparabolicity. This simulator was used to examine the effects of subband engineering on the ballistic performance of nanoscale, SOI transistors.

2. In chapter 3, we discussed the development and implementation of a general real-space simulator for n-MOSFETs in the ballistic limit, using the NEGF formalism. This simulator was used to benchmark a simplified, decoupled mode-space simulation scheme, which was based on an extension of the ideas presented in chapter 2. It was shown that the mode-space simulator is an attractive simulation tool for modeling transport in nanoscale, SOI n-MOSFETs.

3. In chapter 4, we extended the ballistic, decoupled mode-space simulator for SOI n-MOSFETs, to treat the effects of elastic and inelastic scattering rigorously, within the NEGF modeling scheme. This rigorous treatment of scattering was used to benchmark a simpler, phenomenological scattering model which was based on an idea inspired by Büttiker. We demonstrated that our simple model, which can be calibrated to mimic experimentally measured low-field mobility data, captures the essential physics of scattering in nanoscale transistors.

4. In chapter 5, we applied the NEGF formalism to examine the effects of differing device geometries on device performance, with and without scattering effects. Results presented in this chapter indicate that devices with largely differing geometries exhibit near identical characteristics once scattering is turned on, but
exhibit largely differing characteristics once scattering is turned off. A simple procedure to analyze and design single, dual and tri-gate SOI MOSFETs was also presented and verified by rigorous simulation results.

This work can be extended as follows,

1. In order to continue improving silicon device performance, alternate methods of device enhancement are being considered in combination with device scaling. One of these methods involves replacing the silicon substrate itself with strained silicon or with silicon-germanium alloys. Experimental data presented in the literature, indicates that such channel engineering greatly improves the performance of both p and n MOSFETs [71] [72] [73]. This performance improvement has been attributed to an alteration of the silicon bandstructure.

   We could use the techniques outlined in chapter 3, to model such novel devices by extending our approach to treat a generalized $\vec{K}$ dependent Lüttinger Hamiltonian [33] [74]. Such treatment would enable us accurately ascertain the improvement in ballistic device performance as a result of channel engineering. Also, the inclusion of a detailed bandstructure, would enable us model unmodified, silicon p-MOS devices more accurately. Anisur Rahman, in the Purdue Device Simulation Group is exploring this option.

2. Another potential candidate for device scaling is the silicon nanowire [11]. Silicon nanowires can be viewed as 1D conductors, with carriers quantum confined in all the directions that are normal to the direction of transport. The main advantages of silicon nanowires are: 1) Ease of fabrication, 2) Extremely high electron and hole mobilities and 3) Reduced short channel effects due to good gate electrostatics. Therefore, these structures are being aggressively studied recently. All of the techniques that were developed to study electron transport in SOI, n-channel MOSFETs in chapters 3, 4 and 5 can be directly applied to model electron transport in nanowires also. In fact, since transport is one-dimensional, it may also be possible to apply the detailed treatment of
scattering which was presented in chapter 4 to simulate full range current vs. voltage characteristics for short nanowire transistors using the decoupled mode-space solution scheme (chapter 3). Jing Wang, in the Purdue Device Simulation Group is exploring this option.

3. Schottky barrier transistors are being explored by device engineers for possible applications in future VLSI technology. Its relatively simple structure (no S/D extension regions), and low parasitic resistance, make it an ideal candidate for device scaling. Simulation studies of such devices have been based on semiclassical transport models, that use the WKB approximation to capture tunneling effects. Such approaches do not capture the effect of tunneling accurately and do not treat the effects of quantum coherence between the s/D and drain barriers. As device lengths scale to the 10 nm regime, these effects may become important. Our NEGF simulation platform could be applied to simulate such devices and examine their scaling properties wrt. novel silicon transistor structures. Sayed Hasan, in the Purdue Device Simulation Group is exploring this option.

We conclude by noting that the NEGF formalism provides a general simulation platform for modeling quantum transport in nanoscale devices and that there will be several opportunities to apply this formalism as the quest to find a replacement for silicon technology intensifies in the future.
LIST OF REFERENCES


APPENDICES


**Appendix A: The Self-energy calculation for the leads**

To illustrate the self-energy calculation which accounts for the device leads, we consider the effect of coupling the active device Hamiltonian to the drain. The infinite Hamiltonian (Eq. 3.2) and its Green’s function (Eq. 3.6) can be partitioned as follows,

\[
\begin{bmatrix}
G_{\text{device}} & G_{\text{device},D} \\
G_{D,\text{device}} & G_D
\end{bmatrix} =
\begin{bmatrix}
E(k_x, k_z)I - h(x, z)_{\text{device}} & -\beta \\
-\beta & E(k_x, k_z)I - h(x, z)_D
\end{bmatrix}
\]  \hspace{1cm} (A.1)

In Eq. A.1, subscript “D” is used to indicate the infinite block of \(h(x, z)\) and \(G\), representing the drain. The matrix block we are interested in is \(G_{\text{device}}\) as we do not care about the Green’s function within the drain. Using Eq. A.1, \(G_{\text{device}}\) can be expressed in terms of known quantities as,

\[
G_{\text{device}} [E(k_x, k_z)] = [E(k_x, k_z)I - h(x, z)_{\text{device}} - \Sigma_D]^{-1}
\]  \hspace{1cm} (A.2)

where the drain self-energy matrix is,

\[
\Sigma_D = \begin{bmatrix}
0 & 0 & \cdots \\
0 & 0 & \cdots \\
-\beta & 0 & \cdots
\end{bmatrix}
\times \begin{bmatrix}
\cdots & 0 & -\beta \\
\cdots & 0 & 0 \\
\cdots & \cdots & \cdots
\end{bmatrix}^{-1}
\]  \hspace{1cm} (A.3)

Note that for evaluating the matrix product in Eq. A.3, we only need the first block of the inverse of the infinite matrix associated with the drain. Also, note that
the diagonal blocks of this infinite matrix are repeated due to translational invariance within the drain \((\alpha_{N_X} = \alpha_{N_X+1} = \cdots)\). Using this property, and partitioning the matrix as shown in Eq. A.3, a closed form expression for the first block of the inverse (denoted by \(g_D\)) of the infinite matrix, can be obtained as,

\[
I = g_D [E(k_x, k_z)I - \alpha_{N_X+1} - \beta g_D \beta]
\]  

(A.4)

Once \(g_D\) has been solved for, we have,

\[
\Sigma_D = \begin{bmatrix}
0 & \cdots & 0 \\
0 & \cdots & 0 \\
0 & \cdots & \beta g_D \beta
\end{bmatrix}
\]  

(A.5)

Note that, only the last vertical slice of the device couples to the drain. Therefore the self-energy for the drain (Eq. A.5) has a single non-zero block that perturbs the last diagonal block of the device Hamiltonian. To solve Eq. A.4, a basis transformation has to be performed. The eigen vectors of \(E(k_x, k_z)I - \alpha\), diagonalize \(g_D\) simultaneously. Therefore we change basis from 2D real-space to a basis that is composed of the eigenvectors of \(E(k_x, k_z)I - \alpha\) (equivalent to a mode-space transformation at the boundary). This reduces Eq. A.4 to a set of decoupled quadratic equations that can be solved for the diagonal entries \(g_D\), in the transformed representation. It should be noted that each of these equations results in two roots. The root representing outgoing waves is selected as we are ultimately interested in obtaining the retarded Green’s function for the device. An inverse basis transformation is then applied to evaluate \(g_D\) in 2D real-space. A similar procedure is invoked to solve for the self-energy part associated with the source. The final size of the self-energy matrix is \((N_X \times N_Z)^2\) for the real-space solution and \((N_X)^2\) for the decoupled mode-space solution.
Appendix B: From wavefunctions to Green’s functions

In this appendix, we describe the decoupled mode-space approach used in chapter 3 from a wavefunction perspective. Key equations in the wavefunction description are translated into the language of the NEGF formalism and the two methodologies are shown to be equivalent when modeling electron transport in confined transistor geometries. The starting point is Eq. 3.27, which is the decoupled mode-space transformation of the 2D effective mass Hamiltonian for the device shown in Fig. 3.1a. Because the decoupled mode-space solution reduces the 2D problem to a set of 1D problems, one for each subband and transverse energy ($E_{k_j}$ in Eq. 3.27), the solution scheme from a wavefunction perspective is much like that of a true 1D problem.

Figure B.1 is the sketch of a generic subband (labeled $m$) versus position ($x$). Semi-infinite contacts are attached to this subband at the source and drain ends. Because the potential is assumed to be uniform within the S/D, solutions to the 1D wave equation in the S/D are plane waves. If a unit amplitude (Fig. B.1) is injected from the left (source) contact, then some portion reflects from the device and some transmits across and exits the perfectly absorbing right (drain) contact,

$$ \tilde{\Phi}_m(x) = \begin{cases} 1 e^{ik_x x} + r_m e^{-ik_x x} & x < 0 \\ t_m e^{ik_x x} & x > L \end{cases} $$

where $L$ is the length of the active device, $k_x$, is the electron wave-vector along $x$ and $r_m$ and $t_m$ are the reflection and transmission amplitudes for source injection into mode $m$. By solving Eq. 3.27 subject to the boundary conditions, Eqs. B.1, we find the wavefunction due to the injection of a unit amplitude wave from the source. The
corresponding electron density for mode $m$, at each $k_x$ is obtained by summing the contributions from all the transverse modes ($k_j$ in Eq. 3.27).

$$n_m [x, k_x] = \frac{1}{W} \sum_{k_j} |\tilde{\Phi}_m(k_x, x)|^2 f(E - \mu_S)$$  \hspace{1cm} (B.2)$$

where $E$ is the total energy of the electron and the subscript, $S$, refers to injection from the source. The probability that the state at energy $E$ is occupied is given by the source Fermi function ($f$) because we assume that scattering maintains thermal equilibrium in the contacts. Since the width in the transverse direction, $W$, is assumed to be large we can convert the sum over the transverse modes in Eq. B.2 to an integral over the transverse energy. The net 2D charge density due to source injection, at each $x$ position, is then obtained by summing over all the positive $k_x$ states (which are injected), with the sum over $k_x$ similarly converted to an integral over the longitudinal energy ($E_l = E - E_{k_j}$). The final expression for the source injected charge into mode $m$ is,

$$n_S^{m}[x] = \int_{-\infty}^{+\infty} n_S^{m}[E_l]dE_l$$
\[
\frac{1}{a} \sqrt{\frac{2m^*_y k_B T}{\pi \hbar^2}} \int_{-\infty}^{+\infty} F_{-1/2}(\mu_S - E_i) \frac{A_S^m[E_i]}{2\pi} dE_i
\]  
\tag{B.3}
\]

with
\[
\frac{A_S^m[E_i]}{2\pi} = \frac{a}{2\pi} \frac{dk_x}{dE_i} |\Phi_m(x, k_x)|^2
\]  
\tag{B.4}
\]

being the local density of states due to source injection (spin degeneracy is included in Eq. B.3). A similar expression is obtained for the drain injected charge density. The net 2D charge density which is the sum of the source and drain injected components is identical to Eq. 3.11.

The source injected current for mode \( m \), at each longitudinal energy can be expressed as
\[
I_m[E_i] = \frac{q}{W} \sum_{k_j} T_{S-D}^m(E_i) [v_{x}]_{x=0} (E_i) f(E - \mu_S)
\]  
\tag{B.5}
\]

where \( T_{S-D} \) is the current transmission from the source to the drain and \( v_{x} \) is the \( x \) directed velocity for electrons in mode \( m \). The net source injected current is obtained by summing Eq. B.5 over all the positive \( k_x \) states and converting the sums over the wave-vectors to integrals. Note that the factor, \( dk_x/dE_i \), which is used to convert the sum over \( k_x \) to an integral over \( E_l \) is \( 1/[v_x \hbar] \). Therefore, the final expression for the source injected current into mode \( i \) is,
\[
I_S^m = \int_{-\infty}^{+\infty} I_S^m(E_i) dE_i
\]
\[
= \frac{q}{h} \sqrt{\frac{2m^*_y k_B T}{\pi \hbar^2}} \int_{-\infty}^{+\infty} T_{S-D}^m(E_i) F_{-1/2}(\mu_S - E_i) dE_i
\]  
\tag{B.6}
\]

The net current is obtained by subtracting the drain injected component as,
\[
I^m = I_S^m - I_D^m
\]
\[
= \frac{q}{h} \sqrt{\frac{2m^*_y k_B T}{\pi \hbar^2}} \int_{-\infty}^{+\infty} T_{S-D}^m(E_i) \left[ F_{-1/2}(\mu_S - E_i) - F_{-1/2}(\mu_S - qV_{DS} - E_i) \right] dE_i
\]  
\tag{B.7}
\]
Finally, the current transmission coefficient for mode, \( m \), \( T_{S-D}^m \) is

\[
T_{S-D}^m[E_l] = \frac{I_{\text{trans}}}{I_{\text{inc}}} = 1 - |r_m|^2 \quad \text{(B.8)}
\]

which, from Eq. B.1 can be expressed in terms of the computed wavefunction as

\[
T_{S-D}^m[E_l] = 1 - |\tilde{\Phi}_m(x = 0) - 1|^2 \quad \text{(B.9)}
\]

To evaluate the expression for the electron density and current in Eqs. B.3 and B.7, the wavefunction within the device must be known. In order to solve for the wavefunction due to source injection, we discretize Eq. 3.27 on a finite difference grid imposing the boundary conditions, Eq. B.1, to find

\[
\begin{bmatrix} EI - H - \Sigma \end{bmatrix} \tilde{\Phi}_m = -i \gamma_S \quad \text{(B.10)}
\]

where \( H \) is the \( N_X \times N_X \) discretized Hamiltonian operator (Eq. 3.16) and \( \Sigma \) (Eq. 3.18) is the self-energy matrix for mode \( m \). The \( N_X \times 1 \) vector \( \gamma_S \), is a source term that accounts for injection from the source contact. It has one nonzero component, the first

\[
\gamma_S(1) = i \left[ \Sigma(1, 1) - \Sigma^\dagger(1, 1) \right] = [2t_x \sin k_x a]_{x=0} \quad \text{(B.11)}
\]

where \( t_x \) (Eq. 3.5) is the coupling energy between adjacent sites on the finite difference grid and \( a \) is the grid spacing along \( x \). The wavefunction due to source injection can be expressed in terms of the retarded Green’s function as (from Eqs. B.10 and 3.17)

\[
\tilde{\Phi}_m = -iG\gamma_S \quad \text{(B.12)}
\]

In order to express the electron density and current due to source injection in terms of the Green’s function, we consider the following generalization.

\[
\tilde{\Phi}_m^\dagger \tilde{\Phi}_m = G\gamma_S\gamma_S^\dagger G^\dagger \quad \text{(B.13)}
\]

where

\[
\tilde{\Phi}_m = \begin{bmatrix}
\tilde{\Phi}_m^1 \\
\tilde{\Phi}_m^2 \\
\vdots \\
\tilde{\Phi}_m^{N_X}
\end{bmatrix} \quad \text{(B.14)}
\]
is an $N_X \times 1$ vector giving the value of the wavefunction for mode $m$ at each node.

The matrix product in Eq. B.13 is

$$
\tilde{\Phi}_m \tilde{\Phi}_m^\dagger = \begin{bmatrix}
\tilde{\Phi}_m^1 \tilde{\Phi}_m^{* 1} & \tilde{\Phi}_m^1 \tilde{\Phi}_m^{* 2} & \cdots & \tilde{\Phi}_m^1 \tilde{\Phi}_m^{* N_X} \\
\tilde{\Phi}_m^2 \tilde{\Phi}_m^{* 1} & \tilde{\Phi}_m^2 \tilde{\Phi}_m^{* 2} & \cdots & \tilde{\Phi}_m^2 \tilde{\Phi}_m^{* N_X} \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{\Phi}_m^{N_X} \tilde{\Phi}_m^{* 1} & \cdots & \cdots & \tilde{\Phi}_m^{N_X} \tilde{\Phi}_m^{* N_X}
\end{bmatrix}
$$

(B.15)

By generalizing $\gamma_S$ to an $N_X \times N_X$ matrix,

$$
\Gamma_S \equiv \gamma_S = \begin{bmatrix}
[2t_x \sin k_x a]_{x=0} & 0 & \cdots & 0 \\
0 & \cdots & \cdots & 0 \\
0 & \cdots & \cdots & 0 \\
0 & \cdots & \cdots & 0
\end{bmatrix}
$$

(B.16)

we can use Eq. B.13 to generalize the spectral function (Eq. B.4) to

$$
$$

(B.17)

which is an $N_X \times N_X$ matrix. It should be noted that the term $dk_x/dE_l$ in Eq. B.4 is $1/\Gamma_S(1,1)$. On substituting Eq. B.17 into Eq. B.3 we obtain Eq.3.19. An analogous expression for drain injection can be similarly derived.

To find the net current through mode $m$, Eq. B.7 still applies, but the current transmission $T_{S-D}$ needs to be expressed in terms of the Green’s function. Assuming source injection and using the boundary conditions for the same (Eq. B.1) we have

$$
J_m^{inc} \equiv 1v_x^1 \quad \text{and} \quad J_m^{trans} \equiv |t_m|^2 v_x^{N_X}
$$

(B.18)

The velocities at the injection (1) and exit ($N_X$) points are (Fig. B.1)

$$
v_x^1 = \frac{1}{\hbar} \left[ \frac{dE_l}{dk_x} \right]_{x=0} = \frac{2at_x}{\hbar} \left[ \sin k_x a \right]_{x=0} = \frac{a}{\hbar} \Gamma_S(1,1)
$$

$$
v_x^{N_X} = \frac{1}{\hbar} \left[ \frac{dE_l}{dk_x} \right]_{x=L} = \frac{2at_x}{\hbar} \left[ \sin k_x a \right]_{x=L} = \frac{a}{\hbar} \Gamma_D(N_X, N_X)
$$

(B.19)

where, $\Gamma_D$ is the analogue of $\Gamma_S$ (Eq. B.16) for drain injection. Using these results, the current transmission for mode $m$ is

$$
T_{S-D}^m[E_l] = \frac{J_m^{trans}}{J_m^{inc}} = \frac{|t_m|^2 v_x^{N_X}}{v_x^1} = |\tilde{\Phi}_m^{N_X}|^2 \frac{\Gamma_D(N_X, N_X)}{\Gamma_S(1,1)}
$$

(B.20)
Since
\[ \tilde{\Phi}_m^{N_X} = -i G(N_X, 1) \Gamma_S(1, 1) \quad (B.21) \]
we have
\[ T_{m-S-D}^m[E_l] = |G(N_X, 1)|^2 \Gamma_S(1, 1) \Gamma_D(N_X, N_X) \quad (B.22) \]
which when rewritten as
\[ T_{m-S-D}^m[E_l] = \text{Trace} \left[ \Gamma_D G \Gamma_S G^\dagger \right] \quad (B.23) \]
is Eq. 3.12. Note that if scattering were included using Büttiker probes, Eq. B.23, can be generalized to find the transmission between any two probes as
\[ T_{m-P-Q}^m[E_l] = \text{Trace} \left[ \Gamma_Q G \Gamma_P G^\dagger \right] \quad (B.24) \]
which is useful for adjusting the probe Fermi potentials to ensure current conservation.
Appendix C: The relation between the probe self-energy and the low field mobility

In order to relate the probe self-energy to a mobility, we first relate it to a mean free path, $\lambda$. To do so, consider a single mode (we drop the subscript $i$ for convenience) with a uniform potential, which is coupled to S/D contacts and $N$ identical Büttiker probes, each of which has the same coupling strength, $U_p$. The spacing between adjacent probes is $a$. If a unit amplitude is injected from the source (left contact), the net transmission (denoted $T(N)$) from the source to the drain can be expressed as,

$$\frac{1 - T(N)}{T(N)} = N \frac{1 - T_p}{T_p}$$

where, $T_p$ is the net transmission across a single probe. Note that Eq. C.1, has been derived by invoking the additive property of $(1 - T_p)/T_p$ [19]. The net transmission across a single probe, in terms of the transmission into the probe (denoted by $T_p^{in}$) is,

$$T_p = 1 - \frac{T_p^{in}}{2}$$

The reason $T_p$ does not equal $T_p^{in}$, is because each probe isotropically reinjects electrons into the system in order to conserve charge. Substituting Eq. C.2, into Eq. C.1 and solving for $1/T(N)$, we obtain

$$\frac{1}{T(N)} = 1 + N \frac{T_p^{in}/2}{1 - T_p^{in}/2} = 1 + \frac{L}{a} \frac{T_p^{in}/2}{1 - T_p^{in}/2} \sim 1 + \frac{L}{\lambda}$$

where, $L$ is the distance from the source to the drain. The transmission into a single probe is obtained analytically using Eq. 4.16. In the weak scattering limit ($U_p \sim 0$, therefore $T_p^{in} \sim 0$), the final form of the mean free path (from Eq. C.3), in terms of the lattice spacing $a$ is,

$$\lambda = 2a \frac{t_x^2}{U_p^2} \quad \text{where} \quad T_p^{in} = \frac{U_p^2}{t_x^2}$$

(C.4)
Once the mean free path is obtained, it can be related to the diffusion coefficient through Schokley’s relation, which in turn is related to the low field mobility through Einstien’s relation [56]. The relation between the low field mobility and the mean free path is,

\[ \text{mobility} = \lambda \sqrt{\frac{q^2}{2\pi k_B T m_x^*}} \frac{\tilde{\delta}_{-1/2}(\mu - E_{\text{mode}})\tilde{\delta}_{-1}(\mu - E_{\text{mode}})}{\tilde{\delta}_0^2(\mu - E_{\text{mode}})} \] (C.5)

where the arguments to the Fermi functions have been normalized by \( k_B T \) and \( E_{\text{mode}} \) represents the potential energy of electrons in a specific mode. Equations C.4 and C.5, relate the low field mobility to the Büttiker probe strength. Note that the degeneracy factors in Eq. C.5 are position dependent because the mode energy is position dependent. Also note that in our analysis, we assumed single mode occupancy. If several modes are occupied, the mean free path can be interpreted as an average mean free path for all modes and the equivalent mobility as an average low-field mobility for all electrons. The Büttiker probe strength in this case is adjusted differently for each mode (because \( t_x \) is mode dependent) to reflect the same average mean free path for all modes.

In order to verify the validity of our derivation, we simulate a uniformly doped resistor (thickness = 1.5 nm, length = 40 nm, \( N = 161 \) and doping = \( 10^{20} / \text{cm}^3 \)) self-consistently in the linear response region (low \( V_{DS} \)) for different probe strengths (or alternatively, mean free paths evaluated using Eq. C.4). The low bias conductance extracted from numerical simulations is compared against analytical values obtained using Eq. C.3 in Fig. C.1 (Note that the conductance is directly proportional to \( T(N) \)) [19]. Based on our analytical expressions (Eqs. C.3 and C.4), we expect a linear relationship between the inverse of the conductance and the inverse of the mean free path. It is clear from Fig. C.1 that the numerical and analytical values are in close agreement (within 10%) over a wide range of scattering strengths ranging from a mean free path of \( \infty \) (pure ballistic transport), down to 5 nm. This plot clearly validates our interpretation of the Büttiker probe strength and enables us calibrate quantum mechanical parameters to experimental mobility data in a simple
Fig. C.1. The low bias conductance for a 40 nm resistor is plotted as a function of scattering strength, from analytical (solid line) expressions and self-consistent simulations (dashed line). The analytic conductance matches the simulated value to within $\sim 10\%$, indicating that our relation between the Büttiker probe strength and the mean free path (Eq. C.4) is physically correct.

and elegant fashion. The slight discrepancy between the numerical and analytical values is because our analytical expressions are derived in the weak scattering limit.
VITA

Ramesh Venugopal was born on October 20th, 1973 in Chennai, India. He received the B.E. in Metallurgical Engineering from the Government College of Engineering, Pune, India in 1994. After a brief stint in the automobile industry for two years, he went on to receive an M.S. in Metallurgical Engineering from the University of Utah, Salt Lake City, in 1998 followed by a Ph.D. in Electrical Engineering from Purdue University, West Lafayette, Indiana in 2003. Ramesh Venugopal is currently employed by Texas Instruments, Dallas, Texas as a Research Engineer.