NANOSCALE MOSFETS: PHYSICS, SIMULATION AND DESIGN

A Thesis

Submitted to the Faculty

of

Purdue University

by

Zhibin Ren

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

October, 2001

To my family

Acknowledgments

I would like to express my sincere gratitude to my advisor, Professor Mark Lundstrom for his generous help throughout my Ph.D. study at Purdue. Professor Lundstrom is an admirable academic professional. He taught me not only his precious knowledge, but his exceptional professionalism. He impressed me very much by his responsibility and strict attitude in training students. He always provided timely and warm encouragement and support in difficult times. He gave me every opportunity to advertise my work at important conferences and in industry. I especially thank him for his prompt reading and careful critique of my thesis. Throughout my life I will benefit from the experience and knowledge I gained working with Professor Lundstrom.

I am also indebted to professor Supriyo Datta for his valuable guidance and insightful suggestions to help me accomplish my research work. As a well-known device physicist, he was always free in sharing his expertise in applying the Non-equilibrium Green's function approach for semiconductor device simulation.

I wish to thank Professor Dragica Vasileska at University of Arizona for providing me her source code of Schred, spending time in helping me understand the code, and giving me continuous instructions to further develop the program.

It is my pleasure to acknowledge Professors Gerold Neudeck, David Janes and Ron Reifenberger for serving on my examination committee.

Life at Purdue is unforgettable. It was very enjoyable working with my bright fellow students in the device group. We had serious research discussions and pleasant parties.

We established friendships in this big family. I really want to send my thanks and best wishes to them: Mr. Ramesh Venugopal, Dr. Farzin Assad, Dr. Kausar Banoo, Mr. Jung-Hoon Rhew, Mr. Jing Guo, Mr. David Rumsey, Mr. Anisur Rahman, Mr. Prashant Damle, Mr. Steve Bourland, and Mr. Sayed Hasan. Particularly, I wish to thank Mr. Ramesh Venugopal. In the past two years, as we worked closely on some topics, he impressed me by his enthusiasm towards reasearch, and his intelligence to accomplish them. Moreover, I very much appreciated Ramesh for his detailed reading of my thesis. I benefited a lot from his comments and discussions with him.

I would like to thank Dr. Phil Oldiges at IBM for proofreading Chapters 1 and 7 of this thesis.

I also wish to thank the Defense Advanced Research Projects Administration's Advanced Microelectronics Program (DARPA) and the Semiconductor Research Corporation (SRC) for financially supporting this work.

Finally, I owe more than words can describe to my family. This work is dedicated to them.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This thesis discusses device physics, modeling and design issues of nanoscale transistors at the quantum level. The principle topics addressed in this report are 1) an implementation of appropriate physics and methodology in device modeling, 2) development of a new TCAD (technology computer aided design) tool for quantum level device simulation, 3) examination and assessment of new features of carrier transport in nano-scale transistors, and 4) exploration of device design issues near the ultimate scaling limit with the help of the developed tools. We concentrate on the technical issues by investigating a double-gate structure, which has been widely accepted as the ideal device structure for ultimate CMOS scaling. We focus on quantum effects and non-equilibrium, near-ballistic transport in extremely scaled transistors (in contrast to quasi-equilibrium, scattering-dominant transport in long channel devices), where a non-equilibrium Green's function formalism (NEGF) has been used to deal with the quantum transport problem.

1. INTRODUCTION

1.1. Overview of the Problem

CMOS technology has been proven as one of the most important achievements in modern engineering history. In less than 30 years, it has become the primary engine driving the world economy. The secret to the success is very simple: keep delivering more functionality with fewer resources. Device scaling makes this possible. For decades, progress in device scaling has followed an exponential curve: device density on a microprocessor doubles every three years. This has come to be known as Moore's law [1]. The minimum dimension size of a single device for present day technology is about 100 nm in gate-length. Continued success in device scaling is necessary for further development of the semiconductor industry in the years to come. A group of leading companies publishes their projections for the next decade in the most recent International Technology Roadmap for Semiconductors (ITRS-99) [2]. The roadmap projects a device gate-length down to ~30 nm around 2014 [2]. This forecast *promises* us another ten years of brightness. Scaling beyond 30 nm, however, can be much more difficult and different. Remember, we are quite close to the fundamental limits of semiconductor physics. How much further down can we go? It is hard to answer. Nevertheless, without doubt, we are facing numerous challenges, both practically and theoretically. Device simulation requires new theory and approaches to help us understand device physics and to design devices at the sub-30nm scale. Efforts have been put forth in recent years [3-7], but much more is needed. For these purposes, we started a research project in 1997, the results of which make up this thesis.

The principle objectives of this thesis are: 1) to implement appropriate physics and methodology for device modeling, 2) to develop a new TCAD (technology computer aided design) tool for quantum level device simulation, 3) to examine and assess new features of carrier transport in nano-scale transistors, 4) to explore device design issues near the ultimate scaling limit with the help of the developed tools. We address the technical issues by investigating a double-gate structure, which more and more research evidence indicates to be the ideal device structure for ultimate CMOS scaling [4]. We focus on quantum effects and non-equilibrium, near-ballistic transport in extremely scaled transistors (in contrast to quasi-equilibrium, scattering-dominant transport in long channel devices), where a non-equilibrium Green's functions formalism (NEGF) has been used to deal with the quantum transport problem [8-10].

In the remaining parts of this chapter, we will give a quick review of why the double-gate structure is preferred for future device scaling. For comprehensive discussions on device scaling, readers can go to many references, for example [3, 11-12]. We will also present a brief introduction to the NEGF approach before jumping to the extensive discussions in later chapters. For detailed description of this theoretical framework, one may want to read the excellent text books by Datta and Abrikosov *et al*. [10, 13].

1.2 Scaling Devices to Their Limits

There are two primary device structures that have being widely studied and used in CMOS technology. One is the bulk structure, where a transistor is directly fabricated on the semiconductor substrate. The other one 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 device process point of view, and it is still the standard structure in almost all CMOS based products until this day.

For device scaling, we basically try to balance two things: device functionality and device reliability. Both of them have to be maintained at a smaller dimensional size. To accomplish this, we need to suppress any dimension related effects or short channel effects (SCEs) as much as possible. SCEs include threshold voltage (V_{TH}) variations versus channel length, typically V_{TH} rolloff at shorter channel lengths. This effect is usually accompanied by degraded subthreshold swing (S), which causes difficulty in turning off a device. SCEs also include the drain-induced barrier lowering (DIBL) effect. DIBL results in a drain voltage dependent V_{TH} , which complicates CMOS design at a circuit level. As a transistor scales, reliability concerns become more pronounced. Unwanted leakage currents can make the device fail to function properly. Primarily, there are two kinds of leakages, gate tunneling current and junction tunneling current. Both of them result from extremely scaled dimensions and high electric fields.

According to device scaling physics, increasing channel doping concentration (N_B) can effectively suppress SCEs. Frank *et al.* recently published their work quantifying the dependence of the scale length on N_B [14]. To a first order approximation, their theory gives the following equation,

$$\Lambda = W_{dm} + (\varepsilon_{Si} / \varepsilon_I)T_I \ , \tag{1.1}$$

where Λ is the scale length, W_{dm} is the maximum channel depletion depth, T_I is the insulator thickness, and $\varepsilon_{Si} / \varepsilon_I$ is the ratio of dielectric constants of silicon and the insulator. W_{dm} can be directly related to N_B (see for example [15]). Depending on the complexity of the channel doping profile, this theory predicts that the minimum design length L_G lies between Λ to 2Λ . It is quite clear in eqn. (1.1) that high N_B results in reduced W_{dm} , therefore a shorter scale length Λ . Of cause, thinner T_I or higher ε_I also helps device scaling.

Device scaling has come a long way. In the early days, L_G is relatively long, a low uniform N_B can be used providing satisfactory immunity of SCEs. A low N_B gives a small body effect coefficient, which improves the subthreshold swing [15]. As the channel length decreases, a retrograde or ground plane doping profile can be introduced [16-17]. This doping profile has a low doping region near the Si/Oxide interface, but a high doping region underneath. The top region provides better body effect, while the bottom region suppresses SCEs. To achieve even shorter channel lengths, a ground plane profile is not enough, a more complicated doping profile has to be added, namely the super halo [18]. In this case, high gradient halo dopings are formed next to the source/drain junction region. These heavily doped regions can effectively protect the source end of the channel region from the influence due to the electric fields from the drain diffusion region. As the channel length varies around the nominal L_G, a shorter length causes the halo regions to merge, ending up with higher N_B, which resists V_{TH} rolloff. By using the ground plane and halo doping profiles, simulations show that the bulk structure can be scaled down to ~25 nm regime [18]. Beyond that, device scaling of the bulk structure is limited by severe degradation of junction leakage which is caused by the high built-in fields, and can not be avoided in the wake of the super halo engineering.

Partially depleted SOI MOSFETs scale in a very similar manner as the bulk devices do. The buried oxide layer in a SOI device can provide superior electric isolation between the active device region and the substrate region. This property is considered a big improvement over bulk devices. Body isolation, however, also results in charge buildup (majority carriers) within the body region, which gives rise to the unwanted floating body effect (FBE) [19]. A fully depleted SOI MOSFET can help relieve the FBE, but a fully depleted single gate SOI MOSFET is not considered a desired structure for scaling. A single gate SOI device typically has a thick buried oxide layer, which can not terminate any electric lines from the drain end, leaving the source vulnerable to the influence of the drain [17, 20].

All recent studies indicate that the ultra-thin body double gate (DG) SOI MOSFET is the ideal device structure for ultimate scaling [4, 21-22]. In an ultra-thin body DG MOSFET, the second gate electrode can significantly suppress the SCEs. Referring to eqn. 1.1, and noting that W_{dm} can be approximated by $T_{Si}/2$ (T_{Si} is the silicon body thickness), when T_{Si} is scaled to nanometer thicknesses (close to T_I), clearly the scale length will downsize into the nanometer regime. It should be also noticed that high body doping is not needed here, so the band-to-band tunneling junction leakage is no longer a big concern.

Moreover, the use of ultra-thin bodies will result in reduced metallurgical junction perimeter, therefore low junction capacitance. The bodies are typically lightly doped, giving other advantages: 1) there is barely room for the FBE to come into play, 2) the V_{TH} variation due to dopant fluctuations can be eliminated, 3) close-to-ideal subthreshold swing (60 mV/dec) can be achieved, 4) severe mobility degradation due to ion scattering might be avoided.

From a technical point of view, DG MOSFETs are difficult to build. Gate self-alignment is hard to achieve. A misaligned gate will cause high overlap capacitances on one side of the gate, and large underlap junction resistances on the other side of the gate. Recent works show that clever process designs can help get rid of gate misalignment [23-24]. Extension region resistances pose another concern in DG MOSFET design. Due to the use of ultra-thin bodies, these resistances can be very high, limiting device performance. The proposed solution is to use fanned out source/drain regions as close as possible to the channel region [25]. The use of ultra-thin bodies also leaves limited room for adjusting V_{TH} with body doping. Gate stack engineering has to be done to obtain an appropriate V_{TH}, either by employing new contact materials with desirable workfunctions, or maintaining an offset voltage between the two gate electrodes to mimic a different workfunction [24, 26]. Quantum effects (subbband splitting) can become significant as the confinement of carriers becomes stronger within ultra-thin bodies, translating to sensitivity of V_{TH} to the body thickness. This fundamental physics effect poses an additional difficulty to control V_{TH} in ultra-thin bodes. (It is worthwhile to point out that this subband splitting effect will increase the band gap between lowest electron subband and highest hole subband, which may considerably suppress the band-to-band tunneling leakage in ultra-thin silicon bodies.)

Despite the existence of numerous difficulties, the excellent scaling capability demonstrated by the ultra-thin body DG structure can never be underestimated. For this reason, therefore, this thesis will be concentrated on a study of ultra-thin body DG MOSFETs.

1.3 Non-equilibrium Green's Function (NEGF) Formalism

As MOSFETs scale to the nanometer regime, canonical carrier transport theories are no longer capable of describing carrier transport accurately. The canonical theories are basically derived from the Boltzmann transport equation (BTE), with more or fewer approximations being made [27]. These models focus on scattering-dominant transport, which typically occurs in long channel devices. Nanoscale transistors, however, operate in a quasiballistic-transport regime [28]. Simulations using conventional models may either under-predict or over-predict the device performance [29-30].

The BTE (classical version) is a complex integro-differential equation. On one side of the equation, the temporal evolution of a carrier distribution function $f(\mathbf{r},\mathbf{p},t)$ is specified in momentum and real space assuming a Newtonian movement of the particles; on the other side of the equation, modifications to $f(\mathbf{r},\mathbf{p},t)$ due to carrier collision or scattering ($\left.\frac{\partial f}{\partial t}\right|_{Coll}$) are computed within a quantum framework.

Obtaining solutions to the BTE with no approximations can be very difficult. Several Ph.D. theses have been devoted to this topic. Among them, Huster used a response matrix approach [5]. Banoo came up a method of directly solving the six-variable equation in the steady state [6]. Both works contributed significantly to this area. We also note that, however, there are important issues missed in the BTE solutions. As we mentioned earlier, the BTE assumes a classical approach in describing carrier dynamics, so quantum features prevailing in nanoscale devices can never be captured in the solutions. Moreover, in pursuing the solutions, a non-degenerate distribution function is typically used to simplify the scattering operator $\left.\frac{\partial f}{\partial t}\right|_{Coll}$ [6]. This approximation violates Pauli's exclusion principle, which is apparently indispensable in describing the scattering of a highly degenerate carrier gas within an extremely scaled transistor.

To simulate nanoscale devices, the non-equilibrium Green's function formalism (NEGF) provides one of the best frameworks available. (Other approaches include the Pauli master equation method [31-33], and Wigner function method, which is essentially the same as the NEGF [34-37].) The NEGF is a technique to solve the non-equilibrium dynamic equation of the quantum fields. The carriers (*e.g.* electrons and phonons) within semiconductor devices constitute the quantum fields. The Green's functions are defined in terms of field operators, either $<\psi^+(\mathbf{r_2})\psi(\mathbf{r_1})>$ or $<\psi(\mathbf{r_1})\psi^+(\mathbf{r_2})>$. These functions relate the field operator of particles $\psi(\mathbf{r_1})$ at one point in space-time, $\mathbf{r_1}=(\vec{r}_1,t_1)$, to the conjugate field operator $\psi^+(\mathbf{r_2})$ at a different point, $\mathbf{r_2}=(\vec{r}_2,t_2)$. The bracket signifies the need to average over the available states of the system for the nonequilibrium distributions [38]. These functions are used to measure correlations between electrons at two locations denoted by $\mathbf{r_1}$ and $\mathbf{r_2}$, and therefore contain information of the described systems. Self-energy functions are also correlation functions, but particularly related to particle interaction events (for more, see any quantum field theory book, for instance, [39]).

The kinetics of the particle systems are governed by the well-known Dyson's equation, which relates the interacting Green's functions to the non-interacting Green's functions (which can be obtained easily) and self-energy functions [38, 40]. The Fourier transform of the correlation functions with respect to $t_2-t_1\rightarrow E$ will significantly simplify the form of the Dyson's equation. It will also make the correlation functions more physically meaningful. Readers may want to refer to the chapter by Mahan for an enlightening discussion on the significance of the Fourier transform [38]. In the steady state, all quantities can be expressed in terms of $(t_2 - t_1)$. So the Fourier transformed correlation functions become dependent only on $\vec{r}_1$, $\vec{r}_2$ and E.

In pursuing numerical solutions, a matrix representation of the correlation functions is used. Discretization in real space makes $\vec{r}_1$ and $\vec{r}_2$ row or column indices of the matrices. Key kinetic equations describing non-equilibrium transport within a semiconductor device are presented as follows,

$$G(E) = [EI - H_o(E) - \sum(E)]^{-1}, \tag{1.2}$$

$$G^n(E) = G(E)\sum^{in}(E)G^+(E)\text{ , } G^p(E) = G(E)\sum^{out}(E)G^+(E). \tag{1.3}$$

In these equations, G is typically called the retarded Green's function, its Hermitian conjugate, G^+, is called the advanced Green's function. H_o denotes the single-electron effective mass Hamiltonian, in which band structure is incorporated into the effective mass. The Hartree potential for electron-electron interactions is also included in H_o through a scalar potential obtained from the solutions of the coupled Poisson equation. G^n and G^p are correlation functions specifying electron and hole density spectra, respectively. $\sum$, $\sum^{in}$ and $\sum^{out}$ are self-energy functions related to interactions. To see how to derive eqns. (1.2) and (1.3) from the Dyson's equation, one may want to read [40].

The electron density spectrum and the terminal current density spectrum can be evaluated, after self-consistent solutions are obtained for the correlation functions.

$$n(E,m) = \frac{1}{2\pi}G_m^n\text{, and } I_m(E) = \frac{q}{h}Trace[\sum_m^{in} G^p - \sum_m^{out} G^n], \tag{1.4}$$

where $n(E,m)$ is the electron density spectrum for a discretized unit cell, m, G_m^n indicates the mth diagonal entry of G^n, $I_m(E)$ is the current spectrum at terminal m, $\sum_m^{in}$ is the mth diagonal entry of $\sum^{in}$, q is the elementary charge constant, and h is the Plank constant.

For some simple cases, recipes for computing the self-energy functions are available. Several interesting examples are briefly listed below. An extensive study of non-equilibrium transport in SOI MOSFETs using the NEGF approach will be presented in Chapter 3 through Chapter 6 of this thesis.

1) Ballistic transport in MOSFETs

In this scenario, carriers within the active device region are injected either from the source reservoir or the drain reservoir. Both reservoirs are assumed in the equilibrium state, characterized by different Fermi energy levels. The electron-electron interaction is incorporated into the Hartree potential in the single electron Hamiltonian H_o. The interactions between the contact reservoirs and the transport carrier system are measured by $\sum$, $\sum^{in}$ and $\sum^{out}$. No other interactions are included (this is why the transport is ballistic). In this case, $\sum$ can be exactly solved. $\sum^{in}$ and $\sum^{out}$ can be expressed in terms of $\sum$ and the corresponding Fermi energies. For a complete description, see Chapter 3 in this thesis.

2) Büttiker probe models for dissipative transport

In these dissipative transport models, carrier scattering within the device is treated as an interaction between carriers and Büttiker probes [41]. The Büttiker probes are treated as reservoirs similar to the source and drain. $\sum$ for the probes can be mapped onto a macroscopic mobility, a given parameter. Fermi energies for the probes, however, have to be computed self-consistently, to ensure that each Büttiker probe only changes the energy of the carriers and not the total number of carriers in the system. Chapter 4 in this thesis is centered on the Büttiker probe models.

3) Phonon-electron interaction model

In this model, phonons are treated in a harmonic oscillator approximation. The phonon system is assumed in equilibrium, even though the electron system is out of equilibrium. $\sum$, $\sum^{in}$ and $\sum^{out}$ are calculated in the Born approximation, and only the self-energies involving one phonon processes (absorption or emission) are included (higher order processes contribute less to the interaction). Chapter 4 discusses this model in more detail.

1.4 Overview of the Thesis

Chapter 2 is devoted to a 1D simulation study of double-gate MOSFETs. We first describe a simulation tool Schred-2.0, which is a self-consistent Schrödinger-Poisson solver. Using Schred-2.0, we then examine performance related properties of double-gate MOSFETs, such as inversion layer charge, threshold voltage and carrier thermal injection velocity; we also compare device design issues for an asymmetrical (n^+/p^+ polysilicon gate) and a similarly structured symmetrical (n^+/p^+ polysilicon gate) DG MOSFET.

In Chapter 3, we describe the numerical techniques used in developing a 2D simulator for nanoscale double-gate MOSFETs. We implement a quantum ballistic transport model, we also implement a ballistic Boltzmann transport model. We then examine the Poisson equation boundary conditions at the source/drain contacts in ballistic MOSFETs. Finally, we assess the approximations of the mode-space representation in doing 2D quantum simulation.

Chapter 4 is centered on scattering phenomena in ultra-scaled double-gate MOSFETs. We implement, examine, and compare three different scattering models. We use the Green's function formalism in all implementations. The two Büttiker probe based models simulate scattering due to all possible mechanisms (*e.g.* surface roughness, phonon, and impurity *etc.*) as a perturbation represented by the probe's self-energy. These perturbations can be related to conventional low field mobility (μ). The third approach focuses on phonon-electron scattering. Although the phonon-electron interaction model may not generally be used to simulate scattering in MOSFETs, it provides a more rigorous solution to help us better understand specific scattering process.

In Chapter 5 we explore the device physics of nanoscale MOSFETs by 2D numerical simulation of a model transistor. We examine the physics of charge control, source velocity saturation due to thermal injection, and scattering in ultra-small devices. We show in this study that the essential physics of nanoscale MOSFETs can be understood in terms of a conceptually simple model.

In Chapter 6, we examine device design issues for an n-channel double gate MOSFET with a metallurgical gate length of 10 nm. The device structure is engineered to meet the ITRS-99 specifications for the year 2014 transistor generation. First, we outline the procedure to select a combination of silicon film thickness and gate dielectric in order to meet short channel requirements. We then discuss gate stack design in order to meet the threshold voltage and gate leakage requirements. We use the Büttiker probe model to capture mobility degradation due to surface roughness and high doping concentrations in the extremely scaled MOSFET, and we present results that highlight the effects of gate overlap/underlap, S/D extension and quantum contact resistances on the performance of nanoscale transistors. Finally, we use an empirical gate tunneling model to examine the leakage current distribution along the gate and predict the gate leakage current.

In Chapter 7, we summarize the conclusions of this research, and list a few potential directions for future work.

2. 1D SIMULATION OF QUANTUM EFFECTS IN DG MOSFETS

2.1 Introduction

All recent studies of alternative CMOS device structures have reached one common conclusion: the double-gate (DG) device design is ideally suited for ultimate CMOS scaling. A double-gate device structure is composed of a thin Si body (thinner than one half of the gate length) sandwiched between gate stacks (gate contact and gate dielectric). Three different double-gate structures are most commonly used. They are: 1) planar double-gate device, 2) vertical surround-gate device, and 3) FinFet (with a fin-shaped body) [4, 23, 42-44]. Double-gate structures have exhibited numerous advantages over conventional bulk device structures. The presence of two gates significantly reduces short channel effects, improves punch-through properties, permits complete dielectric isolation and reduces junction capacitance [4, 45]. In addition, thin body double-gate MOSFETs also provide nearly ideal subthreshold slope. The reduced junction capacitance and presence of two channels drastically boosts the speed and drive current of double-gate device structures. The following study focuses on a planar double-gate design and summarizing work that has been published by the author [21, 46-47].

This chapter is devoted to a 1D simulation study of double-gate MOSFETs. We first describe a simulation tool Schred-2.0, which is a self-consistent Schrödinger-Poisson solver. Using Schred-2.0, we then examine performance related properties of double-gate MOSFETs, such as, inversion layer charge, threshold voltage and carrier thermal injection velocity. Finally, we present simulation results for an asymmetrical (n^+/p^+ polysilicon gate) and a similarly structured symmetrical (n^+/p^+ polysilicon gate) DG MOSFET aimed at on-current assessment. An asymmetric structure was considered for the following reasons: 1) conventional polysilicon gates can be used to provide the right threshold voltage (this is not true for an analogous symmetric design) and 2) it continues to deliver extraordinarily high on-current with one-predominant channel compared to its dual channel symmetric gate counterpart.

2.2 Description of Schred-2.0

Schred is a self-consistent 1D Schrödinger-Poisson solver originally developed by D. Vasileska of Arizona State University [48-49]. Version 2.0, developed by this author, extends Schred-1.0 to, 1) simulate both bulk MOS (one oxide/silicon interface) and SOI (two oxide/silicon interfaces) device structures, 2) simulate n and p body MOS capacitors, 3) perform quantum simulations of accumulation layers in bulk MOS and 4) calculate ballistic *I-V* characteristics of bulk and SOI MOSFETs based on 1D MOSFET theory.

2.2.1 Overview

The Schrödinger equation is solved assuming an effective mass approximation. The solution scheme used in Schred-2.0 is illustrated using an SOI structure as an example. Figure 2.1 schematically shows the profiles of the conduction and valence bands in the SOI capacitor. Bound-state energies in the quantum solution are also illustrated in Fig. 2.1. Given the body doping concentration (N_D, N_A) and an initial guess for the electron and hole densities (n, p), the 1D Poisson equation,

$$\nabla \cdot [\varepsilon \cdot \nabla V(z)] = -q(p - n + N_D - N_A) \tag{2.1}$$

is solved for $V(z)$, which is the vacuum potential. Once the vacuum level has been calculated, electron and hole densities can either be computed classically or quantum mechanically. The classical approach to evaluating electron and hole densities is based on 3D statistics. A description of this approach can be found in several textbooks (e.g.: [50]). The quantum solution is more involved and merits a detailed explanation.

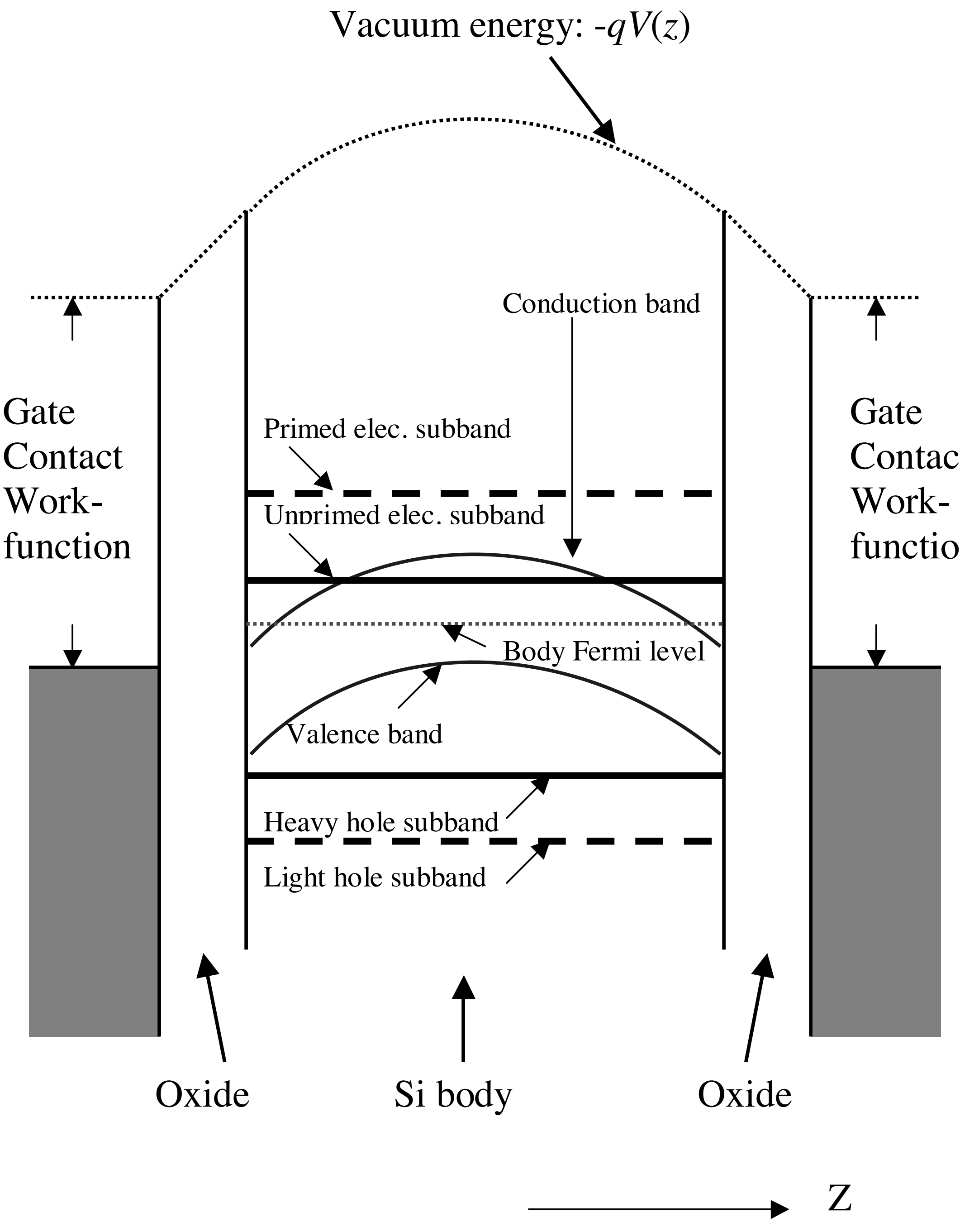


Fig. 2.1. A schematic illustrating band profiles in a double-gate MOS capacitor.

The first step of the quantum solution solves a 1D effective mass equation in the confinement direction (Z in Fig. 2.1). The 1D effective mass equation is,

$$-\frac{\hbar^2}{2m_z^*}\frac{d^2}{dz^2}\psi_i(z) - qV(z)\psi_i(z) = E_i\psi_i(z) \tag{2.2}$$

where E_i is the bound state energy for subband i and$\psi_i(z)$ the corresponding envelope function for subband i. Once the bound state energies and wave functions have been calculated, the carrier density for each bound state is evaluated using 2D statistics as described in the later sections. This 2D carrier density is then distributed using the corresponding envelope function for each bound state to obtain the 3D density. The gate oxide layers are assumed to represent an infinite potential barrier causing the wavefunctions go to zero at the silicon/oxide interfaces (boundary condition to solve the effective mass equation). The calculated carrier density is then fed back to the Poisson equation, which is solved for the new potential profile, until self-consistency is achieved. Schred-2.0, not only simulates 1D electrostatics in an MOS structure, but also computes ballistic *I-V* characteristics of a MOSFET based the 1D electrostatics [28]. A detailed description of the charge and current calculations is presented in the following sections.

2.2.2 Simulation of 1D electrostatics

For the quantum solution, it is assumed that the Si/SiO$_2$ interface is parallel to the (100) plane. The conduction band in bulk silicon can be represented by the six equivalent ellipsoids as shown in Fig. 2.2. When an electric field is applied in the [100] direction these six equivalent minima split into two sets of subbands [51]. The first set of subbands (unprimed) is two fold degenerate and represents those ellipsoids that respond with a heavy effective mass (m_l) in the gate confinement direction while the second set of subbands (primed) is four fold degenerate and represents those ellipsoids that respond with a light effective mass (m_t) in the direction of the applied field. Because of the heavier longitudinal mass, the unprimed subbands have relatively lower bound-state energies, as compared to the primed subbands and are therefore primarily occupied by electrons. The 2D electron density for the unprimed and primed bands is,

$$n_i = n_{2Di} \ln(1 + e^{(\mu - E_i)/k_B T}) \tag{2.3a}$$

where n_{2Di} is a constant with the dimensions of 2D carrier density for subband i (the explicit expressions of n_{2Di} and other subband related constants appearing later in this chapter can be found in Appendix A), and μ the body Fermi energy.

The structure of the valence bands in silicon is complicated and cannot in general be treated analytically. However, to first order one can express the E-k relationship for the heavy and light hole bands around the valence band maxima within an analytic framework based on effective masses quoted in the references [52-53]. The split-off band is usually ignored as the split-off energy in silicon is large resulting in a negligible hole density for these bands. Due to the curvature of the valence bands, it should be noted that holes have negative effective masses, resulting in bound-state energies lower than the valence band maxima when one solves the hole effective mass equation in the confinement direction (Fig. 2.1). Heavy holes have smaller confinement energies as compared to light holes. Therefore the heavy hole subbands are closer to the valence band maxima as compared to the light hole subbands (Fig. 2.1). Holes represent unoccupied states in the valence band. Therefore in calculating the hole density, a distribution function of

$$f_p = 1 - \frac{1}{1 + e^{(E - \mu)/k_B T}} \tag{2.3b}$$

is used. The hole density for subband i is

$$p_i = p_{2Di} \ln(1 + e^{(E_i - \mu)/k_B T}) . \tag{2.3c}$$

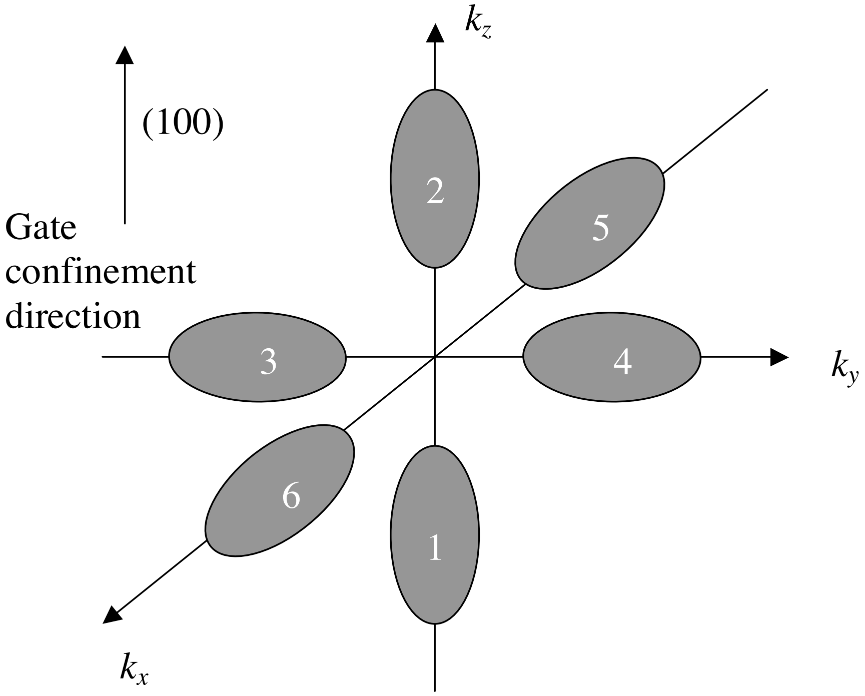


Fig. 2.2. Six equivalent conduction band ellipsoids in bulk silicon. Ellipsoids 1 and 2 respond with a longitudinal effective mass in the gate confinement direction, and give rise to the unprimed subbands, while ellipsoids 3 to 6 respond with a transverse effective mass in the gate confinement direction, resulting in the primed set of subbands.

It should be noted that eqns. 2.3a-c assume Fermi-Dirac statistics. One can also invoke Maxwell-Boltzmann statistics in Schred-2.0. Exchange and correlation corrections to the electrostatic potential as a result of quantum effects, can be accounted for in the local density approximation by invoking the desired options [54]. These corrections decrease the bound-state energies resulting in about 5% increased carrier densities [48].

When performing a quantum simulation of an SOI structure, both electrons and holes have to be treated quantum mechanically. There are two reasons: 1) SOI bodies are usually undoped or lightly doped. Therefore under low bias both electrons and holes are equally important and 2) Quantum confinement due to the two gate dielectrics, affects both electrons and holes. However, when performing quantum simulations of a bulk MOS capacitor, only one type of carrier is quantum mechanically confined for a given bias condition. The confinement is due to electrical fields. The other type of carrier can be treated classically (using 3D statistics). As an example, if a p body MOS capacitor is considered, then in the depletion and inversion regions, electrons have to be treated quantum mechanically, while holes could be treated classically. In the accumulation region, holes have to be treated quantum mechanically, while electrons could be treated classically. A quantum mechanical treatment of the majority carrier in the accumulation regime needs the inclusion of a large number of subbands thus greatly increasing the computational burden. This is because in the accumulation regions, energy bands bend very little, resulting in weak quantum confinement. Therefore the subband energies are closely spaced and a large number of subbands need to be included in order to accurately account for the overall majority carrier concentration. Quantum simulations can capture capacitance degradation effects in accumulation regions. These effects are becoming more important as oxide layer thickness is continuously scaled.

Schred-2.0 can treat both n and p type polysilicon or metal gate contacts. Polysilicon gates are modeled as heavily doped single-crystal silicon. Irrespective of the model used to treat the silicon body (classical or quantum mechanical), electrons and holes in the gate regions are always treated classically assuming 3D statistics. The gate dielectric constant is a user specified quantity as is the gate work function for metal gates. Schred-2.0 also allows different dielectrics/workfunctions for the top and bottom gates in an SOI structure. This enables the study of different gate designs on the performance of MOS capacitors.

2.2.3 *I-V* characteristics simulation based on 1D electrostatics

Current is a constant through a MOSFET and can be evaluated at any point along the channel. Lundstrom pointed out that the current can be easily computed at the source to channel barrier top [28, 55]. For well-tempered MOSFETs, the total areal charge density at this point is based on equilibrium 1D MOS electrostatics and can be expressed as,

$$Q_{inv} = C_{Eff}(V_{GS} - V_{TH}) \ . \tag{2.4}$$

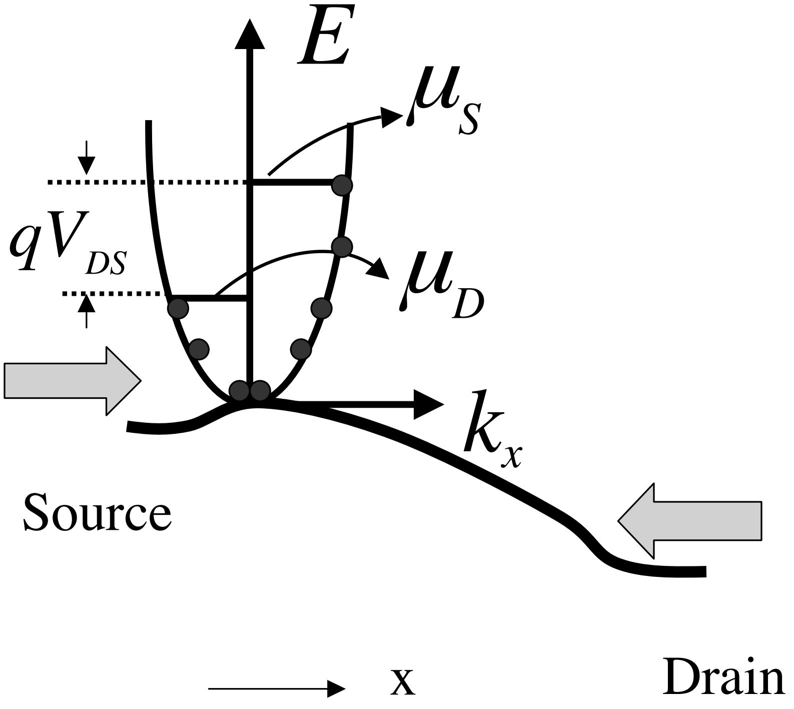


Fig. 2.3. A schematic figure showing the ballistic transport physics in a nMOSFET under low drain bias condition.

In the equilibrium state, the charge distribution in k-space is symmetric resulting in zero net current. However, in the off-equilibrium situation, the charge distribution is no longer symmetric in k-space. This is because in the ballistic limit, the +*k* states are populated according to the source Fermi level while the –*k* states are populated according to the drain Fermi level as pointed out by Natori and Datta [56-57]. The separation between the source and drain Fermi levels is $qV_{DS.}$ Therefore in modeling an off-equilibrium situation, the total charge at the source-to-channel barrier peak has to be correctly apportioned in k-space based on two Fermi levels. It should be noted that while the total charge is still the equilibrium charge, the distribution is no longer an equilibrium distribution. This difference in population between the +*k* and –*k* states results in a net non-zero current which is evaluated using the following expression [28],

$$I_{Di}/W = I_{Oi}\{\Im_{1/2}[(\mu - E_i)/k_B T] - \Im_{1/2}[(\mu - E_i - qV_{DS})/k_B T]\} \qquad (2.5)$$

where I_{Oi} is a constant with the dimensions of current per unit length for subband i, and $\Im_{1/2}$ is the Fermi-Dirac integral of order one-half [58-60]. It should be noted that the Fermi energy appearing in eqn. 2.5 is calculated using a bisection method.

As the drain bias is increased, the –*k* state occupancy is progressively reduced and is totally eliminated at very high drain voltages. Therefore all of the charge at the source-to-channel barrier is a result of source reservoir contributions leading to current saturation at high drain bias. The ballistic transport physics is summarized in Fig. 2.3.

Knowing the ballistic current, we can also compute the conductance of the MOSFET in the linear region of operation (low V_{DS}) as,

$$G_{Di}/W = G_{Oi}\Im_{-1/2}[(\mu - E_i)/k_B T], \tag{2.6}$$

where G_{Oi} is a constant with the dimensions of conductance-length. Equation 2.6 shows that even under the assumption of ballistic transport within the MOSFET, the conductance is finite. This conductance is the quantum contact conductance, and is limited by the number of propagating modes available at the source [10].

At very high V_{DS}, injection from the drain reservoir is completely suppressed. Under such conditions it is possible to calculate a uni-directed thermal injection velocity for source injected carriers. This velocity, which is obtained by dividing the ballistic current with the areal charge density, can be expressed as,

$$\upsilon_{inj}^{i} = \upsilon_T^i \left(\frac{\Im_{1/2}[(\mu - E_i)/k_B T]}{\ln(1 + e^{(\mu - E_i)/k_B T})} \right), \tag{2.7}$$

where υ_T^i is a constant independent of μ and E_i, denoting the non-degenerate limit of υ_{inj}^{i} for electrons on subband i. The uni-directed velocity of source-injected carriers is limited by the source reservoir Fermi energy and can be much higher than the bulk saturation velocity.

In evaluating any internal quantity, contributions from all subbands have to be included. Note that all of the constants used in eqns. 2.3 to 2.7 can be expressed in terms of fundamental constants, and the expressions are described in detail in Appendix A. And also note that eqns. 2.5-2.7 are given for nMOSFETs, but switching the positions of μ and E_i will give the results for pMOSFTEs. Typical outputs from Schred-2.0 simulations are the spatial variation of the conduction-band edge, 3D charge density in the body; 2D surface charge density, and capacitances. The capacitances include inversion layer capacitance C_{inv} and total gate capacitance C_{tot}. In the case of capacitors with poly-silicon gates, Schred-2.0 can also be used to calculate the poly-gate capacitance, C_{poly}. When performing quantum mechanical simulations, Schred-2.0 can provide the subband energies, the subband carrier densities, and wavefunction profiles within the body. In the case of ballistic current calculations one can obtain current as a function of the gate and drain biases, the quantum contact resistance and thermal injection velocity of carriers.

Schred-2.0 is written in Fortran 77. The program is very efficient. On a 167MHz Ultra-1 machine, it typically takes about 10 seconds per bias point for a quantum simulation, and about 5 seconds per bias point for a classical calculation. For quantum simulations in the bulk accumulation regime, it takes a relatively long time (about 2 to 3 minutes) for one bias point because a very large number of subbands need to be treated in order to obtain the correct charge density. A thick body SOI quantum simulation (thicker than 0.1 micron) also involves long computational times for the aforementioned reason. Examples of the application of Schred-2.0 will be presented in the next section, and much more can be found in the following references [21, 47-48]. The user manual for Schred-2.0 is also a good supplement to this document. The user manual is available online. Prospective users may want go to "*http://punch.ecn.purdue.edu/Guest*" and register for membership. Following the instructions to the Schred directory, one can download the manual.

2.3 1D Simulation Study of DG SOI MOSFETs

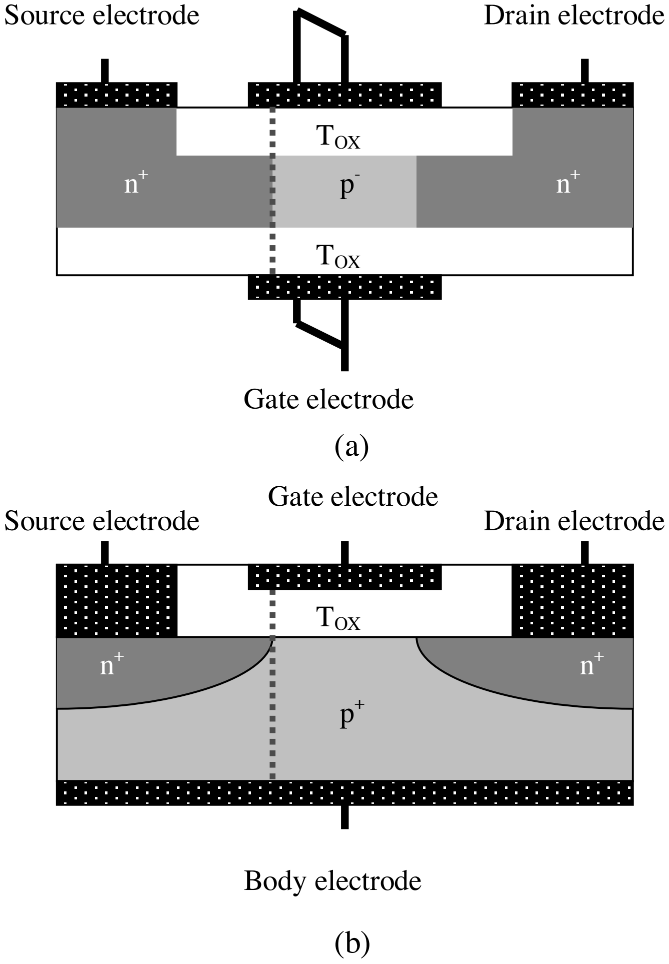
As applications of Schred-2.0, we first examine electrostatics of ultra-thin body double-gate MOS structures. The electrostatic properties in a ultra-thin body to great extent underlie the MOSFET transport characteristics. High mobile charge density and high thermal injection velocity can be achieved in ultra-thin bodies. The effect of electron penetration into the oxide regions is also studied. Inclusion of charge penetration into the oxide regions results in increased effective gate capacitances and reduced threshold voltages. We then present an extensive simulation analysis of a ultra-thin body asymmetrical DG MOSFET. The simulations highlight internal electric quantities. Based on a comparison between symmetric and asymmetric DG MOSFETs, we show that the asymmetrical n^+/p^+ polysilicon gate design can be used to achieve low power applications with extraordinary high on-current.

2.3.1 Fundamental performance factors:

Ultra-thin body SOIs has been demonstrated to result in MOSFETs that are potentially scalable to channel lengths of 10 nm or less [22, 61-62]. It is well known that subthreshold characteristics of MOSFETs are determined by MOS electrostatics. The ultra-thin body is desirable to suppress 2D electrostatics for an improved off-current. In this study, we find that in principle, on-current is also strongly affected by electrostatics. We examine the ballistic limit of ultra-thin body SOIs and show that the on-current of double-gate SOI MOSFETs with ultra-thin bodies can potentially be much higher than twice that of equivalent bulk devices.

Results and discussion:

The simulated ultra-thin device structure is shown in Fig. 2.4a. The simulation domain is limited to a 1D slice indicated in Fig. 2.4. All simulations are performed using the quantum mechanical model in Schred-2.0. The devices have symmetrical gate contacts and insulator layers on both sides of the silicon body. The insulator thickness (T_{OX}) is 1.5 nm. The silicon body thicknesses (T_{Si}) ranges from 1.0 nm to 30 nm, and corresponds to device generations with gate lengths below 50 nm [4, 25]. Intrinsic silicon bodies are used for two reasons: 1) to avoid threshold voltage fluctuations due to variations in dopant distribution and 2) to ensure full body depletion resulting in improved subthreshold swing (*S*). A model n-channel bulk MOSFET is also simulated to provide a basis for comparison (showing in Fig. 2.4b). $N_A = 2\times10^{18}$ cm^{-3} and T_{OX} = 1.5 nm are used following the ITRS specifications for the year 2005 technology generation [2]. For the SOI MOSFETs, hypothetical mid-gap metals ($\phi = 4.66\,\text{eV}$) are assumed for gate contacts. For the bulk device, aluminum ($\phi = 4.10\,\text{eV}$) is used.



(a)

(b)

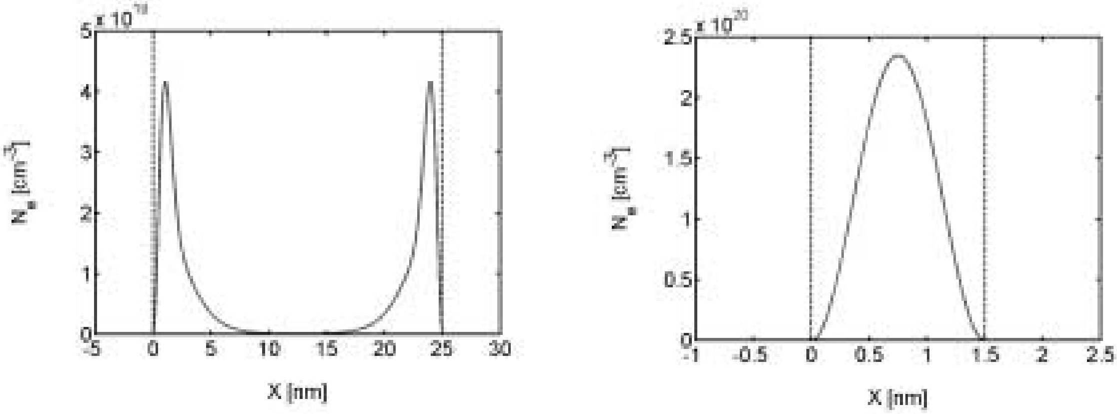
Fig. 2.4. Schematic pictures of model structures: (a) double-gate SOI MOSFET, (b) bulk MOSFET. The dashed lines indicate the Schred-2.0 simulation slices.

The on-current performance of a MOSFET can be expressed in terms of inversion layer carrier density and average injection velocity. Increasing either inversion layer density or injection velocity results in increased on-current. A bulk MOSFET has only one channel, while a typical thick body double-gate MOSFET shows two independent channels resulting in "double channel conduction" [63]. As the body of a double-gate MOSFET is scaled, it has been reported that body inversion occurs, implying that the two independent channels merge together. It is of interest to see if the merged channel can still provide the desired double channel conductivity. In Fig. 2.5a, we show the 2D electron density distribution in a relatively thick silicon body (T_{Si} = 25 nm). The gate bias overdrive is 0.8 V. Most inversion carriers are confined to regions close to the gate/body interfaces. The electron profile is very similar to that would occur in two back to back bulk MOSFETs. The inversion layer electron density can be expressed as

$$n_S = 2C_{Eff}(V_{GS} - V_{TH}), \tag{2.8}$$

where C_{Eff} is the effective oxide capacitance for one gate. C_{Eff} is somewhat degraded from the physical oxide capacitance $\varepsilon_{OX} / T_{OX}$ due to quantum inversion layer thickness. As the silicon film is thinned ($T_{Si} = 1.5\,\text{nm}$), as shown in Fig. 4.5b, the two inversion regions merge to a single inversion layer in the DG structure. This is purely a quantum effect (due to the symmetry of the first subband wavefunction). Note that the electron density peak value in Fig. 2.5b is twice that in Fig. 2.5a. Also note that for extremely thin silicon bodies, the degradation of the oxide capacitance could be lower than in case of a thick body DG SOI capacitor resulting in an integrated 2D charge within a single inversion layer much higher than that of two inversion layers.

Figure 2.6 shows the threshold voltage dependence on the body thickness for DG SOIs. The threshold voltage is obtained by extending the linear region of the charge vs. gate voltage curve to intersect the voltage axis. As can be seen in the figure, V_{TH} rises considerably as the body thickness is below 3nm. This large V_{TH} increase underscores the impact of quantum effects on ultra-thin body devices.



(a) T_{Si} =25 nm

(b) T_{Si} =1.5 nm

Fig. 2.5. Electron distribution profiles in DG SOI structures. Simulations are done at $V_{GS} - V_{TH} = 0.8$ V. The dashed lines represent Si/oxide interfaces.

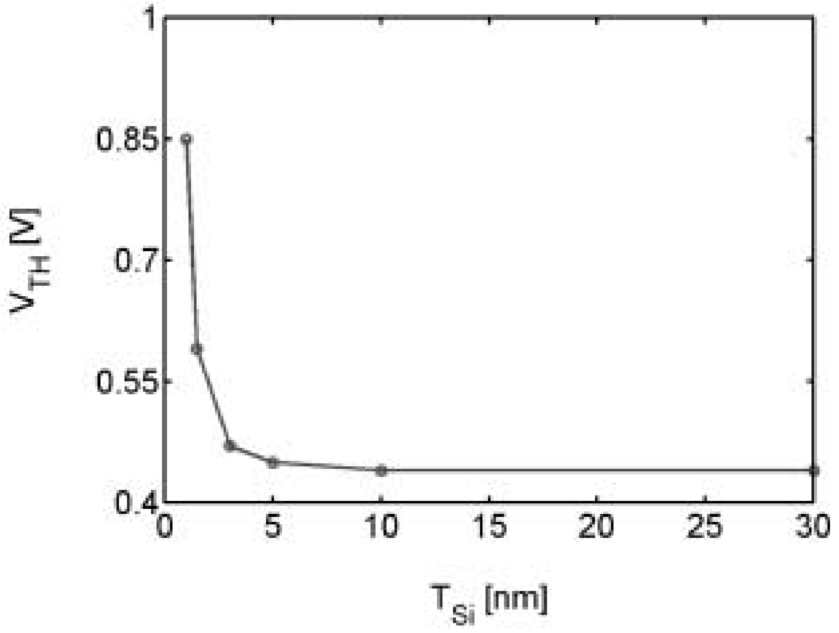
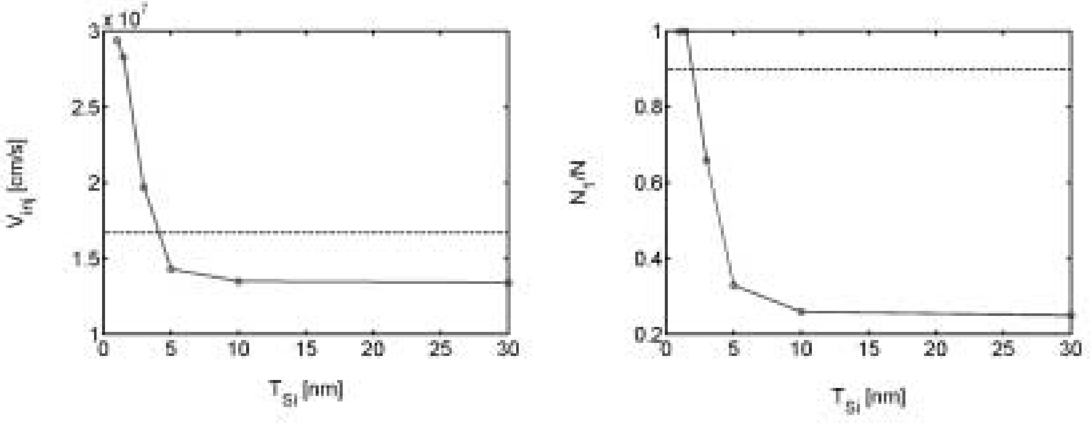


Fig. 2.6. Dependence of threshold voltage on body thickness for DG SOI structures. The threshold voltage is determined by linear extrapolation of gate voltage dependence of the electron charge density.

In Fig. 2.7a, the dependence of carrier injection velocity on body thickness of SOI structures is shown. The gate overdrive is 0.8 V. One interesting result is that the thermal injection velocity of carriers in ultra-thin body DG SOI structures can be boosted to 3.0×10^7 cm/s, which is almost twice the value achieved by carriers in bulk devices. This occurs because 1) strong quantum confinement in ultra-thin structures enlarges the band gap significantly, resulting in single subband occupancy (this is shown in Fig. 2.7b), 2) the 2D electron gas residing in the single subband becomes highly degenerate, giving rise to increased thermal injection velocity. It should be pointed out that all of the results are based on the assumption of a parabolic E-k relation. Non-parabolicity may reduce these predictions to some extent.



(a) (b)

Fig. 2.7. (a) Dependence of electron injection velocity on body thickness of DG SOI structures, (b) the first subband occupation factor versus body thickness of DG SOI structures. The dashed lines indicate the corresponding quantities in the bulk MOSFET. All quantities are evaluated at $V_{GS} - V_{TH} = 0.8$ V.

In Fig. 2.8, we compare the drive current of three model devices in the ballistic limit. Currents are evaluated using the simple 1D transport model as described in Section 2.2. For SOIs, the body thicknesses are 1.5 nm and 25 nm respectively. Gate overdrive voltage is 0.8 V. The model device with the ultra-thin body yields an on-current that is about four times that obtained from a bulk device. (It is normally expected that the DG SOI structure may double the bulk current performance as a result of back channel conduction). We explain this quadrupled current as arising due to doubled inversion layer charge due to the double-gate structure and highly degenerate thermal injection velocity due to the ultra-thin body (see Fig. 2.7a). The thick body double-gate model device shows lower current as compared to the ultra-thin body because the thermal injection velocity is low due to multiple subband occupancy.

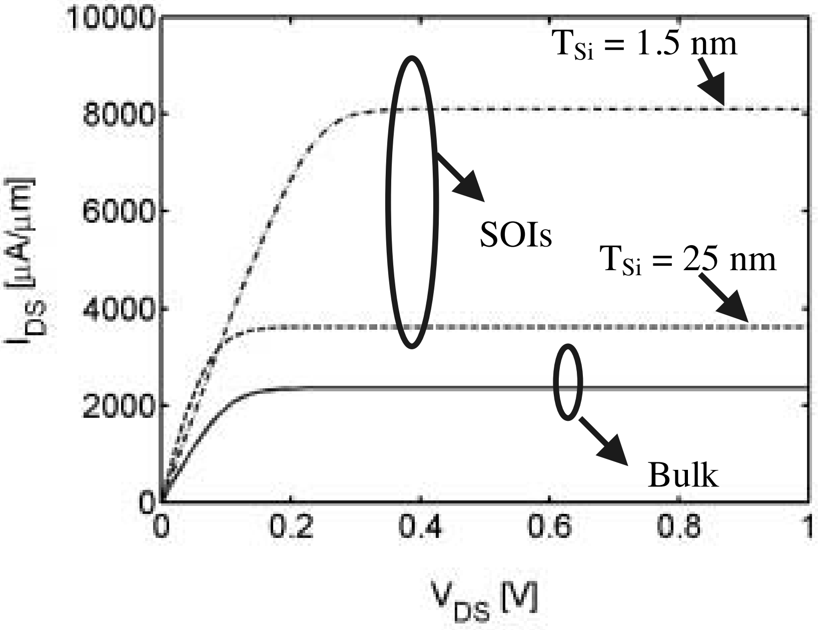


Fig. 2.8. Common source current versus drain voltage for DG SOI and bulk model MOSFETs. $V_{GS} - V_{TH} = 0.8$ V in all three cases.

Conclusion:

In the context of 1D self-consistent Schrödinger-Poisson simulations supplemented by analytical characterizations of carrier transport, we showed that for DG MOSFETs with body thickness below 3 nm, significant increase in threshold voltage is expected. We also showed that, if acceptable threshold voltage can be achieved by gate engineering, ultra-thin body DG MOSFETs demonstrate the capability of delivering remarkably high on-current in the ballistic limit. In addition, we found that a one subband approximation in Schrödinger-Poisson solutions is sufficient for simulating SOIs with body thickness below 3 nm.

2.3.2 Electron penetration into the oxide regions:

Continuous scaling MOSFETs down to the nanometer regime, requires the use of ultra-thin silicon bodies and gate insulators. Quantum effects not only occur in the thin bodies but also in the insulator layers. Up to this point, all of the simulation results assumed that the insulators represented infinitely high potential barriers. In reality, dielectrics have finite band offsets with respect to semiconductors, resulting in quantum tunneling through the insulator regions. Tunneling leakage has been addressed in the literature primarily through the WKB approximation [64-66]. The transmission probability through the gate insulator is evaluated based on the barrier potential profile. Tunneling current is computed by integrating the transmission probability weighted by a Fermi-Dirac factor. This approach, however, is incapable of predicting the effect of charge penetration into the insulator regions. As insulator layers are thinned to around 1.0 nm in physical thickness, charge penetration into dielectrics become more and more important. This penetration effect, enhanced by strong quantum confinement due to ultra-thin bodies in a SOI device or high electric fields in a bulk device, can affect electrostatics in the MOSFET, which in turn, alters its electric characterization. This effect is worth examining through simulations.

Method:

In a p-body SOI MOS structure, the electron penetration effect can be examined by extending the quantum solution domain into the dielectric layers. The Schrödinger equation based on the effective mass approximation still holds. Although the electron density is proportional to $|\psi|^2$, in both semiconductor and insulator regions different effective masses have to assumed different in the two regions. To obtain a Hermitian Hamiltonian, the Schrödinger equation is modified as [67],

$$-\frac{\hbar^2}{2}\nabla\cdot[\frac{1}{m^*(z)}\nabla\psi_i(z)]-qV(z)\psi_i(z)=E_i\psi_i(z)\,, \tag{2.9}$$

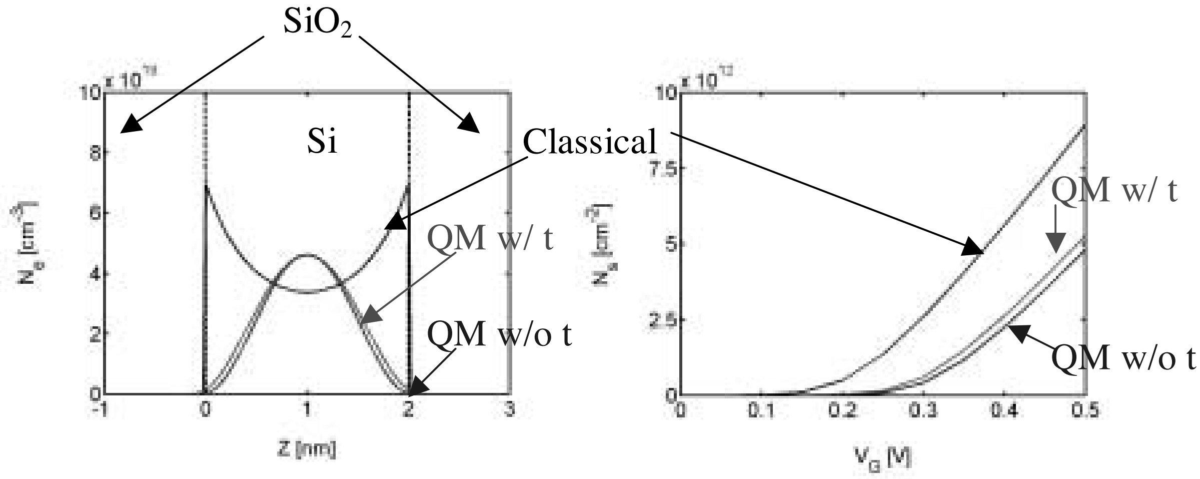
Note that eqn. 2.9 ensures continuity of both electron density ($\propto |\psi|^2$) and current ($\propto \frac{1}{m^*}\nabla\psi$) at the insulator/semiconductor interface. This can be understood by comparing eqn. 2.9 with the Poisson equation $\nabla\cdot[\varepsilon(z)\nabla V(z)] = -\rho(z)$. Note that $\psi(z)$ is analogous to $V(z)$, and $\frac{1}{m^*}\nabla\psi$ is analogous to $\varepsilon\nabla V$. Since $V(z)$ and $\varepsilon\nabla V$ are continuous across all boundaries in a solution to the Poisson equation, it is clear that $\psi(z)$ and $\frac{1}{m^*}\nabla\psi$ will also be continuous in a solution of eqn. 2.9.

The solution boundaries of eqn. 2.9 are moved from the insulator/semiconductor interfaces to the insulator/contact interfaces. The insulator layers although extremely thin, are still assumed to provide minimum electric reliability, meaning that the wave functions decay to negligible values somewhere inside the dielectric layers. Therefore the zero boundary condition for the wave function ($\psi(z)$ =0) can be taken at the insulator/contact boundaries.

Results and discussion:

1D simulations have been performed for a symmetrical DG MOS structure (SiO_2-Si-SiO_2). Two cases were examined: 1) with fixed oxides $T_{OX} = 1.0$ nm, T_{Si} ranging from 1.5 to 5.0 nm, 2) with a fixed silicon body T_{Si} = 2.0 nm, T_{OX} ranging from 1.0 to 5.0 nm. Constant electron effective mass $m^* = 0.4m_e$ is used in the SiO_2 regions [65-66].

Figure 2.9a shows the electron distributions in a 2 nm silicon body. Oxide layers are 1.0 nm thick on each side. Three simulations are compared at the same gate bias: 1) classical, 2) quantum without oxide tunneling, 3) quantum with oxide tunneling. At this dimension the classical model predicts a considerably different charge profile, over estimating electron densities at the SiO_2/Si interfaces. The quantum tunneling effect, resulting in electrons penetrating into the oxide regions, results in a broadened charge distribution. Figure 2.9b further illustrates the differences between the three models through the *Q*-*V* characteristics. The classical simulation indicates an incorrect high effective capacitance (C_{Eff}). Both quantum models indicate positive shifts in threshold voltage, but display a difference in C_{Eff} .

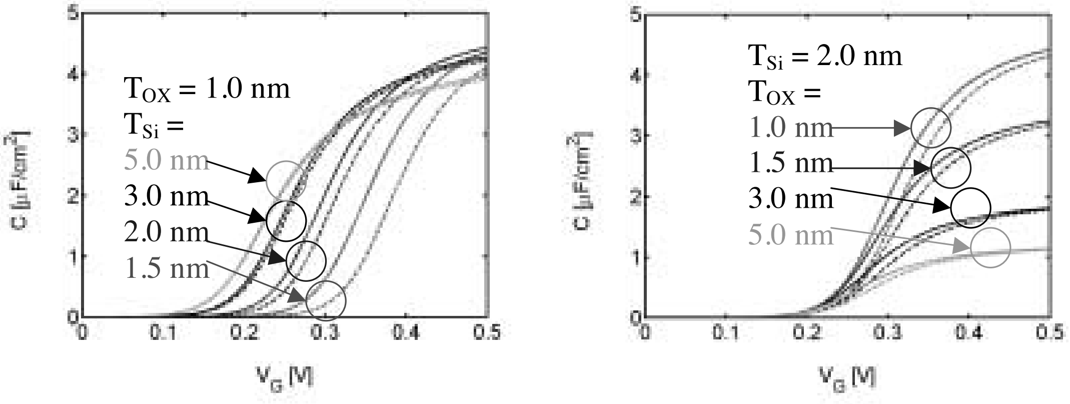


(a) (b)

Fig. 2.9. Comparisons of simulated 3D electron density profiles (a) and 2D density characteristics (b) in three models.

Figure 2.10 is designed to provide an explanation for the difference in the effective oxide capacitance. Quantum mechanically simulated *C*-*V* curves are shown for different T_{Si} with $T_{OX} = 1.0\,\text{nm}$ in Fig. 2.10a. Note that the thin body (1.5 nm) shows the largest split between cases with and without tunneling. The thick body (5.0 nm) shows almost no split. This is because quantum confinement becomes stronger as T_{Si} becomes thinner. Higher confinement energies enable electrons to penetrate deeper into the oxide regions, effectively widening the quantum well (between the insulators). Therefore thinner silicon bodies indicate more evident increases in capacitance and decreases in threshold voltage as compared with a relatively thicker body. Figure 2.10b presents simulated *C*-*V* curves for different T_{OX} with $T_{Si} = 2.0\,\text{nm}$. A discrepancy between the quantum models with and without tunneling effects can still be observed. Since T_{Si} is fixed, all cases have the same level of confinement energy. Therefore the penetration depth into the oxide regions is comparable for different T_{OX} . Devices with thick oxides show relatively unchanged C_{Eff} .

However, in case of thin oxides, electron penetration depths can become large portions of T_{OX} , resulting in considerably increased C_{Eff} .



(a)

(b)

Fig. 2.10. *C-V* characteristics showing the dependences of electron tunneling effect on T_{Si} (a) and T_{OX} (b). The solid lines indicate results assuming electron tunneling, the dashed lines indicate results assuming no tunneling.

Conclusion:

Electron penetration into the gate oxide regions was studied by self-consistently solving the Schrödinger and Poisson equations in a domain containing both semiconductor and insulator regions. Charge penetration was found to remarkably increase the effective gate capacitance and decrease the threshold voltage in devices with T_{OX} , T_{Si} < 3.0 nm. This effect results in a degraded off-state current, which should be considered in addition to the well-known gate tunneling leakage.

2.3.3 Asymmetrical DG MOSFETs

Over the past few years, n^+-p^+ double-gate SOI MOSFETs have been studied for their potential ability of providing well-controlled threshold voltages. With n^+ polysilicon for one gate, and p^+ polysilicon for the other, the gate interaction effect in these asymmetrically-gated devices dynamically tunes the threshold voltage, and provides acceptable V_{TH}'s for both n and p channel transistors [68-69]. Exotic midgap work function materials have to be employed in symmetric-gate MOSFETs in order to achieve a similar threshold voltage. Most recently, Fossum *et al.* further pointed out, the use of the asymmetrical gate architecture may also help suppress Gate Induced Drain Leakage (GIDL) in DG MOSFETs. The reduced GIDL effect was attributed to the weakened electrical field at one of the gates due to the gate asymmetry [70]. An asymmetrical DG MOSFET, however, demonstrates only one predominant conducting channel, so it may be suspected that the on-current performance will be degraded as compared to its symmetrical counterpart. Therefore detailed simulation analyses are demanded for a clarification.

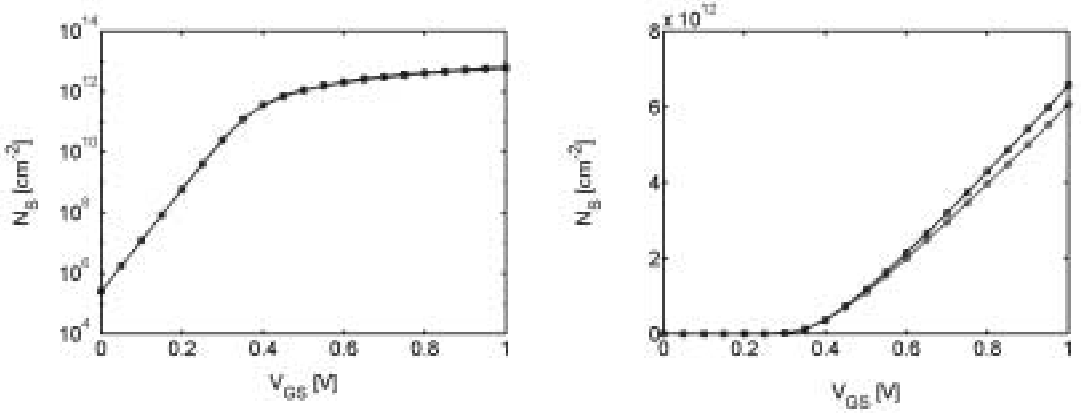
In this work, a simple, clear ballistic simulation study is accomplished to compare and examine the drive current performance of symmetrical and asymmetrical SOI MOSFETs. Simulations performed using Schred-2.0 show that an asymmetric ultra-thin body MOSFET can in fact deliver extraordinary high drive current although it exhibits only one predominant channel. The results are directly derived from solutions to fundamental equations (Schrödinger and Poisson), neglecting channel mobility and parasitic issues. The exceptionally high on-state current in asymmetrical MOSFETs is due to two reasons: 1) superior gate capacitance in thin bodies, 2) superior carrier injection velocities. The former is because of strong gate-gate coupling through ultra-thin silicon bodies, while the latter is because of strong quantum confinement induced band spilt-off. A detailed explanation is presented in the next section. Also note that a classical study conducted by Fossum's group gave rise to a same conclusion [71]. This work is used to validate semiclassical results using a fundamental approach and provide additional insight into the high drive current and overall superiority of asymmetrical MOSFETs.

Analyses and results:

We simulated an L = 50 nm asymmetric and symmetric DG nMOSFETs (refer to Fig. 2.2a for the model device structure). Note that Schred-2.0 is a 1D simulator and that the channel length quoted here can only be viewed as the relevant lateral dimension size with regards to the vertical simulation domain. The Si-film body is lightly doped ($N_A = 1.0 \times 10^{15}$ cm^{-3}) and quite thin (T_{Si} = 10 nm), and the gate oxides are relatively thick ($T_{OX} = T_{OXF} = T_{OXB}$ = 3 nm) for I_{OFF} control. For the asymmetric device, the gates are n$^+$ and p$^+$ polysilicon. However, for the symmetric device, an 'exotic' gate material is assumed in order to yield the same I_{OFF} as in the asymmetric device at V_{DD} = 1.0 V.

The structure is analyzed by the Schred-2.0 self consistently in the z direction. The Schrödinger equation is solved for the inversion layer electron density. In the thin silicon bodies used in our study (10 nm and [1 0 0] oriented), we found that including the first six subbbands (4 from the unprimed subbands, 2 from the primed subbands) in our simulations results in sufficiently high accuracy. The ballistic current (in the x direction), which is used as a measure to benchmark the asymmetric and symmetric DG designs, is obtained indirectly from the 1D Schrödinger-Poisson solutions as described in Sec. 2.2. The on-state current depends on the average carrier injection velocity and available inversion layer electron density. The average carrier injection velocity is the uni-directional thermal velocity ($\sqrt{2k_BT/\pi m^*}$) in the non-degenerate limit. The thermal velocity, which is enhanced by Fermi-Dirac degeneracy at large V_{GS}, can be much higher than the saturation velocity in bulk silicon (1.0×10^7 cm/s) [46]. The factor $\sqrt{2k_BT/\pi m^*}$ in the expression for the degenerate injection velocity indicates that the effective mass of carriers (m^*) in the transport direction is an important factor affecting drive current. Conventional carrier mobility is no longer meaningful. The electrons from the unprimed subbands have a light effective mass (m_t) in the channel direction and display higher injection velocities compared to the electrons from the primed subbands (see Sec. 2.2.3). The total inversion layer electron density assessed at the source injection point, is defined by V_{GS}, and assumed to be independent of V_{DS}. 2D SCEs within these double-gate structures are considered negligible due to the use of thin bodies (10 nm), thin insulators (3 nm) and relatively long channels (50 nm). In this work, the benchmarking is conducted based on the ultimate performance of devices, thus ignoring all parasitics. This technique provides a fair comparison of the symmetric and asymmetric device designs based on fundamental physics models.

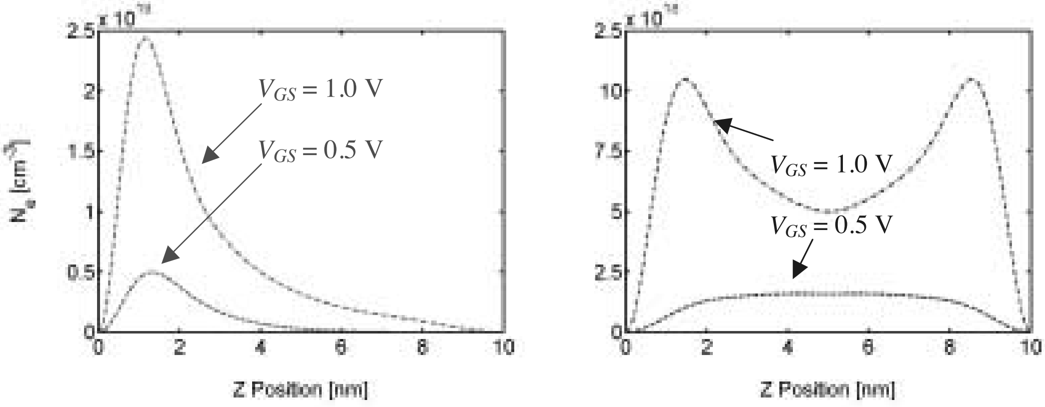
Predicted integrated electron density ($N_S = -Q_{inv} / q$ at the source) versus V_{GS} is plotted for both devices in Fig. 2.11; the predicted *n*(z) distributions across the Si film are shown in Fig. 2.12. Note that the asymmetrical DG MOSFET has only one predominant channel, while its symmetrical counterpart has two channels adjoining the front and back gates. Since the silicon body is only 10 nm thick, it is fully depleted. Ideal subthreshold swing (60 mV/dec at room temperature) is achieved for both devices as shown in Fig. 2.11a. On the semilog scale, with the integrated electron densities calibrated to be the same at V_{GS} = 0.0 V, subthreshold characteristics of the two devices are indistinguishable. From Fig. 2.11b, where the integrated carrier densities are plotted on the linear scale, it can be seen that N_S in the asymmetrical device is comparable to that in the symmetric device at low and moderate V_{GS}, but slightly lower at high V_{GS}. These features are directly related to gate-gate electric coupling in ultra-thin body SOI's as discussed later.



(a) (b)

Fig. 2.11. Schred-2.0 predicted integrated electron density on semilog (a) and linear scales (b). The blue lines with squares represent results of the symmetrical device, the red lines with circles represent results of the asymmetrical device.

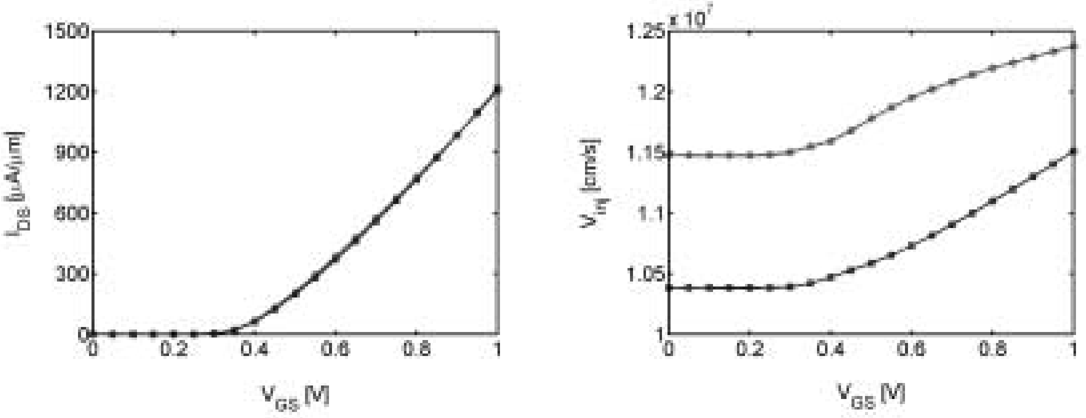
In Figs. 2.12a and b, electron density distributions within silicon bodies are plotted for two gate biases, 0.5 V and 1.0 V. Quantum solutions with infinitely high oxide barrier boundary conditions give rise to zero carrier density at the silicon/oxide interfaces. While body inversion occurs at V_{GS} = 0.5 V in the symmetric device, only the front gate gets inverted in the asymmetric transistor at the same gate bias. Symmetry of the gate configuration results in two inversion peaks in the symmetric device as opposed to a single peak in its asymmetrical counterpart at V_{GS} = 1.0 V, when body inversion occurs in both devices. Note that the peak carrier density value in the asymmetric device is about twice that in the symmetric device. This extra amount of charge compensates the charge loss due to a single channel, resulting in about the same level of inversion electron density at high gate biases as compared to the symmetric device.



(a) (b)

Fig. 2.12. (a) Schred-2.0 predicted $n(z)$ versus V_G across the Si film at the virtual source of the asymmetrical DG nMOSFET. (b) Schred-2.0 predicted $n(z)$ versus V_{GS} across the Si film at the virtual source of the symmetrical DG nMOSFET.

A comparison of linear scale I_{DS}-V_{GS} curves is presented in Fig. 2.13a. The difference between the two curves is almost undetectable. At moderate V_{GS}, the asymmetric MOSFET yields a slightly higher current as compared to the symmetric device, while at high V_{GS} this trend is reversed. Electron injection velocity versus gate bias is compared in Fig. 2.13b. Note that strong transverse electric fields exist in the asymmetrical device. The strong electric fields separate subband energy levels of the 2D electron gas, resulting in increased occupancy of the lower energy subbands. Electrons in these subbands respond with a heavy effective mass in the gate confinement direction, but with a light effective mass in the transport direction, resulting in higher electron injection velocities, as illustrated in Fig. 2.13b. (It should be pointed out that strong transverse electric fields reduce mobility in real devices, but the reduction in mobility may not be important in the quasi-ballistic transport regime ($L \lesssim$ 50 nm).



(a) (b)

Fig. 2.13. (a) Schred-2.0 predicted ballistic current in the asymmetrical and symmetrical DG nMOSFET, V_{DS} = 1.0 V. (b) Schred-2.0 predicted average electron injection velocity in the asymmetrical and symmetrical DG nMOSFET. The blue lines with squares represent results of the symmetrical device, the red lines with circles represent results of the asymmetrical device.

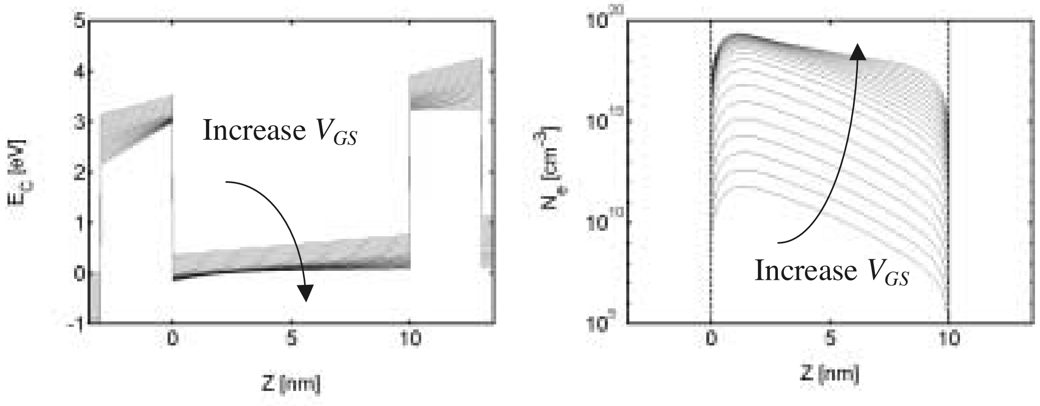
The electron velocities start at the corresponding non-degenerate values at low V_{GS} where the difference between the two curves indicates a larger average transport direction effective mass in the symmetric device (refer to $\sqrt{2k_BT/\pi m^*}$). As the gate bias increases, subband energies are shifted to lower values, but the velocities increase as a result of Fermi statistical degeneracy. The velocity profile for the asymmetric device tends to saturate at large V_{GS} as a result of increased occupancy of the primed subbands. The most interesting result is the fact, that the two device currents are comparable although their internal quantities (N_S , and v_{inj}) look different. To understand the fact, however, one needs to refer to both Figs. 2.11b and 2.13b, since electron densities and injection velocities constitute the device drive current through the expression $I_{DS} = -qN_Sv_{inj}$. Note that the extraordinary high electron density (N_S) presented in the single channel of the asymmetrical device plays a major role in obtaining the exceptional high on-current and merits further discussion.

Discussion:

Gate-to-gate electric coupling through ultra-thin bodies in SOI devices affects charge distributions in the bodies and can translate into improved 2D mobile charge densities. In a thick body partially depleted SOI MOSFET, a neutral silicon block separates the front gate and back gate regions. Therefore the two gates modulate surface charges independently. This situation is more like two back-to-back bulk MOSFETs. Most DG gate SOI MOSFETs are not symmetrically gated (for instance, top gate insulator layers are typically thinner than the bottom gate insulator layers), in which case, one gate may form an inversion layer while the other is still in depletion regime. Thus two threshold voltages are needed to electrically characterize such devices.

In a fully depleted ultra-thin body SOI MOSFET, gates are strongly coupled to each other. Note that in subthreshold operating regimes, there is no screening charge that can protect the body potential from the influence of the gate potentials. So the body potential keeps up almost exactly the pace of the gate potential. It is very similar to what occurs in the base region in a bipolar transistor. This implies $dE_{OXf} / dV_{GS} = dE_{OXb} / dV_{GS} \cong 0$ (E_{OXf} and E_{OXb} are electric fields within the front and back gate oxides) and close-to-ideal $S = 60$ mV/dec. As the gate voltage increases, a ultra-thin body asymmetrically gated SOI device forms only one predominant channel. The channel charges, however, image on both two gates. These channel charges will cause a potential pinning point within the body. In an asymmetrical MOSFET, this point is closer to one gate/silicon interface than the other. The pinning voltage defines the turn-on operating regimes of the device. Figure 2.14a shows the conduction band edge profiles in the ultra-thin body asymmetrically gated SOI as the gate bias is stepped up. The coarsely spaced curves indicate the subthreshold regime, while the closely spaced curves indicate the inversion regime. Note that the potential pinning voltage only depends on the total amount of charge in the body, regardless of gate symmetry (because the channel charge images on both gates). Provided the same amount of mobile charge is obtained at V_{GS} = 0.0 V and the same subthreshold swing (~ 60 mV/dec), an asymmetric DG device will invert at the same voltage as its symmetric counterpart (see Fig. 2.11b).

Once the device is turned on, the amount of charge within the body is characterized by the effective gate capacitance $(C_{Eff}).C_{Eff}$ is the addition of the front capacitance and back capacitances. In an asymmetrical DG MOSFET, the charge centroid is always further away from one of the two gates, giving rise to a decreased C_{Eff}. This is observed in Fig. 2.12 when comparing an asymmetric DG MOSFET to its symmetric counterpart (charges are distributed close to both gates). But the difference is minor in these ultra-thin body devices where charge distribution is limited within the very narrow silicon body regions by insulator layers. Continuously increasing the gate bias beyond the threshold voltage results in a spreading of the mobile charge from the front gate towards the back gate in an asymmetric DG MOSFET. This is shown in Fig. 2.14b, charge density changes very slowly near the front gate but considerably next to the back gate when $V_{GS} > V_{TH}$. The charge spreading effect enhances the back gate capacitance, eventually will eliminate the difference in C_{Eff} between the asymmetrical and symmetrical devices. The comparable C_{Eff} and V_{TH} between the asymmetric and symmetric structures, due to strong gate-to-gate coupling, underlines their similar electrostatic properties.



(a) (b)

Fig. 2.14. (a) Conduction band edge profiles as linearly increasing gate bias in the asymmetrical DG nMOSFET. (b) The log-scaled electron density profiles as linearly increasing gate bias in the asymmetrical DG nMOSFET.

Conclusion:

We conclude that ultra-thin body, asymmetric DG CMOS with n^+ and p^+ polysilicon gates based on conventional technology can provide desired threshold voltages, and more importantly, yield the same or improved performance (I_{ON} and g_m) at low V_{DD} as compared to symmetric DG CMOS when I_{Off} is controlled. The extraordinary I_{ON} in the asymmetric DG MOSFET, is due mainly to two reasons:

1) gate-gate charge coupling resulting in low V_{TH} and high C_{Eff},
2) strong asymmetric electric fields resulting in subband splitting and high υ_{inj}.

2.4 Summary

In this chapter, we first described a 1D Schrödinger-Poisson solver – Schred-2.0. The 1D tool was then used to simulate SOI structures. Important device design parameters such as C_{Eff}, V_{TH} and υ_{inj} were examined. Useful insights illustrating the operation of thin body SOIs were presented. We concluded that ultra-thin body SOI devices can deliver exceptionally high performance compared to thick body SOI and bulk MOSFETs. Electron penetration into the insulator regions was also examined. Electrostatic changes due to the penetration charge noticeably affects performance of devices in nanometer dimensions. Finally, an asymmetric DG MOSFET with n^+ and p^+ polysilicon gates was simulated and compared to its symmetric counterpart. The thin body asymmetric DG device was shown to be able to provide extraordinarily high drive current and desired low threshold voltage.

3. SIMULATION OF QUANTUM EFFECTS IN DG MOSFETS WITH 2D ELECTROSTATICS

3.1 Introduction

As CMOS technology progresses, device dimensions have been scaled into the nanometer regime [4, 18, 25]. Therefore in the future, transistors may operate near their ballistic limit rendering it important to understand ballistic device physics. In most cases, a two dimensional simulator is required to accomplish performance analyses of MOSFETs because of expected 2D short channel effects. The focus of the foregoing chapter was on a critical region near the source where 1D MOS electrostatics was assumed to apply. This treatment is justified when the drain bias is fairly low. In the high drain bias range, important 2D electrostatic effects like DIBL or non-equilibrium transport can not be described directly in the 1D model. In this chapter, we solve the 2D Poisson equation coupled to 1D ballistic transport equations. The 1D ballistic transport is modeled at two different levels: classical level (The Boltzmann transport equation) and quantum level (The Schrödinger equation). The Schrödinger equation is solved using the well-known non-equilibrium Green's function technique [8-9, 40, 72]. The self-consistent solutions enable us directly assess 2D effects and non-equilibrium ballistic transport in MOSFETs. 2D simulations are generally very computationally expensive. In this work we choose ultra-scaled SOI MOSFETs as our model devices and make use of the mode-space representation in expanding the Hamiltonian. The use of extremely thin bodies (T_{Si} < 5 nm) significantly reduces the simulation domain, and use of a mode-space representation greatly reduces the size of the Hamiltonian. Consequently, the computational expense becomes acceptable even on a PC level. The simulation loop consists of two blocks: the Poisson equation which is solved for the potential profile and the transport equation which is solved for charge and current distribution in the device. The finite difference discretization scheme is used in all of the numerical implementation.

Our major objective in this chapter is to describe the numerical techniques used in developing a 2D simulator for nanoscale double-gate MOSFETs. As an application, ballistic transport is examined at a quantum level. Extensive device design simulation studies using the approach developed here can be found in Chapters 5, 6 and [62, 73-74]. The chapter is organized as follows: 1) Section 3.2 describes the 2D Poisson equation solver, 2) Section 3.3 solves the Schrödinger equation in mode-space using the Green's function approach (quantum ballistic transport model), 3) Section 3.4 presents some key simulation results, 4) Section 3.5 examines the potential boundary conditions at the source/drain contacts in ballistic MOSFETs, and 5) Section 3.6 assesses the approximations made in the mode-space representation.

3.2 Solving the Poisson Equation

In Fig. 3.1, we show the model device structure used in our study. The simulation domain and grid mesh are also illustrated.

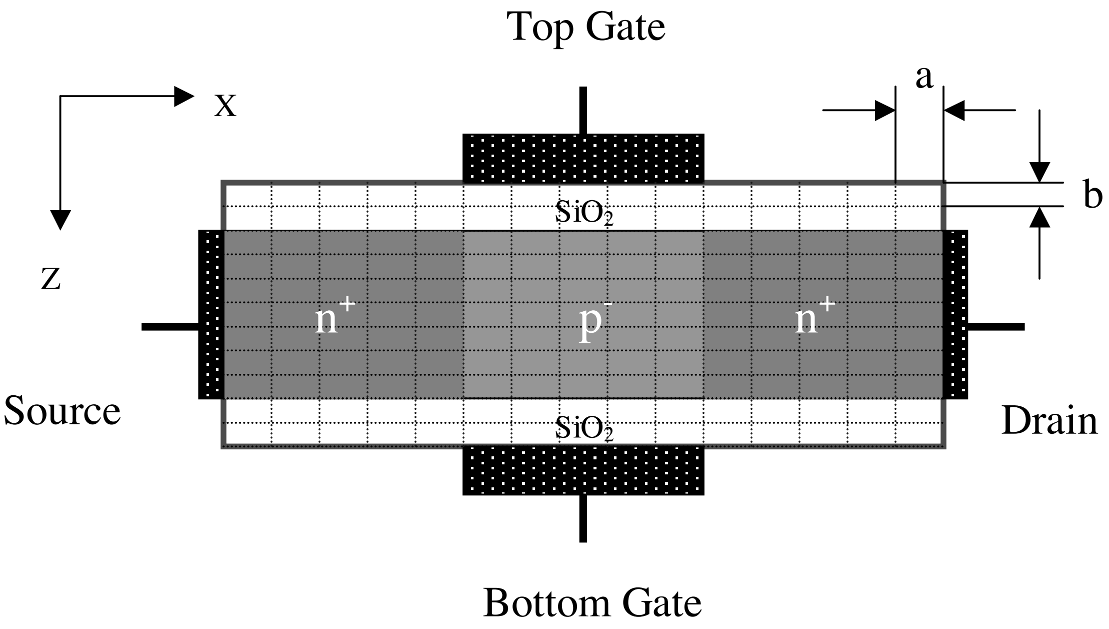


Fig. 3.1. A ultra-thin body double-gate MOSFET structure. The 2D simulation domain is the rectangle enclosed by the solid red line. Uniformly spaced grids are used in both x and z directions, spatial constants are a and b respectively.

The numeric solution to the Poisson equation is obtained by making use of Gauss's law.

$$\oiint[\varepsilon\vec{E}(x,z)]\cdot d\vec{S} = \int_{\Omega} q[p - n + N_D - N_A]d\Omega\ , \qquad (3.1)$$

where $\vec{E}$ is the electric field, p is the hole concentration (which can be neglected in fully depleted ultra-thin body nMOSFETs under consideration in this study), n is the electron concentration, N_D and N_A are donor and acceptor concentrations, q is the elementary charge, ε is the position dependent dielectric constant. The solution domain consists of $N_X \times N_Z$ lattice nodes, where N_X and N_Z the number of nodes in the x and z directions. A 2D numerical solution to the Poisson equation is composed of $N_X \times N_Z$ potential values at each lattice node. To attain the $N_X \times N_Z$ unknowns, the same number of equations is needed. The equations are obtained either by applying eqn. 3.1 at internal nodes (for all internal nodes), or utilizing the boundary conditions (for all boundary nodes). Let's first look at the equations at internal nodes. We choose node [*m*,*n*] (row *m* and column *n*) to illustrate the procedure. Using a central difference approximation for the spatial derivatives, we express $\vec{E}$ in terms of V (vacuum potential). The linearized finite difference form of eqn. 3.1 is

$$\frac{a}{b}V_{m-1,n} + \frac{b}{a}V_{m,n-1} - 2(\frac{a}{b} + \frac{b}{a})V_{m,n} + \frac{b}{a}V_{m,n+1} + \frac{a}{b}V_{m+1,n} = -\frac{ab}{\varepsilon}q(N_D - N_A - n)_{m,n}, \qquad (3.2a)$$

where a and b are mesh spacings in the x and z directions (see Fig. 3.1). The spacing, b, is typically chosen smaller than the spacing, a, to obtain a finer grid in the ultra-thin body or oxide layers for accurate simulations. If node [*m*,*n*] is within the oxide regions or the silicon region, $\varepsilon = \varepsilon_{ox}$, or $\varepsilon = \varepsilon_{Si}$. In the case that the node is positioned at the Si/Oxide interfaces, discontinuity of ε should be accounted for. In such cases eqn. 3.2a becomes

$$\frac{a}{b}V_{m-1,n} + \frac{b}{2a}(1+\frac{\varepsilon_{Bot}}{\varepsilon_{Top}})V_{m,n-1} - (\frac{a}{b}+\frac{b}{a})(1+\frac{\varepsilon_{Bot}}{\varepsilon_{Top}})V_{m,n} + \frac{b}{2a}(1+\frac{\varepsilon_{Bot}}{\varepsilon_{Top}})V_{m,n+1} + \frac{a}{b}\frac{\varepsilon_{Bot}}{\varepsilon_{Top}}V_{m+1,n}$$
$$= -\frac{ab}{\varepsilon_{Top}}q(N_D - N_A - n)_{m,n} \tag{3.2b}$$

where ε_{Top} and ε_{Bot} are dielectric constants for the materials above the interface and below the interface.

Next, we look at the equations for all boundary nodes. At the gate contacts, Dirichlet boundary conditions are specified, meaning $V = V_G$. The gate vacuum potential V_G is determined from the gate bias voltage and workfunction of the contact materials. The numerical equation to be satisfied can be easily written as,

$$V_{m,n} = V_G . \tag{3.2c}$$

At the source/drain contacts, Neumann boundary conditions are imposed, meaning $\vec{n}\cdot\vec{\nabla}V = 0$. These boundary conditions permit contact potentials to float to whatever values are necessary for ensuring charge neutrality at the contact regions. The more common fixed boundary conditions become improper in ballistic transport simulations where non-equilibrium statistics prevail at the source/drain contacts (see Section 3.5 for details). For other boundaries without electrode contacts, the same zero electric field conditions are assumed. These boundary conditions are accomplished numerically by setting

$V_{m,n} - V_{m\pm1,n} = 0$ for the left and right edges,

$V_{m,n} - V_{m,n\pm1} = 0$ for the top and bottom edges,

$2V_{m,n} - V_{m+1,n} + V_{m,n\pm1} = 0$ for the two corner nodes along the top edge, and

$$2V_{m,n} - V_{m-1,n} + V_{m,n\pm1} = 0 \text{ for the two corner nodes along the bottom edge.} \tag{3.2d}$$

Up to this point, we have obtained the $N_X \times N_Z$ equations needed for solving V_{mn}. Given N_D, N_A and n (electron density), eqn. 3.2 represents a set of linear equations that can be solved directly for the vacuum potential. However, when solving a coupled set of equations (the Poisson equation and transport equation), there is a better solution algorithm for solving the Poisson equation [75-77]. This algorithm can provide more efficient convergence in the iteration loop of the Poisson and transport equations. This algorithm involves performing a variable change to n, namely expressing n in terms of the potential and a quasi-Fermi energy, F_n. The quasi-Fermi potential energy is computed based on the old potential,

$$(F_n)_{m,n} = -q(V_{old})_{m,n} + k_B T \cdot \Im_{1/2}^{-1}(\frac{n_{m,n}}{N_C}), \tag{3.3a}$$

where $\Im_{1/2}^{-1}$ stands for the inverse Fermi-Dirac integral of order 1/2 and N_C the effective density of states in the conduction band (a normalization factor). (For the analytical approximation for $\Im_{1/2}^{-1}$, see [85].) The electron density term in eqn. 3.2 now becomes

$$n_{m,n} = N_C \Im_{1/2}[\frac{(F_n)_{m,n} + qV_{m,n}}{k_B T}], \tag{3.3b}$$

With this variable change, eqn. 3.2 now represents a set of nonlinear equations for the potential. The reason for introducing the variable change can be understood as follows: the non-linearity in eqn. 3.2 in fact provides a mechanism damping the updates to V in successive solution iterations of the coupled equation system. Referring to eqns. 3.2a-b and 3.3b, the increase in $V_{m,n}$ on the right side of eqns. 3.2a-b implies an increase in $n_{m,n}$

(through eqn. 3.3b), which, however, will reduce $V_{m,n}$ on the other side of eqns. 3.2a-b. This damping effect helps avoid large changes between V_{old} and V_{new}, therefore making the coupled equations converge stably, and efficiently in terms of computational time. The technique has been used to couple a variety of transport models to the Poisson equation, from drift-diffusion [86], to quantum ballistic [76], to Monte Carlo particle models [77].

Since the Poisson equation is now a nonlinear equation, it is solved by an inner Newton-Raphson loop [78-79]. We denote eqns. 3.2a-d by $F_\alpha(V) = 0$, where the index denoted by α run from 1 to $N_X \times N_Z$. The Jacobian matrix is obtained as

$$F_{\alpha,\beta}(V) \equiv \frac{\partial F_\alpha(V)}{\partial V_\beta}. \tag{3.4a}$$

Given an initial guess or old solution V_{old}, the projected solution is $V_{new} = V_{old} + \Delta V$. Using a Taylor expansion of the first order, we have

$$F_\alpha(V_{new}) \approx F_\alpha(V_{old}) + F_{\alpha,\beta}(V_{old}) \cdot [\Delta V]_\beta = 0. \tag{3.4b}$$

It is very clear that the updates can be obtained as

$$[\Delta V]_\beta = -F_{\alpha,\beta}(V_{old}) \backslash F_\alpha(V_{old}), \tag{3.4c}$$

where the right side of eqn. 3.4c indicates the matrix division of $F_{\alpha,\beta}(V_{old})$ into $F_\alpha(V_{old})$, which is the same as multiplying the inverse matrix of the Jacobian, $[F_{\alpha,\beta}]^{-1}$, to $F_\alpha(V_{old})$ except that it is computed in a different but more efficient way (Gaussian elimination [79]). The process is repeated until the residual of $F_\alpha(V)$ is less than the specified convergence norm. The Newton-Raphson approach provides quadratic convergence, so the number of iterations is small. But the size of the Jacobian is $(N_X \times N_Z)^2$, so the memory and time of conducting Gaussian eliminations may be excessive.

To illustrate the solution process of the Poisson equation, in Fig. 3.2, we present the sparsity pattern of the Jacobian matrix. The pattern shows that the Jacobian is a very sparse matrix. The five diagonal lines indicate that each node is only coupled to its four neighbors in finite difference approximation. This can also be seen in eqns. 3.2a-d. The sparsity of the Jacobian gives rise to large savings in memory and computational time.

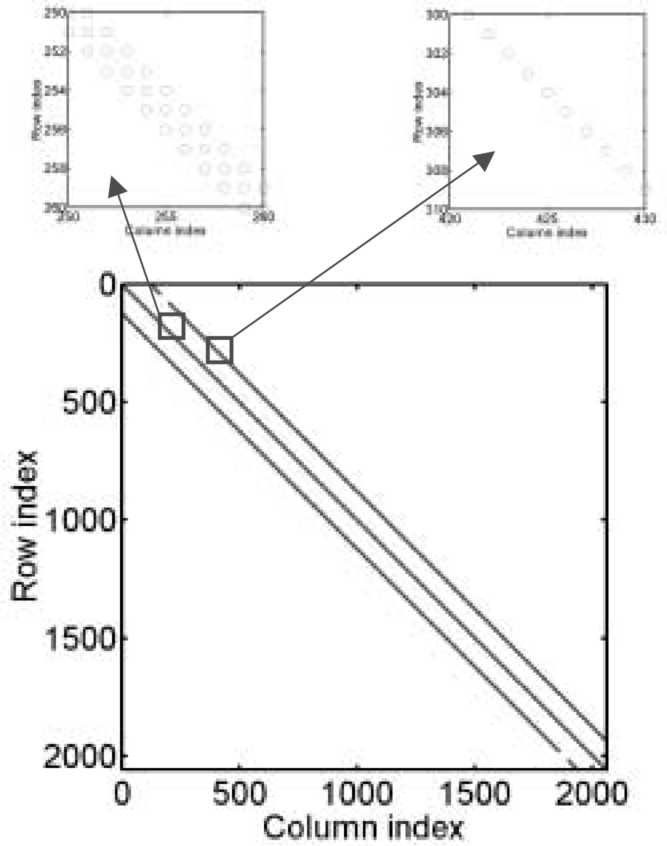
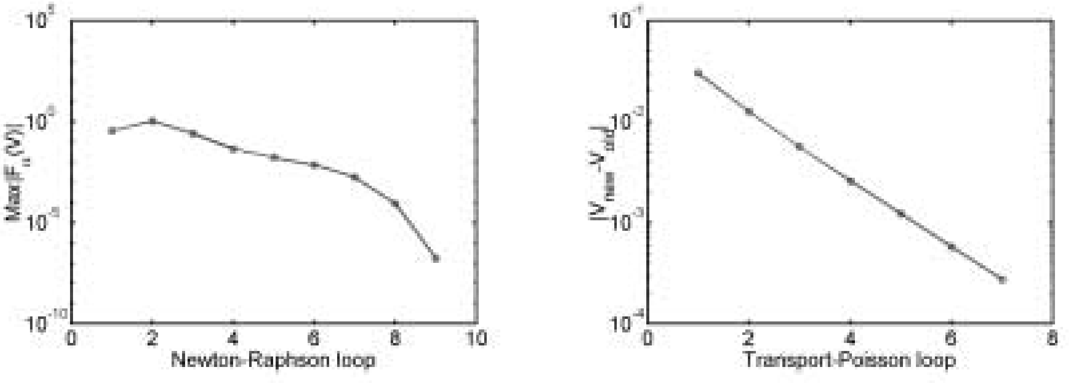


Fig. 3.2. The sparsity pattern of the Jacobian in solving the nonlinear Poisson equation.

Figure 3.3a shows the convergence profile of the Newton-Raphson loop. The maximum residue $F_\alpha(V)$ is plotted versus iteration number. The first several points indicate the stage when the trial potentials are further away from the solution. After that, quadratic convergence is seen. Within small number of iterations, the residue can drop below 10^{-6} V.

Figure 3.3b shows the convergence profile of the outer iteration loop of the transport and Poisson equations. The ballistic Boltzmann transport equation model has been chosen in the simulation (to be described in next section), but the conclusion is general to any model that has been implemented. The maximum difference between the old potential and the new potential is used as the measure of convergence. Due to the use of the aforementioned variable change approach in solving the Poisson equation, the convergence is smooth and monotonic displaying a factor of two decrease in the maximum potential difference for each outer iteration.



(a)

(b)

Fig. 3.3. (a) The maximum potential residue versus the number of Newton-Raphson loop in solving the Poisson equation. (b) The potential correction versus the number of outer iteration loop in solving the coupled Transport-Poisson equation sets.

3.3 Quantum Ballistic Transport

The Green's function is solved to obtain the electron density within the device and current at the terminals in the ballistic limit. Under ballistic conditions, the Green's function method is mathematically equivalent to solving the Schrödinger equation with open boundary conditions. We use the Green's function method because it can be extended to include interactions (i.e. scattering) as discussed in Chapter 4. To solve for the Green's function, a mode space representation is used in the gate confinement direction. This approach greatly reduces the size of the problem and provides good accuracy as compared to full 2D spatial discretization [80-81]. To see how the mode space representation works, we need to refer to Fig. 3.1. The procedure is explained as follows:

I. We first solve the 1D effective mass Schrödinger equation along each z directed slice in the 2D discretization mesh (as we did in the 1D simulations presented in Chapter 2)

$$-\frac{\hbar^2}{2m_z^*}\frac{d^2}{dz^2}\psi_i(x,z) - qV(x,z)\psi_i(x,z) = E_i(x)\psi_i(x,z)\,, \tag{3.5}$$

where m_z^* is the electron effective mass in the z direction. $\psi_i(x,z)$ and $E_i(x)$ are the wavefunction and eigenenergy for subband i at slice x (the eigenvector and eigenvalue of the mode space representation in the gate confinement direction). The envelope wavefunctions of electrons are assumed to be zero at the oxide/Si interfaces if electron penetration into the oxide regions is not accounted for (otherwise, the zero boundary is extended to the contact/oxide interfaces). The width of the slice is chosen less than 3Å. Within each slice, all quantities are assumed to be constant in the x direction.

II. The 3D Hamiltonian for the device is expanded in terms of $\delta(x-x')\psi_i(x,z)$ and $\exp(ik_j y)/\sqrt{W}$. The function, $\exp(ik_j y)/\sqrt{W}$, is the plane wavefunction along the device width (W denotes the device width). The quantum number, k_j, corresponds to the eigenenergy $\frac{\hbar^2 k_j^2}{2m_y^*}$, where m_y^* is the electron effective mass in the y direction. $\delta(x-x')$ is the real space delta function with an eigenvalue x'. It is easy to see that $\delta(x-x')\psi_i(x,z)$ and $\exp(ik_j y)/\sqrt{W}$ form a complete and orthogonal expansion functions set. The Hamiltonian is obtained as

$$H = \begin{bmatrix} H[E_1(x)+E_{k_j}] & 0 & \ldots & 0 & 0 \\ 0 & H[E_2(x)+E_{k_j}] & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & H[E_i(x)+E_{k_j}] & 0 \\ 0 & 0 & \ldots & 0 & \ldots \end{bmatrix}, \qquad (3.6a)$$

where

$$H[E_i(x), E_{k_j}] = \begin{bmatrix} 2t - E_i(1) + E_{k_j} & -t & \ldots & 0 & 0 \\ -t & 2t - E_i(2) + E_{k_j} & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & -t \\ 0 & 0 & \ldots & -t & 2t - E_i(N_X) + E_{k_j} \end{bmatrix}$$

(3.6b)

is the Hamiltonian for subband i, with a planewave eigenenergy E_{k_j}. The subband index, i, runs over all subbands, but in real calculations, including the lowest few subbands provides desired accuracy. E_{k_j} ranges between 0 and $+\infty$ accounting for all possible transverse plane waves. Numbers 1 to N_X in parenthesis replace the position of x because of the discretiztion. The coupling term within each subband is indicated by

$$t = \frac{\hbar^2}{2m_x^* a^2}, \qquad (3.7)$$

where is m_x^* the electron effective mass in the x direction, and a is the finite difference lattice constant. Subband to subband coupling is ignored in this treatment, because it has been shown in full 2D simulations that the band-to-band coupling is negligible in SOI MOSFETs with uniform bodies along the channels [81]. 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 can be treated individually, and the size of Hamiltonian is measured by the nodal number along the channel direction, namely $(N_X)^2$. Therefore, it is very clear the latter approach can provide enormous savings in computational burden.

III. For the subband i, with a planewave eigenenergy E_{k_j}, we write the retarded Green's function relevant to the 1D transport as [72]

$$G(E) = [EI - H[E_i(x), E_{k_j}] - \Sigma]^{-1} = [E_l I - H[E_i(x)] - \Sigma]^{-1}, \tag{3.8}$$

where, we define the longitudinal (x) energy $E_l \equiv E - E_{k_j}$. The third term in the bracket is called the self-energy matrix, which is given as

$$\Sigma = \begin{bmatrix} \Sigma_S & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & ... & 0 \\ 0 & ... & ... & ... & 0 \\ 0 & ... & ... & 0 & 0 \\ 0 & 0 & ... & 0 & \Sigma_D \end{bmatrix}. \tag{3.9a}$$

The two corner entries in $\Sigma(E)$ represent the effects on the finite device Hamiltonian due to the interactions of the device with the contacts [10]. The self-energy concept allows us to eliminate the huge reservoir and work solely within the device subspace whose dimensions are much smaller. $\Sigma(E)$ can be expressed in terms of known quantities [72]. At the source contact,

$$\Sigma_S(E) = -te^{ik_l a}, \text{ where } E = E_{k_j} + E_i(1) + 2t(1 - \cos k_l a), \tag{3.9b}$$

$E_i(1)$ is the subband energy at the contact boundary. $\sum_D(E)$ can be obtained in a similar fashion. It is very important to note that the self-energies are functions of longitudinal energy $E_l \equiv E - E_{k_j}$ as shown in eqn. 3.9b. This allows us to focus on the longitudinal energy in all our calculations. Under ballistic conditions, the transverse mode contributions (planewaves in the y direction) can be treated independently of the longitudinal contribution. The treatment of transverse modes will be explained later in this section.

Once the Green's function is obtained, internal electron density and terminal current of the device under study can be computed [40, 82]. We define a new quantity in terms of self-energies

$$\Gamma \equiv i(\Sigma - \Sigma^{+}) \, . \tag{3.9c}$$

Physically this function determines the electron exchange rates between the source/drain reservoirs and the active device region [10]. But in general it can be viewed as the measure of interaction strength due to any perturbation source. Although the device itself may be in a non-equilibrium state, electrons are injected from the equilibrium source/drain reservoirs. The spectral density functions due to the source/drain contacts can be obtained as

$$A_S = G\Gamma_S G^{+} \text{ and } A_D = G\Gamma_D G^{+} \, , \tag{3.10}$$

where $\Gamma_S \equiv i(\Sigma_S - \Sigma_S^{+})$, and $\Gamma_D \equiv i(\Sigma_D - \Sigma_D^{+})$(For clarity, here we use Γ_S or Γ_D to denote matrices the same size as *G*, with nonzero diagonal entries $i(\Sigma_S - \Sigma_S^{+})$ or $i(\Sigma_D - \Sigma_D^{+})$). Note that the spectral functions are N_X by N_X matrices and the diagonal entries represent the local density-of-states at each node. The source related spectral function is filled up according to the Fermi energy in the source contact, while the drain related spectral function is filled up according to the Fermi energy in the drain contact. The 2D electron density matrix is obtained as

$$n(E_l) = \frac{1}{2\pi a}\int_0^{+\infty}\frac{2}{\pi\hbar}\sqrt{\frac{m_y^*}{2E_{k_j}}}[f(\mu_S - E_l - E_{k_j})A_S + f(\mu_D - E_l - E_{k_j})A_D]\cdot dE_{k_j}\,, \qquad (3.11a)$$

where f is the Fermi-Dirac function, and $\frac{2}{\pi\hbar}\sqrt{\frac{m_y^*}{2E_{k_j}}}$ represents the transverse mode state density (including the spin degeneracy). Since the spectral functions depend on the longitudinal energy only, they can be moved out of the integration sign. Therefore eqn. 3.11a reduces to,

$$n(E_l) = \frac{1}{\hbar a}\sqrt{\frac{m_y^* k_B T}{2\pi^3}}[\Im_{-1/2}(\mu_S - E_l)A_S + \Im_{-1/2}(\mu_D - E_l)A_D]\,, \qquad (3.11b)$$

where the Fermi-Dirac integral of $\Im_{-1/2}$ accounts for all transverse mode contributions (see [59-60] for analytical approximation for $\Im_{-1/2}$, and also note that all quantities appearing as arguments of Fermi-Dirac integrals are normalized to $k_B T$). To obtain the total 2D electron density, we need to integrate eqn. 3.11b over E_l. We also need to sum contributions from every conduction band valley and subband. Finally, we can get a 3D electron density by multiplying the corresponding distribution function $|\psi_i(x,z)|^2$ to the 2D density matrix at each longitudinal lattice node. The 3D electron density is fed back to the Poisson equation solver for the self-consistent solution.

Once self-consistency is achieved, the terminal current can be expressed as a function of the transmission coefficient [10]. The transmission coefficient from the source contact to the drain contact is defined in terms of the Green's function as

$$T_{SD} = Trace[\Gamma_S G \Gamma_D G^+] \,. \tag{3.12}$$

It is straightforward to write the transmitted current as

$$I(E_l) = \frac{q}{h}\int_0^{+\infty} \frac{2}{\pi\hbar}\sqrt{\frac{m_y^*}{2E_{k_j}}}[f(\mu_S - E_l - E_{k_j}) - f(\mu_D - E_l - E_{k_j})A_D]T_{SD}(E_l, E_{k_j})\cdot dE_{k_j}$$

(3.13a)

where the 2 in the numerator is for spin degeneracy. Note that T_{SD} is independent of transverse energy E_{k_j} (refer to eqns. 3.8, 3.9 and 3.10) and can therefore be moved out of the integration sign. Equation. 3.13a then reduces to,

$$I(E_l) = \frac{q}{\hbar^2}\sqrt{\frac{m_y^* k_B T}{2\pi^3}}[\Im_{-1/2}(\mu_S - E_l) - \Im_{-1/2}(\mu_D - E_l)]T_{SD}(E_l) \,, \tag{3.13b}$$

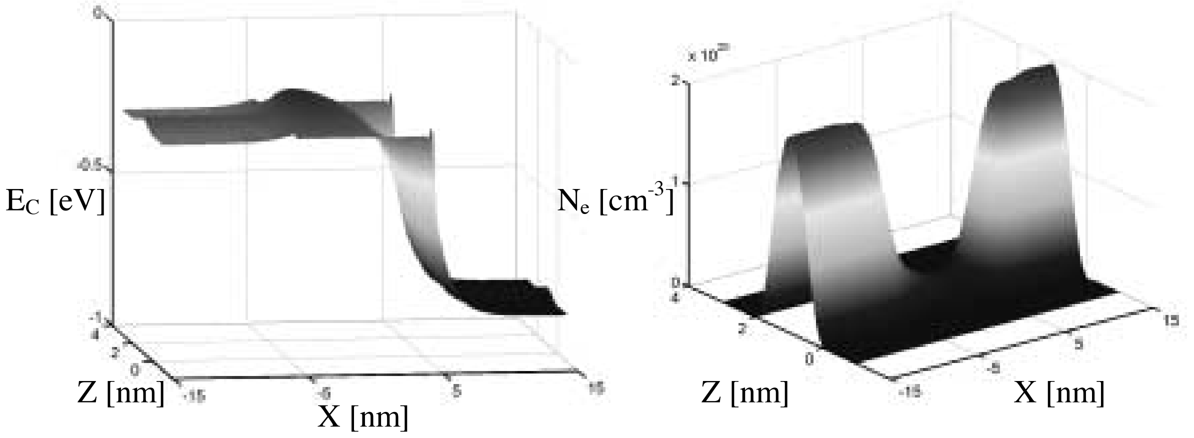
The total current is obtained by integrating over E_l and summing over all valleys and subbands.

For comparison purposes, we also implement a semi-classical approach in solving the ballistic transport in ultra-scaled MOSFETs. This approach is a solution to the Boltzmann transport equation (BTE) solved along the channel direction. But the physics in the direction normal to the channel are treated quantum mechanically in the same way as described earlier in this chapter. The BTE solutions can capture the thermionic emission current, but can never capture any quantum related effects, such as electron tunneling through the source-channel barrier and quantum interference within ballistic devices. The scheme to solve the BTE in the ballistic limit is presented in the Appendix B. The comparison of the quantum versus classical approaches is presented in this section based on simulation results.

3.4 Simulation Results

The schematic diagram of the simulated device is the same as shown in Fig. 3.1. The n-type source and drain regions are doped at10^{20} cm^{-3}. The channel is intrinsic, and the channel junction is abrupt. No gate-to-S/D overlap is assumed. The oxide thickness is 1.0 nm for both top and bottom gates and the silicon film thickness is 2.0 nm. Gate workfunction is 4.4 eV.

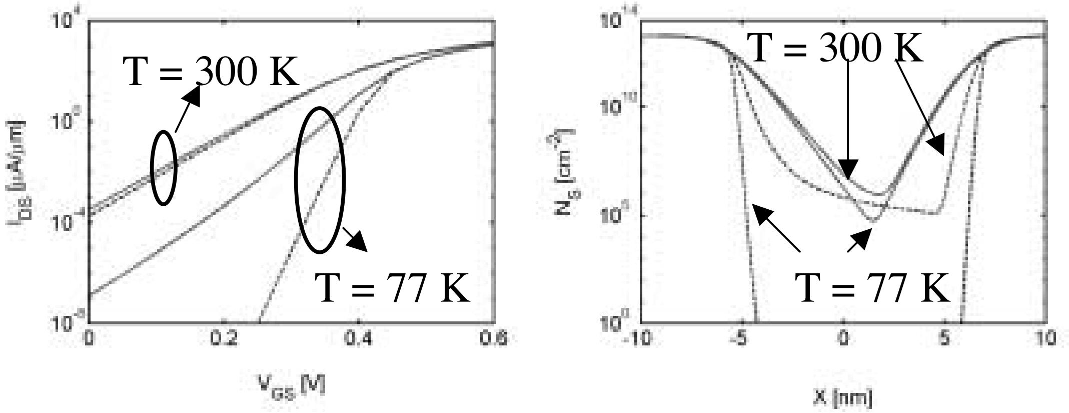
Figures 3.4a and 3.4b show the electron density and potential energy profiles in the x-z cross-section of the model MOSFET as simulated with the Green's function ballistic model (V_{GS} = V_{DS} = 0.6 V). It is observed that, the electron density reduces to zero at the oxide/silicon interfaces due to well-known quantum effects. It is also observed that, there is a barrier region near the source end of the channel. This barrier determines the amount of electrons entering the channel. Its height is modulated by the gate bias.



(a) (b)

Fig. 3.4. Conduction band edge profile (a) and electron distribution (b) within the x-z cross-section of the model MOSFET at V_{GS} = V_{DS} = 0.6 V.

Figure 3.5a shows the simulated I_{DS} versus V_{GS} characteristics of the model MOSFET at room (300 K) and low temperatures (77 K). Both the BTE and Green's function simulation results are presented for comparison. The quantum tunneling current is additional to the thermal emission current, leads to higher off-current for room temperature operations. As the temperature goes down, the thermal emission current is remarkably suppressed by the large channel barrier, the quantum tunneling becomes the only contributor. The tunneling component eventually limits the device scaling. In Fig. 3.5b, the quantum and classical carrier density profiles are also compared. The observed tunneling effects are more evident with regard to the carrier distributions.



(a) (b)

Fig.3.5. (a) Simulated I_{DS} versus V_{GS} characteristics of the model MOSFET at two different temperatures, 300 K and 77 K. V_{DS} = 0.6 V. (b) 2D electron density distributions along the channel at $V_{GS} = V_{DS}$ = 0.6 V. The solid lines are results from the Green's function model, the dashed lines are results from the BTE.

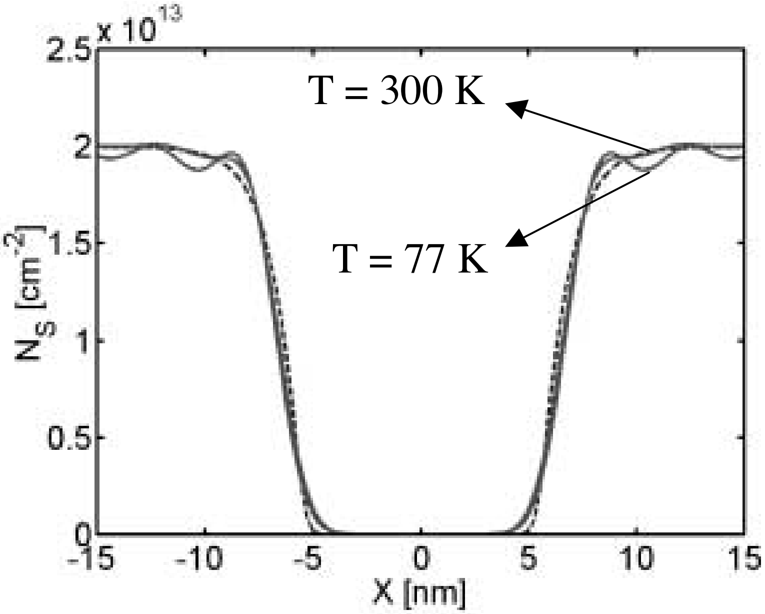


Fig. 3.6. Electron density oscillation due to the quantum interference effect. Solid lines are results from the Green's function model. Dashed line is the result from the BTE model at 300 K. $V_{GS} = V_{DS}$ = 0.0 V.

Quantum interference distinguishes the Green's function and BTE models. Due to the interference between the incident and reflected electron waves, carrier density oscillations can be seen near the channel barrier edges. This effect is illustrated in Fig. 3.6. At high temperatures, the interference effect is somewhat washed out by the statistics; at low temperatures, the oscillation patterns become sharper, indicating strong interference. The BTE simulations do not show any such effect.

3.5 The Floating Boundary Condition

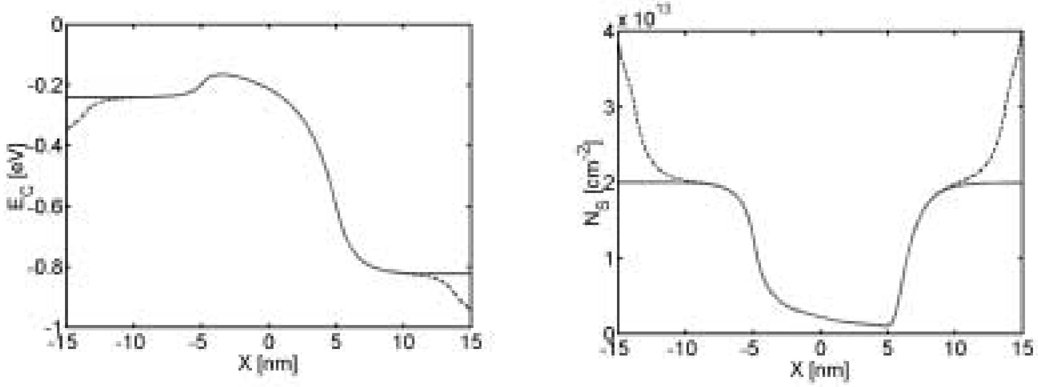
In the ballistic simulations, we impose a floating boundary condition in solving the Poisson equation. This boundary condition is realized by assuming

$$\vec{n} \cdot \nabla V = 0 \,. \tag{3.14}$$

at the source /drain ends of the simulation domain. This boundary condition is different from that being commonly employed in scattering dominated simulations, where a fixed potential boundary is imposed based on equilibrium statistics to obtain charge neutrality. It is of interest to look at how this boundary condition works.

Carriers are injected either from the source or drain, and reflected by the channel barrier. In a real contact, scattering maintains a near-equilibrium distribution. However, in the absence of scattering (ballistic case), at high V_{DS}, the drain-injected carriers are suppressed, leaving the source end partially exhausted. Therefore it is clear that non-equilibrium distribution of electrons emerges as the transport becomes ballistic. For this reason, when solving the Poisson equation, it is inconvenient to fix the potential at the boundaries. A better approach is to impose a zero-field boundary condition at the ends and let the potential float to whatever value it chooses, in order to obtain macroscopic charge neutrality.

Simulated potential profiles and corresponding charge distributions are presented in Figs. 3.7a and 3.7b to illustrate the boundary treatment. As a comparison, results simulated with the fixed-potential boundary are also shown. In the latter case, heavily doped regions ($N_D = 2.0 \times 10^{20}$ cm^{-3}) are intentionally appended to the source and drain of the MOSFET in order to create built-in barriers near the boundaries. These barriers cause strong reflections at the contact regions, giving rise to a near-equilibrium distribution (we note that the heavily doped regions are critical to ensure that the simulations to converge). Except for the boundary regions, the two simulations give identical results. Near the boundaries, the second simulation (with the fixed-potential boundary being imposed) displays a distorted charge and potential profile due to the specially altered source/drain doping. The results indicate that the floating boundary condition becomes more appropriate than the other in performing ballistic simulations.

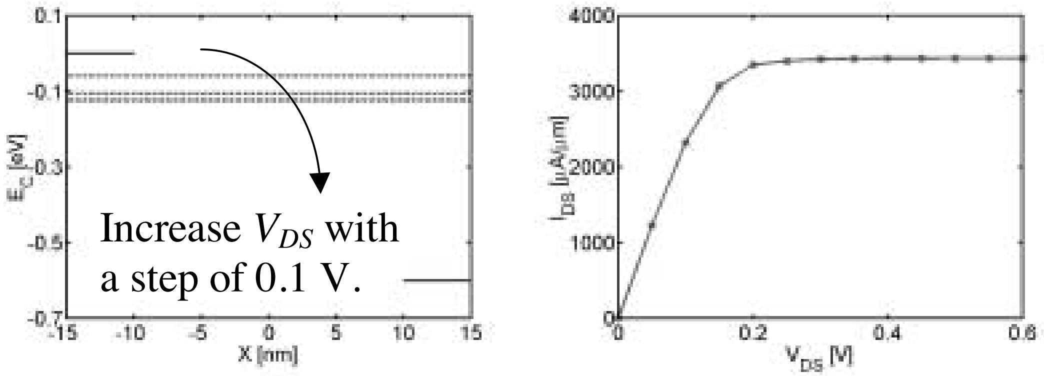


(a) (b)

Fig. 3.7. (a) Conduction band edge profiles along the channel middle line. (b) 2D electron density profile along the channel. $V_{GS} = V_{DS} = 0.6$ V. The solids are results assuming the floating boundary condition in solving the Poisson equation, the dashed lines are results assuming the fixed potential boundary condition.

Note that the boundary under discussion is the simulation boundary within which we assume nonequilibrium ballistic transport, and not the actual contact boundary of the device. Therefore one may find that the potential drop from source-to-drain may not be qV_{DS}. This effect becomes evident when the floating boundary condition is used (shown in Fig. 3.7a). The extra potential drop at the two contacts is the well-known quantum contact potential [83-84]. This contact potential drop decreases as the channel barrier increases, when source and drain regions are closer to equilibrium due to the barrier reflections.

When the barrier in between disappears, the potential will drop entirely at the contacts. This is shown in Fig. 3.8a, where a uniformly doped 2D silicon sheet (mimicking the inversion layer of a MOSFET, N_D = $1.0\times10^{13}\,\mathrm{cm}^{-3}$) is simulated under ballistic conditions. In the absence of a built-in barrier, the simulation shows no potential drop throughout the device. The electron density is constant everywhere, with positive-going electrons in equilibrium with the source Fermi level, and the negative-going electrons in equilibrium with the drain Fermi level. The conduction band edge is positioned at some value so that the charge neutrality is achieved within the device. As the drain bias increase, the drain injected electrons are gradually suppressed. Current saturation occurs (shown in Fig. 3.8b) when the source-injected electron density is equal to the dopant concentration.



(a) (b)

Fig. 3.8. (a) Conduction band edge profiles at different drain biases. The short solid lines indicate the source and drain Fermi potentials at V_{DS} = 0.6 V. The profiles are flat, and stop to change when drain injected electrons are fully suppressed. (b) Simulated I_{DS} versus V_{DS} characteristics of the uniformly doped 2D silicon sheet.

It has been pointed out that in order to reduce parasitic resistances, a fanned-out contact structure must be used [25]. It has also been argued that a ballistic simulation domain must be terminated by such a scattering dominated contact [61]. The concern, then, is that in a scattering dominated contacts, carriers could be backscattering into the source, thereby reducing the on-current. However, MOSFETs are typically wide. Therefore electrons entering the drain are likely to scatter into transverse modes and unlikely to have the longitudinal momentum needed to return to the source. Additionally, the fanned-out contacts used in real devices would further reduce the probability of backscattering into the source. Our ballistic contacts are therefore treated as perfectly absorbing, accepting the non-equilibrium distribution of electrons emerging from the device, and reinjecting a fully thermalized distribution back into the device. The role of backscattering and its relation to contact structure are however topics worthy of additional investigation.

3.6 Discussion on Mode-space Representation

The mode-space approach is employed in solving the Schrödinger equation. It is shown that this approach greatly reduces the size of the problem as compared to a full 2D spatial discretization. It is also important to look at the conditions under which this approach provides good simulation accuracy. In this section, we analytically expand the Schrödinger equation using a mode space representation, and assess the approximation being made in simplifying the Hamiltonian.

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)

$$-\frac{\hbar^2}{2m_x^*}\frac{d^2}{dx^2}\varphi(x,z)-\frac{\hbar^2}{2m_z^*}\frac{d^2}{dz^2}\varphi(x,z)-qV(x,z)\varphi(x,z)=E\varphi(x,z)\ . \qquad (3.15)$$

Left multiplying the mode space eigenvectors (refer to eqn. 3.5) to eqn. 3.15 and doing integrations in real space gives rise to

$$\int [\delta^*(x-x')\psi_i^*(x,z)]\cdot[-\frac{\hbar^2}{2m_x^*}\frac{d^2}{dx^2}\varphi(x,z)]dxdz$$
$$+\int [\delta^*(x-x')\psi_i^*(x,z)]\cdot[-\frac{\hbar^2}{2m_z^*}\frac{d^2}{dz^2}-qV(x,z)]\varphi(x,z)dxdz$$
$$= E\int [\delta^*(x-x')\psi_i^*(x,z)]\cdot\varphi(x,z)dxdz$$

(3.16)

where the * denotes conjugate transforms. Since $[\delta(x-x')\psi_i(x,z)]$ is a real function, so its conjugate is equal to itself. By the definition of the delta function, the third term in eqn. 3.16 becomes

$$E\int \psi_i^*(x',z)\cdot\varphi(x',z)dz = E\tilde{\varphi}_i(x')\,. \tag{3.17a}$$

Note that $\tilde{\varphi}_i(x')$ is the expansion coefficient of $\varphi(x',z)$ with regard to the mode space eigenvector $\psi_i(x',z)$ as defined by

$$\varphi(x',z) = \sum_{i=1}^{\infty}\tilde{\varphi}_i(x')\psi_i(x',z)\text{ , and }\int \psi_i^*(x',z)\psi_j(x',z)dz = \delta_{ij}\,, \tag{3.17b}$$

where δ_{ij} is the usual Kronecker delta. Again making uses of the properties of the delta function and eqns. 3.5, 3.17b, we can rewrite the second term in eqn. 3.16 as

$$\int \psi_i^*(x',z)\cdot[-\frac{\hbar^2}{2m_z^*}\frac{d^2}{dz^2}-qV(x',z)]\varphi(x',z)dz = E_i(x')\tilde{\varphi}_i(x')\,. \tag{3.17c}$$

Finally let’s look at the first term in eqn. 3.16. It can be expressed as

$$\int \delta(x-x')\psi_i^*(x,z)[-\frac{\hbar^2}{2m_x^*}\frac{d^2}{dx^2}\varphi(x,z)]dxdz =$$

$$\int \delta(x-x')\{-\frac{\hbar^2}{2m_x^*}\frac{d^2}{dx^2}[\psi_i^*(x,z)\varphi(x,z)]\}dxdz$$

$$-\int \delta(x-x')\varphi(x,z)[-\frac{\hbar^2}{2m_x^*}\frac{d^2}{dx^2}\psi_i^*(x,z)]dxdz$$

$$-\int \delta(x-x')[-\frac{\hbar^2}{m_x^*}\frac{d}{dx}\varphi(x,z)\frac{d}{dx}\psi_i^*(x,z)]dxdz.$$

(3.17d)

Note that the second term in eqn. 3.17d reduces to $-\frac{\hbar^2}{2m_x^*}\frac{d^2}{dx'^2}\tilde{\varphi}_i(x')$ after integration. *If we assume*

$$\frac{d}{dx}\psi_i^*(x,z) = 0\,, \tag{3.18}$$

Equation (3.16) becomes

$$-\frac{\hbar^2}{2m_x^*}\frac{d^2}{dx'^2}\tilde{\varphi}_i(x') + E_i(x')\tilde{\varphi}_i(x') = E\tilde{\varphi}_i(x')\,. \tag{3.19}$$

Equation 3.19 is the mode space transformation of eqn. 3.15 under assumption of eqn. 3.18. Equation 3.19 is indeed a 1D partial differential equation, so the original 2D problem is greatly reduced. Using the finite difference scheme we can easily obtain eqn. 3.6a.

The assumption is valid when the vertical (in the gate confinement direction) potential profile variations along the channel direction are negligible. For instance, if $V(x,z)$ takes the same shape but different values at different x, the eigenfunctions

$\psi_i(x,z)$ are the same along the channel, although the eigenvalues $E_i(x)$ are different. As a result, eqn. 3.18 holds exactly in this scenario. Note that eqn. 3.19 indicates subbands are independent, labeled by index i, so the big Hamiltonian can be divided into uncoupled sub-blocks, in which only the three diagonals are nonzero.

For SOI MOSFETs with uniform thin bodies, there is little possibility for the potential to vary vertically. Therefore eqn. 3.18 is a fairly good approximation. This has been verified by the full 2D spatial discretization approach [81]. However, 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. So the mode space approach becomes inappropriate for bulk device simulations. SOI MOSFETs with nonuniform bodies along the channel can also invalidate the use of the mode space approach for the same reason.

3.7 Summary

The accomplishments of this chapter are twofold: 1) we described the numerical techniques used in simulating 2D double-gate MOSFETs, 2) we explored the important device physics of MOSFETs operating in the ballistic limit. The numerical implementation of the self-consistent solution loop consists of two components, namely, the Poisson equation solver, and the transport equation solver. We first presented the Newton-Ralphson approach of solving the Poisson equation. This nonlinear solution approach can provide desired convergence efficiency. We then introduced the Green's function based solution scheme to the Schrödinger equation (The solution scheme to the BTE was also presented in Appendix B). We showed that by invoking a variable change to the Poisson equation, stable coupling between the transport equation and Poisson equation can be obtained, resulting in monotonic convergence down to small errors of 10^{-4} V. Using highly simplified device structures, we simulated double-gate MOSFETs with the full quantum Green's function model and semiclassical BTE model. The results indicated that, strong quantum tunneling through the channel barrier occurred in sub-10 nm MOSFETs. This tunneling effect can eventually limit device scaling. Quantum interference was also observed in our simulations. These interference effects became more evident as simulation temperatures were decreased.

4. TREATMENT OF SCATTERING IN NEGF MOSFET SIMULATIONS

4.1 Introduction

In the previous chapter, we discussed electron transport in the ballistic limit. Real devices operate below this limit. Therefore this chapter is dedicated to describing scattering mechanisms in MOSFETs. Dissipative transport can be due to many reasons. Microscopically, electrons are confined within a very thin channel in a MOSFET. The channel is either sandwiched by gate insulators (in thin body SOI structures), or by gate insulators and silicon substrates (in bulk structures). In principle, the insulator surfaces can never be perfectly smooth, and semiconductor lattices are never defect free. Also, both channel carrier densities and impurity concentrations are typically very large and devices typically work at relatively high temperatures (much greater than 300 K). All these factors contribute to carrier scattering. The important scattering mechanisms affecting carrier transport in transistors are:

1) surface roughness scattering,
2) electron-electron scattering,
3) impurity-electron scattering,
4) phonon-electron scattering.

Mobility (μ) is a measure of scattering in conventional device simulation tools with low mobilities indicating a high level of scattering. In quantum mechanical simulators, scattering is characterized by the electron state lifetime (τ) as a scattering event implies the end of an existing electron state and the start of a new one. In general, mobility can be related to electron lifetime through an expression

$$\mu = \frac{q < \tau >}{< m^* >}, \tag{4.1}$$

where, q is the elementary charge and m^* the effective mass in the transport direction of the scattered electron, and <> stands for ensemble average. The averages are needed because the lifetime may be energy dependent, and multiple subbands with different effective masses may be occupied.

Irrespective of the techniques used to simulate scattering, the essential physics of scattering have to be captured. Lundstrom's 1997 paper illustrates the essential physics of scattering. The key points in his paper can be summarized by considering the on-state of a MOSFET with electrons thermally injected from the source, undergoing scattering in the channel and being collected by the drain. *Scattering can occur anywhere inside the device, but only those scattering events occurring in the low field region near the source significantly affect the on-state current* [29, 55]. Scattering in this low field region is important for two reasons: 1) electrons leaving the big reservoir retain high longitudinal energies (transport relevant energies); 2) backscattered electrons need only a small amount of longitudinal energy to overcome the potential drops and reach the source. Electrons in the high field region near the drain can be very energetic, but electrons scattering near the drain end are forced by the high electric fields to leave the channel, having little chance making it back to the source. Therefore scattering in this region does not degrade the on-state current directly.

This chapter is focused on scattering phenomena in ultra-scaled double-gate MOSFETs. Three different scattering models have been implemented, examined, and compared. The study of different scattering models focuses on the essential physics of scattering occurring in nanoscale MOSFETs (see Lundstrom [55]). The two Büttiker probe based models simulate scattering due to all possible mechanisms (*e.g.* surface roughness, phonon, and impurity etc.) as a perturbation represented by the probe's self-energy [40-41, 72]. These perturbations can be related to conventional low field mobility (μ). The third approach focuses on phonon-electron scattering [82, 87-89]. Although the phonon-electron interaction model may not generally be used to simulate scattering in MOSFETs, it provides a more rigorous solution to help us better understand specific scattering process. The Green's function formalism has been used in the implementation of all of the aforementioned scattering models. This formalism is very general and provides the framework necessary to treat the complicated transport processes in nanoscale devices.

The chapter is organized as follows: Section 4.2 presents the theoretical framework, Section 4.3 compares different approaches using simulation results (note that applications of the scattering models in device simulations are presented in Chapters 5, 6 and [26, 74, 90]), the last section provides an extended discussion on the phonon-electron scattering model.

4.2 Theory

In Chapter 3, we described the procedure to solve the 2D Schrödinger equation self-consistently with the 2D Poisson equation in order to model ballistic transport in nanoscale transistors. The overall procedure is very much the same when we solve these equations in the dissipative transport regime. We briefly review the steps common to both ballistic and dissipative transport model solutions:

1) A 2D solution to the Schrödinger equation is obtained by solving two 1D problems, one in the direction normal to the channel, which yields the vertical electron concentration and subband profiles, and the other, along the channel direction based on the subband profiles yielding 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 corresponding subband wavefunctions.

3) The 2D Poisson equation is solved using this electron density to obtain a new potential profile. This potential profile is used to resolve the 2D Schrödinger equation and the calculation cycle is repeated till self-consistency is achieved.

4.2.1 Büttiker probe based scattering models

We begin our discussion of dissipative transport by describing the two Büttiker probe based scattering models [41]. In the ballistic limit, there are only two reservoirs connected to a device, namely, the source and drain contacts. These contacts couple the nanoscale system to the infinite surroundings and are treated as a perturbation to the intrinsic Hamiltonian of the system. Contacts inject carriers into and extract carriers from the active device region while conserving the current through the device (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 scattering centers in the MOSFET. These Büttiker probes perturb the Hamiltonian of the system in a manner similar to the source and drain contacts and can also be viewed as reservoirs coupled to the system. Each probe (representing an isolated scatterer within the system) interacts with the system through a self-energy. The self-energy of a Büttiker probe is conceptually equivalent to that of the source/drain (see Chapter 3 for the source/drain self-energies). This self-energy of the Büttiker probes introduces incoherencies in the system thus modeling the effect of scattering. To correctly model real scattering phenomena, it is necessary to ensure that Büttiker probes perturb just the carrier energy/momentum and not the total number of carriers in the system. This implies that one can view a Büttiker probe as extracting carriers from the device system, perturbing the energy/momentum of those carriers and reinjecting an equal number back into the system with a different energy/momentum distribution. The interaction between a probe and the system results in a broadening of the local density of states and is characterized by a parameter η (which can be related to a self-energy). This parameter can be related to the dephasing time of a quantum state through the following relation [72],

$$\tau = \frac{\hbar}{2\eta} \tag{4.2}$$

The dephasing time (τ) can be interpreted as the time during which a carrier's (electron in our case) initial state is fully destroyed by a scattering event. Therefore, τ can be mapped onto an equivalent macroscopic mobility through eqn. 4.1.

The Fermi energy characterizes how a reservoir exchanges carriers with the device system. Since Büttiker probes extract and inject carriers into the system, they have an associated Fermi energy that should be adjusted to achieve carrier conservation within the device. Carrier conservation at each scattering center (zero probe current) guarantees current continuity through the transistor.

To implement the Büttiker probe based scattering models, we start with the 1D Hamiltonian in the transmission direction (longitudinal direction) for a particular subband, i. This Hamiltonian can be expressed as (refer to Chapter 3),

$$H_l = \begin{bmatrix} 2t - E_i(1) & -t & \ldots & 0 & 0 \\ -t & 2t - E_i(2) & -t & \ldots & 0 \\ 0 & -t & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & -t \\ 0 & 0 & \ldots & -t & 2t - E_i(N) \end{bmatrix}, \tag{4.3}$$

where $t = \hbar^2 / 2m_{li}^* a^2$, a is the lattice spacing, m_{li}^* the effective mass in the transport direction of electrons in the subband i, N the total number of lattice nodes in the transmission direction and E_i the potential energy for the subband under consideration. To account for the source/drain contacts and all of the Büttiker probes, a self-energy matrix is needed. This matrix is given by,

$$\Sigma = \begin{bmatrix} \Sigma_{source} & 0 & 0 & 0 & 0 \\ 0 & \Sigma_S(1) & 0 & ... & 0 \\ 0 & ... & ... & ... & 0 \\ 0 & ... & ... & \Sigma_S(n) & 0 \\ 0 & 0 & ... & 0 & \Sigma_{drain} \end{bmatrix}. \tag{4.4}$$

where, Σ_{source} and Σ_{drain} are the source and drain contact self energies (calculated as explained in Chapter 3) and $\Sigma_S(1)$, $\Sigma_S(n)$etc., are the self-energies of the Büttiker probes. $\Sigma_S(1)$ represents the self energy of the first probe, while $\Sigma_S(n)$ that of the *nth* probe. The self-energy for a Büttiker probe is related to the broadening parameter η, by the relation, $\Sigma_S = -i\eta$ [72]. Büttiker probes represent isolated scattering events, therefore the self-energy matrix is diagonal and has nonzero entries only at those points along the device where a probe has been introduced. The Green's function for the device in the transmission direction is,

$$G = [E_l I - H_l - \Sigma]^{-1}, \tag{4.5}$$

where E_l is the longitudinal energy. The state spectral functions due to source/drain contacts and all of the Büttiker probes are obtained using,

$$A_n = G\Gamma_n G^+, \tag{4.6}$$

where, $\Gamma_n \equiv i(\Sigma_n - \Sigma_n^+)$, with n running over the source/drain contacts and all of the Büttiker probes. (For clarity, here we use Γ_n to denote a matrix the same size as G, with a nonzero diagonal entry $i(\Sigma_n - \Sigma_n^+)$.) A_n is a matrix, and its diagonal terms constitute local density-of-states,

$$D_n(E_l, m) \equiv (A_n)_{mm} = \sum_{\sigma,\rho}[G_{m\sigma}\Gamma_n\delta_{\sigma n}\delta_{n\rho}G^+_{\rho m}] \; . \tag{4.7}$$

On simplification, eqn. 4.7 reduces to,

$$D_n(E_l, m) = \left|G_{mn}\right|^2 \Gamma_n \tag{4.8}$$

where, D_n is a column matrix representing the local spectral function due to perturbation n. Conceptually, the spectral function is proportional to perturbation strength Γ_n and propagates through the entire domain according to $\left|G_{mn}\right|^2$. Since G_{mn} (with a running index *m*) is the *nth* column of G, one does not need to calculate the entire G (which is computationally expensive) in order to obtain the spectral function. Only those columns corresponding to source/drain contacts or Büttiker probe positions need to be calculated. Therefore, if scattering is assumed occurring in a small portion of the entire device regions (Büttiker probes are placed only at the corresponding portion), this approach can considerably save computational resources. In the case of Büttiker probe at each node, this approach is equivalent to calculating the entire Green's function.

Transmission between any two reservoirs can be calculated in a manner similar to the calculation of the source to drain transmission described in Chapter 3. The transmission between reservoirs *m* and *n* can be expressed as,

$$T_{mn}(E_l) = Trace[\Gamma_m G \Gamma_n G^+] \tag{4.9}$$

Since Γ_m and Γ_n are diagonal matrixes with one entree, T_{mn} can be simplified,

$$T_{mn}(E) = \sum_i [\Gamma_m G \Gamma_n G^+]_{ii} = \sum_{i,\sigma,\rho,\gamma}[\Gamma_m\delta_{im}\delta_{m\rho}G_{\rho\sigma}\Gamma_n\delta_{\sigma n}\delta_{n\gamma}G^+_{\gamma i}]_{ii} \; . \tag{4.10}$$

On simplification, eqn. 4.10 reduces to,

$$T_{mn}(E_l) = \Gamma_m |G_{mn}|^2 \Gamma_n \tag{4.11}$$

Since G_{mn} is just one element of matrix G, one does not need to calculate the entire G in order to obtain the transmission. Also note that in calculating the spectral functions, those columns corresponding to source/drain contacts or Büttiker probe positions have to be calculated, G_{mn} can be obtained from the corresponding column.

Knowing the spectral function, the 2D electron density at node *m* including the effect of all scattering centers as well as the source and drain reservoirs is,

$$n(E_l, m) = \frac{1}{\hbar a}\sqrt{\frac{m_{ti}^* k_B T}{2\pi^3}} \sum_n \Im_{-1/2}(\mu_n - E_l) \cdot D_n(E_l, m) \tag{4.12}$$

where, the summation over *n* represents contributions from all reservoirs, D_n the local density-of-states due to source *n*, and μ_n the Fermi potential of reservoir *n*. The Fermi-Dirac integral of $\Im_{-1/2}$ accounts for all transverse modes contributions (its arguments being normalized to $k_B T$), and the factor in front of the summation sign is a 2D density constant representing the particular subband *i*, with an electron effective mass m_{ti}^* in the transverse direction. (for details see Chapter 3).

The total current at reservoir *m* is given by,

$$I_m(E_l) = \frac{q}{\hbar^2}\sqrt{\frac{m_{ti}^* k_B T}{2\pi^3}} \sum_n [\Im_{-1/2}(\mu_m - E_l) - \Im_{-1/2}(\mu_n - E_l)] \cdot T_{mn} \tag{4.13}$$

where, summation over n accounts for contributions from all sources to reservoir m and T_{mn} the transmission from source m to n. The factor in front of the summation sign is a current density constant representing the subband i.

Equation 4.13 can be rearranged to obtain a compact expression of the form,

$$I_m(E_l) = \frac{q}{\hbar^2}\sqrt{\frac{m_{ti}^* k_B T}{2\pi^3}}[\tilde{T}_m \cdot \Im_{-1/2}(\mu_m - E_l) - \sum_n T_{mn} \cdot \Im_{-1/2}(\mu_n - E_l)]$$

$$= \frac{q}{\hbar^2}\sqrt{\frac{m_{ti}^* k_B T}{2\pi^3}}\sum_n [\tilde{T}_m \delta_{mn} - T_{mn}] \cdot \Im_{-1/2}(\mu_n - E_l)$$

$$= \frac{q}{\hbar^2}\sqrt{\frac{m_{ti}^* k_B T}{2\pi^3}}\sum_n \mathbf{T}_{mn} \cdot \Im_{-1/2}(\mu_n - E_l),$$

(4.14)

where,

$$\tilde{T}_m(E_l) = \sum_n T_{mn}(E_l)\,, \tag{4.14a}$$

$$\mathbf{T}(E_l)_{mn} \equiv \tilde{T}_m(E_l)\delta_{mn} - T_{mn}(E_l)\,. \tag{4.14b}$$

The requirement that the net current at each scatterer m, equals zero i.e.

$$\int_{-\infty}^{+\infty} I_m(E_l) dE_l = 0, \tag{4.15}$$

imposes a set of constraints on the μ_n's (Fermi-potential of scatterers). This set of constraining equations can be solved for the μ_n's using different schemes. The schemes will be described in a later section. Note that this far, we presented the equations for the case of single subband occupancy. To account for contributions from all of the subbands, a summation over all subbands has to be performed.

The important equations are summarized as follows. At longitudinal energy E_l, the local density-of-states including all subbands is

$$\mathbf{D}_n(E_l,m) = \sum_i \left[\frac{1}{\hbar a} \sqrt{\frac{m_{ti}^* k_B T}{2\pi^3}} \cdot D_{ni}(E_l,m) \right], \tag{4.16}$$

where $D_{ni}(E_l,m)$ is the spectral function for subband, i, as given by eqn. 4.8. The net transmission due to all subbands is

$$\mathbf{T}_{mn}(E_l) = \sum_i \left[\frac{q}{\hbar^2} \sqrt{\frac{m_{ti}^* k_B T}{2\pi^3}} \cdot \mathbf{T}_{mni}(E_l) \right], \tag{4.17}$$

where $\mathbf{T}_{mni}(E_l)$ is the transmission between reservoir m and n for subband, i, as given by eqn. 4.14b. The total 2D electron density is

$$n(E_l,m) = \sum_n \Im_{-1/2}(\mu_n - E_l) \cdot \mathbf{D}_n(E_l,m), \tag{4.18a}$$

and the net current at reservoir m is,

$$I_m(E_l) = \sum_n \Im_{-1/2}(\mu_n - E_l) \cdot \mathbf{T}_{mn}. \tag{4.18b}$$

Equation 4.15 has to be satisfied by each scatterer (Büttiker probe) but not necessarily at each energy in order to conserve current. Therefore, the Fermi energy of each Büttiker probe has to be adjusted to ensure that the total probe current is identically zero.

Two different types of Büttiker probe models are explored. The first Büttiker probe model assumes that each scatterer is described by a single, position-dependent Fermi-potential. Therefore electrons from all subbands (including transverse modes) are fully thermalized at each probe according to the corresponding probe Fermi-potential and temperature. The total current at Büttiker probe *m* over the entire longitudinal energy spectrum is $\mathbf{I}_m \equiv \int_{-\infty}^{+\infty} I_m(E)dE$. This total current is a function of the Fermi potential of all Büttiker probes and the source and drain contacts. Newton's method is employed in order to search for the Fermi potentials of all Büttiker probes (μ_n 's) such that the total current at each Büttiker probe is zero. The Jacobian used in Newton's method can be evaluated numerically as

$$A_{mn} = \frac{\partial \mathbf{I}_m}{\partial \mu_n} = \frac{\mathbf{I}_m(\mu_1,..,\mu_n + \Delta\mu_n,...)}{\Delta\mu_n}, \tag{4.19}$$

and, corrections to the μ_n 's during the solution searching iterations are (also refer to eqn. 3.4c in Chapter 3),

$$d\mu_n = \sum_m [-A^{-1}{}_{mn}\mathbf{I}_m]\,. \tag{4.20}$$

Because the Fermi potentials of the source and drain are known, indexes *m* and *n* in eqns. 4.19 and 4.20 run over Büttiker probes only.

In this first Büttiker probe model, electrons are scattered among all transverse modes and subbands based on a unique Fermi-potential of the scatterer at each real-space position. Both the energy and momentum of scattered electrons are fully thermalized, thus approximating the essential physics of scattering. In a real MOSFET, scattering strongly relaxes the transverse energy. In this approximate model, the total energy is relaxed.

The second Büttiker probe method satisfies eqn. 4.15 by constraining the probe current to be zero at each longitudinal energy, $I_m(E_l) = 0$. This is accomplished by allowing each scatterer Fermi-potential to be both position and longitudinal energy dependent. (Of course, this means that the electron distribution is not a Fermi-Dirac distribution. For mathematical convenience only, we describe the distribution by an energy-dependent Fermi-level.) Electrons with specific longitudinal energies from all subbands (including transverse modes) are "thermalized" at each probe according to the corresponding probe Fermi potential. In this model, we need to solve for the Fermi potentials of all Büttiker probes at each longitudinal energy. Newton's method can be still used, but will be very inefficient due to the large number of longitudinal energy steps. A much more efficient procedure to solve the current conservation problem is to directly search for $\Im_{-1/2}(\mu_n - E_l)$. Based on eqn. 4.18b, it is quite obvious that a linear equation can be solved for $\Im_{-1/2}(\mu_n - E_l)$,

$$\sum_n \Im_{-1/2}(\mu_n - E_l)\mathbf{T}_{mn} = I_m(E_l), \quad I_m(E_l) = 0, \text{for any Büttiker probes.} \tag{4.21}$$

Note that this approach of solving a linear equation cannot be used in the first model due to its nonlinearity as a result of integration over longitudinal energy.

In this second scattering model, electrons are scattered among the various transverse modes and subbands based on a unique Fermi-potential of the scatterer at each real-space position and for each longitudinal energy. It appears that the overall longitudinal energy (transport related energy) of scattered electrons is conserved. However their longitudinal momentum could be randomized. It seems that this model captures the phase-breaking process in the transport direction in a MOSFET, but the longitudinal energy relaxation as a result of scattering along the channel is not properly captured. This implies that the model could over-estimate the channel reflection actually occurring in a MOSFET. The solution searching procedure for both models is illustrated using a flow chart.

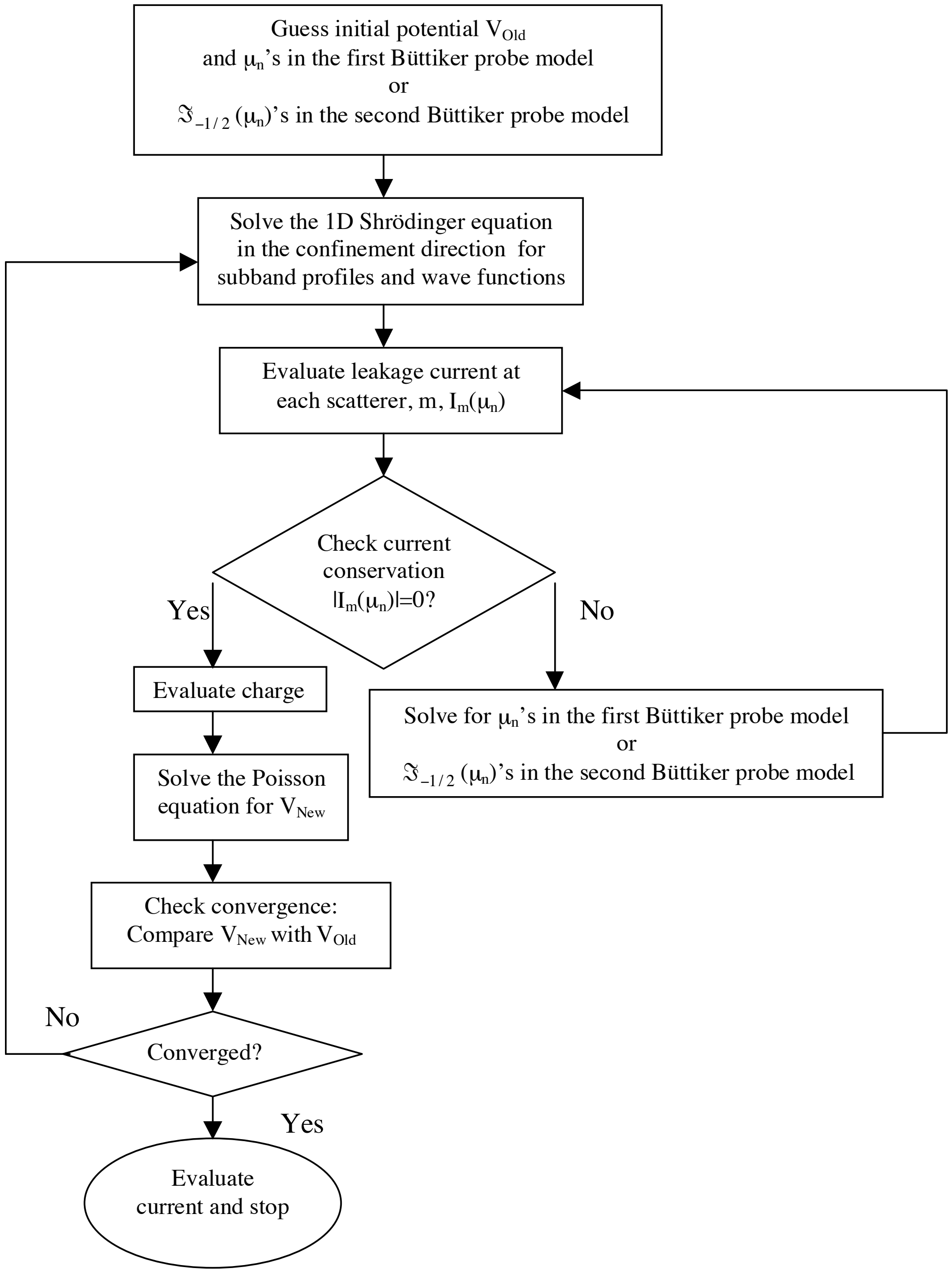

Guess initial potential V_{Old}
and μ_n's in the first Büttiker probe model
or
$\Im_{-1/2}(\mu_n)$'s in the second Büttiker probe model
Solve the 1D Shrödinger equation in the confinement direction for subband profiles and wave functions
Evaluate leakage current at each scatterer, m, $I_m(\mu_n)$
Check current conservation $|I_m(\mu_n)|=0$?
Yes
No
Evaluate charge
Solve for μ_n's in the first Büttiker probe model
or
$\Im_{-1/2}(\mu_n)$'s in the second Büttiker probe model
Solve the Poisson equation for V_{New}
Check convergence: Compare V_{New} with V_{Old}
No
Converged?
Yes
Evaluate current and stop

Fig. 4.1. Flow chart illustrating the solution procedure for the Büttiker probe treatment of scattering.

To this point, we have presented the basic ideas of using the Büttiker probes concept to model dissipative transport. Referring to the general Green's function formalism outlined in Chapter 1, we can see that the phenomenological approach calculates in-scattering and out-scattering functions as

$$\Sigma_n^{in} = \Gamma_n f(\mu_n) \text{ and } \Sigma_n^{out} = \Gamma_n[1 - f(\mu_n)], \tag{4.22}$$

where $f(\mu_n)$ is the Fermi-Dirac occupancy factor for the probe n. The scattering functions and Green's function are solved self-consistently, to achieve current conservation. The simple model does help us understand the dissipative transport, but the treatment lacks physical rigor. For more rigorous calculations, the scattering should be included based on the exact type of interactions and the level of approximation. The scattering function should be solved from the density functions. As an example, we describe the electron-phonon scattering in the self-consistent Born approximation in the next section.

4.2.2 A simple phonon-electron interaction model

We now introduce a scattering model based on the phonon-electron interaction [10, 82, 87-89]. In order to succinctly explain how phonon scattering is treated within the Green's function formalism, we assume single subband occupancy and 1D transport in the longitudinal direction. The method outlined in this section can then be generalized to the multiple subband case including the effect of transverse modes (2D transport). In a manner similar to the Büttiker based models, the phonon-electron interaction introduces a perturbation to the device Hamiltonian through a corresponding self-energy. The Green's function can be expressed as,

$$G(E_l) = [E_l I - H_l - \Sigma]^{-1}, \tag{4.23}$$

where, E_l is the longitudinal energy, H_l the longitudinal Hamiltonian, and $\sum$ the self-energy matrix accounting for the source and drain contacts and all of the phonon-electron interaction centers. In this work, the phonon-electron interaction is assumed to be a local interaction. Therefore $\sum$ is a diagonal matrix as in the cases of the two Büttiker probe models. Invoking the self-consistent first-order Born approximation, the scattering functions for in and out scattering due to phonon-electron interactions can be related to the electron density at energy E_l, $E_l + \hbar\omega_0$ and $E_l - \hbar\omega_0$ as [10, 82, 89],

$$\sum^{in}(E_l) = D_0[(N+1)\cdot G^n(E_l + \hbar\omega_0) + N \cdot G^n(E_l - \hbar\omega_0)], \tag{4.24a}$$

$$\sum^{out}(E_l) = D_0[(N+1)\cdot G^p(E_l - \hbar\omega_0) + N \cdot G^p(E_l + \hbar\omega_0)]. \tag{4.24b}$$

where, D_0 is a constant representing the interaction strength felt by electrons due to one phonon, N is the number of optical phonons of energy $\hbar\omega_0$ in equilibrium with the lattice at a specific temperature, G^n and G^p are the density functions for electrons and holes respectively. The phonon number at any temperature is given by the Bose-Einstein statistics,

$$N = \frac{1}{e^{\hbar\omega_0 / k_B T} - 1}. \tag{4.24c}$$

$\sum^{in}(E_l)$ is related to the electron in-scattering rate as $\sum^{in}(E_l) \equiv \hbar / \tau_{in}$, while $\sum^{out}(E_l)$ is related to the electron out-scattering rate as $\sum^{out}(E_l) \equiv \hbar / \tau_{out}$.

Equation 4.24a can be explained through the pictorial representation of scattering shown in Fig. 4.2a. Electrons can be scattered into empty states at energy E_l from filled states at energy $E_l + \hbar\omega_0$ by the emission of an optical phonon with energy $\hbar\omega_0$. The in-scattering rate for this mechanism is proportional to the interaction strength, the electron density at energy $E_l + \hbar\omega_0$, and the number of phonons plus one (*N*+1) [87]. Electrons with energy $E_l - \hbar\omega_0$ can also be scattered into empty states at energy E_l due to phonon absorption. The in-scattering rate for this mechanism is again proportional to interaction strength, the electron density at energy $E_l - \hbar\omega_0$, and the number of phonons (*N*). The total in-scattering rate $\sum^{in}(E_l)$ is the sum of the two contributions. $\sum^{out}(E_l)$, which is the net out-scattering rate at energy E_l, can be understood in a similar manner through Fig. 4.2b.

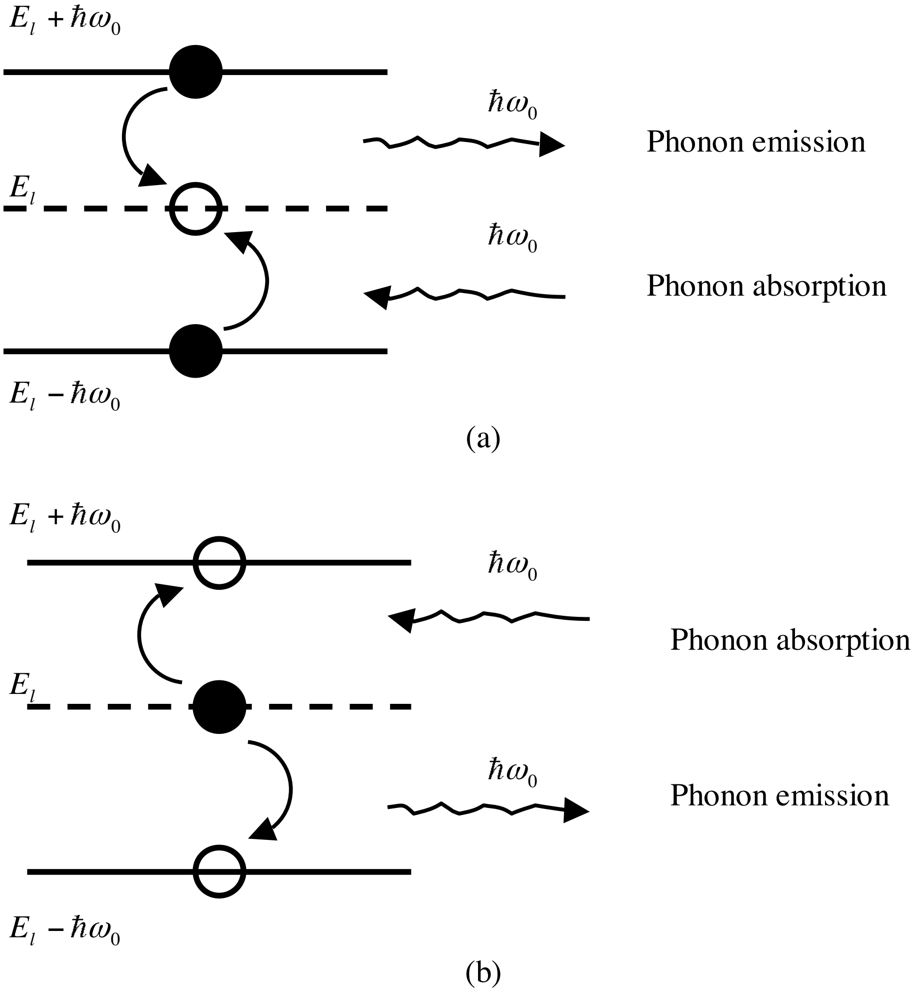


Fig. 4.2. (a) Pictorial diagram illustrating electrons scattered into an empty state at energy E_l. (b) Pictorial diagram illustrating electrons scattered out of a filled state at energy E_l.

Γ determines the net carrier exchange rate between the system and surroundings and is a sum of the in and out scattering functions [10, 82]

$$\Gamma = (\Sigma^{in} + \Sigma^{out}) , \tag{4.25a}$$

Also note that (see Chapter 3),

$$\Gamma \equiv i(\Sigma - \Sigma^{+}) . \tag{4.25b}$$

On simplification, one obtains,

$$\Sigma = -\frac{i}{2}(\Sigma^{in} + \Sigma^{out}) . \tag{4.26}$$

This is the Σ, that is used to compute the Green's function in eqn. 4.23. Note that we have ignored the real part of Σ in eqn. 4.26. Strictly speaking, Σ is not purely imaginary; its real part consists of two components. One of them arises because of the fact that the interaction must be causal in time. The other corresponds to the Fock exchange potential (the Hartree potential is included in the Poisson equation) [10, 89]. These terms are very minor as compared to the Hartree potential and the potential due to impurity. These terms are not directly related to dissipative transport. Therefore they are left out in our calculations.

The electron density function, $G^{n}(E_l)$ is related to the in-scattering function by [40],

$$G^{n}(E_l) = G\Sigma^{in} G^{+} , \tag{4.27a}$$

and the hole density function, $G^{p}(E_l)$ (states unoccupied by electrons) is related to the out-scattering function by

$$G^p(E_l) = G\sum^{out} G^+ . \tag{4.27b}$$

The scattering functions arise from two different sources, namely, the interactions with the leads and the phase-breaking interactions within the device. As shown in eqn. 4.22, one can view the in-scattering function due to any interactions as a product of Γ and a Fermi-Dirac occupancy factor f . For the interactions with the leads, the factor represents the statistics of the real contact reservoir, while for the internal scattering interactions, the factor is just a parameter characterizing the virtual reservoir. Making use of the Fermi-Dirac occupancy factor, we can rewrite eqn. 4.27a as,

$$G^n(E_l) = \sum_m [f_m G\Gamma_m G^+] , \tag{4.27c}$$

with subscript *m* indicating the interaction with source *m*. Expressed in this form, the expression for electron density becomes the same as that calculated for the Büttiker probe models. Referring to eqns. 4.25a and 4.27b, we see that the hole density is,

$$G^p(E_l) = \sum_m [(1 - f_m) G\Gamma_m G^+] \tag{4.27d}$$

as expected.

Using the above identities and definitions, the phonon-electron interaction model is implemented by iteratively solving eqns. 4.23, 4.24 and 4.27. The solution scheme ends when a converged value of $\sum^{in}$ and $\sum^{out}$ is attained. Boundary conditions have to be specified for $\sum^{in}$ or $\sum^{out}$ at the source and drain contacts. Since the two contacts are treated as reservoirs, equilibrium statistics are imposed at the contacts, resulting in the following boundary conditions,

$$\Sigma^{in}_{source} = \Gamma_{source} f(\mu_{source} - E), \tag{4.28a}$$

and

$$\Sigma^{in}_{drain} = \Gamma_{drain} f(\mu_{drain} - E), \tag{4.28b}$$

The Poisson equation is coupled to this transport model in order to obtain self-consistent solutions. Current at the source or drain contacts is evaluated using [40, 82],

$$I_{source} = \frac{q}{h}\int_{0}^{\infty} Trace[\Sigma^{in}_{source} G^{p} - \Sigma^{out}_{source} G^{n}]dE_{l}\,. \tag{4.29a}$$

Note that the spin degeneracy is not included in eqn. 4.29a. Using eqns. 4.27c and 4.27d, one can rewrite the expression for current to have a similar form as that in eqn. 4.13.

In this model, the current at each scattering center automatically becomes zero (current conserved throughout the device). This can be seen by looking at the current flowing into a conceptual reservoir (a scattering center), for example, the reservoir at node *m*,

$$I_{m} = \frac{q}{h}\int_{-\infty}^{+\infty} Trace[\Sigma^{in}_{m} G^{p} - \Sigma^{out}_{m} G^{n}]dE_{l}\,. \tag{4.29b}$$

For simplicity, we can assume that the phonon energy is zero. Referring to eqns. 4.23 and 4.27, we obtain

$$I_{m} \propto \int_{-\infty}^{+\infty} Trace[G^{n}_{m} G^{p} - G^{p}_{m} G^{n}]dE_{l}\,, \tag{4.29c}$$

where G_m^n and G_m^p are electron and hole density at node m. It is very clear that the local phonon-electron interaction implies the trace in eqn. 4.29c being zero at any energy values. So there is no particle and energy exchanges at any virtual scattering centers (analogous to the second Büttiker model). In general cases where phonon energies are nonzero, the trace at each specific energy value may not be zero, but the integration over the entire energy range is still zero. This allows electrons or holes to exchange energies with the semiconductor lattice by emitting or absorbing phonons, but total particle number is conserved (analogous to the first Büttiker model). The solution scheme for phonon-electron interaction model is summarized through a flowchart shown below.

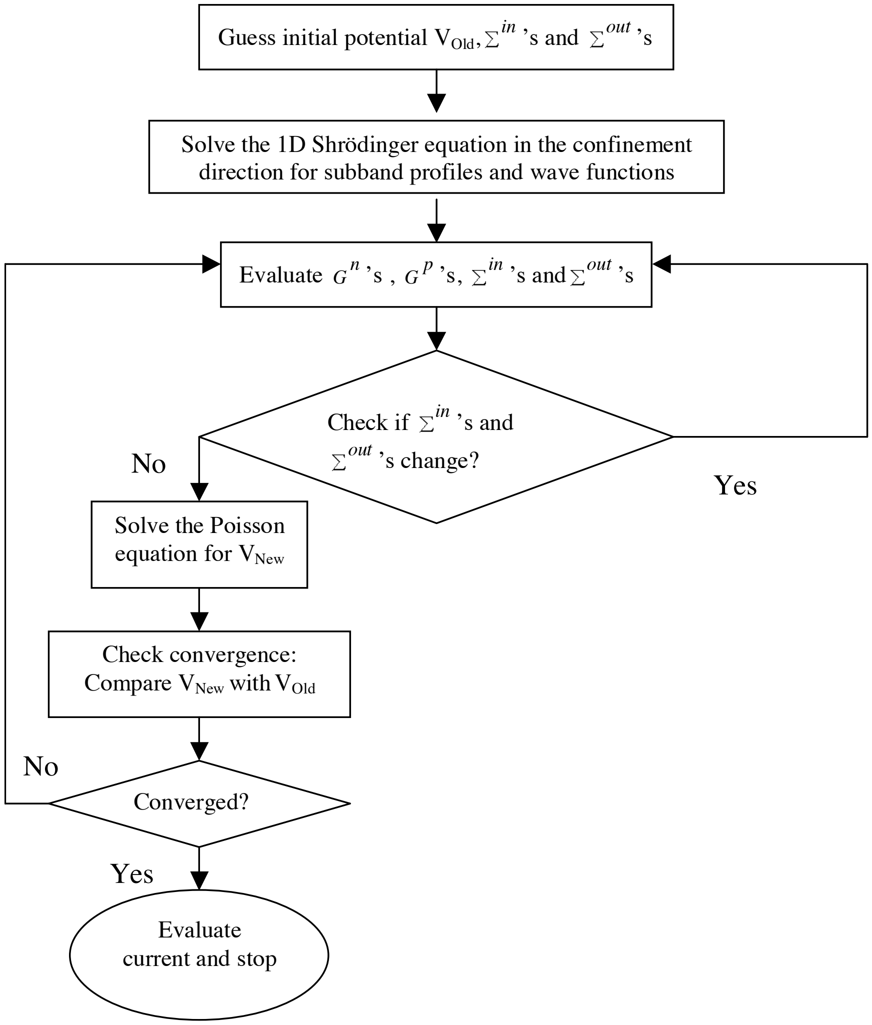

Guess initial potential V_{Old}, Σ^{in}'s and Σ^{out}'s
Solve the 1D Shrödinger equation in the confinement direction for subband profiles and wave functions
Evaluate G^n's , G^p's, Σ^{in}'s and Σ^{out}'s
Check if Σ^{in}'s and Σ^{out}'s change?
No
Yes
Solve the Poisson equation for V_{New}
Check convergence: Compare V_{New} with V_{Old}
No
Converged?
Yes
Evaluate current and stop

Fig. 4.3. Flowchart illustrating the solution scheme for simulating phonon-electron interaction.

This far, our focus has been restricted to a one subband treatment and a 1D transport scheme. A MOSFET is typically very wide. Therefore contributions from a large number of transverse modes have to be included. However, accurately accounting for the transverse modes is computationally very difficult. Therefore we resort to using a very simple treatment of the transverse modes in order to capture the essential physics of phonon scattering while reducing the computational burden. A more strict treatment will be outlined in the discussion section of this chapter.

The transverse mode contributions can be integrated into Fermi-Dirac factors as in the cases of the two Büttiker probe models. For the longitudinal energy, we can define a net occupancy factor at the source and drain contacts as,

$$f(E_l - \mu_{source}) = \frac{\Im_{-1/2}(\mu_{source} - E_l)}{\Im_{-1/2}(\mu_{Max} - E_{Lowest})} \tag{4.30a}$$

$$f(E_l - \mu_{drain}) = \frac{\Im_{-1/2}(\mu_{drain} - E_l)}{\Im_{-1/2}(\mu_{Max} - E_{Lowest})} \tag{4.30b}$$

where, $\frac{2}{\hbar}\sqrt{\frac{m_t^* k_B T}{2\pi}}\Im_{-1/2}(\mu_{source} - E_l)$ represents the total number of transverse modes that contribute to a state with longitudinal energy E_l due to the source contact (μ_{source}) and $\frac{2}{\hbar}\sqrt{\frac{m_t^* k_B T}{2\pi}}\Im_{-1/2}(\mu_{drain} - E_l)$ represents the total number of transverse modes that contribute to a state with longitudinal energy E_l due to the drain contact (μ_{drain}). $\frac{2}{\hbar}\sqrt{\frac{m_t^* k_B T}{2\pi}}\Im_{-1/2}(\mu_{Max} - E_{Lowest})$ represents the maximum number of transverse modes that could contribute to transport in the longitudinal direction. μ_{Max} is the maximum of μ_{source} and μ_{drain}, and limits the number of transverse modes that contribute to transport in the longitudinal direction. E_{Lowest} denotes the lowest longitudinal energy level, below which transmission is practically suppressed. The 1D transport model when coupled with the boundary conditions defined in this section, can be used to calculate all quantities of interest. The final results from the 1D model based on the longitudinal energies is multiplied by the 2D factor $\frac{2}{\hbar}\sqrt{\frac{m_t^* k_B T}{2\pi}}\Im_{-1/2}(\mu_{Max} - E_{Lowest})$, to approximately handle transport through transverse modes (2D transport).

4.3 Results

Up to this point, we have discussed Büttiker probe and phonon scattering models. We have gone through the formalism of each model and detailed their numerical implementation. In this section, we present simulation results from each of the aforementioned models. The simulations are designed to compare the models, and assess them against the scattering theory of MOSFETs. We first present results generated using the Büttiker probe models and examine the difference between energy-relaxing and phasing-breaking scatterings, as modeled by the first and second Büttiker probe models respectively. We then illustrate the effect of the two scattering processes on electron transport in a MOSFET. Following that, we address the issue of uncontrollable tunneling leakage occurring within the Büttiker probe framework. Finally, we present simulation results based on a simplified treatment of phonon-electron scattering. The objectives of this section are, 1) to present a general picture of how different scattering models work, and 2) to illustrate how these models capture the essential features of dissipative transport in a MOSFET. A simplified double gate structure is assumed in all of the simulations.

4.3.1 Description of model device

Most of the simulations are done using a model double-gate SOI MOSFET with a body thickness of 3 nm. In the 3 nm body, multiple subbands are occupied by electrons, so subband-to-subband scattering must be included. For studying subthreshold regime leakage, a 1.5 nm body is used. This is because strong vertical direction quantum confinement in this extremely thin body gives rise to single subband occupation and tunneling current can be assessed without interference from different subbands. In both devices, the channel and source/drain extension lengths are 10 nm with the source and drain regions being doped at 1.0×10^{20} cm^{-3}. P bodies are doped at 1.0×10^{16} cm^{-3}. Both top and bottom gate insulators (SiO$_2$) are 1.5 nm thick. Gate contacts are midgap workfunction metals ($\phi = 4.66$ eV). A uniformly spaced 2D mesh is used in the simulations. The lattice constant for the spatial grid is 2.5 Å.

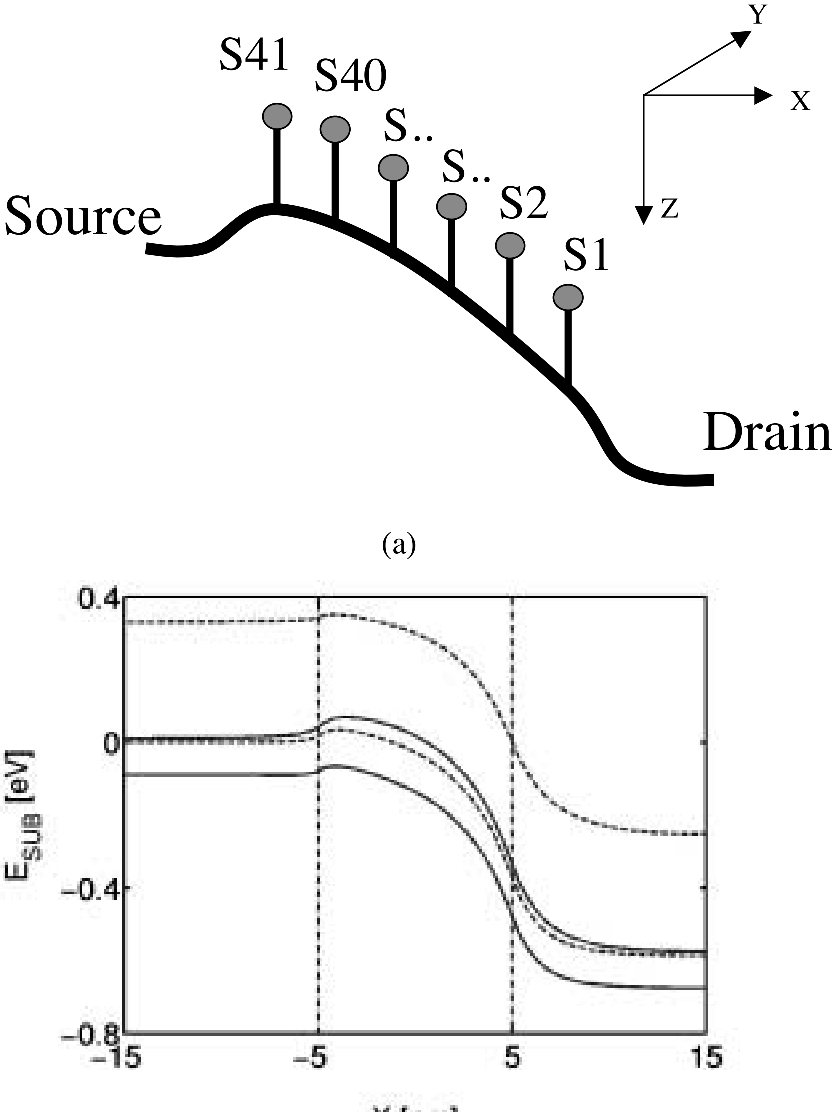


Fig. 4.4. (a) An energy band profile illustrating the placement of Büttiker probes. (b) The profiles of multiple subbands used in the simulations. The channel region is between the vertical dashed lines.

4.3.2 Phase-breaking and energy-relaxing scattering

We first compare the two Büttiker probe models. Figure 4.4a shows that how the probes are placed within the device channel region (note that the study focuses on scattering in channels, so source and drain regions are treated ballistically). We place probes starting from the drain end of the channel, and gradually increase the probe number towards the source end. The probe number within the channel varies between 0 (fully "ballistic") and 41 (fully "dissipative"). The energy broadening parameter (η) characterizing the probe perturbation is set to 30 meV, corresponding to a low field mobility (μ) of 100 cm^2/V-s. The correlation between η and μ has been confirmed by simulating a 1D n$^+$ semiconductor bulk. A conductance is calculated from the *I-V* curve in the low V region and mobility is then obtained from the conductance, assuming a drift-diffusion transport. The mobility is in good agreement with that predicted by eqn. (4.2). Figure 4.4b shows the multiple subband profiles from the source to drain. The profiles are generated by doing a self-consistent ballistic simulation with the 3 nm body device. Bias conditions are $V_{DS} = V_{GS} = 0.6$ V. The solid lines represent unprimed subbands for electrons with heavy effective mass in the gate confinement direction. The dashed lines represent primed subbands for electrons with light effective mass in the gate confinement direction (see also Chapter 2 for more descriptions on the primed and unprimed subbands). Note that the first three subbands shown in Fig. 4.4b have low energies, therefore are primarily occupied by electrons. Also note that these subbands are closely spaced in energy, resulting in scattering between subbands.

Figure 4.5 shows the self-consistently calculated current spectra versus electron longitudinal energy at biases of $V_{DS} = V_{GS} = 0.6$ V. Figures. 4.5a and 4.5b present the results generated with the first Büttiker probe model. In this model scattered electrons are fully shuffled around in both energy and momentum space (energy-relaxing). Figures. 4.5c and 4.5d are results by the second Büttiker probe model, where, scattered electrons are fully shuffled around in momentum space, but their longitudinal energy is conserved (phase-breaking). The second case is more like the transport within a 1D quantum wire.

In the ballistic limit, electrons enter the device from the source and leave through the drain. The number and energy of electrons are both conserved throughout the device. This is shown in Fig. 4.5a, where the current spectra due to the source and drain are symmetric. The source-injected current is positive for electrons entering the device, the drain-collected current is negative for electrons leaving the device. The two peaks observed indicate the contributions from different subbands. In Fig. 4.5b, electron energies are relaxed during the transport. It can be seen that the source-injected current is reduced in magnitude, and the drain-collected current no long mirrors the source current in shape. It can also be seen that electrons leave the device with lower longitudinal energies because they lose their energy as a result of scattering. Figure 4.5c shows the ballistic current spectra for the second Büttiker probe model. This is identical to that shown in Fig. 4.5a, since different dissipative transport models should behave the same in the non-dissipative limit. Figure 4.5d displays current spectra in the presence of phase-breaking scatterings. Electrons may change moving direction when scattered, but will retain their longitudinal energy. Therefore the source and drain current spectra versus *longitudinal energy* are identical in shape and opposite in sign. The current magnitude decreases as compared to the ballistic case because scattered electrons lose their directed momenta as a result of scattering.

Figure 4.6 shows the effect of the scatterer placement on device performance. We use the subband profiles shown in Fig. 4.4b for the simulations. Firstly, Büttiker probes (representing scattering regions) are placed at the drain end of the channel. We can see that the two models give the same current when no scattering is assumed in the device (zero probes). As the scatterer number increases, the first model shows no change in current at the beginning, and a linear decrease after that. This can be understood by looking at Fig. 4.4b. When scatters are placed near the end of the channel, electrons scatter near the drain and lose much of the longitudinal energies, making it difficult for them to get back to the source. Therefore the first Büttiker probe model predicts that scattering near the drain end does not reduce the source injected current by an appreciable amount. When scatterers are placed very close to the source injection point, scattering in this region can easily cause electrons to reflect back into the source before they dissipate too much longitudinal energy. As a result, increasing the number of scatterers towards the source injection end will dramatically decrease the current being collected at the drain.

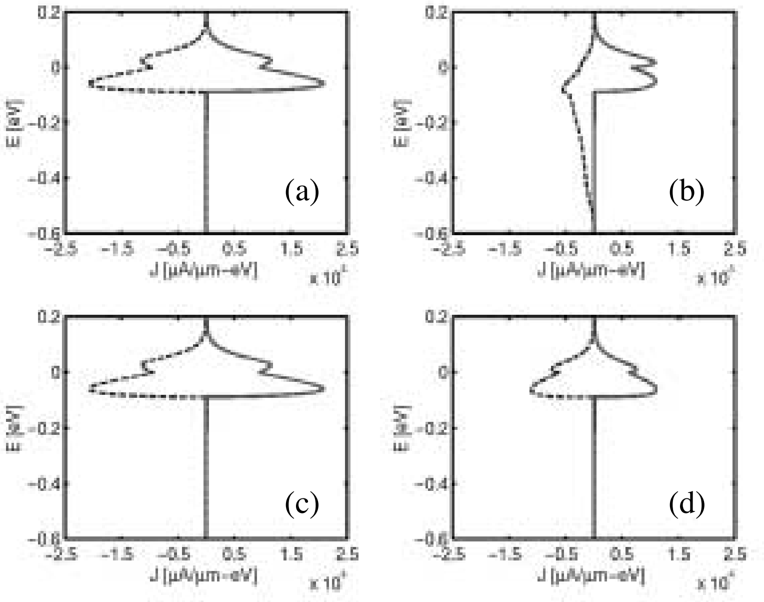


Fig. 4.5 The red lines stand for the source-injected current, the blue lines stand for the drain-collected current. The source Fermi level is zero on the energy scale.

(a) Ballistic current spectra (using the first Büttiker probe model).
(b) Current spectra assuming scatterers placed throughout the entire channel region (using the first Büttiker probe model).
(c) Ballistic current spectra (using the second Büttiker probe model).
(d) Current spectra assuming scatterers placed throughout the entire channel region (using the second Büttiker probe model).

The second Büttiker probe model shows distinctly different results. The transmitted current goes down almost linearly along with the number of scatterers placed in the channel, regardless of their position. In this model, scattered electrons can change transport direction, without dissipating their longitudinal energy. Therefore scattering anywhere in the channel can cause electrons to be reflected back into the source. We also note that, if only one subband is included in the simulation, the dependence is close to perfectly linear, while, with multiple subbands are included, subband-to-subband scattering makes the dependence tend towards the one predicted by the first model. This can be explained by referring to subband profiles shown in Fig. 4.4b. Near the drain, some electrons may transfer from lower energy subbands to higher energy subbands without losing their longitudinal energy. This causes a large amount of kinetic related longitudinal energy to be transferred to potential energy thereby reducing the probability of such electrons back scattering to the source. When the two Büttiker probe models are assessed against Lundstrom's scattering theory of MOSFETs, the first model shows very good agreement. This model indicates that a small region near the source end is critical to electron transport in a MOSFET and that the length of this critical region can be much shorter than the actual channel length [46, 55]. This phenomenological approach although simple physically, does seem capable of simulating dissipative transport in MOSFETs operating in the on-state. The second model, when compared to the first, seems improper in its conserving the longitudinal energy, therefore is not a good model for simulations of dissipative transport in MOSFETs.

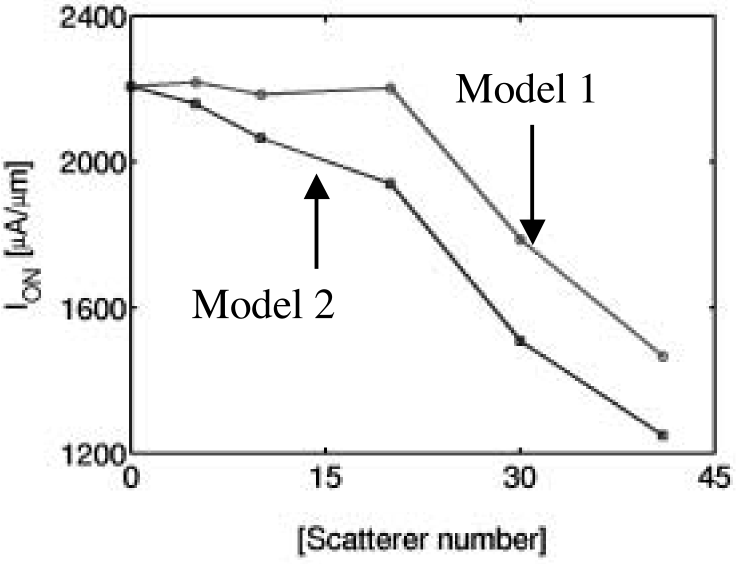


Fig. 4.6. On-current dependence on the number of scatterers as placed continuously from the drain end of the channel to the source end of the channel.

4.3.3 Unphysically high leakage current

In Section 4.3.2, we examined the two Büttiker probe models by looking at their predicted on-state current. In this section, our focus is on simulating off-state leakage. The Büttiker probe models are found to predict unphysically high tunneling leakage in the off-state. Fig. 4.7 shows simulation results using the first Büttiker probe model (red dashed line with symbols) and compares them to the pure ballistic results (blue solid line). The ballistic transport model is physically rigorous and can be used as a basis for comparison. Scattering influences the drift current (on-state), but not the diffusion current (off-state). The large tunneling leakage is related to the manner in which we implement the Büttiker probe model. When the probe perturbation strength (η) is assumed constant (energy independent), a large transmission even in the classically forbidden regions (below the channel barrier peak) is observed.

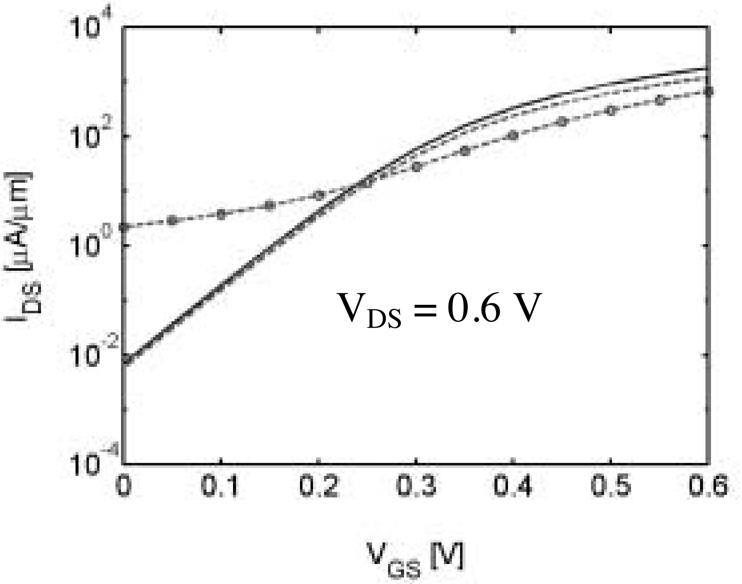


Fig. 4.7. The blue solid line indicates the ballistic simulation result. The red dashed line indicates the result using the first Büttiker probe model assuming truncated self-energy in the classical forbidden regions. The red dashed line with symbols indicates the result using the first Büttiker probe model assuming constant self-energy.

Note that scattering will introduce spurious energy states in the forbidden regions, therefore may increase the tunneling current. Use of constant perturbation strength in the

Büttiker probe model, however, causes too much state broadening, which gives rise to unreasonably high source-barrier tunneling (this unwanted tunneling current is a negligible part of the on-state current) as shown in Fig. 4.8. Fig. 4.8a plots the off-state conduction band profile and Fig. 4.8b shows the current spectra obtained using the Büttiker probe model. It is clear that the off-state current is overwhelmingly dominated by the tunneling component (spectra below the channel barrier peak). The ballistic current spectra (shown in Fig. 4.8c) will become invisible if the same current scale is used in plotting the same.

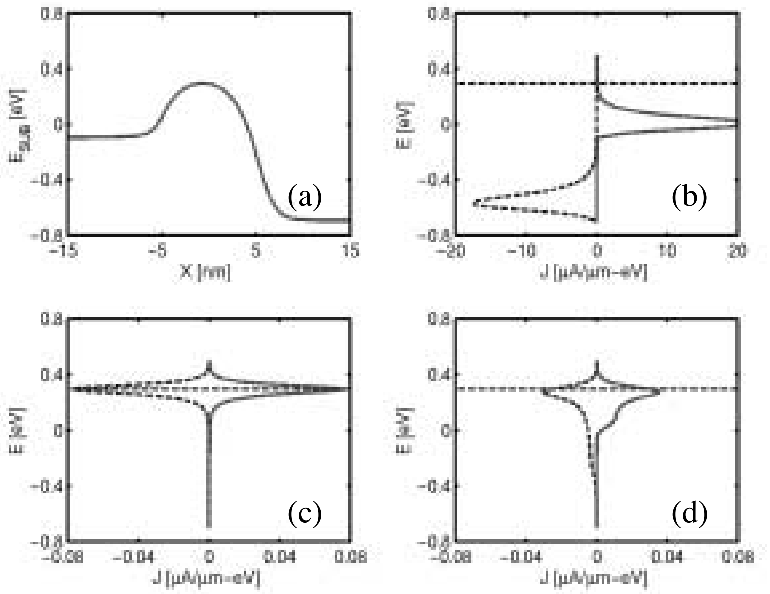


Fig. 4.8. The dashed line indicates the channel barrier height, the red lines indicate the source-injected current, the blue lines indicate the drain-collected current. The first Büttiker probe model is used in the simulations. Also note that the source Fermi level is zero on the energy scale.

(a) Off-state subband profile of a MOSFET, $V_{DS} = 0.6$ V, $V_{GS} = 0.6$ V.
(b) Current spectra showing the abnormally large tunneling current (assuming constant self-energy).
(c) Ballistic current spectra.
(d) Current spectra showing the reduced tunneling current (assuming truncated self-energy in the classical forbidden regions).

The issue can be resolved through the use of an energy dependent η. From the standpoint of scattering physics, the perturbation strength should diminish as the particle density involved in the interactions diminishes. This is automatically achieved within the phonon-electron scattering model where the interactions are treated strictly through the physics (results to be shown in the next section). Within the Büttiker probe models, the same physics also need to be retained. This can be accomplished by forcing the interactions to be off when the amount of electrons that could participate in the interactions is suppressed by the subband energy (*e.g.* in the regions far below the subband energy edge). We assume that

$$\eta(E) = \frac{\eta_0}{1 + e^{(E_{SUB} - E)/k_B T}}, \tag{4.31}$$

where E_{SUB} is the local subband edge (This treatment has no exact physics behind, the only goal is to cut off η--the interaction coefficient below the conducting subband). Equation 4.31 makes η go smoothly from a constant η_0 (when beyond the band edge) to zero (when deeply below the band edge) [91-92]. By doing this, the unphysical tunneling current can be removed. Results using this model for the probe energy are shown in Fig. 4.7 (the red dashed line). The off-state current spectra are also presented in Fig. 4.8d. The modified η eliminates the unwanted states in the barrier region, making the current comparable to that obtained from ballistic simulations. Although the abnormal off-state current can be rectified in this way, the treatment is too arbitrary. The Büttiker models may not be capable of predicting accurate off-state performance of MOSFETs. But as off-state characteristics are mostly determined by electrostatics and not by the transport. Green's function ballistic models should provide a fairly accurate picture of the subthreshold features of a transistor.

4.3.4 Phonon-electron interaction

Finally, we simulate phonon-electron interaction within the simplified 1D framework (see Section 4.2.2). In these simulations, the phonon-electron interaction is treated using the self-consistent first Born approximation, which means that only one-phonon scattering is included, but it is included exactly (to all orders in the language of perturbation theory). Note that in this model, the interaction strength (or scattering function) is calculated self-consistently from the density matrix. This can be understood by referring to eqns. 4.24 and 4.27. The scattering rates are proportional to the densities of the particle being scattered (see eqn. 4.24), and the interactions broaden the density matrix at a specific energy level (see eqn. 4.27). The interaction strength (D_0) felt by electrons due to one phonon is set to $10\times(\hbar\omega_0)^2$ (see Datta or paper by Ando and Flores for more on D_0 [10, 38, 67, 88]). The self-consistent calculation loop guarantees that the interaction strength goes to zero as the state density goes to zero. In Fig. 4.9, we show the simulated I_{GS} vs. V_{GS} at high V_{DS}. In contrast to the results from the Büttiker models, the phonon-electron scattering model predicts normal subthrehold characteristics. The straight dashed line in the figure indicates the ideal subthreshold swing of 60 mV/dec at room temperature.

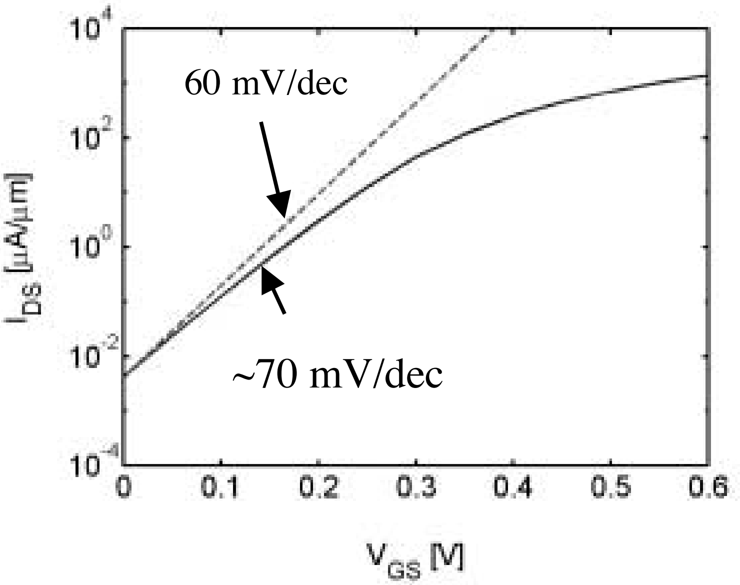


Fig. 4.9. The phonon-electron interaction model predicted I_{DS} versus V_{GS} on semilog scale (blue solid line, V_{DS} = 0.6 V). The red dashed line shows the ideal subthreshold swing.

In Fig. 4.10, we present the results showing the effect of phonon-electron scattering on on-state transport in a MOSFET. Multiple subbands are included in the simulations (Fig. 4.10a). The dependence of I_{ON} versus the number of phonon-electron scattering centers placed within the device is plotted in Fig. 4.10b. Optical phonon scattering in MOSFETs is very weak at room temperature as the phonon energy of ~ 30 mV. This can be seen clearly from eqn. 4.24. The scattering rates depend on the phonon numbers (N or N+1, depending on either the absorption process or emission process), which are very small in the above case. Therefore, I_{ON} may not be strongly degraded even though in the case that phonon-electron scattering is assumed occurring throughout the entire channel region. Figure 4.10c shows the current spectra when no phonon-electron scattering assumed in the simulation. As expected, the results go back to the spectra predicted by the ballistic model. Figure 4.10d shows the current spectra when phonon-electron scattering is included occurring throughout the entire channel. In this case, the energy spectrum of the current shifts downwards as we go from the source contact to the drain contact. Electrons emit phonons during the scattering (for simplicity, only the phonon-emission process is assumed in the simulation), therefore their energies are relaxed as indicated.

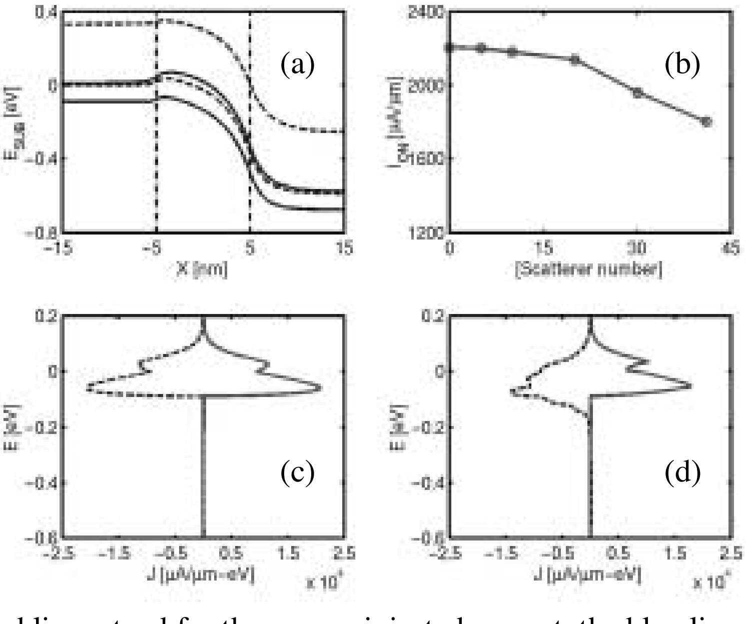


Fig. 4.10. The red lines stand for the source-injected current, the blue lines stand for the drain-collected current. The phonon-electron interaction model is used in the simulations. The source Fermi level is zero on the energy scale.

(a) Subband profile used in the simulations.
(b) Dependence of on-current on the number of scattering centers as placed continuously from the drain end of the channel to the source end of the channel.
(c) Current spectra when no phonon-electron scattering assumed in the channel.
(d) Current spectra when scattering assumed throughout the channel.

This simplified model does provide us useful insights into phonon-electron interactions in a MOSFET: 1) the treatment is strict in theory (in the 1D sense), so this model will not predict abnormal off-state characteristics, 2) optical phonon-electron scattering is weak at room temperature, so the model can not generally be used to simulate dissipative transport in on-state of MOSFETs.

4.4 Discussion on Phonon-electron Scattering

Phonon-electron scattering has been studied essentially within a 1D framework, focusing on the longitudinal component of electron transport. The transverse modes are treated based on their average contributions to overall scattering, which are included in Fermi-Dirac integrals. A rigorous treatment is very important in characterizing 2D transport in MOSFETs. In the section below, we present an approach explicitly accounting for transverse mode contributions to channel transport. The discussion is primarily centered on the theoretical formulation. Numerical implementation of this approach requires an extremely large computational capability, and has not been attempted in this work.

In order to account for effect of transverse mode scattering on electron transport, the scattering rates have to be expressed in terms of both transverse and longitudinal energies of electrons. A typical SOI MOSFET has a thin but very wide body, and electrons can be modeled as a 2D gas. Because of the large width, the system shows translational invariance along the width direction. These analyses lead us to assume that, 1) the scattering rates (τ^{in}, τ^{out}) or scattering functions ($\sum^{in}, \sum^{out}$) should be expressed in terms of total-energy of scattered electrons, $e.g. E_l + E_t$ (instead of the longitudinal energy, E_l as has been done in the simple treatment, here E_t represents the transverse mode energy);

2) the scattering rates should be functions of the longitudinal position x but not transverse position y due to the translational invariance (Note that the second assumption implies plane waves are the transverse eigenfunctions of the Hamiltonian even in presence of the scattering). This treatment in spirit is similar to that used by Lake and Datta in their work of simulating resonant-tunneling diodes [82].

To compute the total energy dependent $\Sigma^{in}(E)$ and $\Sigma^{out}(E)$, we have to start with the total Hamiltonian of the device. The Hamiltonian is expanded using plane wave functions $\exp(ik_j y)/\sqrt{W}$ in the transverse direction, using position delta functions $\delta(x-x')$ in the longitudinal direction, and using mode space eigenfunctions $\psi_i(x,z)$ in the gate confinement direction (see also Chapter 1). The expansion gives rise to

$$H = \sum_{i,k_j} H[E_i(x), E_{k_j}] \ , \tag{4.32}$$

where $E_i(x)$ is the energy profile for subband i, and $H[E_i(x), E_{k_j}]$ represents a longitudinal Hamiltonian for the subband i with a transverse energy $E_t = E_{k_j}$, given as

$$H[E_i(x), E_{k_j}] = \begin{bmatrix} 2t - E_i(1) + E_{k_j} & -t & \ldots & 0 & 0 \\ -t & 2t - E_i(2) + E_{k_j} & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & -t \\ 0 & 0 & \ldots & -t & 2t - E_i(N) + E_{k_j} \end{bmatrix}. \tag{4.33}$$

Once the Hamiltonian is obtained, the scattering functions for in and out scattering due to phonon-electron interactions can be calculated following the procedure discussed in Section 4.2.2. To keep the discussion complete, the key equations are again listed here. The scattering functions due to phonon emission and absorption are given as

$$\sum^{in}(E) = D_0[(N+1)\cdot G^n(E+\hbar\omega_0) + N\cdot G^n(E-\hbar\omega_0)]\,, \tag{4.34a}$$

$$\sum^{out}(E) = D_0[(N+1)\cdot G^p(E-\hbar\omega_0) + N\cdot G^p(E+\hbar\omega_0)]\,. \tag{4.34b}$$

$\sum^{in}(E) = \hbar/\tau_{in}$ is the net electron in-scattering rate at total energy E. For phonon emission, the rate is proportional to the 2D electron density at energy $E+\hbar\omega_0$, for phonon absorption, it is proportional to the 2D electron density at energy $E-\hbar\omega_0$. A similar analysis holds for the electron out-scattering rate $\sum^{out}(E) = \hbar/\tau_{out}$. (Referring to Section 4.2.2 for detailed explanations of other terms). Note that the scattering function computed here are functions of total energy (E_l + E_t) of scattering electrons.

Obtaining the 2D electron density function, G^n, takes a little more effort. The calculations need to include contributions of all transverse modes, which are not explicitly accounted for in the previous 1D treatment. Similarly, G^p also needs to be evaluated for the out scattering rates (note that G^p represents the states that are not occupied by electrons). The procedure can be illustrated by focusing on a single energy, E. Analogous to eqn. 4.27a, the 2D electron density spectrum is expressed as

$$G^n(E) = \int_0^\infty [\frac{2}{\pi\hbar}\sqrt{\frac{m_y^*}{2E_{k_j}}} G(E,E_{k_j})\sum^{in}(E)G^+(E,E_{k_j})]dE_{k_j}\,, \tag{4.35a}$$

and

$$G^p(E) = \int_0^\infty [\frac{2}{\pi\hbar}\sqrt{\frac{m_y^*}{2E_{k_j}}} G(E,E_{k_j})\sum^{out}(E)G^+(E,E_{k_j})]dE_{k_j}\,, \tag{4.35b}$$

where $G(E,E_{k_j})$ is the Green's function as calculated from eqn. 4.33, namely

$$G(E,E_{k_j}) = [EI - H[E_i(x) - E_{k_j}] - \Sigma(E)]^{-1} = [E_l I - H[E_i(x)] - \Sigma(E)]^{-1}\,. \qquad (4.36)$$

It should be noticed that $E_l \equiv E - E_{k_j}$ has been used in obtaining eqn. 4.36, and $H[E_i(x)]$ is the same as eqn. 4.3. The integration is done to include contributions from transverse modes, and $\frac{2}{\pi\hbar}\sqrt{\frac{m_y^*}{2E_{k_j}}}$ represents the state density related to transverse modes. $\Sigma^{in}(E)$ and $\Sigma^{out}(E)$ need to be self-consistently computed along with $G^n(E)$ and $G^p(E)$. In numerical simulation practice, integration limits in eqns. 4.35a and 4.35b can be reduced to a finite value, namely $E_{k_j}(\max)$ which ensures the longitudinal direction state density, $G(E,E_{k_j})\cdot\Sigma^{in}(E)\cdot G(E,E_{k_j})$ negligible when E_{k_j} is greater than $E_{k_j}(\max)$.

In order to complete the self-consistent iteration, $\Sigma(E)$ in eqn. 4.36 has to be expressed in terms of known quantities, and boundary constrains also need to be specified. The self-energy $\Sigma(E)$ arises from two different sources. One is due to the contacts,

$$\Sigma_C(E) = -te^{ik_l a}\,, \text{ where } E = E_{k_j} + E_C + 2t(1-\cos k_l a)\,. \qquad (4.37a)$$

where E_C is the subband energy at the corresponding contact end, and the other is due to internal phonon-electron interactions,

$$\Sigma_S(E) = -\frac{i}{2}[\Sigma_S^{in}(E) + \Sigma_S^{out}(E)]\,. \qquad (4.37b)$$

The boundary constraints require that equilibrium statistics prevail at the contacts, namely, the source and drain. Referring to eqn. 4.23, we can express the in-scattering function at the contact as

$$\Sigma_C^{in}(E) = f(\mu_C - E)\Gamma_C(E, E_{k_j}), \qquad (4.38a)$$

where, $f(\mu_C - E)$ is the Fermi-Dirac statistics factor given as

$$f(\mu_C - E) = \frac{1}{1 + e^{(E-\mu_C)/k_B T}}, \qquad (4.38b)$$

and $\Gamma_C = -2\,\mathrm{Im}(\Sigma_C)$. Note that equilibrium statistics is reflected in $f(E)$, and the energy involved in $f(\mu_C - E)$ is the total energy as opposed to the longitudinal energy. To determine the out-scattering function at the contact, we begin with the electron density function, $G^n = fG\Gamma G^+$, we also note that the hole density function is $G^p = A - G^n$, where $A = G\Gamma G^+$ is the total density spectrum function. So Σ_C^{out} at the contact can be expressed as

$$\Sigma_C^{out}(E) = (1 - f)\Gamma_C(E, E_{k_j}). \qquad (4.38c)$$

Up to this point, the self-consistent loop is formed. The iteration process starts with an initial guess of $\Sigma_S^{in}(E)$ and $\Sigma_S^{out}(E)$, then makes use of eqn. 4.37 to obtain $\Sigma(E)$. Provided $\Sigma(E)$, equations 4.35 and 4.36 are used to get $G^n(E)$ and $G^n(E)$. Finally $\Sigma_S^{in}(E)$ and $\Sigma_S^{out}(E)$ are recalculated through eqn. 4.34, until the updates to the self-energies decrease below the specified convergence criteria. The terminal current is evaluated as

$$I_C = \frac{q}{h}\int_{-\infty}^{+\infty}\left[\int_0^{\infty}\frac{2}{\pi\hbar}\sqrt{\frac{m_y^*}{2E_{k_j}}}Trace[\Sigma_C^{in}(E,E_{k_j})G^p(E,E_{k_j}) - \Sigma_C^{out}(E,E_{k_j})G^n(E,E_{k_j})]dE_{k_j}\right]dE \qquad (4.39)$$

once convergence is achieved. If more than one subband is included in the simulation, a summation over all involved subbands should be accomplished to get the total electron density and current.

In principle, this process is just an extended and rigorous version of the 1D model. But to implement the 2D model involves several difficulties that need to be addressed. First, the computational requirements increase hundreds of times. We are now dealing with the total-energy of electrons. Although the energy range of interest is practically limited by the subband energy at the lower end, and Fermi energy at the higher end, for each total-energy E, evaluation of Hamiltonian needs to be repeated for hundreds of transverse energies (E_{k_j}). The integrations have to be numerically accomplished by doing summations over all E_{k_j}'s involved. In contrast, in the 1D model, the Hamiltonian is calculated once for each E_l, and integrations over the transverse modes can be obtained analytically. Therefore, to efficiently simulate 2D scatterings, powerful computational resources are needed.

Second, singularities exist in the integrations of eqns. 4.35a and 4.35b. The singularity is due to the 1D state density corresponding to plane-wave states in the device width direction. The singularity poses big threats to the self-consistent calculations because the singular energy point cannot be traced during the iterations. Special treatment is needed for accurate and timely convergences.

It is worthwhile to point out that the phonon-electron scattering model can be generalized to account for the phase-breaking elastic scattering process. In such a process, electrons are shuffled around in momentum space, but intact in energy space. So this process can be simulated with phonon-electron interactions assuming zero phonon energy. The interaction strength coefficient D_0 becomes a tunable quantity for desired scattering intensities. The generalized model is analogous to the first Büttiker probe model. But this model is more rigorous in theory. It should be able to avoid the uncontrollable tunneling leakage in the subthrehold regime of an operating MOSFET, and it should also be able to capture transverse mode scattering properly.

4.5 Summary

In Chapter 4, we described dissipative transport in MOSFETs within the Green's function framework. We numerically implemented two Büttiker probe models and assessed the models against the MOSFET scattering theory. We also presented a simplified 1D phonon-electron scattering model. Furthermore, we theoretically derived and discussed the rigorous 2D treatment of phonon-electron interactions. We noted that large computational resources are required for precise 2D simulations of scattering. We also noted that the first Büttiker probe model captures the physics of dissipative transport in MOSFETs, and this model is computationally manageable at a PC level. Therefore the first Büttiker model may be a useful simulation technique to model dissipative transport in small MOSFETs.

5. ESSENTIAL PHYSICS OF CARRIER TRANSPORT IN NANOSCALE MOSFETS

5.1 Introduction

Scaling MOSFET's to their limits is a key challenge now faced by the semiconductor industry. Physically detailed simulations which capture the off-equilibrium transport (e.g. velocity overshoot) [93-95] and the quantum mechanical effects that occur in these devices [80] can complement experimental work in addressing these challenges. Also needed, however, is a simple conceptual view of the nanoscale transistor — to help interpret detailed simulations and experiments and to guide experimental work. Such a model has recently been outlined [55, 96]. Our objective in this chapter is to assess and discuss this basic view through the use of numerical simulations. As a vehicle for these studies, we use a model 10 nm double-gate MOSFET, but we expect the conclusions to apply to nanoscale MOSFET's more generally. We use a semiclassical approach, because recent work shows that MOSFET's operate essentially classically down to channel lengths of about 10 nm [61-62]. We also restrict our attention to the steady-state current vs. voltage characteristics, which are relevant to the high-speed operation of digital circuits [15].

Figure 5.1 summarizes the essential physical picture that will be discussed in this chapter. We adopt a transmission view of the device [10, 57] in which carriers are injected into the channel from a thermal equilibrium reservoir (the source), across a potential energy barrier whose height is modulated by the gate voltage, into the channel, which is defined to begin at the top of the barrier. The beginning of the channel is populated by carriers injected from the thermal equilibrium source (and, under low drain bias, from the thermal equilibrium drain as well). The density of carriers at the top of the barrier is controlled by MOS electrostatics so that the charge in the semiconductor balances that in the gate. Under equilibrium conditions (V_{DS} = 0 V) in an electrostatically well-tempered device, equilibrium, 1D MOS electrostatics apply at this point, so the inversion layer density can be computed as for a 1D MOS capacitor. Above threshold,

$$Q_{inv}(0) = qn_S(0) \approx C_{EFF}(V_{GS} - V_{TH}), \tag{5.1}$$

where C_{EFF} is the effective oxide capacitance (as influenced by quantum mechanical confinement, polysilicon depletion, etc. [15]). We will show that a type of "gradual channel approximation" applies at this point, so that the inversion layer density at the source end of the channel remains nearly equal to its equilibrium value even when a drain bias is applied.

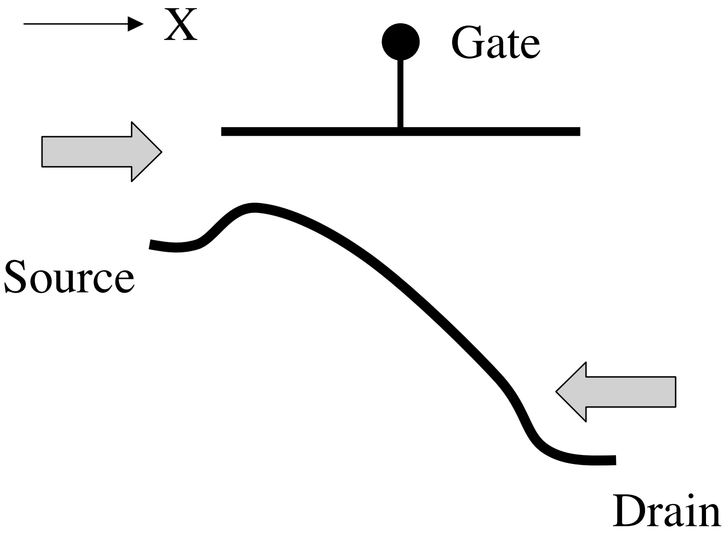


Fig. 5.1. The conduction subband edge versus position from the source to the drain of a nanoscale MOSFET under high gate and drain bias. Also shown are the thermal injection fluxes from the source and drain.

Some fraction of the carriers injected from the source into the channel backscatter and return to the source, others flow out the drain and comprise the steady-state drain current, I_{DS}. (For a high drain bias, carriers injected from the drain do not need to be considered.)

Assuming current continuity, I_{DS} may be evaluated at the beginning of the channel where the carrier density is known from MOS electrostatics to find

$$I_{DS} = WQ_{inv}(0) < \upsilon(0) > \approx WC_{OX}(V_{GS} - V_{TH}) < \upsilon(0) > , \tag{5.2}$$

where $< \upsilon(0) >$ is the average velocity of carriers at the beginning of the channel. The maximum value of $< \upsilon(0) >$ is approximately the equilibrium uni-directional thermal velocity, $\tilde{\upsilon}_T$, because the positive velocity carriers at the beginning of the channel were injected from the thermal equilibrium source [55]. Backscattering from the channel determines how close to this upper limit the device operates. Under high drain bias, the average velocity at the beginning of the channel can be related to a channel backscattering coefficient, *r*, according to [55]

$$< \upsilon(0) > \approx \left(\frac{1-r}{1+r}\right)\tilde{\upsilon}_T , \tag{5.3}$$

where $0 < r < 1$ is a backscattering coefficient in the spirit of McKelvey [103-104]. (Note that when eqn. 5.3 is inserted into eqn. 5.2, we get a result presented earlier [55]. Also note that the backscattering coefficient, *r*, depends on the scattering physics and on the self-consistent potential within the channel, so *r* is a function of the gate and drain biases.) The importance of the source velocity is, of cause, well-known [107], we relate it to a channel backscattering coefficient to clarify the source velocity limit.

Because of the high electric field and strong velocity overshoot, carrier transport through the drain end of the channel is rapid. As a result, the D.C. current is controlled by how rapidly carriers are transported across a short low-field region near the beginning of the channel. Carriers diffuse across the beginning of the channel in much the same way that they diffuse across the base of a bipolar transistor, and they are collected by the high-field portion of the channel much as in the collector of a bipolar transistor [108]. We refer to the critical, low field region near the beginning of the channel as the "kT-layer" because it is roughly the distance over which the channel potential drops by k_BT/q. Scattering within the kT-layer limits the steady-state drain current; scattering near the drain end of the channel has only an indirect effect. This is analogous to the well-known Bethe condition for thermionic emission in a forward-biased metal-semiconductor diode [109], except that in a MOSFET the flow of carriers is down the potential barrier rather than up. For well-designed MOSFET's, the length of the kT-layer (which is set by 2D electrostatics as influenced by velocity overshoot within the channel [106]) is about one-mean-free path, which means that transport across the kT-layer is quasi-ballistic.

In the following sections, we use detailed, numerical simulations to confirm this basic physical picture and to expand upon it. Note that in presenting the basic, physical picture, we have made several simplifying assumptions. For example, we assumed high drain bias, although a full range expression can be developed [90]. We also assumed non-degenerate carrier statistics; degeneracy increases the average thermal velocity, causes the average velocities of the positive and negative halves of the distribution at the top of the barrier to differ, and influences the length of the critical region (i.e. the criterion of a k_BT/q potential drop must be generalized for degenerate statistics). Some of these issues will be discussed further in this chapter, but our intent is to present the basic, physical picture in simple form, so a full discussion must be deferred to later publications. The following specific issues will be addressed in this chapter:

1) injection velocity limits at the source end of the channel
2) the off-equilibrium distribution function at the source
3) charge control in a nanoscale MOSFET
4) the role of scattering and the generalized Bethe Condition for a MOSFET
5) the role of velocity overshoot in the channel
6) the magnitude of the quantum contact resistance in nanoscale MOSFETs

To examine these effects, we numerically simulated the simple, model MOSFET shown in Fig. 5.2. The device is a double-gate (DG) MOSFET with an exceptionally thin (1.5 nm)

Si body, a 1.5 nm SiO_2 gate oxide, and an L_G = 10 nm. A hypothetical mid-gap workfunction gate material was assumed. The device is assumed to be wide in the z-direction (out of the page), so that many transverse modes are occupied. Also note that the idealized metal contacts in Fig. 5.2 represent the actual contacts where dissipative scattering would dominate and maintain a thermal equilibrium carrier distribution. (Real contacts would also flare out to reduce series resistance.)

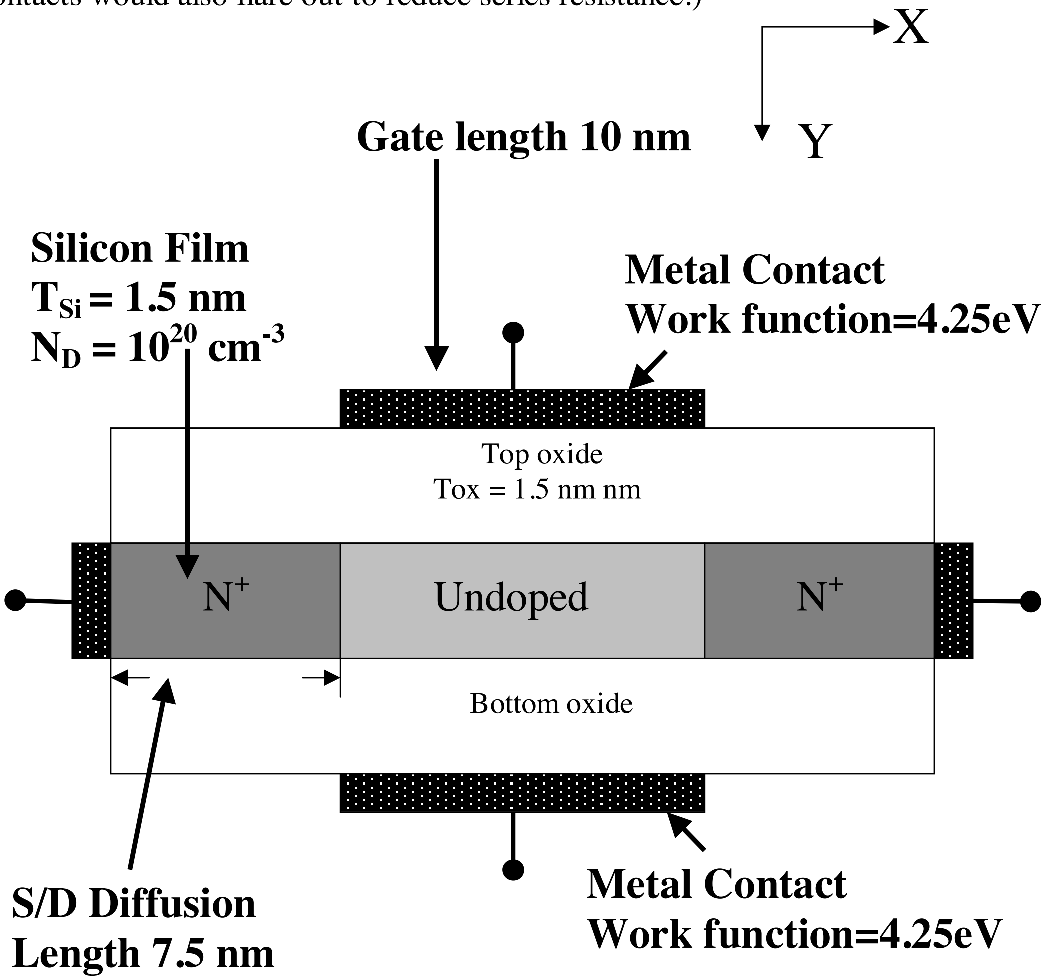


Fig. 5.2. Structure of the L_G = 10 nm double gate MOSFET with T_{OX} = 1.5 nm, T_{Si} = 1.5 nm, and V_{DD} = 0.6 V. This device was simulated with a 2D solution to Poisson equation coupled to a 1D transport solution [62].

Our simulations treated electrostatics two-dimensionally, but transport is essentially one-dimensional in this geometry, so a simplified, 1D transport model was used [62]. Quantum confinement effects in the direction normal to the Si film were treated in the one subband approximation. Several different approaches were used to describe transport along the channel. In the ballistic case, both a semiclassical (Boltzmann) solution and a quantum transport using a Green's function approach [72] were used. The Green's function solution was discussed in Chapter 3 and the Boltzmann solution in Appendix B. Quantum transport in the presence of phase breaking scattering was treated using a simple generalization of the Büttiker probe concept [72]. (As discussed in Chapter 4, we verified this approach captured the essential features of scattering observed in semiclassical approaches.) Conventional drift-diffusion and energy-transport models were also available.

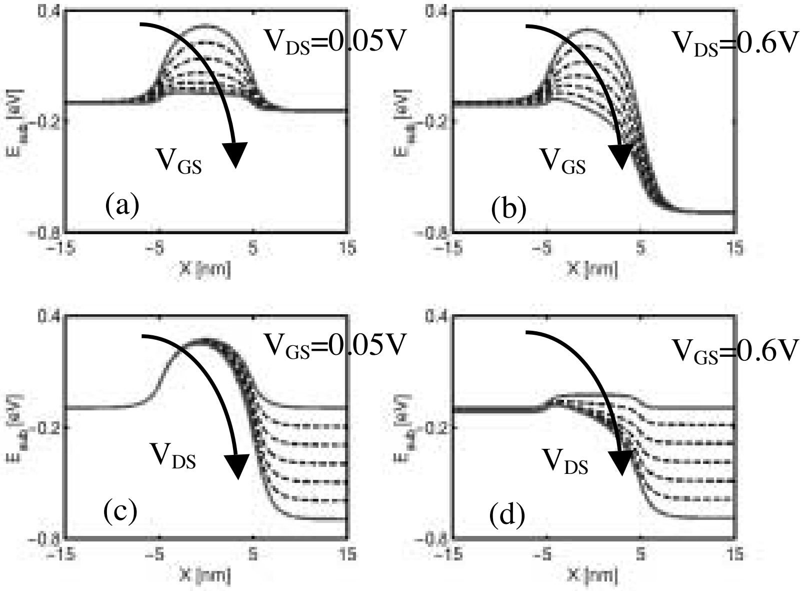


Fig. 5.3. The computed self-consistent conduction subband edge versus position for DG MOSFET of Fig. 5.2. (a) V_{DS} = 0.05 V and V_{GS} from 0.0 V to 0.6 V. (b) V_{DS} = 0.6 V and V_{GS} from 0.0 V to 0.6 V, (c) V_{GS} = 0.05 V and V_{DS} from 0.0 V to 0.6 V, and (d) V_{GS} = 0.6 V and V_{DS} from 0.0 V to 0.6 V.

The simplified device geometry and the ultra-thin body help to clarify the device physics to be explored in this study, but the conclusions of this study are born out by full 2D simulations of thicker body devices. Those results, however, are clouded by multi-subband conduction and stronger two-dimensional electrostatics (e.g. DIBL). Although the model device is a double gate MOSFET, we expect that the general conclusions of the study will apply to bulk MOSFET's as well. Figure 5.3 shows the computed self-consistent conduction subband profiles vs. position under a variety of bias conditions. (The program used to perform these simulations is available [110], and more extensive simulations of the same device have been reported in [62].)

5.2 The Ballistic MOSFET

The physical picture presented in Section 5.1 is most easily examined in the ballistic limit, and since present-day devices operate relatively close to this limit [46, 99], there is also a practical motivation to examine the ballistic MOSFET. For this purpose, we numerically simulated the model MOSFET of Fig. 5.2 using a semiclassical, ballistic transport model coupled to a two-dimensional solution to Poisson equation [73].

In Section 5.1, we argued that the maximum average carrier velocity at the beginning of the channel was the equilibrium, uni-directional thermal velocity. Assuming that only one subband is occupied, it can be shown that [28, 56]

$$\tilde{\upsilon}_T = \sqrt{\frac{2k_BT}{\pi m_t^*}}\left\{\frac{\Im_{1/2}(\eta)}{\ln(1+e^{\eta})}\right\} = \upsilon_T\left\{\frac{\Im_{1/2}(\eta)}{\ln(1+e^{\eta})}\right\}, \tag{5.4}$$

where $\eta = (E_F - E_1)/k_BT$, and the factor in brackets accounts for carrier degeneracy and approaches unity for a nondegenerate gas. (More generally, when multiple subbands are occupied, Schrödinger-Poisson simulations are needed [28].) Figure 5.4 shows the equilibrium $\tilde{\upsilon}_T$ vs. n_S characteristic computed from eqn. 5.4. Note that below threshold, $\tilde{\upsilon}_T \approx \upsilon_T \approx 1.2\times 10^7$ cm/s, but that above threshold, the carriers become degenerate, and the thermal injection velocity increases. Finally, note that the degenerate thermal injection velocity is the average velocity of all the carriers, while the Fermi velocity, υ_F, refers to the velocity of carriers at the Fermi level. The two are related by

$$\tilde{\upsilon}_T(\eta \to \infty) = \left(\frac{4}{3\pi}\right)\upsilon_F \tag{5.5}$$

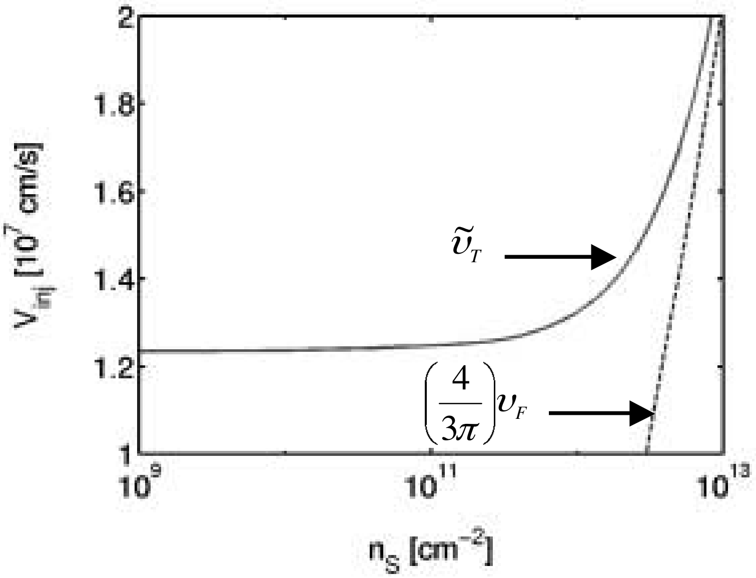


Fig. 5.4. The equilibrium thermal injection velocity, $\widetilde{\upsilon}_T$, versus inversion layer density, n_S, for the DG SOI MOSFET as evaluated from eqn. 5.4. Also shown is υ_F, the Fermi velocity.

We assert that the equilibrium, uni-directional thermal velocity is the maximum velocity that can be observed at the source end of the channel. The maximum source velocity exceeds the saturated velocity in the bulk, but the origin of this high velocity is much different than that of the conventional velocity overshoot that occurs in steep electric field gradients [27]. These high source velocities will, however, not be achieved unless the velocity within the channel is even higher (e.g. unless strong velocity overshoot within the channel).

The simulations displayed in Figs. 5.5 and 5.6 confirm the assertions made in the previous paragraph. Figure 5.5 is a plot of $< \upsilon(0) >$ versus drain bias as obtained by simulating the ballistic device of Fig. 5.2. (The location, $x = 0$, is taken as the top of the source-to-channel barrier, which changes with bias.) Under low bias, the average velocity is nearly zero because the negative velocities of carriers injected from the drain nearly cancel the positive velocities of those injected from the source. When the drain bias exceeds a few k_BT/q, then the negative velocity carriers injected from the drain are suppressed, and the average velocity approaches the equilibrium thermal velocity, $\tilde{\upsilon}_T$. Figure 5.6 shows the average velocity vs. position profiles at different drain to source voltages. As expected in this ballistic transistor, the velocity near the drain increases without limit (band structure limits have not been included). Under high drain voltages, however, the velocity at the top of the barrier saturates at the value displayed in Fig. 5.4. These results confirm the assumption made in Section 5.1 and earlier [55]. They show that velocity saturations occurs in a ballistic MOSFET, but it is the velocity at the top of the barrier that saturates at the thermal limit as opposed to the high-field velocity saturation in a bulk semiconductor which occurs because of scattering.

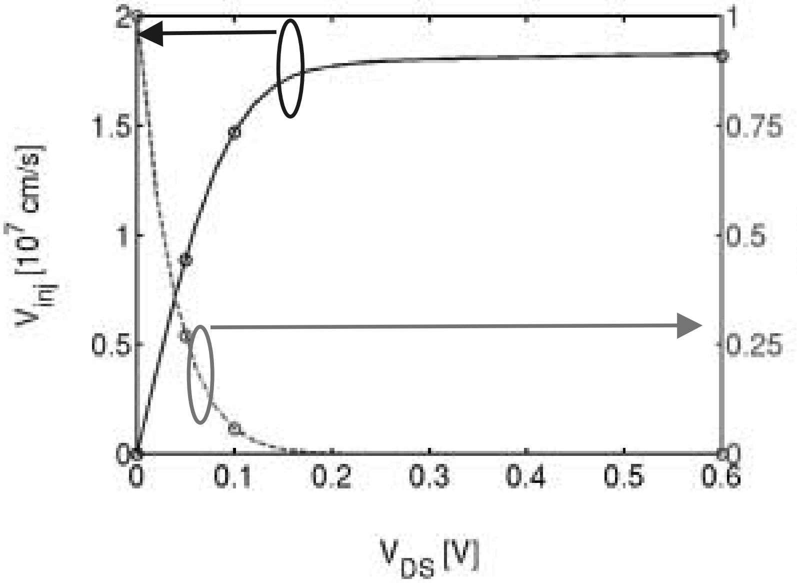


Fig. 5.5. The average velocity at the beginning of the channel versus V_{DS} for the device of Fig. 5.2 under ballistic conditions. For the gate voltage used $n_S \approx 5\times10^{12}\,\mathrm{cm}^{-2}$. Also shown is the ratio, J^- / J^+, (negatively-directed flux to the positively-directed flux), which is a measure of the anisotropy of the distribution (dashed line). Note that the velocity at the beginning of the channel saturates at the thermal equilibrium injection velocity as given by eqn. 5.4 when the negative half of the distribution is suppressed $J^- / J^+ = 0$). The large dots identify the four voltages examined in Fig. 5.7.

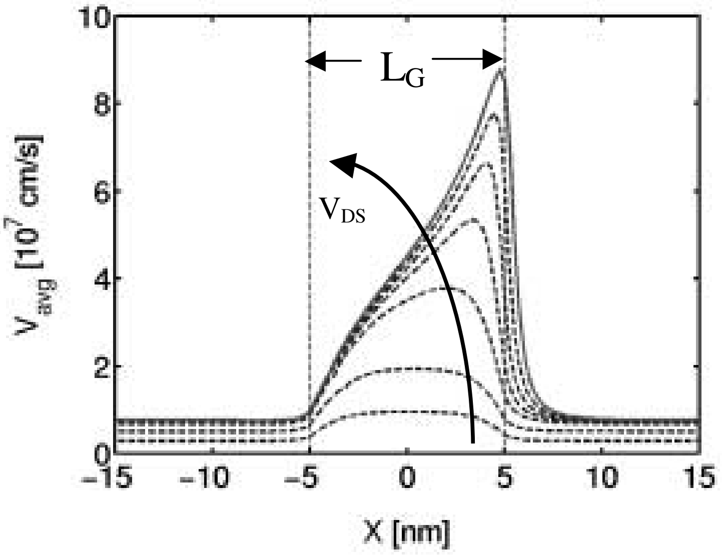


Fig. 5.6. The average velocity versus position for the device of Fig. 5.2 under ballistic conditions. For the gate voltage (V_{GS} = 0.6 V) used $n_S \approx 5\times10^{12}\,\text{cm}^{-2}$. Results for several different drain voltages are shown (V_{DS} = 0.0 V, 0.05 V, and 0.1 V to 0.6 V, with a step of 0.1 V).

In the ballistic MOSFET, a special kind of equilibrium exists; **k**-states are in equilibrium with the contact from which they were populated [10]. The overall carrier distribution, however, can have a highly off-equilibrium shape. For example, under high drain bias, the carrier distribution at x = 0 assumes a hemi-Fermi-Dirac, shape. This is suggested by the dashed line in Fig. 5.5, which shows the ratio, J^- / J^+, of the negative flux to the positive flux vs. drain bias. This ratio approaches zero when the drain bias is large enough to suppress the injection of negative-velocity carriers from the drain. The net velocity then saturates at $\widetilde{\upsilon}_T \approx 1.8\times10^7$ cm/s, which is 5% higher than the equilibrium thermal injection velocity shown in Fig. 5.4 (the difference is due to two-dimensional electrostatics). These effects are shown directly in Fig. 5.7, which plots the computed ballistic distribution functions at the top of the barrier for the four different voltages noted in Fig. 5.5. For low V_{DS} the velocity distribution is nearly symmetrical about $\upsilon_x = 0$. (In a long channel device, this symmetry is a result of carrier scattering, but in the ballistic

MOSFET, the positively-directed carriers are injected from the source and the negatively-directed carriers from the drain.) As the drain bias increases, the magnitude of the negative-velocity component decreases. Note, however, that although the overall velocity distribution has a highly nonequilibrium shape, each half is in equilibrium with its respective contact.

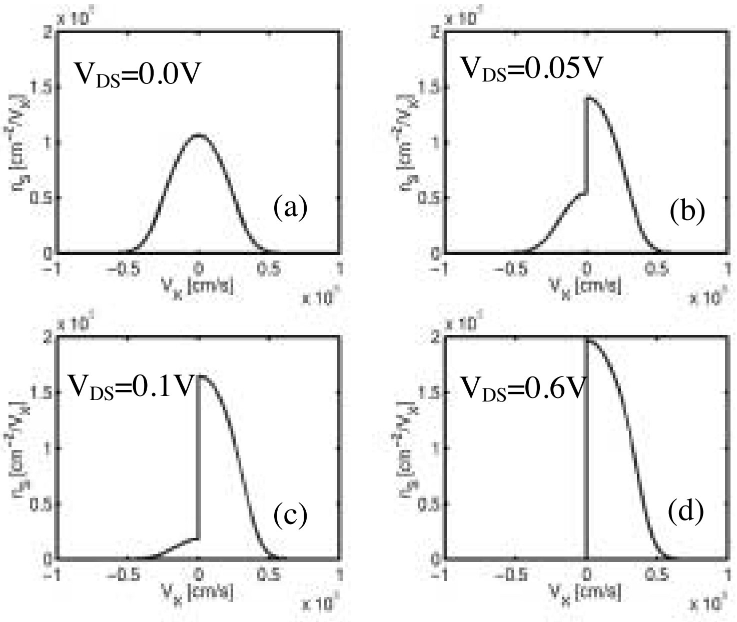


Fig. 5.7. 2D electron density versus longitudinal velocity as computed at the top of the source to channel barrier under ballistic conditions. The results are for V_{GS} = 0.6 V and (a) V_{DS} = 0.0 V, (b) V_{DS} = 0.05 V, (c) V_{DS} = 0.1 V, (d) V_{DS} = 0.6 V.

We turn now to the issue of charge control in the ballistic nanotransistor. Because the carrier distribution at the top of the barrier approaches a hemi-Fermi-Dirac distribution under high drain bias, it might be expected that under high bias, $n_S(0)$ would be one-half of its equilibrium value, eqn. 5.1. Figure 5.8, however, shows that this is not the case — $n_S(0)$ is approximately constant with drain bias. This occurs because MOS electrostatics demands that the charge on the gate balance that in the semiconductor, so that as V_{DS} increases, the conduction band is pushed down, more electrons are injected from the source, and $n_S(0)$ is maintained approximately at the value given by eqn. 5.1. This barrier lowering mechanism was also seen in Fig. 5.3d. The plot of $n_S(0)$ versus V_{DS} confirms that in a "well-tempered MOSFET," which is designed to electrostatically isolate the drain from the source [97], MOS electrostatics maintains the inversion layer charge at the beginning of the channel at an approximately constant value. Although the velocity distribution is highly nonequilibrium in shape, the charge density is maintained at approximately its equilibrium value. The same effect has also been observed in 2D Monte Carlo simulations [98].

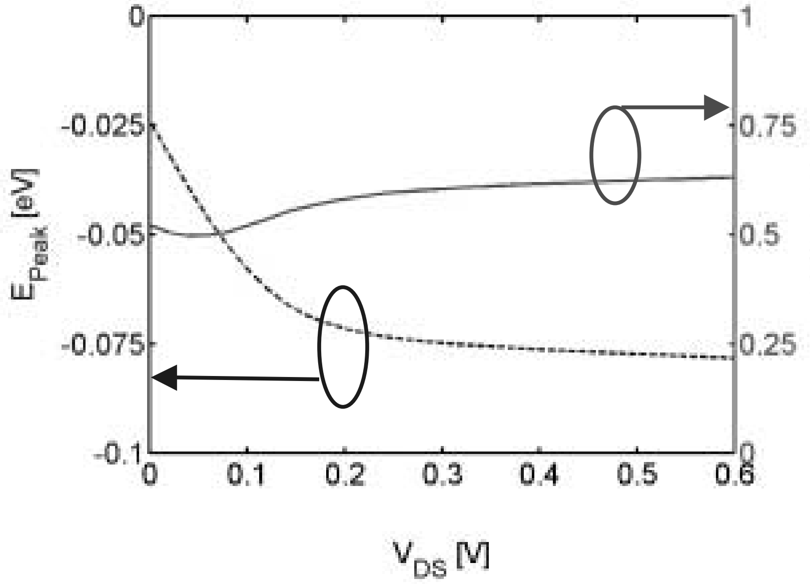


Fig. 5.8. Illustration of the charge control mechanism for the device of Fig. 5.2 under ballistic conditions. Solid line: the carrier density at the beginning of the channel versus V_{DS} for the device. Dashed line: The source to channel barrier height vs. V_{DS}. Figure 5.5 showed that as the ratio, J^- / J^+, decreases from 1 to 0, the average velocity increased. This figure shows that $n_S(0)$ remains essentially constant and that the source-to-channel barrier height decreases with increasing V_{DS} to maintain a constant carrier density at the top of the barrier. (The small increase is attributed to DIBL.)

Because the physics of the ballistic MOSFET is rather simple, a compact model is readily developed. Using the approach of [28] and assuming single subband occupation, one can show that [90]

$$\frac{I_{DS}}{W} = Q_{inv}(V_{GS})\tilde{\upsilon}_T \left[\frac{1 - \Im_{1/2}(\eta - U_{DS}) / \Im_{1/2}(\eta)}{1 + \ln(1 + e^{\eta - U_{DS}}) / \ln(1 + e^{\eta})} \right], \tag{5.6}$$

where $Q_{inv}(V_{GS})$ is the inversion layer charge (approximately $2C_{EFF}$ $(V_{GS}$ - $V_{TH})$ above threshold) and U_{DS} is V_{DS} normalized to k_BT/q. (Under nondegenerate conditions, the Fermi-Dirac integrals are replaced by exponentials, and under high drain bias, the term in brackets approaches unity.) Under high gate bias $Q_{inv} \approx 2C_{EFF}(V_{GS}$ - $V_{TH})$, so eqn. 5.6 reverts to eqn. 5.2 .

Conventionally, a MOSFET's channel resistance is proportional to its channel length, but there is also a ballistic component independent of channel length that may be important in nanoscale MOSFETs [28]. For low drain bias, eqn. 5.6 gives the ballistic conductance as

$$\frac{G_{DS}}{W} = \frac{I_{DS}}{W} = Q_{inv}(V_{GS})\left(\frac{\tilde{\upsilon}_T}{2k_BT/q}\right)\cdot\left[\frac{\Im_{-1/2}(\eta)}{\Im_{1/2}(\eta)}\right]. \tag{5.7}$$

As discussed in [28], under fully degenerate conditions, eqn. 5.7 reduces to $G_{DS} = M\left(e^2/2h\right)$, where M is the number of occupied transverse modes.

In Fig. 5.9 we compare the ballistic *I-V* characteristics as computed by direct numerical simulation and by the analytical expression, eqn. 5.6. The agreement is good – except for the output conductance, a two-dimensional effect not treated by the 1D analytical model. The channel resistance, R_{DS} , of this nano-MOSFET, as computed from the slope of the simulated characteristic in Fig. 5.9 or from eqn. 5.7, is about 60 Ω-μm. For comparison, we also show the simulated I_{DS} vs. V_{DS} characteristic for the transistor including a simple model for scattering (to be discussed in the next section). With scattering included, the channel resistance increases to about 200 Ω-μm. This value includes the conventional channel resistance, which is proportional to channel length, L_G, and the quantum contact resistance, which is given by eqn. 5.7 and is independent of L_G. Note that the ballistic channel resistance is about 35% of the total channel resistance.

Depending on the channel length and inversion layer mobility, this length-independent component to R_{DS} may be important.

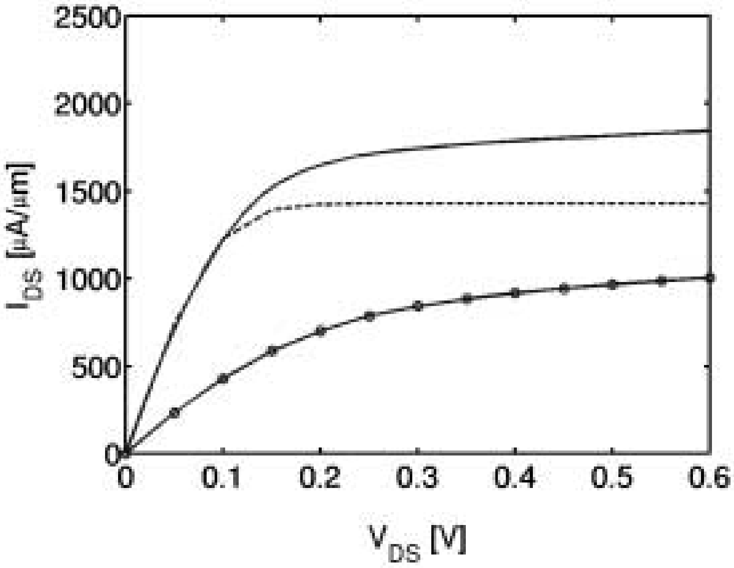


Fig. 5.9. Comparison of the simulated I_{DS} - V_{DS} characteristics of the ballistic device with the analytical model of eqn. 5.6. Solid line: simulated ballistic I_{DS} vs. V_{DS} for a gate voltage of 0.6 V. Dashed line: analytical I_{DS} vs. V_{DS}. Solid line with symbols: simulated I_{DS} vs. V_{DS} including the effects of scattering. (An inversion layer mobility of 100 cm^2/V-s was assumed.)

5.3 Scattering

In a ballistic MOSFET, the positive-velocity carriers at the top of the barrier are injected from the source and negative-velocity carriers from the drain, but scattering mixes the two streams. The result is that the carrier distribution at the top of the barrier does not approach a hemi-Fermi-Dirac distribution under high drain bias; $<\upsilon(0)>$ is less than $\tilde{\upsilon}_T$ under on-current conditions. When $V_{DS} >> k_BT/q$, so that all negative-velocity carriers at the top of the barrier arise from backscattering, eqn. 5.3 applies. Well-designed MOSFETs currently operate with $r \approx 0.4$ [46], so from eqn. 5.3 $<\upsilon(0)>$ is about one-third of its limit (but devices with $r \approx 0.2$ have been recently reported [99]). Figure 5.9 illustrates how scattering reduces device performance with respect to the ballistic limit; the channel resistance increases to several times the ballistic resistance, the on-current is reduced to about one-half of the ballistic limit, the drain saturation voltage increases, and the output conductance increases.

In this section, we examine two issues in detail: 1) charge control in the presence of scattering, and 2) the issue of why the channel backscattering coefficient is sensitive to backscattering very near the source end of the channel but relatively insensitive to scattering deep within the channel. For these studies, we use a Green's function method with a simple, Büttiker probe model of scattering, which we tested to ensure that it captures the essential physics of scattering in a MOSFET. As shown in [62], device operation is essentially classical (except for the strong quantum confinement effects); the quantum transport model was used because it was available and had been extensively tested on this device [62]. The broadening parameter, η, in the scattering model (see [72]) was set to 30 meV, which results in an inversion layer mobility of 100 cm^2/V-s for a long channel device. See Chapter 4 for a discussion of the formalism and solution methods.

Figure 5.10, which compares the self-consistent conduction subband profiles under on-current conditions with and without scattering, shows that the source-to-channel barrier is higher in the presence of scattering. This can be understood in terms of the self-consistent electrostatics of the MOSFET. For a given gate voltage, we expect the same inversion layer charge density at the top of the barrier – in the presence or absence of scattering. For the ballistic case, the carrier distribution is a hemi-Fermi-Dirac distribution, and the barrier height is established to provide the necessary inversion layer density. In the present of scattering, the carrier distribution function at the top of the barrier is more nearly symmetric in υ_x; so a higher barrier results in the same inversion layer density.

Figure 5.11 displays the simulated average velocity and carrier density at the top of the barrier vs. V_{DS} with a high gate voltage applied. The corresponding results for the ballistic case (from Figs. 5.5 and 5.8) are also displayed. Note first of all, that the inversion layer density at the top of the barrier, is nearly equal to its equilibrium value in the presence or absence of scattering (this is a simple consequence of self-consistent electrostatics and is relatively insensitive to the specific transport model). Note also that the maximum velocity at the top of the barrier does not saturate as clearly as for the ballistic case and that it is well below the thermal injection limit. Still, one can identify a drain saturation voltage of $V_{DSAT} \approx 0.3$ V, which is significantly greater than the ≈ 0.2 V in the ballistic case. It's clear that the mechanism for velocity saturation at the top of the barrier is different in the case of scattering and that it does not involve suppression of carrier injection from the drain as it in the ballistic case.

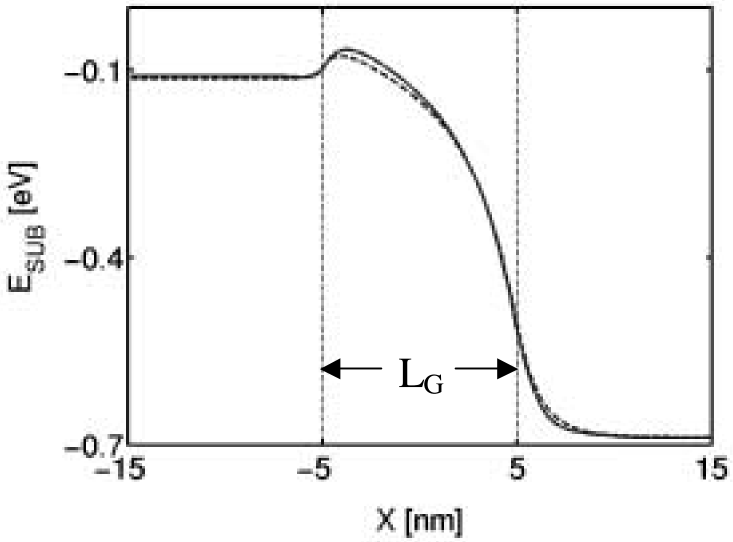


Fig. 5.10. Illustration of the effects of scattering on the self-consistent potential within the device of Fig. 5.2 under a bias of $V_{DS} = V_{GS} = 0.6$ V. Solid line: the lowest conduction subband energy vs. position in the presence of scattering. Dashed line: the same plot in the absence of scattering. The key difference is a slightly lower source-to-channel energy barrier in the presence of scattering.

In the presence of scattering, velocity saturation at the beginning of the channel occurs because of the self-consistent electrostatics in the device. As shown in Fig. 5.3d, for V_{DS} greater than about 0.2 V, most of the additional applied drain voltage is dropped across the drain end of the channel, and conditions near the source are relatively constant. From eqn. 5.3, one can estimate that $r \approx 0.3$ at $V_{DS} \approx V_{DSAT}$. Below V_{DSAT}, the electric field near the source varies directly with V_{DS}, but above V_{DSAT}, the source electric field increases slowly with increases in V_{DS}. The slow rise in $< \upsilon(0) >$ with V_{DS} beyond the saturation voltage occurs because of the slow increase in electric field, which slowly decreases r. Since I_{DS} is the product of $< \upsilon(0) >$ and $Q_i(0)$, which is approximately constant, these observations also explain the I_{DS} vs. V_{DS} characteristic displayed in Fig. 5.9.

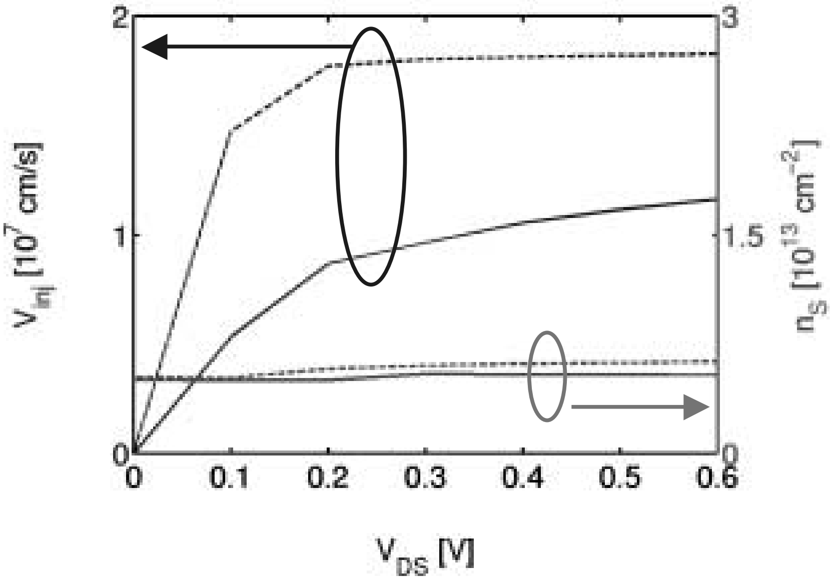


Fig. 5.11. Illustration of the effects of scattering on the average velocity and charge at the top of the barrier for the device of Fig. 5.2 with V_{GS} = 0.6 V. The carrier density at the beginning of the channel vs. V_{DS} (right vertical axis). The average velocity at the beginning of the channel vs. V_{DS} (left vertical axis). The solid lines include scattering, and the dashed lines are the corresponding results for ballistic conditions (from Figs. 5.5 and 5.8).

Given the central role of the backscattering coefficient, r, in the operation of a MOSFET, we should examine the physics that controls it. The backscattering coefficient is determined by both carrier scattering and by the potential drop within the channel. Figure 5.12 schematically illustrates a stream of carriers injected into the channel from the quasi-equilibrium point at the top of the barrier. The fraction that backscatters and returns to the source is defined as r. If backscattering occurs beyond a certain critical distance (denoted as ℓ in Fig. 5.12), then it is unlikely that the carrier will have sufficient longitudinal energy to surmount the barrier and exit into the source. More likely, it will be reflected by the channel potential, perhaps undergo several scattering electric field reflections, and exit from the drain. These scattering events will increase the carrier density in the channel and through Poisson equation, the self-consistent electric field throughout the entire channel, but they do not contribute directly to *r* as we have defined it. To understand why this occurs, one must realize that Fig. 5.12 is a plot of *longitudinal* energy ($m^* \upsilon_x^2 / 2$) not *total* energy ($m^* \upsilon^2 / 2$). For the typical case of a wide MOSFET, there is a continuous distribution of transverse modes. Only a small fraction of the carriers will backscatter directly at the source and possess sufficient longitudinal energy to surmount the barrier. Note that this argument applies to both elastic and inelastic scattering. Finally, note that If this were a quantum wire MOSFET in which the only degree of freedom was the x-axis, then *r* would be sensitive to backscattering through out the entire channel.

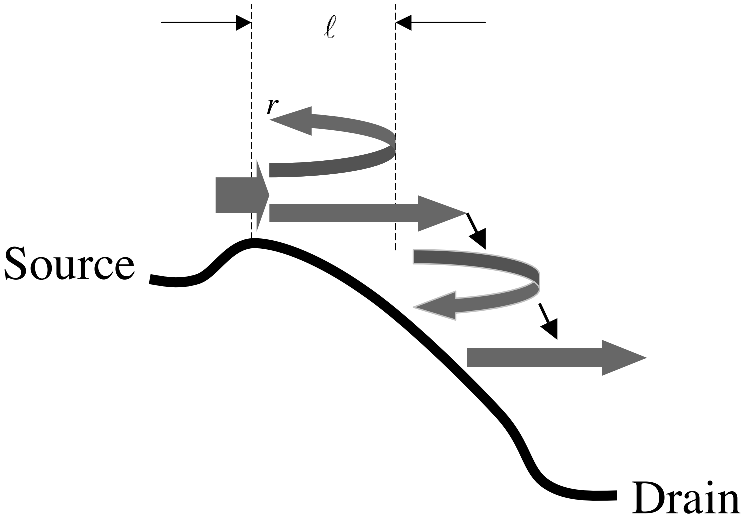


Fig. 5.12. Illustration of carrier backscattering in a MOSFET under high drain bias. If a carrier backscatters beyond a critical distance, ℓ, from the beginning of the channel, then it is likely to exit from the drain and likely to return to the source.

From the argument presented above, we conclude that the steady-state drain current is limited only by backscattering that occurs within a critical distance, ℓ, from the beginning of the channel. The existence of such a critical distance was first noted by Price, who observed in performing Monte Carlo simulations of carrier transport down a potential barrier, that if carriers penetrated only a very short distance into the potential drop, then even if they did scatter, they were unlikely to return to their injection point at the top of the barrier [100]. Price used a detailed balance argument to relate his "down the potential" simulations to "up the potential" transport. Recognizing the close connection between transport up or down the barrier, we can make use of the well-known Bethe condition for a metal-semiconductor junction to establish ℓ. Bethe showed that currents near the thermionic (i.e. ballistic) limit occurs when the first k_BT/q of potential drop at the junction, occurs over a distance much less than the mean-free-path. Since this critical distance (known as the kT-layer [101]) is a small fraction of the barrier width, the thermionic emission typically applies. From Price's detailed balance argument, we recognize a close connection between transport with and against the barrier, which suggests that the critical layer for the MOSFET is also the distance over which the first k_BT/q of channel potential drops, typically a small fraction of the channel length.

By identifying the critical distance, ℓ, with the k_BT layer, the expression for the backscattering coefficient for a field free semiconductor slab of length, L, [10, 27],

$$r = \frac{L}{L + \lambda} \tag{5.8}$$

can be generalized to [55]

$$r = \frac{\ell}{\ell + \lambda} . \tag{5.9}$$

Since the critical backscattering occurs in a region where the carriers have gained little energy from the channel field, the appropriate mean-free-path to use in eqn. 5.9 is λ_o, the near-equilibrium mean-free-path for backscattering, which can be obtained from the measured mobility of a long-channel MOSFET. A comparison of the simple expression, eqn. 5.9, with a rigorous evaluation of *r* by direct Monte Carlo simulation, shows good agreement [55]. Note also that the key result, eqn. 5.9, need not be postulated; it can be derived by scattering theory (see Chapter 9 of [27] for an introduction to scattering theory).

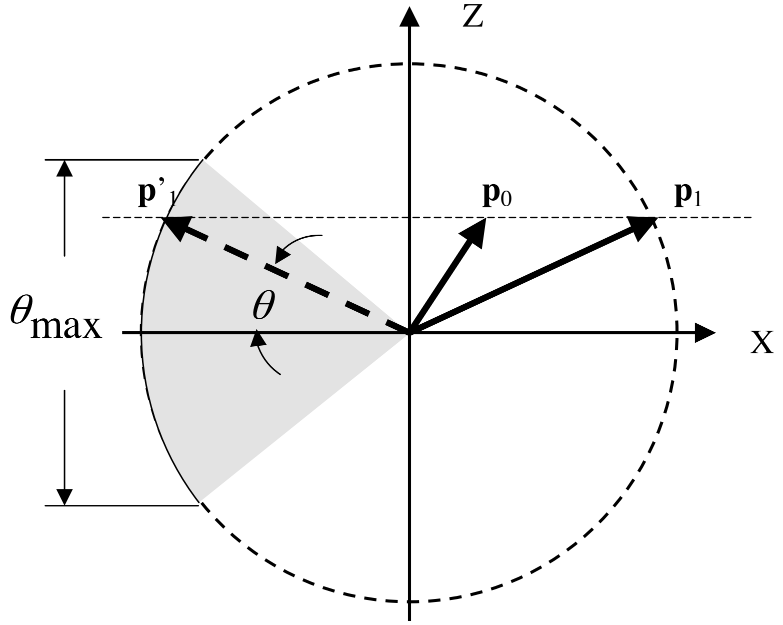


Fig. 5.13. Illustration of backscattering and how it contributes to *r*. A 2D confined carrier is injected into the source with momentum $\mathbf{p_0}$. It propagates down the potential drop towards the drain gaining an energy ΔE with corresponding momentum, $\mathbf{p_1}$. It then scatters to momentum $\mathbf{p_1}$**'** (since we assume elastic scattering, $\mathbf{p_1}$ = $\mathbf{p_1}$**'**). Only carriers within the shaded region have sufficient longitudinal momentum to cross the barrier and enter the source.

Calculating the channel backscattering coefficient (even under the simplifying assumptions that lead to eqn. 5.9) is non-trivial, but a simple argument explains why the importance of backscattering decreases from source to the drain. Consider a charge carrier injected from the source into the channel with momentum $\mathbf{p}_0 = (p_{xo}, p_{zo})$. (Because of the quantum confinement in the y-direction, the electron has two degrees of freedom.) If this injected carrier gains an energy, ΔE, by acceleration in the longitudinal electric field, then its momentum is $\mathbf{p}_1$, where $p_1^2 = \left(p_{xo}^2 + 2m\Delta E\right) + p_{zo}^2$. Assume that the electron then backscatters elastically to momentum, $\mathbf{p}_1'$ (see Fig. 5.13). If the backscattered electron propagates ballistically to the beginning of the channel, what is the probability that it can cross the barrier, and, therefore, contribute to *r*? To do so, requires sufficient longitudinal kinetic energy,

$$\frac{p_{x1}^{'2}}{2m^*} = \frac{p_1^2}{2m^*}\cos^2\theta \geq \Delta E \,. \tag{5.10}$$

Equation 5.10 defines a maximum angle, $\theta_{\max}$, for backscattered carriers that contribute to *r*,

$$\theta_{\max} = \cos^{-1}\left(\sqrt{\frac{\Delta E}{\Delta E + E_0}}\right). \tag{5.11}$$

(see Fig. 5.13). Finally, the fraction of the scattered carriers that contribute to *r* is the fraction with $|\theta| < \theta_{\max}$ or

$$F = \frac{\theta_{\max}}{\pi} = \frac{\cos^{-1}\left(\sqrt{\frac{\Delta E}{\Delta E + E_0}}\right)}{\pi} \,. \tag{5.12}$$

Figure 5.14 is a plot of *F* versus $\Delta E / E_0$; it shows that when the carriers have traveled down the potential drop by an amount equal to the injection energy, E_0 (k_BT for a non-degenerate, 2D carrier gas), then even if they do scatter, only 50% of them have a chance to contribute to *r*. As carriers travel further down the potential drop, the probability that a scattering event will contribute to the channel backscattering coefficient, *r*, steadily decreases.

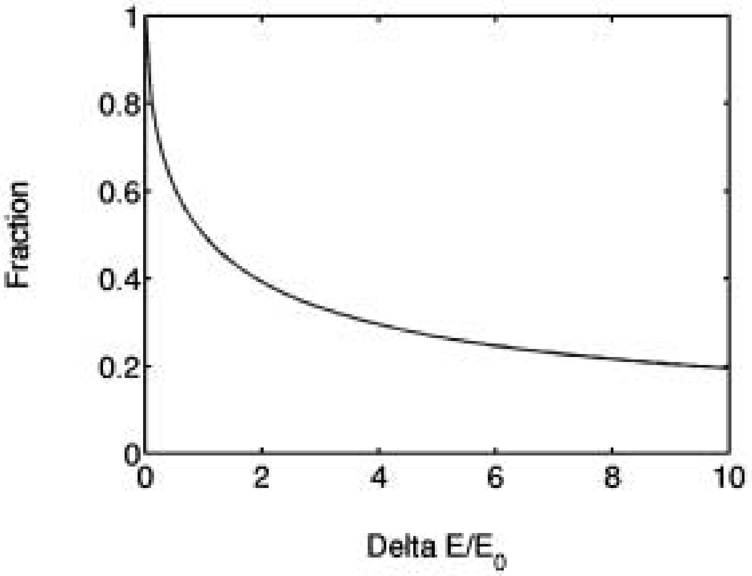


Fig. 5.14. The fraction of the scattered electrons that contribute to the channel backscattering coefficient, *r* (i.e. the shaded region in Fig. 5.13). The curve is evaluated from eqn. 5.12 assuming that a carrier gains an energy, ΔE, before isotropically scattering, then propagates back to the barrier without scattering again.

The simple argument presented above explains why scattering near the source controls the backscattering coefficient, *r*. In practice, the critical region is even more weighted towards the beginning of the channel than suggested by Fig. 5.14. There are two reasons; first, as the backscattered carrier propagates towards the source, it may be scattered again, and second, as the injected carrier penetrates deeper into the channel, its energy increases and so does the probability of scattering by phonon emission, which lowers its energy and makes it less likely to return to the source.

5.4 Discussion

Transport in a nanoscale MOSFET is nonlocal; the average carrier velocity does not depend on the local electric field. A mobility can be precisely defined, but since it depends on an essentially unknown distribution function, it is not a useful parameter [102].) Mobility is, however, well-defined parameter in a long channel MOSFET. From the near-equilibrium mobility, which is readily measured in a long-channel MOSFET, the near-equilibrium mean-free-path for backscattering, which is the important transport parameter for a nanoscale MOSFETs, can be determined. In this case, one can say that mobility is a meaningful parameter for nanoscale MOSFETs. (There are, of course, complicating factors that have to be dealt with, such as the use of halo implants which can result in different channel dopings for long and short-channel devices and, therefore, different mobilities.)

Shockley used scattering theory to relate the near-equilibrium diffusion coefficient, D_o, to the mean-free-path for backscattering as [103-104]

$$D_o = \upsilon_T \lambda_o / 2 \,. \tag{5.13}$$

(See Chapter 9 in [27] for an alternative derivation of this result.) Since near-equilibrium conditions prevail, the Einstein relation may be invoked and the result is a simple relation between the near-equilibrium mobility and the near-equilibrium mean-free-path for backscattering. Finally, we note that eqn. 5.13 assumes nondegenerate carrier statistics, but this assumption fails above threshold. In the more general case, the relation between λ_o and μ_o becomes more complex. Note also, that defining the width of the critical region from the k_BT/q potential drop also assumes nondegenerate carrier statistics. Our use of nondegenerate statistics establishes the central ideas simply.

We have been careful to refer to λ_o as the mean-free-path for *backscattering*, but we have not defined it precisely. The relation of the mean-free-path for backscattering that we use and the mean-free-path itself is analogous to the relation between the momentum relaxation time, τ_m, and the mean time between scattering events, τ. This mean-free-path can be precisely defined in terms of the transition rate per unit time for scattering from state **k** to **k'** S(**k**, **k'**) as [105]

$$\frac{1}{\lambda_o} \equiv \sum_{k_x^{'}>0,k_z^{'}} \frac{S(\mathbf{k},\mathbf{k}^{'})}{\upsilon_x(\mathbf{k})}, \tag{5.14}$$

where we have assumed nondegenerate carrier statistics.

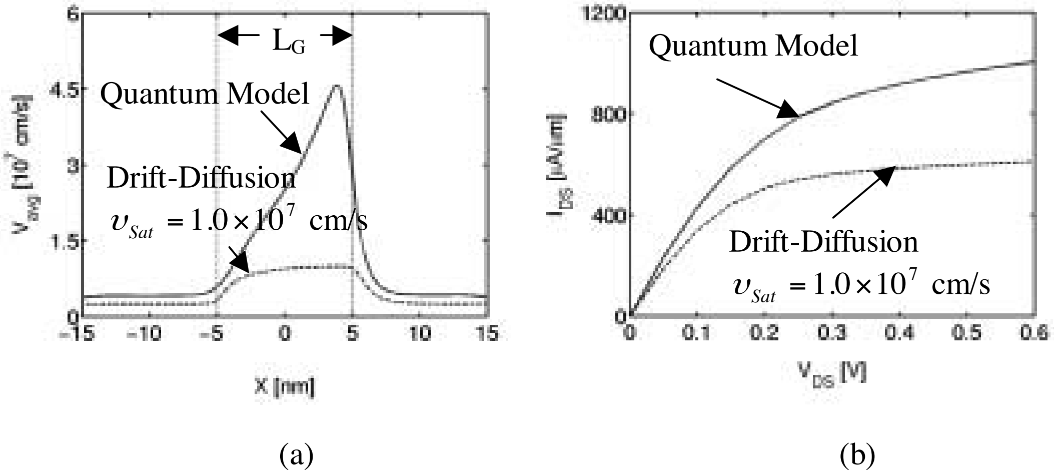


Fig. 5.15. (a) Average velocity vs. position at $V_{GS} = V_{DS} = 0.6$ V for the 10 nm DG SOI-MOSFET. (b) I_{DS} vs. V_{DS} for $V_{GS} = 0.6$ V for the 10 nm DG-SOI MOSFET.

For the past decade, much of the modeling and simulation work has focused on accurately describing velocity overshoot within the channel, but in the view presented in this chapter, velocity overshoot is considered to play a secondary role. It can, however, have significant effects on devices [106]. We should note first that to achieve a velocity at the source that approaches the thermal limit, the velocity within the channel must be even higher. When the source velocity is well below the thermal limit, it is possible for a velocity saturated simulation to get the velocity at the source correct, but it will erroneously clamp the velocity near the drain at an unphysically low value. The inversion layer density in the channel will be too high near the drain, which will lead to errors in the self-consistent channel potential. These carriers will screen the source from charges on the drain, so we should expect a unphysically low output conductance from a velocity-saturated model. These effects are shown in Fig. 5.15. Figure 5.15a compares the channel velocity vs. position profiles under on-current conditions for a velocity-saturated drift-diffusion transport model and for the Green's function method. We observe that the

Green's function method captures the velocity overshoot that occurs near the drain. Figure 5.15b compares the simulated I_{DS} vs. V_{DS} characteristics for the two transport models. Note that the output conductance is considerably higher when velocity overshoot is included. Bude has observed that the effect can be as large as 40% for nanoscale bulk MOSFETs [106].

5.5 Summary

A conceptual view of the essential physics of carrier transport in nanoscale MOS transistors was presented and confirmed by numerical simulation. Key results are: 1) that the source velocity saturates and that its limit is set by thermal injection, 2) that the carrier density at the top of the source to channel barrier is fixed by MOS electrostatics (in an electrostatically well-designed MOSFET), 3) that scattering in very short region near the beginning of the channel limits the on-current, and 4) that the role of velocity overshoot is primarily an indirect one based on its influence on the self-consistent potential throughout the channel. The results show that the physics that determines the steady-state current of a MOSFET can be understood in terms of a simple model. This view of nanoscale MOSFET device physics should provide a useful guide for experimental and theoretical work, for developing compact models, and for interpreting detailed simulations.

6. COMPUTATIONAL STUDY OF L_G = 10 NM DOUBLE-GATE MOSFETS

6.1 Introduction

In this chapter we explore the device design and physics issues for transistors near their ultimate scaling limits using a simulation tool, nanoMOS [62, 110]. The International Technology Roadmap (ITRS-99) specifications for the year 2014 transistor generation, equivalent oxide thickness (0.6 nm), off-state leakage (160 nA/μm) and power supply voltage (0.6 V) have been followed [2]. The device structure we examine is a double gate (DG) n-channel MOSFET with a metallurgical gate length of 10 nm. The issues addressed are:

1) *Device design*: First, we outline the procedure to select the correct combination of silicon film thickness and gate dielectric in order to meet short channel requirements. We then discuss the technique used to engineer the gate stack in order to meet the threshold voltage and gate leakage requirements.

2) *I-V Characteristics in the Ballistic Limit*: We present simulation results from both classical and quantum ballistic transport models and discuss quantum effects in this nanoscale transistor.

3) *I-V Characteristics with Scattering*: A simplified quantum mechanical scattering model is used to treat the effect of surface roughness and to capture mobility degradation due to high doping concentrations. Using this model, we present results that highlight the role of gate overlap/underlap, source/drain extension and quantum contact resistance on the performance of nanoscale transistors. A study of the expected on-current as a function of channel mobility is also presented.

4) *Gate tunneling*: Assuming an empirical gate tunneling model, we examine the leakage current distribution along the gate and predict the gate leakage current.

Our analysis enables us estimate the performance of L_G = 10 nm generation transistors. This analysis also helps identify important design issues that need to be considered in order to achieve the performance targets specified by ITRS-99, as critical transistor dimensions are scaled down in the future.

6.2 Theory

The Schrödinger equation is solved in two dimensions in order to generate the results presented in this work. As discussed in Chapter 3, a 2D solution to the Schröedinger equation is obtained by solving two 1D problems, one in the direction normal to the channel, which yields the vertical electron concentration and subband profiles, and the other, along the channel direction based on the subband profiles yielding the electron concentration in the transmission direction. Two different approaches are used to treat the physics of carrier transport along the channel direction. The first approach is a solution to the Boltzmann equation in the ballistic limit (only the thermionic emission part is captured) while the second approach, which is more general, uses the Green's function formalism described in Chapter 3 (tunneling through the source to channel barrier is captured) to simultaneously capture the physics of both ballistic and dissipative transport in a full quantum framework. A 2D Poisson solver is coupled to each of the transport models to provide self-consistent solutions. Note that electrostatic effects due to penetration of the electron wavefunctions into the oxide regions are accounted for in our simulations by extending the quantum solution domain to include the these regions. Gate leakage current calculations, however, are based on an empirical model and are a post processing operation. In the following section, we provide a brief description of the procedure used to calculate the charge and current distributions in case of ballistic and dissipative transport. Details of the calculation scheme were discussed in Chapters 3 and 4.

6.2.1 Ballistic transport

When the SiO$_2$/Si interface is parallel to the (100) plane, the six equivalent conduction band minima of the silicon body split into two sets of subbands due to different effective masses in the confinement direction [51] (the so called unprimed and primed subbands).

Within the Boltzmann framework, carriers are injected into the channel from a thermal equilibrium reservoir (the source), over a subband energy barrier whose height is modulated by the gate voltage. The spatial subband profile therefore can be divided into two regions: points to the left of the peak subband energy and points to the right of the peak subband energy. The 2D electron density for these sets of points is (the detailed derivation is given in Appendix B),

$$n_{left}(x) = \sum_i n_{2Di}\left\{\ln(1+e^{\mu_S}) + \left[\frac{1}{\sqrt{\pi}}\int_0^{E_{Pi}}\frac{dE}{\sqrt{E}}\Im_{-1/2}(\mu_S - E) + \frac{1}{\sqrt{\pi}}\int_{E_{Pi}}^{\infty}\frac{dE}{\sqrt{E}}\Im_{-1/2}(\mu_D - E)\right]\right\} \tag{6.1a}$$

$$n_{right}(x) = \sum_i n_{2Di}\left\{\ln(1+e^{\mu_D}) + \left[\frac{1}{\sqrt{\pi}}\int_0^{E_{Pi}}\frac{dE}{\sqrt{E}}\Im_{-1/2}(\mu_D - E) + \frac{1}{\sqrt{\pi}}\int_{E_{Pi}}^{\infty}\frac{dE}{\sqrt{E}}\Im_{-1/2}(\mu_S - E)\right]\right\} \tag{6.1b}$$

where, n_{2Di} is a constant with the dimension of areal carrier density for subband, i, E_{Pi} the peak energy for subband i, and μ_S and μ_D, the source and drain contact Fermi energies. $\Im_{-1/2}$ is the Fermi integral of order –1/2 [60]. It should be noted that all energies are specified relative to the local subband potential and are normalized to the thermal energy ($k_B T$).

Quantum mechanically, there is no division of points depending on their spatial location along the channel. The 2D electron density, calculated based on the Green's function formalism is given by (see also Chapter 3),

$$n(x) = \sum_i n_{Oi} \left\{ \int_{-\infty}^{+\infty} dE \left[\Im_{-1/2}(\mu_S - E) D_{Si}(E,x) + \Im_{-1/2}(\mu_D - E) D_{Di}(E,x) \right] \right\} \tag{6.2}$$

where, n_{Oi} is a constant with the dimensions of 2D carrier density, D_{Si} and D_{Di} are the source and drain contributions to the local density function for subband, i.

The ballistic current can be evaluated either at the source or the drain contact as (see also Chapter 3 and Appendix B)

$$I_D / W = \sum_i I_{Oi} \int_{-\infty}^{+\infty} T_{SDi}(E) [\Im_{-1/2}(\mu_S - E) - \Im_{-1/2}(\mu_D - E)] dE \tag{6.3}$$

where, I_{Oi} is a constant with the dimensions of current/length for subband i, and T_{SDi} is the source–to-drain transmission as a function of electron energy. In the classical case, $T_{SDi}(E)$ is 1 above the subband barrier maximum, and 0 below it. Therefore the integral over energy in eqn. 6.3 can be computed analytically. In the quantum model, $T_{SDi}(E)$ includes contributions from above and below the subband barrier maximum [61-62]. Therefore current contributions at every energy, have to be evaluated numerically. This distinctive treatment of channel transmission provides a way of assessing quantum tunneling against the thermionic limit.

6.2.2 Dissipative transport

Scattering in MOSFETs is treated through the Green's function formalism using a very simple Büttiker-probe model discussed in Chapter 4. Scattering centers are viewed as reservoirs similar to the source and drain. However, they differ from the source and drain reservoirs as they can only change the energy of the carriers and not the total number of carriers in the system. This model captures the essential physics of scattering as can be seen from our simulation results and results presented in [72, 74, 111]. Each scattering center is modeled as a Büttiker probe with a perturbation strength characterized by a position dependent self energy, η . The self energy can be related to a dephasing time which can be interpreted as the time within which a carrier's (electron in our case) initial state is fully destroyed by a scattering event [72]. Therefore it is possible to map the dephasing time onto an equivalent mobility. The 2D electron density including the effect of all scattering centers as well as the source and drain reservoirs is given by eqn. 4.18a (refer to Chapter 4 for detailed derivations), it is also presented here for readers' convenience,

$$n = \sum_i n_{Oi} \sum_j \int_{-\infty}^{+\infty} dE[\Im_{-1/2}(\mu_j - E) D_{ji}(E,x)] \tag{6.4}$$

where, summation over i accounts for contributions from all subbands, the summation over j represents contributions from all reservoirs, D_{ji} the local density for subband i due to source j , and μ_j the Fermi potential of reservoir j .

Electrons reaching a scattering center are fully thermalized according to the scatterer's Fermi potential. Because scattering processes lead to energy relaxation while simultaneously conserving the number of particles, the net current at each scattering center must be identically equal to zero. The total current at reservoir j has been derived in detail in Chapter 4, the expression is given as (see also eqn. 4.18b)

$$I_{ej} = \sum_i I_{Oi} \sum_k \int_{-\infty}^{+\infty} T_{jki}(E)[\Im_{-1/2}(\mu_j - E) - \Im_{-1/2}(\mu_k - E)]dE \tag{6.5}$$

where, summation over i accounts for contributions from all subbands, summation over k accounts for current due to all sources and T_{jki} is transmission from source k to j in subband i. The requirement that the net current at each scatterer j, equals zero ie: I_{ej} = 0, imposes a set of constraints on μ_j's (Fermi-potential of scatterers). This set of constraining equations can be solved for the μ_j's using a non-linear Newton's method as discussed in Sec. 4.2.1. Note, that the scatterer's Fermi potentials are position and not energy dependent in our treatment. Therefore electrons are thermalized (or scattered) among all transverse modes and subbands based on a unique Fermi-potential of the scatterer at each real-space position, thus capturing the essential physics of scattering as described in [111].

Our work extends recent pioneering work [61] by using the Green's function approach to obtain quantum solutions in both the ballistic and dissipative regimes (without invoking the WKB approximation). Based on results from our ballistic and dissipative transport models we critically examine device physics and design issues for future generation double gate MOSFETs.

6.3 Results

The device structure we examine is shown in Fig. 6.1. This device is a simplified and ultra-scaled version of a DG transistor fabricated at Purdue [23]. The gate length is 10 nm and so are the lengths of the source/drain extension regions. An intrinsic body ($N_A = 10^{16}$ /cm^3) is assumed in order to eliminate threshold voltage fluctuations due to variations in body doping and to reduce mobility degradation due to ionized impurity scattering. Both abrupt as well as graded junctions have been examined in this simulation study using an effective oxide thickness as specified by ITRS-99 (year 2014 generation transistor). In this section we analyze the issues outlined in the Introduction.

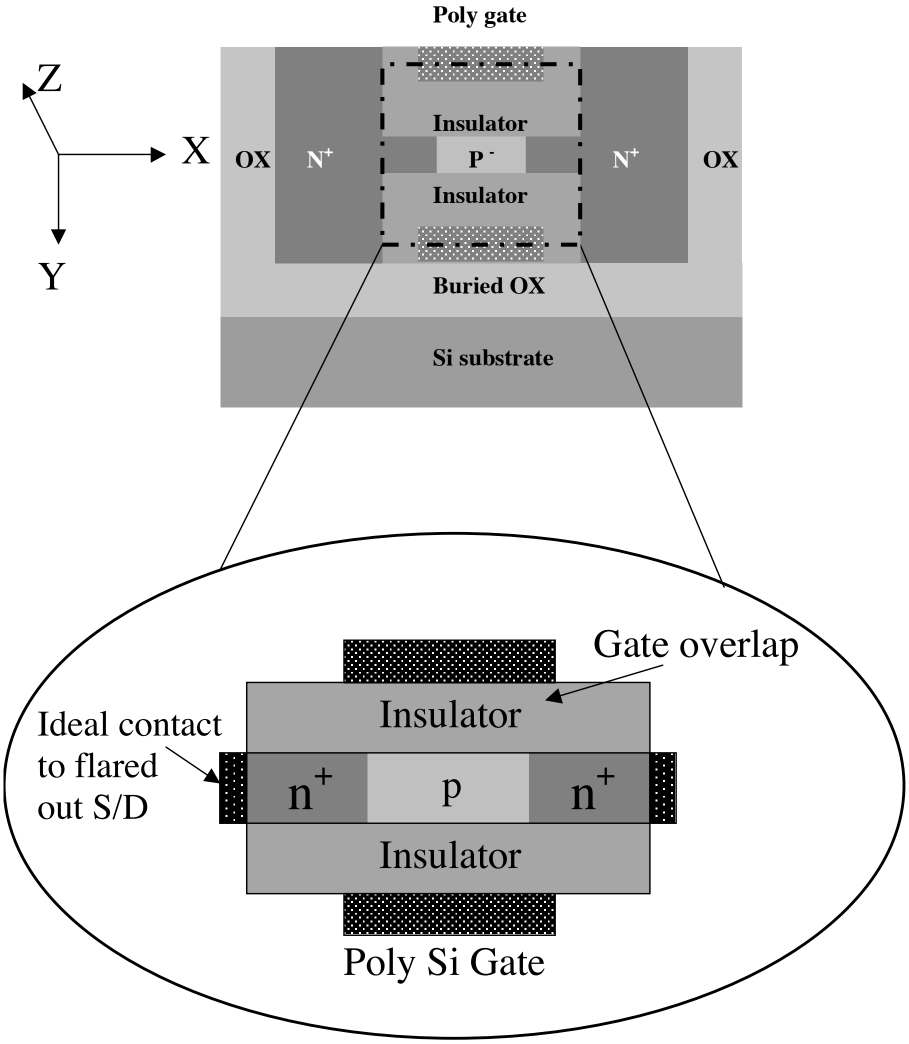


Fig. 6.1. Sketch of the double-gate MOSFET fabricated by Tai-Chi Su et al. [23] and the idealized device structure used in our study. Current flow is along the x-direction, the gate confinement is in the y-direction, and the width of the device is along the z direction.

6.3.1 Device design

Figure 6.2 is a plot of the threshold voltage sensitivity to channel length around L_G = 10 nm, the threshold voltage is defined at a current value of 1 μA/μm when $V_{DS} = V_{DD}$ = 0.6 V. This quantity, dV_{TH} / dL_G, is plotted as a function of silicon film thickness and gate dielectric constant κ (The equivalent oxide thickness is 0.6 nm in all cases). The maximum acceptable variation in the threshold voltage is chosen to be 50 mV. Use of a high κ gate material permits a thicker physical insulator which helps meet the gate leakage requirement but degrades short channel immunity thus reducing the maximum silicon film thickness that can be used [4, 112]. Mobility in thin silicon films could be extremely low [113]. Therefore, our aim is to try to maximize κ and T_{Si} simultaneously while trying to minimize threshold voltage rolloff and gate leakage. The best combination of κ and T_{Si} is κ =19.5 (physical insulator thickness T_{ins} (*physical*) = 3 nm) and T_{Si} = 3 nm. Figure 6.3, which shows the subthreshold swing vs. T_{Si} for different κ materials yields a similar conclusion. A subthreshold swing of 80 mV/decade is selected in order to efficiently turn off the device in case of reduced threshold voltage due to fluctuations.

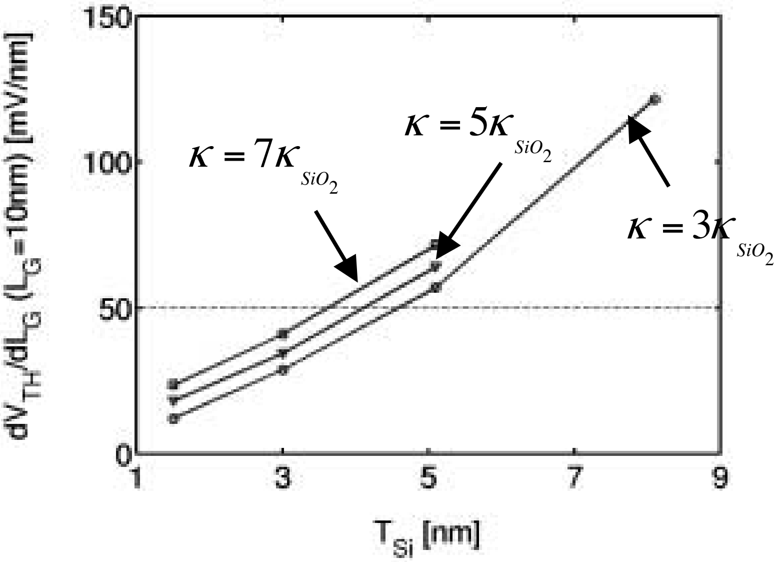


Fig. 6.2. Threshold voltage (defined at a current value of 1 μA/μm when V_{DS} = V_{DD} = 0.6 V) variation for fluctuations in the gate length around L_G = 10 nm is plotted vs. silicon body thickness. Results for three different gate dielectric constants have been shown. The dashed line indicates the maximum permissible V_{TH} variation to turn off the device in the worst case (10% L_G variation, one lattice constant variation in T_{Si}).

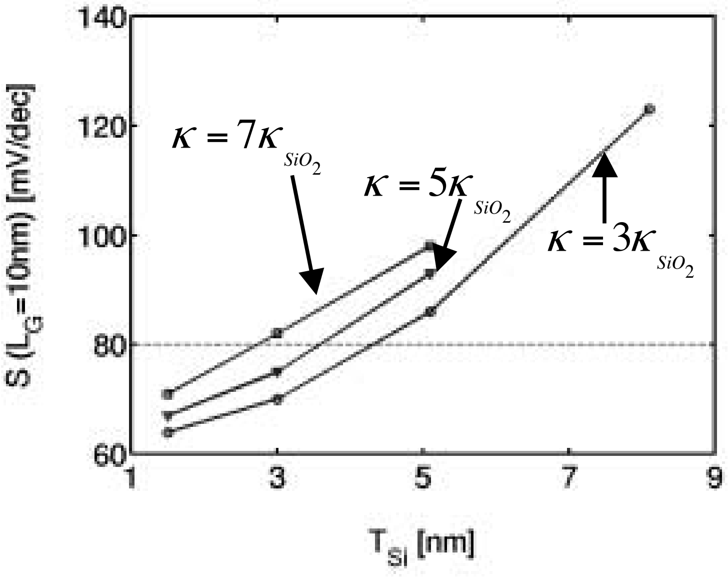


Fig. 6.3. Subthreshold swing vs. T_{Si} for different κ materials. The dashed line indicates the maximum permissible subthreshold swing to turn off the device in the worst case. Both Fig.6.2 and Fig.6.3 provide a similar conclusion for the best combination of κ and T_{Si} needed to achieve the desirable SCE for the L_G =10 nm MOSFET under study.

It is also important to look at the threshold voltage variation as a function of the silicon film thickness because fluctuations in thickness will produce a spread in V_{TH} . A one monolayer (~0.5 nm for silicon (100) planes) variation in the film thickness as a result of fabrication uncertainties could cause a 20% variation in body thickness for the 3 nm body. Figure 6.4, is a plot of ΔV_{TH} vs. T_{Si} around T_{Si} = 3nm from 1D non-self-consistent (empirical expression based on a rectangular quantum well model [25]), 1D self-consistent, and 2D self-consistent solutions. Both, non-self-consistent 1D and self-consistent 1D solutions under-predict the threshold voltage variation with body thickness as they do not account for 2D short channel effects. However, self-consistent 2D simulations show that the threshold voltage change is within 50 mV for a 20% variation in T_{Si} about T_{Si} = 3nm, which is the assumed nominal body thickness.

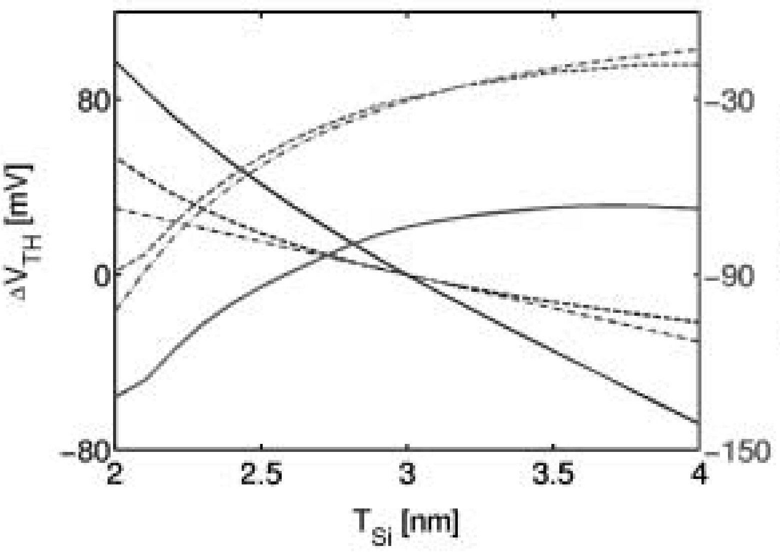


Fig. 6.4. Threshold voltage sensitivity vs. body thickness around T_{Si} = 3nm. Comparison between 1D analytical predictions (based on ideal rectangular well confinement, dashed lines with dots), 1D self-consistent MOS simulations (dashed lines), and 2D self-consistent device (solid lines) simulations is shown. V_{TH} fluctuations are shown on the left, and V_{TH} sensitivity is shown on the right.

With the silicon film thickness, T_{Si}, and physical insulator thickness, T_{ins} selected based on the aforementioned analysis, we look for suitable gate materials that would provide a nominal threshold voltage of ~ 0.15 V ($0.25 \times V_{DD}$). This threshold voltage was selected to account for subthreshold degradation due to channel length and silicon film thickness variations as discussed in previous paragraphs. Figure 6.5 shows the charge density vs. gate voltage relation for five different gate designs: (I) symmetric N^+-N^+ poly-germanium, (II) symmetric N^+-N^+ poly-silicon, (III) asymmetric N^+-P^+ poly-germanium, (IV) asymmetric N^+-P^+ poly-silicon, and (V) symmetric N^+-N^+ poly-silicon with a back gate bias of –0.9 V. It should be noted that all of the results in Fig. 6.5, are from 1D Schrödinger-Poisson simulations and do not account for 2D short channel effects that lower the threshold voltage. Thus the threshold voltage predictions in Fig. 6.5 have to be corrected to account for these 2D effects. On making the corrections (about 0.1 V), it is seen that due to the use of an intrinsic body and a thin effective oxide, the symmetric N^+-N^+ poly-silicon and poly-germanium gates yield low threshold voltages. The asymmetric

N^+-P^+ design, however, with its strong quantum confinement results in a very high threshold voltage for both poly-silicon and poly-germanium gates [47]. It seems impossible to obtain an appropriate gate material that would provide the desired threshold voltage. Therefore the threshold voltage requirement of ~ 0.15 V is met by using a symmetric N^+-N^+ poly-silicon gate with a back gate bias (-0.9 V) increasing circuit design complexity. It should be noted that this threshold voltage meets the off-current requirement (160 nA/μm) in the worst case when both, channel length and body thickness exhibit maximum variation (see Table 6.2).

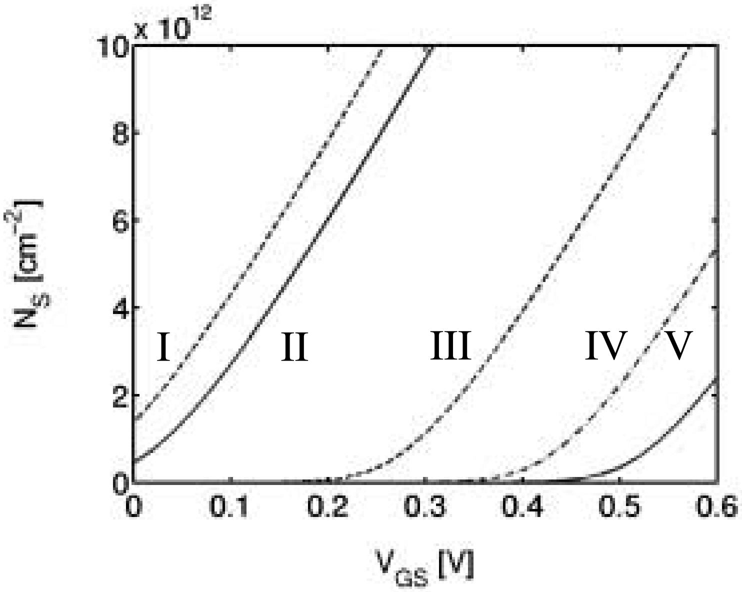


Fig. 6.5. Inversion electron density vs. gate bias for five different gate designs. T_{Si} = 3nm, and T_{OX} (*Phys*) = 3nm (T_{OX} (*Eff*) = 0.6 nm) for both top and bottom gates. I) N^+/ N^+ poly-germanium gate, II) N^+/ N^+ poly-silicon gate, III) N^+/P^+ poly-germanium gate, IV) N^+/P^+ poly-silicon gate, and V) N^+/N^+ poly-silicon gate with a back gate bias of -0.9V. Note that Fig. 6.5 has to be corrected to account for 2D SCE degradations. On making the corrections, threshold voltage requirement of ~ 0.15 V can only be met by using a symmetric N^+ /N^+ poly-silicon gate with a back gate bias (-0.9 V).

It has been pointed out in the literature that the use of an asymmetric gate design, especially in case of thick body DG MOSFETs degrades short channel immunity [114]. Channel charge is primarily confined to a region near one gate and does not effectively screen the drain electric field from penetrating the source. However, the use of a thin body in our device structure ensures that despite using an asymmetric gate and shifting the charge centroid towards one gate, we still have enough body charge to effectively screen the effect of the drain. This ensures that we obtain the same degree of short channel immunity with our asymmetric device structure as we would if a symmetric device structure had been used instead.

Due to the use of a very thin effective gate oxide, it may be expected that poly-depletion will strongly degrade the total gate capacitance. As shown in Fig. 6.6, which plots the inversion layer density vs. gate voltage for different levels of polysilicon doping, this degradation is not pronounced (~ 10%) for two reasons: 1) The use of an asymmetric structure ensures that the bottom gate is accumulated and contributions to poly depletion are from the top gate alone and 2) The degradation of the total gate capacitance is primarily due to the silicon film rather than the poly or the oxide because of the use of an intrinsic body and the quantum mechanical nature of the charge distribution within the body. Thus, although the effective oxide thickness is very small, poly-depletion effects are not pronounced. It should be noted that different levels of poly-silicon doping result in different work functions for the gate. Therefore in order to assess the effect of poly-depletion alone, the plots in Fig. 6.6 were corrected to account for differences in threshold voltage. The effective gate capacitance (C_{EFF}) as extracted from the slope of N_S versus V_{GS} in Fig. 6.6 is only ~50% of the oxide capacitance (C_{OX}) as defined by $\kappa\varepsilon_o / T_{OX}$. C_{EFF} is the overall gate capacitance of a series combination of C_{inv} and C_{OX}, as the oxide and inversion capacitances become comparable, C_{EFF} is degraded considerably. This result is consistent with that reported in [28]. Also note that for such a thin body T_{Si} = 3 nm, the gate symmetry has little effect on the electron distribution. The electron centroid is very close to the body middle line. Even a symmetrical gate design (with midgate workfunction gate contacts) can not make appreciable difference in C_{inv}, therefore on C_{EFF}.

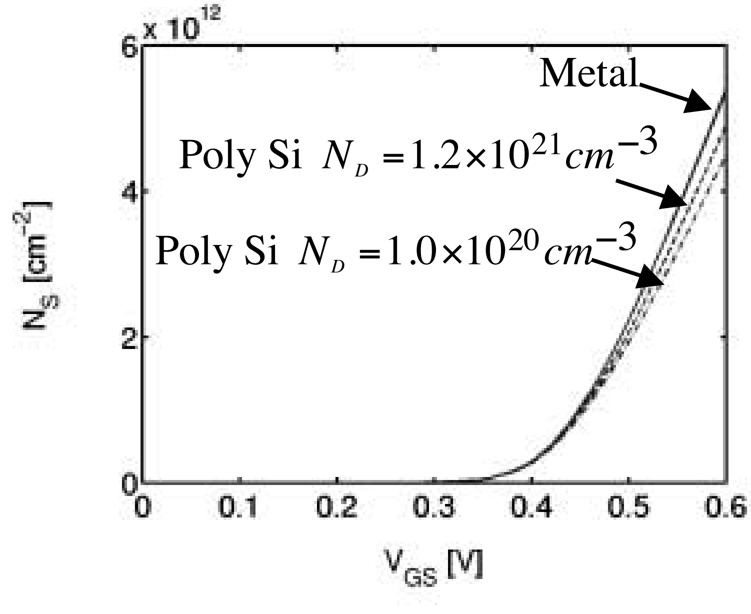


Fig. 6.6. Poly-depletion effects on inversion electron density are shown through 1D MOS simulations. Compared to metal gates (solid line), 1×10^{20} cm^{-3} poly-silicon gates (dashed line) and 1.2×10^{21} cm^{-3} poly-silicon gates (dashed line) show ~20% to ~10% reductions in the effective gate capacitance.

Table 6.1
Structural specifications of the simulated model double gate MOSFET.

Parameter	Value
L_G (*metallurgical*)	10 nm
T_{Si}	3.0 nm
T_{ins} (*physical*)	3.0 nm
T_{OX} (*effective*)	0.6 nm
L_{SD}	10 nm
UL	0.0 nm
N_{SD}	10^{20} cm^{-3}
N_B	10^{16} cm^{-3}
σ_{SD}	1.0 nm/dec

In Table 6.1, $T_{ins}(physical)$ is the physical gate insulator thickness, while T_{OX} (*effective*) is the equivalent SiO_2 thickness. A dielectric constant of 19.5 is assumed for the gate oxide. N_{SD} and N_B are source/drain and channel doping concentrations, respectively. L_{SD} and UL are source/ drain extension and gate underlap lengths, σ_{SD} is the gradient of Gaussian source/drain profile. Discussions on UL and σ_{SD} can be found in Section 3.3.

6.3.2 *I-V* characteristics in the ballistic limit

We now focus on the physics of ballistic transport for the device structure presented in Table 6.1. Figure 6.7 shows the I_{DS} vs. V_{GS} characteristics for the nominal device in the ballistic limit at different drain voltages from both Boltzmann (semiclassical) and Green's function (quantum) simulators. The semiclassical solution accounts for quantum effects in the confinement direction alone, while the quantum solution accounts for quantum effects in both the confinement and the transmission directions. The predicted off-current from the quantum simulator is higher as a result of tunneling through the source barrier, an effect that not captured by the semiclassical simulator. The current characteristics indicate that transport in this nanoscale transistor is essentially classical. (Quantum confinement effects normal to the channel lead to substantial V_{TH} shifts, but quantum effects along the channel are weak). Quantum effects do not affect device electrostatics as seen from the same degree of DIBL observed in both quantum and semiclassical solutions.

Figure 6.8 shows the I_{DS} vs. V_{DS} characteristics in the ballistic limit at two different gate voltages. It can be seen that at low V_{GS} the quantum solution always predicts a lower current as compared to its semiclassical counterpart. However, at high V_{GS} the quantum currents that begin lower actually become higher than their corresponding semiclassical values as V_{DS} is increased. These trends can be understood when one looks at the areal electron density and injection velocity at the top of the potential barrier at high V_{GS} (Fig.

6.9). The 2D charge at the top of the barrier consists of both tunneling as well as thermionic components. Thus, although the quantum simulator predicts a higher charge at the top of the potential barrier due to source tunneling, these carriers have low velocities. Therefore the current from quantum simulations is initially low (at low V_{DS}). However, as we continue to increase V_{DS}, the low source barrier also shrinks in width. Therefore the amount of tunneling charge increases. Thus although these carriers have low velocities, the tunneling charge contribution becomes significant. This finally results in a higher quantum current in the on-state for this device.

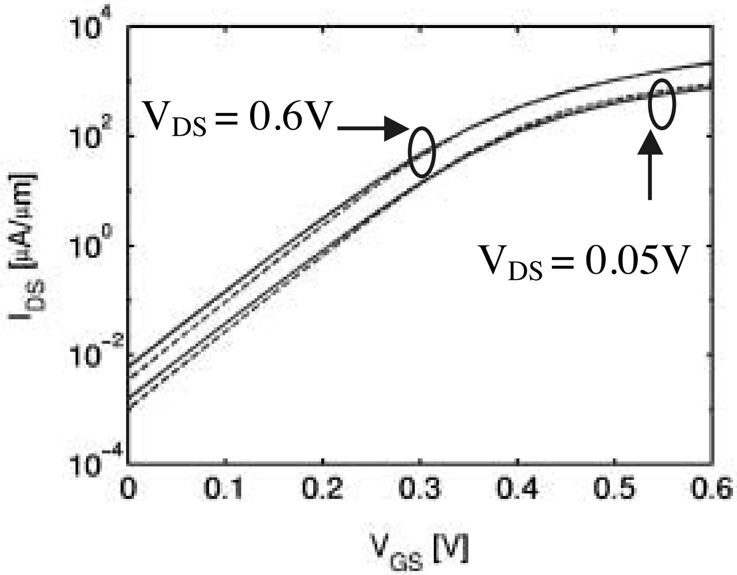


Fig. 6.7. I_{DS} vs. V_{GS} based on the classical (dashed lines) and quantum (solid lines) ballistic simulations for the L_G = 10nm, T_{Si} = 3nm device. Biased back gate is used to obtain V_{TH} of ~0.15 V. Drain biases are 0.05 V and 0.6 V, respectively. Quantum simulations show similar amount of DIBL (75mV/V) as compared to classical simulations, but larger subthreshold swing and higher off-state current due to quantum tunneling through the source-to-channel barrier.

Note that in a previous study [62], the source to channel barrier was relatively high. The tunneling charge contribution was not as significant as we show here, resulting in a quantum on-current that was lower than the classical result. This is similar to the low V_{GS} case (V_{GS} = 0.5 V) as seen in Fig. 6.8. It should also be noted that the quantum charge density shows some oscillations due to interference effects. However, when Poisson equation is solved for the potential, these local charge oscillations are washed out, resulting in a smooth potential profile. Electron transmission (current), which is a function of the overall potential profile varies smoothly with increasing V_{DS}. Thus the quantum current does not reflect the observed local charge oscillations.

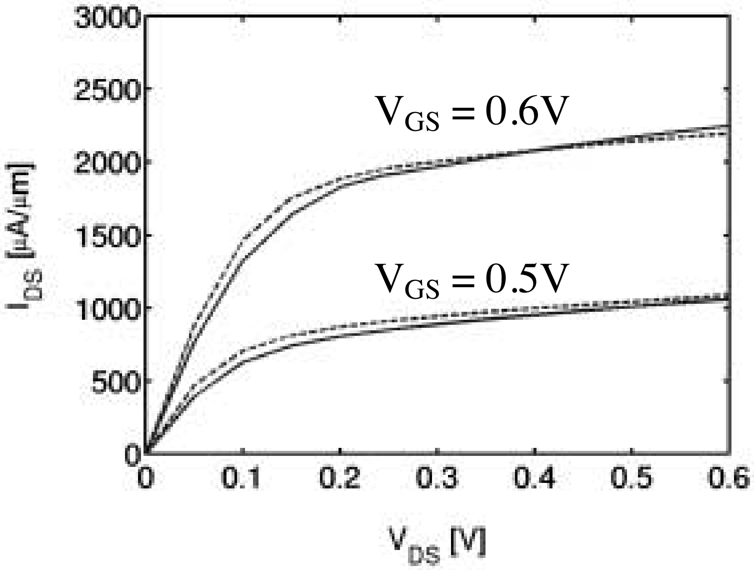


Fig. 6.8. I_{DS} vs. V_{DS} using classical (dashed lines) and quantum (solid lines) ballistic simulations for the L_G = 10 nm, T_{Si} = 3 nm device. Gate biases are 0.5 V and 0.6 V, respectively.

Figure 6.10 shows the energy profiles along the channel for both the primed and the unprimed subbands and Fig. 6.11 their corresponding charge contributions. In the channel region, the charge density is relatively low compared to the source/drain regions. Therefore the first unprimed subband accounts for almost all of the channel charge. This results in the high injection velocities observed in Fig. 6.9. However, due to the high donor density in the source/drain, we see that the higher subbands also contribute significantly to the total 2D charge in those regions, thus rendering a single subband treatment inadequate. From a semiclassical point of view, the four primed subbands are electrostatically equivalent although electrons in two of these subbands respond with a heavy mass while those in the other two respond with a light mass in the channel direction. This is because the four primed subbands, within the semiclassical framework, have the same density of states effective mass. However, it can be seen from the quantum charge density profiles, that the four primed subbands are actually not equivalent. Carriers in those primed subbands that respond with a heavy mass behave in a classical manner as compared to those that respond with a light mass in the transmission direction. The charge density contributions from the light subbands (primed) are pushed away from the source barrier along the channel and also exhibit a higher degree of tunneling into the forbidden energy regions due to quantum effects along the channel direction. Other researchers also reported this effect, and they believed that this effect could improve short channel effects because the charge repulsion from the channel barrier makes the channel length effectively longer [115]. We do not observe the improvement, and actually find that electron tunneling into the channel forbidden energy regions increases the off-current.

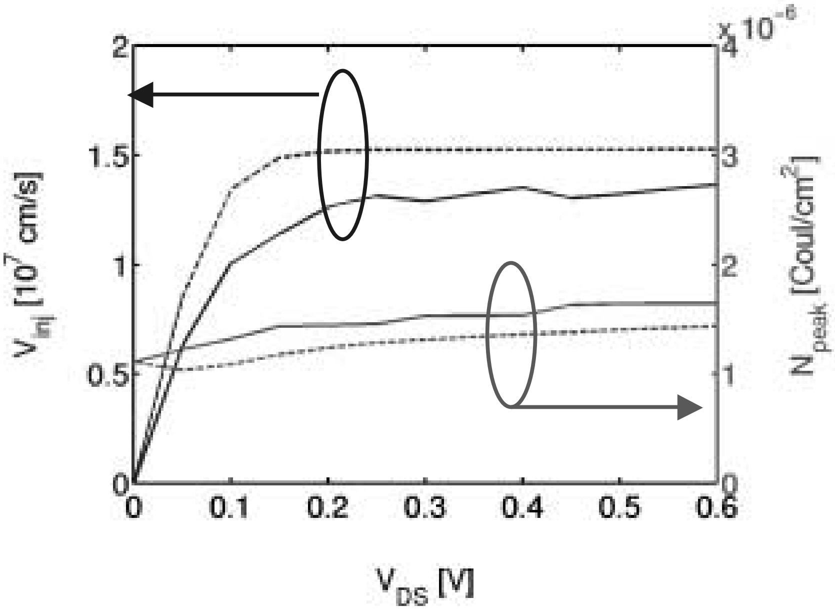


Fig. 6.9. Electron injection velocity (on the left) and areal density (on the right) at the channel beginning region are plotted vs. V_{DS}. V_{GS} is 0.6 V. The solid lines indicate results from quantum ballistic simulations, while the dashed lines represent results from classical ballistic simulations.

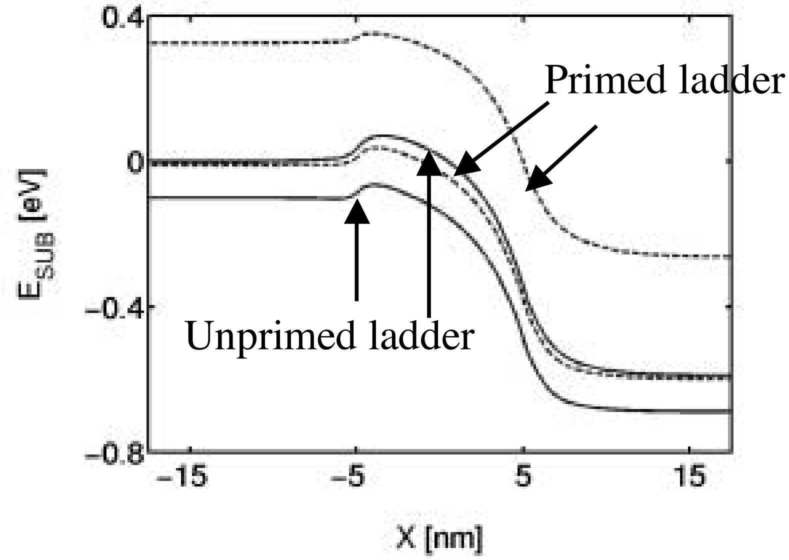


Fig. 6.10. Subband profiles along the channel simulated using the quantum ballistic model at $V_{GS} = V_{DS} = V_{DD}$. Solid lines represent the unprimed ladder of subbands (heavy mass in the confinement direction) and dashed lines represent the primed ladder of subbands (light mass in the confinement direction).

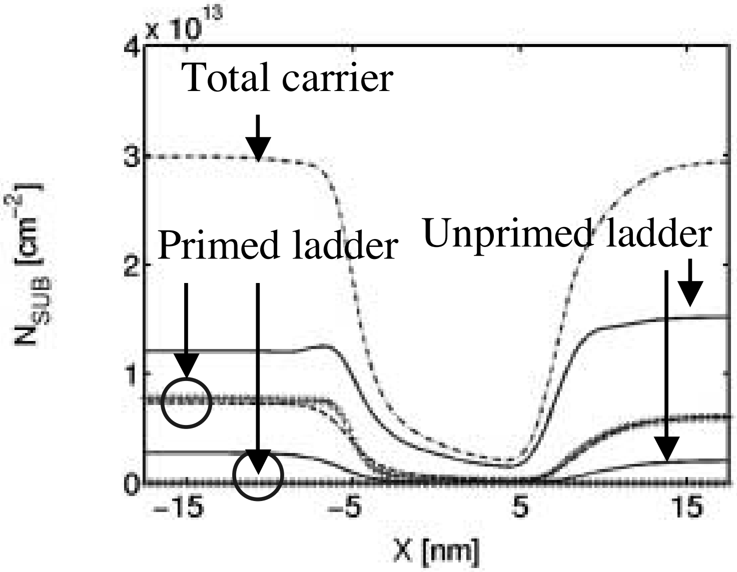


Fig. 6.11. 2D charge density profiles along the channel simulated using the quantum ballistic model at $V_{GS} = V_{DS} = V_{DD}$. The total 2D electron density in the source/drain regions equals the total dopant density ensuring charge neutrality. Due to different effective masses in the channel direction, primed subband charge distributions exhibit varying degrees of quantum effect (crossed and dashed lines).

6.3.3 *I-V* characteristics with scattering

The first step in simulating dissipative transport through Büttiker-probes is to calibrate the state lifetime in the quantum simulations to an equivalent mobility. We assume that the primary scattering mechanisms in our model device are due to ionized impurities (μ_I) and surface roughness (μ_{SR}). The resultant mobility is obtained using Mathissen's rule,

$$\frac{1}{\mu} = \frac{1}{\mu_I} + \frac{1}{\mu_{SR}}. \tag{6.6}$$

In the channel region, the doping concentration is low, and the use of an ultra-thin body causes the charge centroid to be very close to the Si/SiO$_2$ interface. Therefore channel mobility is expected to be primarily determined by surface roughness scattering. In order to quantify surface scattering in our device, we performed 1D simulations in bulk MOS capacitors to evaluate the distance (T_{inv}) of the charge centroid from the oxide/silicon interface at different average vertical electric fields (E_{eff}). The effective surface roughness scattering mobility (μ_{SR}) is computed using the Bell labs mobility model in the universal region [116]. Since surface roughness scattering is primarily characterized by T_{inv}, and both T_{inv} and μ_{SR} are functions of E_{eff}, we can directly relate μ_{SR} to T_{inv}. Although this derived mobility applies to a bulk MOSFET, we assume that it provides a rough estimate of carrier mobility in our ultra-thin body in regions where surface roughness scattering dominates. It is worth noting that the computed μ_{SR} vs. T_{inv} curves do not show strong universality for different levels of bulk doping concentrations.

The curve shown in Fig. 6.12 is simulated using a doping concentration of 5×10^{18} cm^{-3} and should be viewed as an average estimation. Using this approach, we estimate the electron mobility due to surface roughness scattering in the T_{Si} = 3 nm double gate MOSFET to be 200 cm^2/V-s. This mobility value was extracted at a high gate bias (V_{GS} =

0.6 V) which corresponds to the on-state of device operation. The dashed line in Fig. 6.12 indicates the centroid of electrons from the top Si/SiO$_2$ interface at $V_{GS} = 0.6$ V.

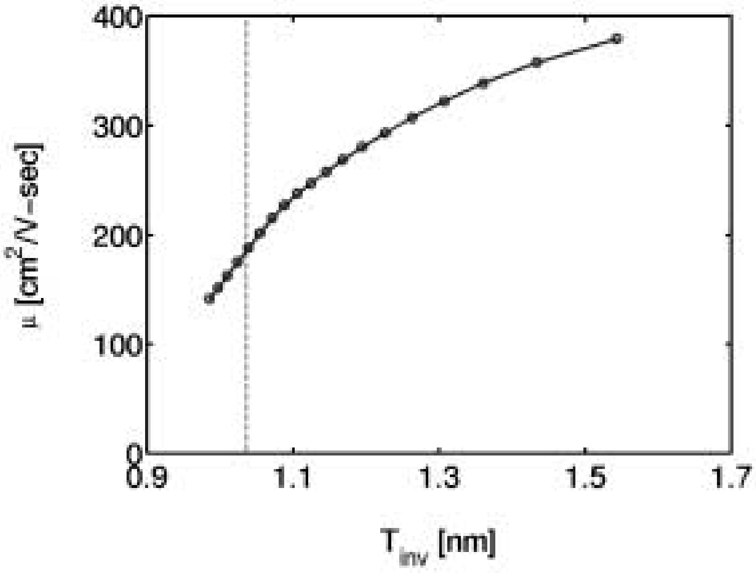


Fig.6.12. Mobility vs. electron inversion layer thickness is plotted by simulating a 1D bulk MOS capacitor using the Bell Labs model [116]. The dashed line indicates the average distance of inversion layer electrons from the top Si/SiO$_2$ interface in our model device at $V_{GS} = V_{DD}$. The estimated channel mobility in our device is 200 cm^2/V-s.

In the source/drain regions, the doping concentration is high. The Caughey-Thomas model was used to obtain mobility as a function of doping density [117]. High doping concentrations in the source/drain regions and the use of a thin body, result in mobility degradation (mobility as low as ~50 cm^2/V-s) due to surface roughness as well as ionized impurity scattering in these regions. Thus the use of a position dependent mobility enables us model the source/drain extension and tip resistances accurately.

It has been pointed out in the literature that source/drain abruptness and gate overlap are important device design parameters in case of end-of-the-roadmap generation transistors [22, 112]. Therefore, we quantify the effect of gate overlap and source/drain junction abruptness on the performance of our model device using the scattering model outlined in Sec. II. A source/drain abruptness of 1-2 nm/decade using Gaussian doping profiles has been examined. The gate overlap is varied from –2 nm to +4 nm (negative overlap implies that the gate actually underlaps the channel region) with the metallurgical junction length fixed at 10 nm. It is clearly seen from Fig. 6.13, that a more abrupt junction (1 nm/decade) gives a better on-current performance (~ 20%) due to reduced parasitic resistances (mainly the tip resistance). However, the on-current variation with gate overlap does not exhibit a monotonic trend. When the gate underlaps the channel region, it cannot effectively modulate the source-to-channel barrier. This implies an increased tip resistance that accounts for the observed decrease in on-current. As the gate overlap is increased, we find that the on-current that initially increases, begin to decrease at high gate overlap values. This trend can be explained from an electrostatic point of view. The effect of increased gate overlap, is to flatten the potential profile over a longer distance as we move from source to drain. This reduces the effective channel electric field at the beginning of the channel thus increasing the channel reflection coefficient leading to a reduced on-current [55].

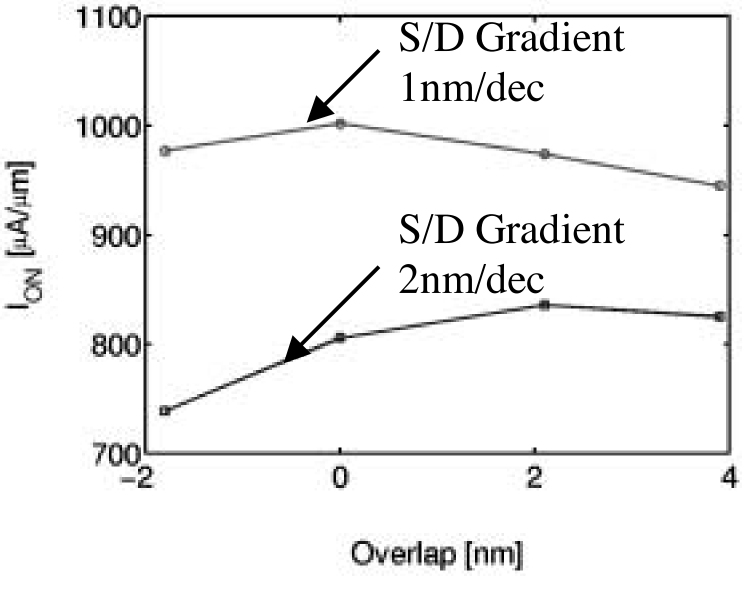


Fig.6.13. Device performance vs. gate overlap/underlap is plotted for two different source/drain junction gradients. Steep gradients improve device performance significantly. Optimized gate alignment (neither too much overlap or underlap) is also important for better performance.

Figure 6.14, is a plot of the sheet resistivity vs. position along the channel. Sheet resistivity is derived at low V_{DS} using the following expression [15],

$$\rho_{sh} = \frac{\partial V / \partial x}{I_{DS} / W} \tag{6.7}$$

where, *V(x)* is the Fermi potential at a point x along the channel and I_{DS} is a constant independent of x as required by current continuity and evaluated in the linear region (V_{GS} = 0.6 V, V_{DS} = 0.01 V). The non-uniform sheet resistivity in the channel direction can be divided into the following regions: 1) quantum contact resistance region, 2) source/drain extension resistance region, 3) tip resistance region and 4) channel resistance region. In the case of a graded junction, it can be seen that the tip resistance is higher and spreads to a greater degree into the source/drain extensions as compared to an abrupt junction. This enables the gate voltage to modulate the tip resistance to a greater extent in case of abrupt junctions. Thus junction abruptness is necessary to obtain high on-currents (Fig. 6.13).

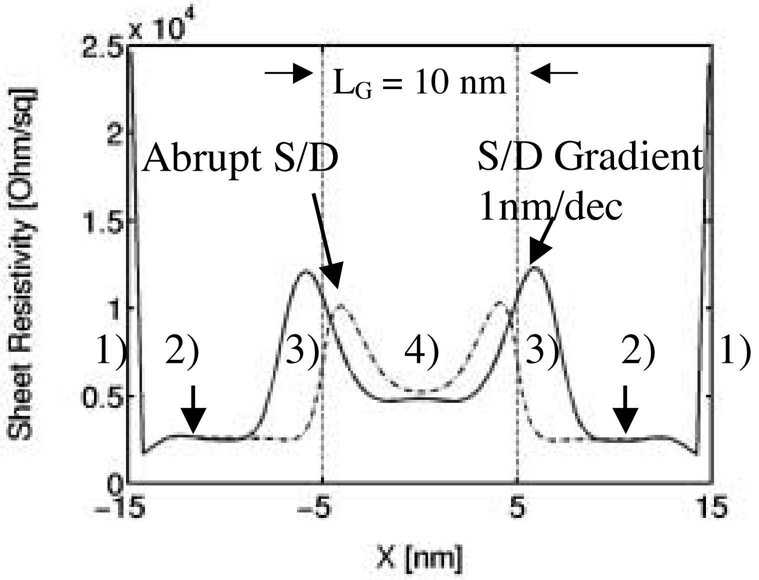


Fig.6.14. Sheet resistivity is plotted along the channel in the linear region ($V_{GS} = V_{DD}$, V_{DS} = 10 mV). The various components of the device resistance are: 1) quantum contact resistance 2) source/drain extension resistance 3) tip resistance and 4) channel resistance. It can be clearly seen that tip resistance contributes significantly to the overall device resistance for the L_G = 10 nm model device.

It should be noted that non-equilibrium conditions prevail at the interface between the contacts and the active device region in order to maintain current flow through the device. This implies a discontinuity in the Fermi potential at the contact/device interface thus resulting in the observed quantum contact resistance [28]. In the presence of strong scattering within the device, the quantum contact resistance represents a minor contribution to the overall device resistance. However, as the active device regions become more and more ballistic, contributions from the quantum contact resistance become significant. In the ballistic limit, the quantum contact resistance is the only resistance that is present and accounts for the finite ballistic current [28].

Source/drain extension resistances pose important design issues for the future generation MOSFETs. In our model device the use of a thin body and high doping concentration in the source/drain extensions, results in a low mobility (~50 cm^2/V-s) and high resistance in these regions as compared to a bulk device with deeper source/drain junctions. Also, the tip resistance is a significant fraction of the overall device resistance (~50%). Therefore, unlike long channel devices, the channel resistance does not dominate the I-V characteristics of ultra-scaled transistors. The total resistance in the linear region for our model device is ~160 Ω-μm. This value is 30% of V_{DD}/I_{ON} (ITRS-99 target is 10% of V_{DD}/I_{ON}). It is clear that source/drain and tip doping engineering is necessary to reduce the linear region resistance. Note that in all of our simulations this far, there is no real metal-semiconductor contact resistance. Inclusion of this resistance would further reduce the performance of our device.

The applied voltage is equal to the Fermi level offset between the source and drain contacts. This voltage is dropped across the various resistance components mentioned in the previous section. Due to the high channel resistance at low V_{GS}, most of this applied voltage is dropped in the channel region as shown in Fig. 6.15. The voltage drop in the source/drain extensions and at the contacts is relatively low. However, as the gate voltage is increased, the channel conductivity increases. Therefore the voltage dropped in the channel region is reduced. This leads to a flattened Fermi potential profile in the channel and large voltage drops in the contact, source/drain and tip regions. In the ballistic limit, there is no mechanism responsible for an internal voltage drop and all of the applied voltage is dropped across the contact/device interface resulting in a finite ballistic current.

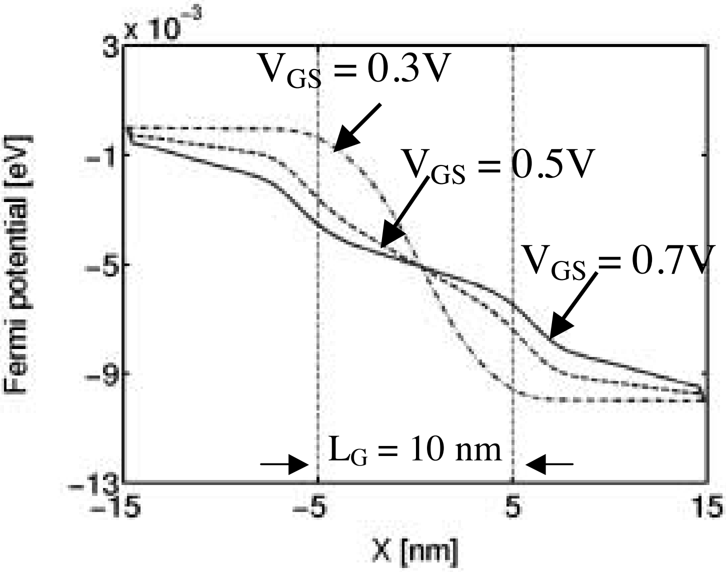


Fig.6.15. Fermi energy along the channel is plotted at different gate voltages. The source/drain junction gradient is 1nm/decade and the drain voltage is 10 mV. As the gate voltage is increased, channel resistance is reduced and contact resistance is increased.

Real devices operate below the ballistic limit because of carrier scattering. Mobility, in our device is estimated using eqn. 6.6. In the channel region, carrier mobility is primarily determined by the extent of surface roughness scattering as the body doping is extremely low. However, in the heavily doped source/drain extensions, carrier mobility is a function of both surface roughness and ionized impurity scattering. The I_{DS} vs. V_{DS} characteristics, assuming a channel mobility of 200 cm^2/V-s and a source/drain junction gradient of 1nm/decade is shown in Fig. 6.16.

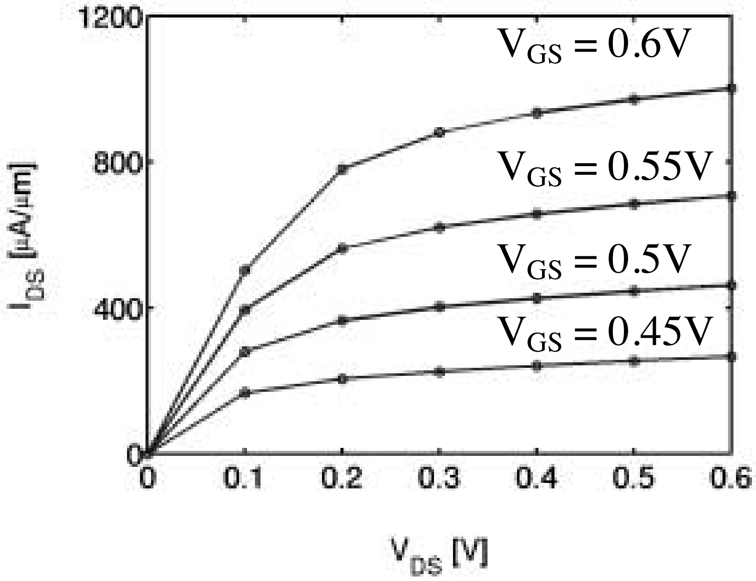


Fig. 6.16. The I_{DS} vs. V_{DS} characteristics for the L_G = 10 nm, T_{Si} = 3 nm device, assuming a channel mobility of 200 cm^2/V-s and a source/drain junction gradient of 1nm/decade.

It is in our interest to examine the on-current performance of our device vs. channel mobility (see Fig. 6.17), assuming that superior interface engineering could actually reduce the extent of surface roughness scattering in the channel region. The pure ballistic result (no scattering anywhere within the device region) is also plotted in Fig. 6.17 for comparison purposes. It can be seen the on-current initially increases linearly with increase in channel mobility. However, as channel mobilities increase beyond 100 cm^2/V-s, carrier transport in the channel region becomes quasi ballistic and insensitive to the channel mobility. The on-current in this regime, is limited by the parasitic source/drain and tip resistances and clamped around 1000 μA/μm (no metal–semiconductor contact resistance included in the simulations). These results indicate that the ITRS-99 on-current target of 1500 μA/μm for double gate MOSFETs cannot be met even under the assumption of ballistic transport in the channel region of the L_G = 10 nm MOSFET. It seems that for future generation transistors, engineering the source/drain, tip and contact regions to reduce parasitic resistances is more important as compared to channel engineering. A possible solution would be the use of an extremely small extension region whose length would be ultimately limited by parasitic gate-to-source/drain capacitance requirements, by reducing the source/drain junction abruptness to less than1 nm/decade, and by employing big fanned out source/drain regions for large contact areas.

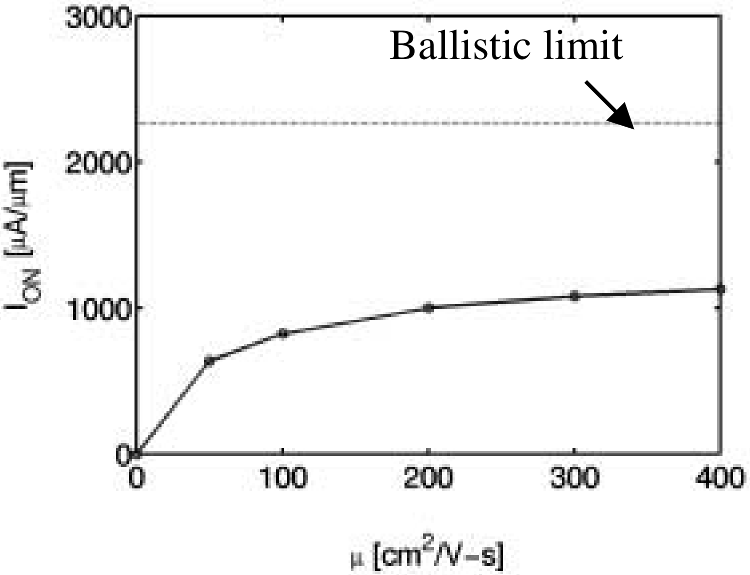


Fig.6.17. Device performance vs. channel mobility including the effect of parasitic resistances is compared to the ballistic limit (no contact resistance). It is clear that parasitic resistances limit device performance because very high channel mobilities yield on-currents much below the ultimate limit.

6.3.4 Gate tunneling

Quantum tunneling into the insulator affects the electrostatics within the device. and generates gate leakage current. From the electrostatic point of view, electron wavefunction penetration into the insulator region, relaxes the body quantum confinement resulting in a reduced threshold voltage. This reduction in threshold voltage is more pronounced in case of ultra-thin bodies because higher electron confinement energies in these ultra-thin bodies can cause significant penetration (in our model device with a body thickness of 3 nm, this threshold voltage shift was not significant). This threshold voltage reduction increases the off-current. In the on-state, electron penetration effects are stronger due to the high gate voltage resulting in an increased effective body thickness (wider quantum well). Therefore, the separation between subbands is reduced, increasing the carrier concentrations in high energy subbands. The 2D charge density in the inversion layer is increased but the average injection velocity in the channel direction is reduced compared to a device exhibiting idealized body confinement. Since the on-current is a product of the 2D charge density and the average injection velocity, the overall effect of electron penetration is a relatively unchanged on-current.

The constant field scaling method maintains the same V_{DD}/T_{OX} (physical) for successive technology generations to maintain device reliability. As the physical oxide thickness is reduced while maintaining a constant electric field, the gate leakage current increases exponentially. It has been reported in the literature that SiO_2 cannot be used as a gate insulator when the physical oxide thickness is less than 1 nm [112]. A high κ gate material has to be used instead. An acceptable level of gate leakage density for a technology generation is estimated based on the acceptable off-current [2]. It has been recently reported that a gate leakage density of 100 A/cm^2 is acceptable to meet the performance requirements for nanoscale MOSFETs [112]. We use this current density level as a reference to assess gate leakage within our model device. The use of a high κ material in our device structure, will cause the band off-set between the silicon film and the gate insulator to be reduced. A high κ material also enables us to use a thick insulator. Due to the increased thickness and reduced barrier height of the gate insulator, we assume that Fowler-Nordheim tunneling will be the primary mechanism inducing gate leakage [15, 66]. The conduction band offset between the semiconductor and insulator was calibrated to obtain the reference leakage current density of 100 A/cm^2.

Fig. 6.18 illustrates the leakage current distribution along the gate for our model device generated using the Fowler-Nordheim model as a post processing operation. It is clear from Fig. 6.18 that the leakage current distribution is highly non-uniform. The current density is negligible over most of the channel region, but attains very high values near the drain side of the device. This is because vertical electric fields are strongest near the drain side due to the high gate to drain bias in the off-state. Therefore, estimating the total gate current assuming a uniform leakage current density is erroneous. In fact the gate leakage current assuming a uniform leakage current density of 100 A/cm^2 is 10 nA/μm, while our 2D calculations that capture the high current density contributions near the drain actually yield a value as high as 40 nA/μm. One might reduce gate leakage by engineering the insulator at the drain end of the device. Such engineering would leave the source end unaffected, thus resulting in relatively unchanged on-characteristics.

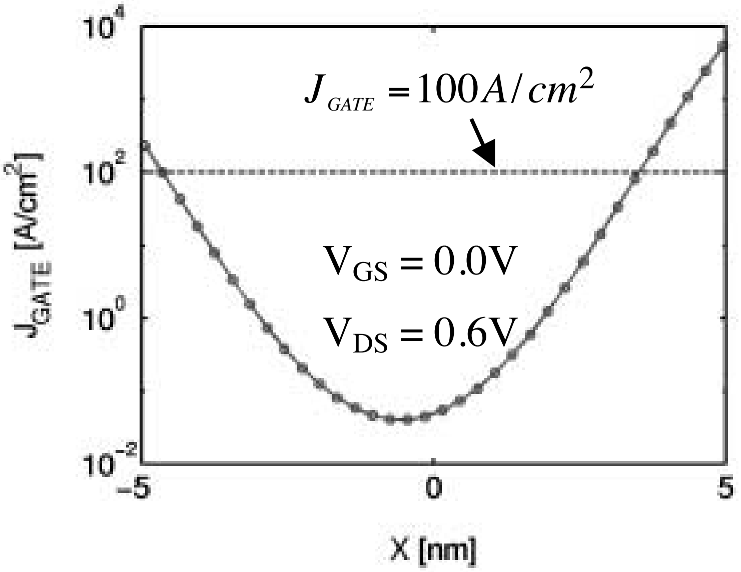


Fig.6.18. Off-state gate leakage is plotted along the channel assuming a leakage density of 100 A/cm^2. It should be noted that the current distribution is non-uniform with most of the leakage occurring at the drain side of the device.

The maximum permissible off-current for the year 2014 generation transistor specified by ITRS-99 at an operating temperature of 25 ^{0}C is 160 nA/μm. The off-current could be much higher at 100 ^{0}C, due to a degradation of subthreshold characteristics with increasing temperature. For the model device under study the maximum permissible off-currents have been presented at 25 ^{0}C and 100 ^{0}C in Fig. 6.19. We compare the total gate leakage current against I_{OFF} at various gate leakage current density values for our model device. It can be seen from Fig. 6.19 that a gate leakage density of 100 A/cm^2 provides a tolerable level of leakage current (40 nA/μm) compared to I_{OFF} (160 nA/μm). Thus, a gate leakage density of 100 A/cm^2 is feasible from a design point of view provided insulator reliability is guaranteed. Figure 6.19 also shows that a gate leakage density of

1000 A/cm^2 results in a leakage current which is significantly less than I_{OFF} for operating temperatures of 100 ^{0}C.

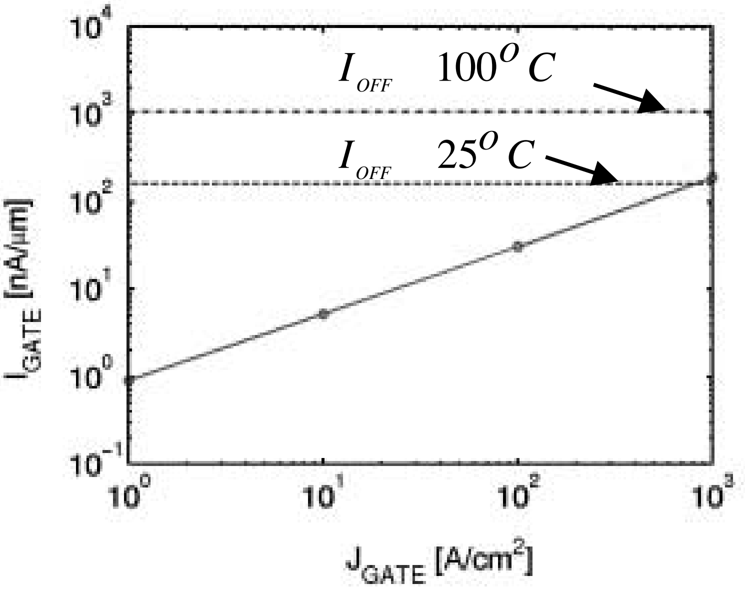


Fig.6.19. Estimated gate leakage current values are plotted vs. gate leakage density in the off-state, V_{GS} = 0.0 V, V_{DS} = 0.6 V. ITRS-99 specifications for the year 2014 transistor generation is also shown (dashed lines) for two operating temperatures. A leakage density of 100 A/cm^2 provides an acceptable degree of gate leakage in our model device.

6.4 Discussion

It is clear from the results presented in the previous sections that several issues need to be considered when designing an L_G = 10 nm DG MOSFET. The final performance of our model device is summarized in Table 6.2. In order to achieve good short channel characteristics at such small channel lengths, a thin silicon body needs to be used. It has been reported in the literature that a 5 nm body can meet the ITRS-99 specification of I_{OFF} for a channel length of 10 nm [22]. In our study we find that if fabrication uncertainties are considered, the permissible value of body thickness needed to meet the off-current requirement is reduced to 3 nm. For this choice of silicon film thickness, a 10% variation in the gate length along with a single monolayer variation in the body thickness results in a net threshold voltage degradation of ~80 mV. This degradation is ~50% of the nominal threshold voltage. Had a thicker body been used, this variation would be more pronounced as a result of reduced short-channel immunity. It appears that a very high degree of fabrication accuracy will be required to produce such small transistors.

Table 6.2 also indicates that for a 3 nm body with a nominal threshold voltage chosen to be ~0.15V, the nominal off-current at room temperature could be as low as 6 nA/μm. However, in the worst case (10% reduction in channel length and one monolayer increase in body thickness) the off-current could be as high as 130 nA/μm which is still below the ITRS-99 requirement.

It has been shown that an asymmetric N^+-P^+ polysilicon gate design provides an acceptable threshold voltage for a silicon film thickness of 10 nm [47]. However, as the silicon film thickness is reduced quantum confinement effects increase the threshold voltage. Our study indicates that as the body thickness is reduced, it is no longer possible to use an N^+-P^+ polysilicon gate, and one would have to resort to using exotic gate materials to meet the threshold voltage requirement. Searching for exotic gate materials as the body thickness is varied seems to be a very difficult task. Incorporating a new material into existing technology is even more practically formidable. An alternate solution is to employ a back-gate bias in order to adjust the threshold voltage. This introduces an additional power supply complicating the layout and circuit design unless a more tractable solution is obtained in the future.

The on-current in our model device is severely degraded as a result of quantum contact, source/drain extension and tip parasitic resistances. No contact resistance has been included in our study this far. A commercial simulator can be used to estimate the contact resistance by simulating the fanned out region of the contact as illustrated in Fig. 6.20. We assume an optimistic value of the contact resistance of 10^{-8} Ω-cm^2 as specified by the ITRS-99 for the year 2014 technology generation. The length of the fanned out region is chosen to be equal to the source/drain extension lengths. The calculation indicates that the contact resistance on either side of the device is 100 Ω-μm, thus resulting in a total parasitic source resistance of 180 Ω-μm. The contact resistance reduces the effective gate voltage resulting in a drop in the on-current. This reduced on-current can be estimated from the results plotted in Fig. 6.16 (includes all resistances except the contact resistance) using a bisection scheme. The maximum on-current we finally obtain from our model device is 650 μA/μm. Based on this on-current we find that the parasitic source resistance in our device is ~ 20% of V_{DD}/I_{ON}. Note that in making this estimate, we have assumed a very optimistic value of $\rho_c = 10^{-8}$ Ω-cm^2.

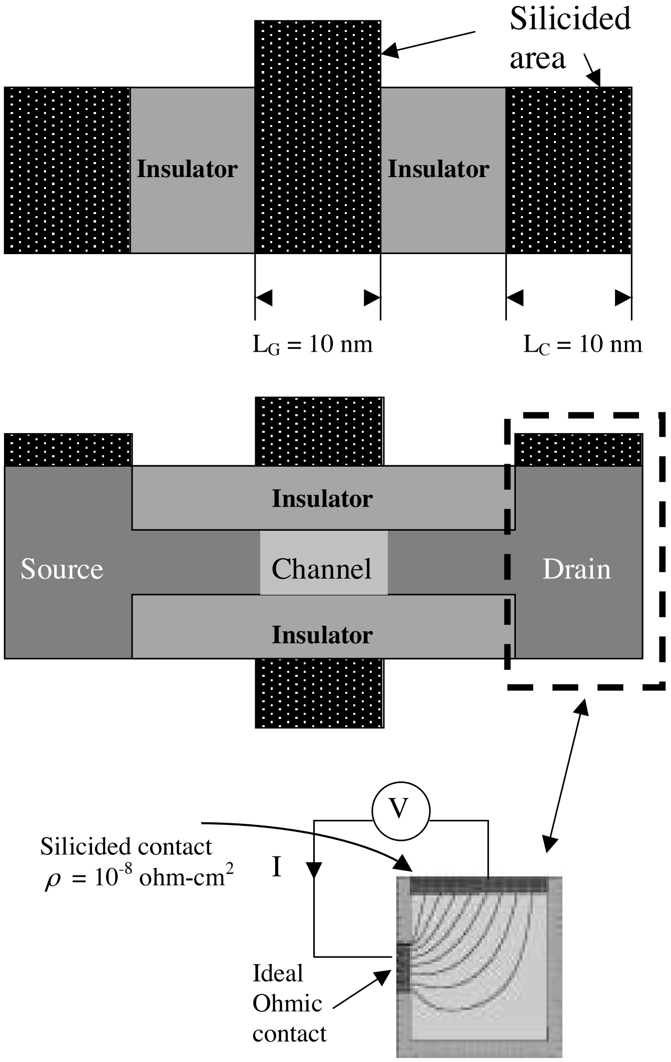


Fig. 6.20. Top and side views of the double-gate MOSFET followed by the simulation domain used to estimate the contact resistance. Current flow lines are indicated within the simulation domain.

One could reduce the source/drain extension resistances by shrinking the length of the source/drain regions. Such a solution, however, would introduce a large gate-to-source/drain capacitance degrading circuit performance significantly. Moreover the tip and contact resistances are unaffected by the reduction in source/drain extension lengths and will continue to degrade the on-current performance in any case. The parasitic resistance limits the final on-current to just ~30% of the ballistic limit in our model device. Furthermore, dimensional uncertainty also degrades the on-current, which in the worst case (10% increase in the channel length and one monolayer decrease in body thickness) could be as low as 500 μA/ μm. It seems that the critical issues affecting device performance of future generation transistors are device parasitics, which degrade the on-current and dimensional stability, which affect both the on and the off-currents.

Table 6.2
Computed performance of the model double-gate MOSFET.

Parameter	Nominal	Worst Case
V_{DD}	0.6 V	
V_{TH}	0.15 V	± 80 mV
S	75 mV/V	
DIBL	80 mV/dec	
C_{eff} / C_{OX}	~0.5	
I_{ON}	650 μA/μm	500 μA/μm
B = I_{ON} / I_{BALL}	30%	
I_{OFF} @ $25^{o}C$	6 nA/μm	130 nA/μm
I_{Gate} @ J_{Gate} = 100 A/cm^2	40 nA/μm	
R_{Par} @ *source*	180 Ω-μm	
$R_{Ch} \equiv V_{DD} / I_{ON}$	900 Ω-μm	

6.5 Summary

In this chapter we examined the device design issues for an n-channel double gate MOSFET with a metallurgical gate length of 10 nm. The device structure was engineered to meet the ITRS-99 specifications for the year 2014 transistor generation. Our simulations show that,

- Ultra-thin bodies (T_{Si} = 3 nm) and extremely thin effective oxides (T_{OX} (*effective*) = 0.6 nm) are needed in order to suppress short channel effects and maximize channel inversion charge density.

- 10% fluctuations in L_G and T_{Si} will lead to ~50% fluctuations in V_{TH} (as compared to ~15% for present-day technology).

- The quantum mechanical nature of the charge distribution in the body causes the inversion layer thickness to be comparable to the ultra-thin effective oxide thickness thus significantly reducing the effective gate capacitance ($C_{eff}\,/\,C_{OX}$ ~0.5).

- Use of an ultra-thin silicon body increases surface roughness scattering, which in turn reduces the channel mobility (< 200 cm^2/V-s).

- It is difficult to find a suitable gate material in order to achieve the right threshold voltage for this transistor generation. A possible solution would involve the use of a back gate bias, which complicates circuit design.

- Device performance is limited by parasitic resistances. In long channel devices, channel transport significantly affects device performance. However, at the length scale considered in our study, even the occurrence of quasi ballistic transport in the channel resulted in an on-current far below the ultimate ballistic limit.

- The junction tip and contact resistances provide the most significant contribution to the overall device resistance. Extremely abrupt source/drain junctions and large flared out contacts would be needed to minimize parasitic resistances. Accurate gate alignment would also help improve device performance.

- With V_{DD} = 0.6 V it is difficult to simultaneously achieve the on and off-current targets.

- The gate leakage current distribution is extremely non-uniform with most of the leakage occurring over a small region near the drain. Gate oxide engineering at the drain end will present an important design consideration as dielectric dimensions are scaled in the future.

In summary, we have performed a comprehensive simulation study of double gate MOSFETs which shows that channel length at the 10 nm scale should be feasible. The device was idealized in number of ways, notably the assumption of a high-κ gate dielectric and an especially low metal-semiconductor contact resistance. Even with these optimistic assumptions, however, our study shows that it will be extremely difficult to achieve the desired device performance targets. The study identified issues that will need to be addressed. New channel materials with high mobility would be especially helpful, but appears that these issues will also to be addressed by circuit design and system architecture.

7. CONCLUSION

7.1 Summary

This thesis addressed device physics, modeling and design issues of nanoscale transistors at the quantum levels. The device structures studied were double gate MOSFETs, with extremely scaled channel lengths (less than 30 nm) and body thicknesses (less than 5 nm). To accomplish the objectives, simulation tools were developed [49, 110]. The fundamental physics equations that were solved include the Poisson equation, which dictates the electrostatics in the devices, and the Schrödinger equation, which describes the transport and distribution of carriers in the devices.

The first stage of the work focused on simulations of 1D MOS structures, in which the 2D carrier gas was studied in the equilibrium state. The study of the 2D carrier gas revealed important quantum effects related to gate confinement in MOSFETs. We found that for DG MOSFETs with ultra-thin bodies ($T_{Si} < 3.0$ nm): i) sensitivity of V_{TH} to T_{Si} becomes significant, increasing the difficulty to control V_{TH}, ii) electron penetration into gate oxide layers becomes considerable, affecting both V_{TH} and C_{EFF}, iii) a one subband approximation is satisfactory in simulating devices, and iv) occupation degeneracy may strongly enhance υ_{inj}. We also found that for the n^+/p^+ asymmetric MOSFETs (T_{Si} ~10 nm), gate charge coupling can provide the desired V_{TH} and extraordinarily high C_{Eff}.

The second stage of the work constitutes the primary portion of this thesis. We developed a 2D simulator for nanoscale double-gate MOSFETs (nanoMOS) [110]. The program solves open-boundary transport problems using a non-equilibrium Green’s function (NEGF) formalism. We were one of the first reported groups to apply the NEGF in simulating MOSFETs. We examined both ballistic transport (Chapter 3) and dissipative transport (Chapter 4) in MOSFETs. In the former case, we focused on quantum transport features occurring in extremely scaled MOSFETs by contrasting quantum solutions to semiclassical solutions (Appendix B). In the latter case, we implemented and compared different scattering models in the effort of pursuing an appropriate dissipative transport modeling. We began the dissipative transport study by describing the Büttiker probe based scattering models where scattering centers are treated as reservoirs that change the energy or momentum of the carriers and not the total number of carriers in the system. Each scattering center is modeled through a perturbation strength characterized by a position dependent self-energy, which can be mapped onto an equivalent mobility. We then described a simplified phonon-electron scattering. The phonon-electron scattering model provides a more theoretical sound benchmark, helping us better understand a specific scattering process.

Using a model 10 nm double-gate MOSFET as a vehicle, we conducted extensive device physics and design simulation studies (see Chapter 5 and Chapter 6) with the approaches developed in this work. Important conclusions are: i) MOSFETs essentially operate as classical devices until the channel length shrinks below about 10 nm, when quantum tunneling through the channel barrier becomes significant, limiting device scaling, ii) solving the Green's function in a mode space representation can greatly reduce the size of the problem and provides good accuracy as compared to full 2D spatial discretization, iii) the Büttiker probe model captures the physics of dissipative transport in MOSFETs, and is computationally affordable at a PC level, iv) future devices may intrinsically operate very close to the ballistic limit, but their extrinsic performance will be limited by device parasitics and the desired level of dimensional control, rather than channel mobility engineering.

7.2 Future work

There could be three immediate extensions to this work, we describe them one by one as follows.

1) The NEGF approach can be used to simulate gate oxide leakage characteristics of 1D MOS structures. Currently, the widely used method in modeling gate leakage is the WKB approximation [65, 118]. In this method quantum transmissions through the oxide layers are evaluated approximately in a post-process operation. The electrostatic profile is obtained separately, without considering the effects on the charge distribution of the tunneling current. In addition, incident electrons are represented by plane-waves, so the effects of 2D carriers in the inversion layer on the gate tunneling can never be properly assessed. The NEGF method, however, can exactly solve the transport problem, when coupled to the Poisson equation, providing a way of self-consistently assessing the gate leakage and electrostatics profile. As was illustrated in Chapter 4, this method also allows us to examine the energy spectrum of the leakage current, which distinguishes the contributions from the 2D discrete states and 3D continuous states.

2) In Chapter 4, we described a methodology for simulating phonon-electron scattering in MOSFETs within the NEGF framework. The treatment was essentially one-dimensional however, focusing on the longitudinal component of the transport. Scattering related to the transverse modes was not explicitly accounted for. Real devices are typically wide, involving a large numbers of transverse modes. It is important to incorporate a more rigorous method of treating the transverse mode contribution. In the discussion section of Chapter 4, we presented a theoretical recipe for characterizing 2D scattering in MOSFETs, but numerical solutions to the 2D problem have not been attempted in this work. Numerical implementation of this approach requires an extremely large computational capability, in addition to precious experiences with nanoscale device physics. Venupopal in the Device Simulation Group at Purdue is currently working on this project, generating some promising results [92].

3) Schottky Barrier MOSFETs recently have caught the attention of device engineers for their possible applications in future VLSI technology [119]. A Schottky Barrier

device has a relatively simple structure (no source and drain extension regions), and therefore shows promises of low parasitic resistance. Reported simulation studies of such devices have been primarily based on the WKB approximation or other empirical models in obtaining thermionic emission and quantum tunneling current components through the Schottky barriers [119-121]. These conventional approaches raise similar concerns to those we addressed in part 1, namely that current evaluation has to be done as a post-process operation, and neither tunneled charges within the barrier region nor quantum coherence between the source and drain barriers can be included in the semiclassical framework. As device channel lengths scale, these quantum effects may appreciably impact device performance. Therefore it is desirable to apply the NEGF formalism in the replacement of semiclassical approximations. Assuming ballistic transport within devices (refer to Chapter 3 in this thesis), Guo in the Device Simulation Group at Purdue is currently extending the NEGF method in modeling nanoscale Schottky barrier transistors [122].

The NEGF approach is a very powerful mathematical tool for addressing how a quantum-state evolutes temporally under a varieties of interactions within any tiny system (or quantum level device). As device scaling continues, novel structures/designs must eventually take over the role currently played by semiconductor-based transistors. Some recent works have brought carbon-tube and molecule-cluster based device structures into focus [123-126]. These new areas provide us plenty of opportunities to apply the NEGF approach at theoretical research levels. In principle, the NEGF formalism not only enables us to understand the microscopic phenomena, but also enable us to exploit the potential of them. Moreover, it should be noted that this approach is applicable in all non-equilibrium systems, not limited to electron devices (see the introduction of [9, 127]).

LIST OF REFERENCES

[1] G.E. Moore, “Progress in digital integrated electronics,” *IEDM Tech. Digest*, pp. 11-13, 1975.

[2] *International Technology Roadmap for Semiconductors*, Semiconductor Industry Association, CA, 1999.

[3] Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, S.-H. Lo, G. Sai-Halasz, R. Viswanathan, H.-J. C. Wann, S. Wind and H.-S. Wong, “CMOS scaling into the nanometer regime,” *Proc. IEEE*, **85**, pp. 468-504, 1997.

[4] H.-S. Wong, D. Frank and P. Solomon, “Device Design Considerations for Double-Gate, Ground-Plane, and Single-Gated Ultra-Thin SOI MOSFET’s at the 25 nm Channel Length Generation,” *IEDM Tech. Digest*, pp. 407-410, 1998.

[5] K. Huster, *Two-dimensional scattering matrix simulations of Si MOSFETs*, Ph.D. dissertation, Purdue University, West Lafayette, IN, 2000.

[6] K. Banoo, *Direct Solution of the Boltzmann Transport Equation in Nanoscale Si Device*, Ph.D. dissertation, Purdue University, West Lafayette, IN, 2000.

[7] D.K. Ferry, R. Akis, D. Vasileska, “Quantum effects in MOSFETs: use of an effective potential in 3D Monte Carlo simulation of ultra-short channel devices,” *IEDM Tech. Digest*, pp. 287 -290, 2000.

[8] L.V. Keldysh, “Quantum transport equations for high electric fields,” *Sov. Phys.-JETP* **20**, pp. 1018, 1965.

[9] J. Rammer and H. Smith, “Quantum field-theoretical method in transport theory of metals,” *Rev. Mod. Phys.* **62**, pp. 323-359, 1986.

[10] S. Datta, *Electronic Transport in Mesoscopic Systems*, Cambridge University Press, Cambridge, UK, 1997.

[11] B. Davari, R.H. Dennard and G.G. Shahidi, "CMOS scaling for high-performance and low-power-the next ten years," *Proc. IEEE*, **89**, pp. 595-606, 1995.

[12] H.-S. Wong, D. Frank and P. Solomon, C. H.-J. Wann and J. Welser, "Nanoscale CMOS," *Proc. IEEE*, **87**, pp. 537-570, 1999.

[13] A.A. Abrikosov, L.P. Gorkov and I.E. Dzyaloshinski, *Quantum Field Theoretical Methods in Statistical Physics*, 2nd ed., Pergamon, New York, 1965.

[14] D.J. Frank, Y. Taur and H.-S. P. Wong, "Generalized Scale Length for Two-Dimensional Effects in MOSFETs," *IEEE Electron Dev. Lett.*, **10**, pp. 385-387, 1998.

[15] Y. Taur and T. Ning, *Fundamentals of VLSI Devices*, Cambridge University Press, Cambridge, UK, 1998.

[16] Y.-C. Sun, Y. Taur, R.H. Dennard and S.P. Klepner, "Submicron-Channel CMOS for Low-Temperature Operation," *IEEE Trans. Electron Devices*, **34**, pp. 19-27, 1987.

[17] S.J. Wind, D.J. Frank, and H.-S. Wong, "Scaling silicon MOS devices to their limits," *Microelectronic Engineering*, **32**, pp. 271-282, 1996.

[18] Y. Taur, C.H. Wann and D. Frank, "25 nm CMOS Design Considerations," *IEDM Tech. Digest*, pp. 789-792, 1998.

[19] A. Wei, M.J. Sherony and D.A. Antoniadis, "Effect of Floating-Body Charge on SOI MOSFET Design," *IEEE Trans. Electron Dev.,* **45**, pp. 430-438, 1998.

[20] L.T. Su, J.B. Jacobs, J.E. Chung and D.A. Antoniadis, "Deep-Submicrometer Channel Design in Silicon-on-Insulator (SOI) MOSFET's," *IEDM Tech. Digest*, pp. 183-186, 1994.

[21] Z. Ren, S. Bourland, S. Lee, J. Denton, M.S. Lundstrom and R. Bashir, "Ultra-thin Body SOI by Controlled Oxidation of Thin Si Membranes," presented at *IEEE Silicon Nanoelectronics Workshop*, Honolulu, Hawaii, June 11-12, 2000.

[22] L. Chang, S. Tang, T. King, J. Bokor and C. Hu, "Gate Length Scaling and Threshold Voltage Control of Double-Gate MOSFETs," *IEDM Tech. Digest*, pp. 719-722, 2000.

[23] G. Neudeck, T.-C. Su and J. Denton, "Novel Silicon Epitaxy for Advanced MOSFET Devices," *IEDM Tech. Digest*, pp. 169-172, 2000.

[24] M.-K Ieong, E.C. Jones, T. Kanarsky, Z. Ren, O. Dokumaci, R.A. Roy, L. Shi, T. Furukawa, Y. Taur, R.J. Miller, H-S Wong, "Deep-Submicrometer Channel Design in Silicon-on-Insulator (SOI) MOSFET's," to appear in *IEDM*, 2001.

[25] D.J. Frank, S. Laux and M. Fischetti, "Monte Carlo Simulation of a 30 nm Dual-Gate MOSFET: How Short Can Si Go?" *IEDM Tech. Digest*, pp. 553-556, 1992.

[26] Z. Ren, R. Venugopal, S. Datta and M.S. Lundstrom, "Examination of Design and Manufacturing Issues in a 10 nm Double Gate MOSFET using Nonequilibrium Green's Function Simulation," to appear in *IEDM*, 2001.

[27] M.S. Lundstrom, *Fundamentals of Carrier Transport*, 2nd ed., Cambridge University Press, Cambridge, UK, 2000.

[28] F. Assad, Z. Ren, D. Vasileska, S. Datta, and M.S. Lundstrom, "On the performance limits for Si MOSFET's: A theoretical study," *IEEE Trans. Electron Dev.*, **47**, pp. 232-240, 2000.

[29] Z. Ren and M.S. Lundstrom, "Simulation of nanoscale MOSFETs: A scattering theory interpretation," *Superlattices and Microstructures*, **27**, pp. 177-189, 2000.

[30] K. Banoo, J.-H. Rhew, M.S. Lundstrom, C.-W. Shu, and J.W. Jerome, "Simulating Quasi-Ballistic Transport in Si Nanotransistors," presented at *the 7th International Workshop on Computational Electronics* (*IWCE*), University of Glasgow, Glasgow, UK, May 22-25, 2000.

[31] M.V. Fischetti, "Theory of electron transport in small semiconductor devices using the Pauli master equation," *J. Appl. Phys.*, **83**, pp. 270-291, 1998.

[32] M.V. Fischetti, "Master-equation approach to the study of electronic transport in small semiconductor devices," *Phys. Rev. B*, **59**, 1999, pp. 4901-4917, 1999.

[33] F. Castella, "From the Von-Neumann equation to the Quantum Boltzmann equation in a deterministic framework," *J. Statistical Phys.*, **104**, pp. 387-447, 2001.

[34] W.R. Frensley, "Wigner-function model of a resonant-tunneling semiconductor device," *Phys. Rev. B*, **36**, pp.1570-1580, 1987.

[35] N.C. Kluksdahl, A.M. Kriman, D.K. Ferry and C. Ringhofer, "Self-consistent study of resonant-tunneling diode," *Phys. Rev. B*, **39**, pp.7720-7735, 1987.

[36] F.A. Buot and K.L. Jensen, "Lattice Weyl-Wigner formulation of exact many-body quantum-transport theory and application to novel solid-state quantum-based devices," *Phys. Rev. B*, **42**, pp.9429-9457, 1990.

[37] Z. Han, N. Goldsman and C.-K. Lin, "2-D Transport Device Modeling by Self-Consistent Solution of the Wigner and Poisson Equations," *Conf. Proc., SISPAD 2000, Intern. Conf. on Simulation of Semiconductor Processes and Devices*, pp. 62-65, Seattle, WA, September, 6-8, 2000.

[38] *Quantum Transport in Semiconductor,* edited by D.K. Ferry and C. Jacoboni, Plenum Press, New York, 1991.

[39] F. Mandl and G. Shaw, *Quantum Field Theory*, Revised ed., John Wiley & Sons, UK, 1993.

[40] S. Datta, "A simple kinetic equation for steady-state quantum transport," *J. Phys. Condens. Matter*, **2**, pp. 8023-8052, 1990.

[41] M. Büttiker, "Four-Terminal Phase-Coherent Conductance," *Phys. Rev. Lett.* **57**, pp.1761-1764, 1986.

[42] S. Miyano, M. Hirose, and F. Masuoka, "Numerical Analysis of a Cylindrical Thin-Pillar Transistor (CYNTHIA)," *IEEE Trans. Electron Dev.*, **39**, pp. 1976-1881, 1992.

[43] L. Risch, W.H. Krautschneider, F. Hofmann, H. Schafer, T. Aeugle, and W. Rosner, "Vertical MOS Transistors with 70 nm Channel Length," *IEEE Trans. Electron Dev.*, **43**, pp. 1495-1498, 1996.

[44] X. Huang, W.C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y.K. Choi, K. Asano, V. Subramanian, T.J. King, J. Bokor, and C. Hu, "Sub 50-nm FinFET: PMOS," *IEDM Tech. Digest*, pp. 67-70, 1999.

[45] C.H Wann, K. Noda, T. Tanaka, M. Yoshida, and C. Hu, "A Comparative study of Advanced MOSFET Concepts," *IEEE Trans. Electron Dev.,* **43**, pp. 1742-1753, 1989.

[46] F. Assad, Z. Ren, S. Datta, M.S. Lundstrom, and P. Bendix, "Performance Limits of Si MOSFET's," *IEDM Tech. Digest*, pp. 547-549, 1999.

[47] J.G. Fossum, Z. Ren, K. Kim and M.S. Lundstrom, “Extraordinarily High Drive Currents in Asymmetrical Double-Gate MOSFETs,” *Supperlattices and Microstructures*, **28**, pp. 525-530, 2000.

[48] D. Vasileska, D.K. Schroder and D.K. Ferry, “Scaled silicon MOSFET’s: degradation of total gate capacitance,” *IEEE Trans. Electron Devices*, **44**, pp. 584-587, 1997.

[49] *http://punch.ecn.purdue.edu/Member/CeHub/Program/Schred/*

[50] S.M. Sze, *Physics of Semiconductor Devices*, John Wiley and Sons, New York, 1981.

[51] T. Ando, A.B. Fowler, and F. Stern, “Electronic properties of two-dimensional systems,” *Rev. of Mod. Phys.,* **54**, pp. 437-672, 1982.

[52] C.Y. Hu, S. Banerjee, K. Sadra, B.G. Streetman and R. Sivan, “Quantization effects in inversion layers of PMOSFETs on Si (100) substrates,” *IEEE Electron Dev. Lett.*, **17**, pp. 276-278, 1996.

[53] S.-i. Takagi, M. Takayanagi and A. Toriumi, “Characterization of Inversion-Layer Capacitance of Holes in Si MOSFETs,” *IEEE Trans. Electron Dev.*, **46**, pp. 1446-1450, 1999.

[54] P. Hohenberg and W. Kohn, “Inhomogeneous electron gas,” *Phys. Rev.*, **136**, pp. B864-B871, 1964.

[55] M.S. Lundstrom, “Elementary scattering theory of the MOSFET,” *IEEE Electron Dev. Lett.*, **18**, pp. 361-363, 1997.

[56] K. Natori, “Ballistic metal-oxide-semiconductor field effect transistor,” *J. Appl. Phys.*, **76**, pp. 4879-4890, 1994.

[57] S. Datta, F. Assad, and M.S. Lundstrom, “The Si MOSFET from a transmission viewpoint,” *Superlattices and Microstructures*, **23**, pp. 771-780, 1998.

[58] J.S. Blakemore, “Approximations for the Fermi-Dirac integrals, especially the function, $\Im_{1/2}[\eta]$, used to describe electron density in a semiconductor,” *Solid-State Electron.*, **25**, pp. 1067-1076, 1982.

[59] P. Van Halen and D.L. Pulfrey, “Accurate, short series approximations to Fermi-Dirac integrals of order -1/2, 1/2, 1, 3/2, 2, 5/2, 3, and 7/2,” *J. Appl. Phys.*, **57**, pp. 5271-5274, 1985.

[60] M. Goano, “Series Expansion of the Fermi-Dirac Integral $\Im_j(x)$ Over The Entire Domain of Real *j* and *x*,” *Solid-State Electron.*, **36**, pp. 217-221, 1993.

[61] Y. Naveh and K.K. Likharev, “Modeling of 10-nm-Scale Ballistic MOSFET’s,” *IEEE Electron Dev. Lett.*, **21**, pp. 242-244, 2000.

[62] Z. Ren, R. Venugopal, S. Datta, M.S. Lundstrom, D. Jovanovic, and J.G. Fossum, “The Ballistic Nanotransistor: A Simulation Study,” *IEDM Tech. Digest*, pp. 715-718, 2000.

[63] J.P. Colinge, *Silicon-on-Insulator Technology*: *Materials to VLSI*, Kluwer Academic Publishers, New York, 1991.

[64] W.A. Harrison, *Phys. Rev.*, **123**, pp. 85, 1961.

[65] J. Cai and C.-T. Sah, “Gate tunneling currents in ultrathin oxide metal-oxide-silicon transistors,” *J. Appl. Phys.*, **89**, pp. 2272-2285, 2001.

[66] W.C. Lee and C. Hu, “Modeling Gate and Substrate Currents due to Conduction-and Valence-band Electron and Hole Tunneling,” *Symp. VLSI Tech. Digest*, pp. 198-199, 2000.

[67] S. Datta, *Quantum Phenomena*, Cambridge University Press, Cambridge, UK, 1989.

[68] T. Tanaka, H. Horie, S. Ando, and S. Hijiya, “Analysis of p^+ poly Si double-gate thin-film SOI MOSFETs” *IEDM Tech. Digest*, pp. 683-686, 1991.

[69] K. Suzuki and T. Sugii, “Analytical models for n^+-p^+ double-gate SOI MOSFET’s,” *IEEE Trans. Electron Dev.*, **42**, pp. 1940-1948, 1995.

[70] J.G. Fossum, K. Kim, and Y. Chong, “Extremely Scaled Double-Gate CMOS Performance Projections, Including GIDL-Controlled Off-State Current,” *IEEE Trans. Electron Dev.*, **46**, pp. 2195-2199, 1999.

[71] K. Kim and J.G. Fossum, “Optimal double-gate MOSFETs: Symmetrical or asymmetrical gates?” *IEEE International SOI CONF.*, pp. 98-99, 1999.

[72] S. Datta, “Nanoscale Device Modeling: the Green’s Function Method,” *Superlattices and Microstructures*, **28**, pp. 253-278, 2000.

[73] J.-H. Rhew, Z. Ren and, M.S. Lundstrom, “Numerical simulation of a Ballistic Nano-MOSFET,” submitted for publication, 2001.

[74] J. Guo, Z. Ren and M. Lundstrom, "A computation exploration of lateral channel engineering to enhance MOSFET performance", presented at *the 8th International Workshop on Computational Electronics* (*IWCE*), University of Illinois, Urbana Champaign, USA, October 15 - 18, 2001

[75] G. W. Brown and B. W. Lindsay, “The numerical solution of Poisson’s equation for two-dimensional semiconductor devices,” *Solid-State Electron.*, **19**, pp. 991-992, 1976.

[76] T. Conklin, T. Naugle, S. Shi, S. Frimel, S.M. Roenker, K.P. Kumar, T. Cahay and M.M. Stanchina, “Inclusion of tunneling and ballistic transport effects in an analytical approach to modeling of NPN InP based heterojunction bipolar transistors,” *Superlattices and Microstructures*, **18**, pp. 21-32, 1995.

[77] F. Venturi, R.K. Smith, E.C. Sangiorgi, M. Pinto, and B. Riccó, “A General Purpose Device Simulator Coupling Poisson and Monte Carlo Transport with Applications to Deep Submicron MOSFETs,” *IEEE Trans. Electron Dev.*, **8**, pp. 360-369, 1989.

[78] D.J. Rose and R.E. Bank, “Global approximate Newton methods,” *Numerische Mathematik*, pp. 279-295, 1981.

[79] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes*: *the art of scientific computing*, Cambridge University Press, Cambridge, New York, 1989.

[80] D. Jovanovic and R. Venugopal, presented at *the 7th International Workshop on Computational Electronics* (*IWCE*), University of Glasgow, Glasgow, UK, May 22-25, 2000.

[81] R. Venugopal, *Preliminary Report*, to be published.

[82] R.K. Lake and S. Datta, “The Non-equilibrium Green’s function method applied to double barrier resonant tunneling diodes,” *Phys. Rev. B*, **45**, pp. 6670-6685, 1992.

[83] R. Landauer, “Conductance determined by transmission: probes and quantised constriction resistance,” *J. Phys. Condens. Matter*, **1**, pp. 8099-8109, 1989.

[84] M.C. Payne, “Electrostatic and electrochemical potential in quantum transport,” *J. Phys. Condens. Matter*, **1**, pp. 4931-4938, 1989.

[85] W.B. Joyce and R.W. Dixon, “Analytic approximations for the Fermi energy of an ideal Fermi gas,” *Appl. Phys. Lett.*, **31**, pp. 354-356, 1977.

[86] H.K. Gummel, "A Self-consistent iterative scheme for one-dimensional steady state transistor calculations," *IEEE Trans. Electron Dev.*, **11**, pp. 455-465, 1964.

[87] F.S. Khan, J.H. Davies and J.W. William, "Quantum transport equations for high electric fields," *Phys. Rev. B*, **36**, pp. 2578-2577, 1987.

[88] E.V. Anda and F. Flores, "The role of inelastic scattering in resonant tunneling heterostructures," *J. Phys. Condens. Matter*, **3**, pp. 9087-9101, 1991.

[89] R.K. Lake, G. Klimeck, R.C. Bowen and D. Jovanovic, "Single and multiband modeling of quantum electron transport through layered semiconductor devices," *J. Appl. Phys.* **81**, pp. 7845-7869, 1997.

[90] A. Rahman, Z. Ren, J.-H. Rhew and M.S. Lundstrom, "A Compact Scattering Model for Nanoscale Double-Gate MOSFET," presented at *the 4th International Conference on the Modeling and Simulation of Microstructure* (*MSM*), Hilton head, SC, March 19-21, 2001.

[91] D. Jovanovic, personal communication, 2001.

[92] R. Venugopal, personal communication, 2001.

[93] S.E. Laux, and M.V. Fischetti, "Monte Carlo simulation of submicrometer Si n-MOSFET's at 77 and 300 K," *IEEE Electron Dev. Lett.*, **9**, pp. 467-469, 1988.

[94] M.R. Pinto, E. Sangiorgi, and J. Bude, "Silicon transconductance scaling in the overshoot regime," *IEEE Electron Dev. Lett.*, **14**, pp. 375-378, 1993.

[95] C. Jungemann, S. Keith, M. Bartels, and B. Meinerzhagen, "Efficient full-band Monte Carlo simulation of silicon devices," *IEICE Trans. on Electronics*, **E82**, pp. 870-879, 1999.

[96] M.S. Lundstrom, Z. Ren, and S. Datta, "Essential Physics of Carrier Transport in Nanoscale MOSFETs," *Conf. Proc.*, *SISPAD 2000*, *Intern. Conf. on Simulation of Semiconductor Processes and Devices*, pp. 1-5, Seattle, WA, September, 6-8, 2000.

[97] H. Hu, J.B. Jarvis, L.T. Su, and D.A. Antoniadis, "A study of deep-submicron MOSFET scaling based on experiment and simulation," *IEEE Trans. Electron Dev.*, **42**, pp. 669-677, 1995.

[98] J. Bude, personal communication, Dec., 1999.

[99] G. Timp, J. Bude, K.K. Bourdelle, J. Garno, A. Ghetti, H. Gossmann, M. Green, G. Forsyth, Y. Kim, R. Kleiman, F. Klemens, A. Kornblit, C. Lochstampfor, W. Mansfield, S. Moccio, T. Sorsch, D.M. Tennant, W. Timp, R. Tung, "The Ballistic Nanotransistor," *IEDM Tech. Digest*, pp. 55-58, 1999.

[100] P.J. Price, "Monte Carlo calculation of electron transport in solids," *Semiconductors and Semimetals*, **14**, pp. 249-334, 1979

[101] F. Berz, "Diffusion near an absorbing boundary," *Solid-State Electron.*, **15**, pp. 1245-1255, 1974.

[102] S. Bandyopadhyay, C.M. Maziar, M.E. Klausmeier-Brown, S. Datta, and M.S. Lundstrom, "Rigorous Technique to Couple Monte Carlo and Drift-Diffusion Models for Computationally Efficient Device Simulation," *IEEE Trans. Electron Dev.*, **34,** pp. 392-399, 1987.

[103] J.P. McKelvey, R.L. Longini, and T.P. Brody, "Alternative approach to the solution of added carrier transport problems in semiconductors, " *Phys. Rev.*, **123**, pp. 51-57, 1961.

[104] W. Shockley, "Diffusion and drift of minority carrier in semiconductors for comparable capture and scattering mean free paths," *Phys. Rev.,* **125**, pp. 1570-1576, 1962.

[105] M.A. Alam, M.A. Stettler, and M.S. Lundstrom, "Formulation of the Boltzmann Equation in Terms of Scattering Matrices, *Solid-State Electron.*, **36**, pp. 263-271, 1993.

[106] J. Bude, "MOSFET Modeling in the Ballistic Regime," *Conf. Proc.*, *SISPAD 2000, Intern. Conf. on Simulation of Semiconductor Processes and Devices*, pp. 23-26, Seattle, WA, September, 6-8, 2000.

[107] P. M. Solomon, "A Comparison of Semiconductor Devices for High-Speed Logic," *Proc. IEEE*, **70**, pp. 489-509, 1982.

[108] E.O. Johnson, "The Insulated-Gate Field-Effect Transistor-A Bipolar Transistor in Disguise," *RCA Review*, **34**, pp. 80-94, 1973.

[109] F. Berz, "The Bethe condition for thermionic emission near an absorbing boundary," *Solid-State Electron.*, **28**, pp. 1007-1013, 1985.

[110] *www.ece.purdue.edu/celab/*

[111] M.S. Lundstrom and Z. Ren, "Essential Physics of Carrier Transport in Nanoscale MOSFETs," to appear in *IEEE Trans. Electron Devices*.

[112] T. Ghani, K. Mistry, P. Packan, S. Thompson, M. Stettler, S. Tyayi, and M. Bohr, "Scaling Challenges and Device Design Requirements for High Performance Sub-50 nm Gate length Planar CMOS Transistors," *Symp. VLSI Tech. Digest*, pp. 174-175, 2000.

[113] J.-H. Choi Y.-J. Park and H.-S. Min, "Electron Mobility Behavior in Extremely Thin SOI MOSFET's," *IEEE Electron Dev. Lett.*, **16**, pp.527-529, 1995.

[114] M.-K. Ieong, *et al.*, *Conf. Proc.*, *SISPAD 2000*, *Intern. Conf. on Simulation of Semiconductor Processes and Devices*, pp. 100-103, Seattle, WA, September, 6-8, 2000.

[115] Y. Fu, M. Karlsteen and M. Willander, "Energy band structure of quantum-size metal-oxide-semiconductor field effect transistor," *Supperlattices and Microstructures*, **22**, pp. 405-410, 1997.

[116] M. Darwish, J. Lentz, M. Pinto, P. Zeizoff, T. Krutsick and H. Vuong, "An Improved Electron and Hole Mobility Model for General Purpose Device Simulation," *IEEE Trans. Electron Dev.*, **44**, pp. 1529-1538, 1997.

[117] D.B.M. Klaassen, "A Unified Mobility Model for Device Simulation-I. Model Equations and Concentration Dependence," *Solid-State Electron.*, **35**, pp. 953-959, 1992.

[118] E.M.Vogel, K.Z. Ahmed, B. Hornung, W.K. Henson, P.K. McLarty, G. Lucovsky, J.R. Hauser and J.J. Wortman, "Modeled Tunnel Currents for High Dielectric Constant Dielectrics," *IEEE Trans. Electron Devices*, **45**, pp. 1350-1354, 1998.

[119] J. Kedzierski, P. Xuan, E.H. Anderson, J. Bokor, T.-J. King, C. Hu, "Complementary silicide source/drain thin-body MOSFETs for the 20 nm gate length regime," *IEDM Tech. Digest*, pp. 57-60, 2000.

[120] J.R. Tucker, C. Wang and P.S. Carney, "Silicon field-effect transistor based on quantum tunneling," *Appl. Phys. lett.*, **65**, pp. 618-620, 1994.

[121] C.-K. Huang, W.E. Zhang and C.H. Yang, "Two-Dimensional Numerical Simulation of Shottky Barrier MOSFETs with Channel Length to 10 nm," *IEEE Trans. Electron Dev.*, **45**, pp. 842-848, 1998.

[122] J. Guo, personal communication, 2001.

[123] N.C. Greenham and R.H. Friend, “Semiconductor Device Physics of Conjugated Polymers”, *Solid State Physics* **49**, ed. H. Ehrenreich and F. Spaepen, Academic Press, New York, 1995.

[124] C. Dekker, “Carbon Nanotubes as Molecular Quantum Wires,” *Phys. Today* **52**, pp. 22-28, May, 1999.

[125] M.A. Reed, “Molecular-scale electronics,” *Proc. IEEE*, **87**, pp. 652–658, 1999.

[126] J. Taylor, H. Guo and J. Wang, “*Ab initio* modeling of quantum transport properties of molecular electronic devices,” *Phys. Rev. B*, **63**, 245407, pp.1-13, 2001.

[127] D.C. Langreth and J.W. Wilkins, “Theory of Spin Resonance in Dilute Magnetic Alloys,” *Phys. Rev. B*, **6**, pp.3189-3227, 1972.

[128] S. Datta, personal communication, 2000.

[129] *Medici*, *Two-Dimensional Device Simulation Program*, *User’s manual*, Avant! Corporation, CA, 2000.

[130] Caughey, D. M and Thomas, R. E., “Carrier mobilities in silicon empirically related to doping and field”, *Proc. IEEE*, vol. 55, pp. 2192-2193, 1967.

Appendix A
Expressions for Constants Appearing in Equations in Chapter 2

In Chapter 2, we described a simulation tool, Schred-2.0, which solves the Schrödinger-Poisson equation set self-consistently in 1D MOS structures. We presented there a list of formulae of how the program computes interesting quantities, such as carrier density, ballistic current and ballistic limit channel conductance, etc. To keep the chapter concise, some parameters in those formulae were not explicitly expressed. This appendix provides the complete expressions.

We assume that the SiO_2/Si interface is parallel to the (100) plane, the channel transport direction is along [100]. We account for six valleys in the conduction band and heavy hole and light hole valleys in the valance band (see Fig. 2.1). The six valleys in the conduction band split into two sets of subbands, namely the unprimed set and primed set (see Chapter 2 for the definitions). We use simple parabolic E-k relations in all derivations. So there are two effective masses m_l and m_t in characterizing the conduction band valleys, where

$$m_l = 0.98 m_e \text{ and } m_t = 0.19 m_e \text{ , } m_e = 0.91 \times 10^{-30} \text{ kg.} \tag{A.1a}$$

There are four effective masses m_{l_t}, m_{ll}, m_{h_t} and m_{hl} in characterizing the valance band valleys. For the light hole valley, we use m_{l_t} to denote the effective mass responding in the transverse direction (or the gate confinement direction), and m_{l_t} to denote the effective mass responding in the longitudinal direction (parallel to the gate surface). For the heavy hole valley, m_{h_t} and m_{hl} assume the similar meanings [52-53].

$$m_{lt} = 0.20 m_e\,,\ m_{ll} = 0.169 m_e\,,\ m_{ht} = 0.29 m_e \text{ and } m_{hl} = 0.433 m_e\,. \tag{A.1b}$$

In terms of these effective masses, the density-of-state effective mass, m_C, and conductivity effective mass, m_D can be defined. For the unprimed set electrons, they are

$$m_C = 4m_t \text{ and } m_D = 2m_t\,. \tag{A.2a}$$

For the primed set electrons, they are

$$m_C = 4(\sqrt{m_t} + \sqrt{m_l})^2 \text{ and } m_D = 4\sqrt{m_t m_l}\ . \tag{A.2b}$$

For heavy holes,

$$m_C = m_D = m_{hl}\,. \tag{A.3a}$$

For light holes,

$$m_C = m_D = m_{ll}\,. \tag{A.3b}$$

Carrier density

In the carrier density expressions (see eqn. (2.3a) for electrons, and eqn. (2.3c) for holes),

$$n_{2D} \text{ or } p_{2D} = \frac{m_D k_B T}{\pi \hbar^2}\,. \tag{A.4}$$

Current density

In the current density expression, eqn. (2.5),

$$I_O = \frac{q}{\hbar^2}\sqrt{\frac{m_C}{2}}\left(\frac{k_B T}{\pi}\right)^{3/2}, \tag{A.5}$$

where q is the elementary charge constant.

Ballistic channel conductance

In the conductance expression, eqn. (2.6),

$$G_O = \frac{q^2}{\pi^{3/2}\hbar^2}\sqrt{\frac{m_C k_B T}{2}}. \tag{A.6}$$

In the expression of uni-directed thermal velocity of source-injected carriers, eqn. (2.7),

$$\upsilon_T = \sqrt{\frac{2k_B T m_C}{\pi m_D^2}}. \tag{A.7}$$

Appendix B

1D Solution to Boltzmann transport equation in ballistic limit

A 1D solution to the Boltzmann Transport Equation (BTE) in the ballistic limit can be obtained directly [128]. In our treatment, the quantum effect in ultra-thin body double gate MOSFETs is accounted for by solving the Schrödinger equation in the gate confinement direction. The solutions give rise to discrete subbands within which carriers (electrons in our case) are constrained. Since the device widths are assumed to be infinite and translational invariance holds, subband potential energies can only vary in one direction in response to different bias conditions. Therefore, a 1D solution is sufficient to describe the electron transport within the MOSFETs. (More precisely speaking, this 1D solution is actually a charge-sheet description.) The solution procedure can be illustrated through Fig. B1.

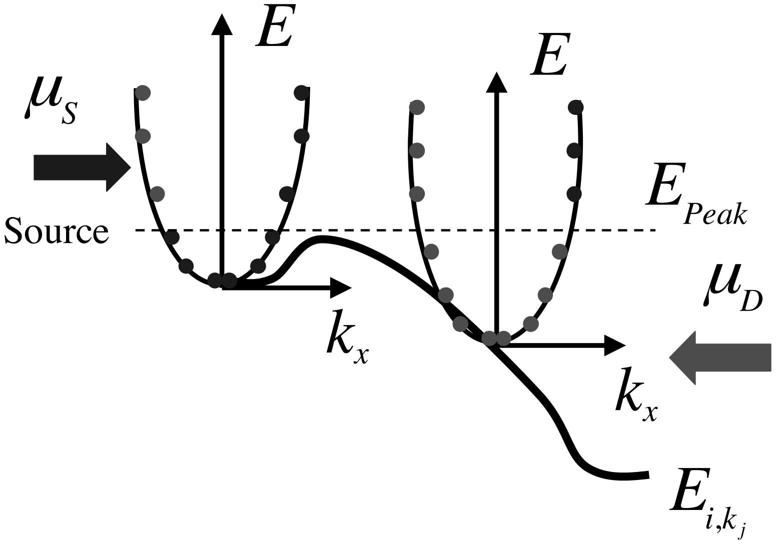


Fig. B1. Real space and k-space distribution of electrons within a 1D subband in the ballistic transport limit.

The energy profile, E_{i,k_j}, illustrated in Fig. B1 represents subband, i, with a transverse direction (the y direction) quantum number, k_j, as discussed in Chapter 3. This profile can be spatially divided into two regions: points to the left of the peak subband energy (region 1) and points to the right of the peak subband energy (region 2). In region 1, electrons with energy lower than E_{Peak} are in equilibrium with the source reservoir, electrons with energy higher than E_{Peak} are either coming from the source reservoir (right-going electrons) or coming from the drain reservoir (left-going electrons). In region 2 electrons can be sorted in the similar way. The two reservoirs are characterized by two Fermi potential energies μ_S and μ_D. Based on the analyses given above, electron density along the subband, and current density at source/drain terminal can be easily obtained.

1) Electron density

The electron spatial density in region 1 of subband, E_{i,k_j}, can be written as,

$$\begin{aligned} n_{left}(x,E_{k_j}) = &\int_0^\infty [\frac{1}{\pi\hbar}\sqrt{\frac{m_x^*}{2E_x}}\frac{1}{1+e^{(E_x+E_i+E_{k_j}-\mu_S)/k_BT}}]dE_x \\ &+\int_0^{E_{Peak}} [\frac{1}{\pi\hbar}\sqrt{\frac{m_x^*}{2E_x}}\frac{1}{1+e^{(E_x+E_i+E_{k_j}-\mu_S)/k_BT}}]dE_x \ , \\ &+\int_{E_{Peak}}^\infty [\frac{1}{\pi\hbar}\sqrt{\frac{m_x^*}{2E_x}}\frac{1}{1+e^{(E_x+E_i+E_{k_j}-\mu_D)/k_BT}}]dE_x \end{aligned} \quad \text{(B.1a)}$$

where the subscript *left* stands for region 1. Fermi-Dirac statistics are assumed, and spin degeneracy is also included. To account for the contributions from all transverse modes, integration over E_{k_j} is needed, which gives

$$n_{left}(x) = \int_0^\infty \frac{1}{\pi\hbar}\sqrt{\frac{m_y^*}{2E_{k_j}}}[n_{left}(x,E_{k_j})]dE_{k_j}$$

$$= n_{2Di}\left\{\ln(1+e^{\tilde{\mu}_S}) + \frac{1}{\sqrt{\pi}}\int_0^{\tilde{E}_{Peak}} \frac{d\tilde{E}_x}{\sqrt{\tilde{E}_x}}\Im_{-1/2}(\tilde{\mu}_S - \tilde{E}_x) + \frac{1}{\sqrt{\pi}}\int_{\tilde{E}_{Peak}}^\infty \frac{d\tilde{E}_x}{\sqrt{\tilde{E}_x}}\Im_{-1/2}(\tilde{\mu}_D - \tilde{E}_x)\right\},$$

(B.1b)

where the hat ~ means that all quantities are specified relative to the local subband potential $E_i(x)$ and normalized to the thermal energy k_BT. The areal electron density factor, n_{2Di}, for subband, i, is $\frac{\sqrt{m_x^* m_y^*}}{\pi\hbar^2}\frac{k_BT}{2}$. To numerically accomplish the first integral in eqn. B.1b, a variable change, $x = \sqrt{\tilde{E}_x}$, can be made to avoid the divergence difficulty at $\tilde{E}_x$=0.

In the same way, the electron density in region 2 can be obtained,

$$n_{right}(x) = n_{2Di}\left\{\ln(1+e^{\tilde{\mu}_D}) + \frac{1}{\sqrt{\pi}}\int_0^{\tilde{E}_{Peak}} \frac{d\tilde{E}_x}{\sqrt{\tilde{E}_x}}\Im_{-1/2}(\tilde{\mu}_D - \tilde{E}_x) + \frac{1}{\sqrt{\pi}}\int_{\tilde{E}_{Peak}}^\infty \frac{d\tilde{E}_x}{\sqrt{\tilde{E}_x}}\Im_{-1/2}(\tilde{\mu}_S - \tilde{E}_x)\right\}.$$

(B.1c)

The 2D density is then distributed according to the corresponding wavefunctions in the gate confinement direction (due to the quantum effect) to give the 3D density profile. Finally, summation over all relevant subbands has to be done for the total electron density. Note that in eqns. B.1b and B.1c, only a single conduction band valley has been considered. To account for contributions from all valleys, the expressions should be multiplied by the valley degeneracy factor, which is 4 for the unprimed subbands and 2 for the primed subbands (refer to Chapter 2 for definitions of the sets of subbands).

2) Terminal current

Following a very similar procedure, the terminal current can be computed. Since the current is conserved throughout the entire device, it can be evaluated at any cross section normal to the current flow direction. To keep the result general, we assume the cross section is located at *x*, which is to the left of the channel barrier peak (readers will see in a moment that the result is independent of *x*). Again, starting with the energy band, E_{i,k_j}, the current is given as

$$\begin{aligned} J(E_{k_j}) = & \int_{0}^{E_{Peak}} [\sqrt{\frac{2E_x}{m_x^*}}\frac{q}{\pi\hbar}\sqrt{\frac{m_x^*}{2E_x}}\frac{1}{1+e^{(E_x+E_i+E_{k_j}-\mu_S)/k_BT}}]dE_x \\ & + \int_{E_{Peak}}^{\infty} [\sqrt{\frac{2E_x}{m_x^*}}\frac{q}{\pi\hbar}\sqrt{\frac{m_x^*}{2E_x}}\frac{1}{1+e^{(E_x+E_i+E_{k_j}-\mu_S)/k_BT}}]dE_x \\ & - \int_{0}^{E_{Peak}} [\sqrt{\frac{2E_x}{m_x^*}}\frac{q}{\pi\hbar}\sqrt{\frac{m_x^*}{2E_x}}\frac{1}{1+e^{(E_x+E_i+E_{k_j}-\mu_S)/k_BT}}]dE_x \\ & - \int_{E_{Peak}}^{\infty} [\sqrt{\frac{2E_x}{m_x^*}}\frac{q}{\pi\hbar}\sqrt{\frac{m_x^*}{2E_x}}\frac{1}{1+e^{(E_x+E_i+E_{k_j}-\mu_D)/k_BT}}]dE_x \end{aligned} ,$$

(B.2a)

where $\sqrt{2E_x / m_x^*}$ represents the current related velocity derived from the parabolic E-k relationship. q is the elementary charge constant. Because of the opposite signs in electron velocity, the contribution from the first integral cancels that from the third integral. So eqn. B.2a becomes

$$J(E_{k_j}) = \frac{q}{\pi\hbar}\int_{E_{Peak}}^{\infty} [\frac{1}{1+e^{(E_x+E_i+E_{k_j}-\mu_S)/k_BT}} - \frac{1}{1+e^{(E_x+E_i+E_{k_j}-\mu_D)/k_BT}}]dE_x \,. \qquad \text{(B.2b)}$$

Integrating over E_{k_j} gives

$$J = \int_0^\infty \frac{1}{\pi\hbar}\sqrt{\frac{m_y^*}{2E_{k_j}}} J(E_{k_j}) dE_{k_j} = \frac{q}{\hbar^2}\sqrt{\frac{m_y^*}{2}}\left(\frac{k_B T}{\pi}\right)^{3/2} [\Im_{1/2}(\tilde{\mu}_S - \tilde{E}_{Peak}) - \Im_{1/2}(\tilde{\mu}_D - \tilde{E}_{Peak})] , \tag{B.2c}$$

where the hat ~ assumes the same meaning as that in eqn. B.1b. The total current is the sum of contributions from all relevant valleys and subbands.

Appendix C
nanoMOS2.0: A 2D-Simulator for Double-gate MOSFETs

This brief document explains the procedure to run the program, nanoMOS2.0, on the Purdue Simulation Hub [110]. This program is a self-consistent (Poisson with a transport model) 2D-simulator for thin body (less than 5 nm), fully depleted, double gated, n-MOSFETs (no holes). nanoMOS2.0 provides a choice of four transport models: classical ballistic, quantum ballistic, drift-diffusion, and quantum dissipative. It should be noted that each of these models accounts for quantum effects in the confinement direction exactly, and the names indicate the method to treat transport in the channel direction.

nanoMOS2.0 extends nanoMOS1.0 by adding a new quantum dissipative transport model. The new model treats scattering in MOSFETs through the Green's function formalism using a simple Büttiker-probe model [41]. Scattering centers are treated as reservoirs similar to the source and drain except that they only change the energy of the carriers and not the total number of carriers in the system. Each scattering center is modeled through a perturbation strength characterized by a position dependent self-energy, η, which can be mapped onto an equivalent mobility. For detailed discussion, reader may want to refer to Chapter 4 of this report.

1) Running nanoMOS_2.0:

In order to run nanoMOS2.0, an input file has to be specified. Creation of an input file is most easily achieved by copying an example file from the "EXAMPLES" directory to the working directory and modifying the example input file to simulate the desired problem. Each input file is divided into several directives. These directives are used to organize the input assignment statements into groups that characterize the device geometry, transport model choice etc. For example, the format of the "device" directive looks like,

device **nsd**=1e20, **nbody**=0, **lgtop**=10, **lgbot**=10, **lsd**=7.5, **overlap**=0,

+ **tsi**=1.5, **tox_top**=1.5, **tox_bot**=1.5, **temp**=300

Each directive begins on a new line. First the directive name is specified, followed by the various assignment statements. If a directive statement is longer than one line, the continuation symbol "+", must appear in the first column for the following lines. Commas or blanks are assumed to be separators between two assignments. *However, neither tabs nor upper case letters may not be used anywhere in the input deck.* Also, an assignment statement cannot contain any blanks. For example

nsd = 1e20

is not allowed. The assignment should read,

nsd=1e20.

Comments can be added to the input deck by preceding the comment with a "$" sign. After creating the input deck, typing nanamos at the Matlab prompt results in a request for an input file. Specifying the input file name results in the start of a simulation.

2) Description of parameters in the input file:

<u>Device</u>

nanoMOS2.0 is a 2D-simulator for thin body, fully depleted, double gated n-MOSFETs. A Gaussian distribution doping profile is assumed in the silicon body [129], and the gate is modeled as a metal with a user-specified work function. The parameters needed to generate the device structure and the coordinate system used in nanoMOS2.0 is illustrated in Fig. C1.

The device structure is symmetric about the x-axis and the top oxide silicon interface represents the y=0 plane. The various parameters in the device directive are,

nsd: Source/Drain doping concentration (cm^{-3})

nbody: Body doping concentration (cm^{-3})

lgtop: Length of the top gate (nm)

lgbot: Length of the bottom gate (nm)

lsd: Length of the Source/Drain (nm)

overlap: Source/Drain extension length (nm)

xchar: Horizontal characteristic length of the source/drain Gaussian distrubution (nm)

tox_top: Top insulator thickness (nm)

tox_bot: Bottom insulator thickness (nm)

tsi: Silicon film thickness (nm)

temp: Lattice temperature (K)

Grid

The grid is specified through two parameters

hx: Horizontal node spacing (nm)

hy: Vertical node spacing (nm)

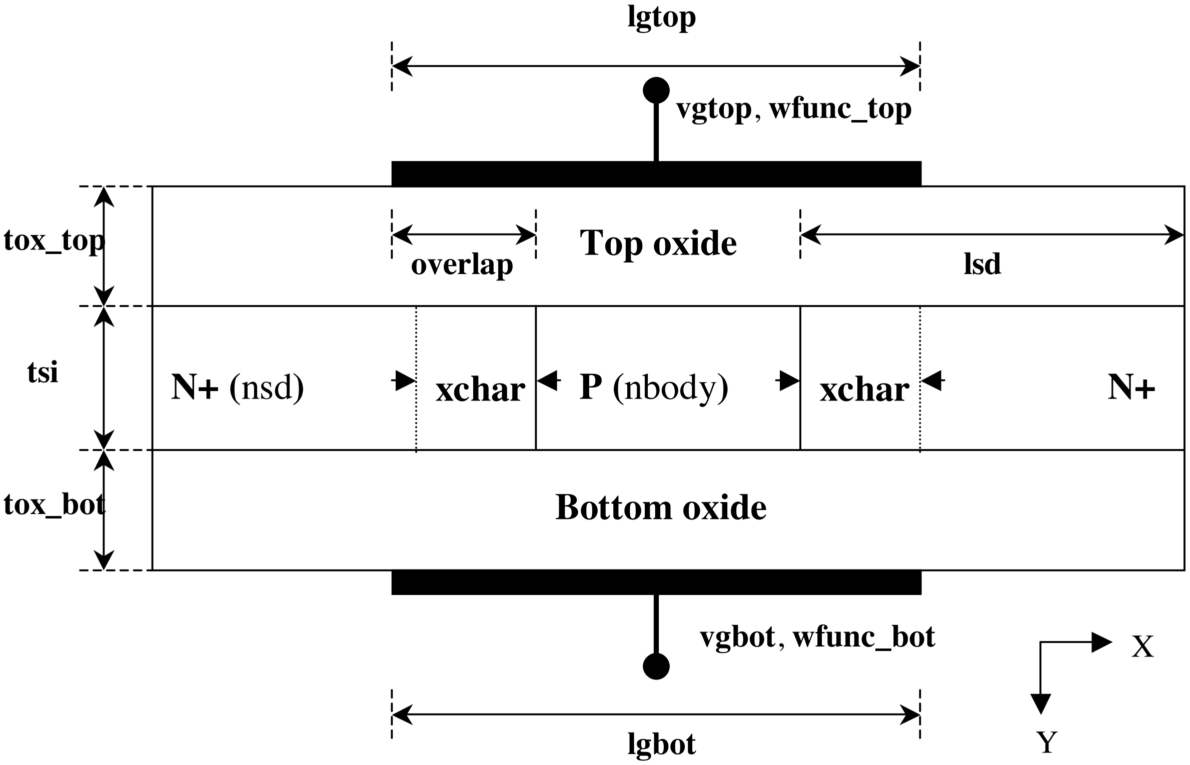


Fig. C1. Parameters to specify the device structure.

Transport model

Classical Ballistic Transport model (clbte): In this model, quantum effects in the y-direction (threshold voltage shift) are treated exactly by solving a 1D Schrödinger equation at every "x" (transmission direction) point within the device. This results in a set of subband profiles (ESUB(x)). Carrier transport in each subband is then modeled to account for the thermionic emission component alone. Quantum tunneling through the source-channel barrier is assumed to be zero. For more detailed description, please read Chapter 3 and Appendix B in this report.

Quantum ballistic Transport model (qbte): In this model, quantum effects in the y-direction (threshold voltage shift) are treated exactly by solving a 1D Schrödinger equation at every "x" point as in the case of model 1, to yield a set of subband profiles (ESUB(x)). Carrier transport in each subband is then treated by solving a 1D Schrödinger equation in the transmission (x) direction using the non-equilibrium green's function method [62, 72]. This approach accounts for quantum tunneling through the source-channel barrier, which was ignored in model 1. For more detailed description, please read Chapter 3 in this report.

Drift Diffusion (dd): This model is a quantum corrected drift diffusion model, where quantum effects in the y-direction (threshold voltage shift) are accounted for exactly. The y-direction is treated quantum mechanically by solving a 1D Schrödinger equation at every "x" point as in the case of model 1, to yield a set of subband profiles (ESUB(x)). Carrier transport in each subband is then treated by solving a 1D drift-diffusion equation in the transmission direction (x). A field dependent mobility model (Caughey-Thomas [130]) is used in the drift-diffusion solution and the model parameters **mu_low**, **beta** and **vel_sat** (saturation velocity) are user specified quantities,

$$\mu(E_{//}) = \frac{\text{mu\_low}}{[1+(\frac{E_{//}\cdot\text{mu\_low}}{\text{vel\_sat}})^{\textbf{beta}}]^{\textbf{1/beta}}} \qquad \text{(C.1)}$$

where is $E_{//}$ the electric field in the transport direction.

Quantum Dissipative Transport model (qdte): In this new model, quantum effects in the y-direction (threshold voltage shift) are treated exactly by solving a 1D Schrödinger equation at every "x" point as in the case of model 1, to yield a set of subband profiles (ESUB(x)). Dissipative transport in MOSFETs is treated through the Green's function formalism using a simple Büttiker-probe model [41]. Scattering centers are treated as reservoirs similar to the source and drain except that they only change the energy of the carriers and not the total number of carriers in the system. Each scattering center is modeled through a perturbation strength characterized by a position dependent self-energy, η. The user specified parameter, **mu_low**, can be mapped onto an equivalent η. For detailed discussion, please read Chapters 4 and 5 of this report.

<u>Options</u>

There are five assignments under the "options" directive. Three of them are flags that take on a value of "true" or "false", while the remaining pertain to the subband and valley information.

ox_penetrate: If set to "true", when solving the Schrödinger equation in the vertical (y) direction, the electron wave function is allowed to penetrate into the oxide regions, but the gate leakage current is NOT computed. If set to "false", the Schrödinger equation is solved in the y-direction assuming an infinite potential barrier at the oxide-silicon interface.

dg: If set to "true", both the top and bottom gate voltages are ramped in voltage simultaneously. If set to "false", the bottom gate voltage is fixed at **vgbot** while the top gate voltage, **vgtop** is ramped.

fermi: If set to "true", Fermi-Dirac statistics are used in solving the transport models. If set to "false", Maxwell-Boltzmann statistics are used. Note that this flag only applies to the drift-diffusion model. For the other three models, Fermi-Dirac statistics are always assumed.

valleys: There are 3 sets of valleys as shown in Fig. C2. Electrons in one set respond to the vertical (y) confining potential with a heavy effective mass (m_l, the so-called unprimed valleys) while those in the other two sets of valleys respond with a light effective mass (m_t, the so-called primed valleys) Thus **valleys** can be set to "unprimed" or "all". If set to "unprimed", only those electrons in valleys for which m_y is equal to $0.98m_e$ are treated in the simulation. If set to "all", electrons in all valleys are included in the simulation.

num_subbands: Each set of valleys yields a corresponding set of subbands when a 1D Schrödinger equation is solved in the confinement direction. However, it is not necessary to consider all of the subbands when solving for electron transport, as only the lowest few from each valley are occupied by electrons. Higher subbands are unoccupied and do not contribute to carrier transport. The parameter **num_subbands,** is thus the number of subbands that needs to be considered for each of the 3 valleys (*e.g*: num_subbands=2 (2 from each valley), valleys=all (3 valleys) implies that a total of 6 subbands will be used in the simulation). We suggest that the users do a Schred-2.0 simulation to determine how many subbands to use.

Bias

Bias information is specified through the following parameters.

vgtop: Top gate voltage

vgbot: Bottom gate voltage

vs: Source contact voltage

vd: Drain contact voltage. It is recommended that a very high drain bias not be used to run Full Quantum simulations as the tight binding bandstucture at high energies may not be accurate despite using a fine real space grid.

vd_initial: Drain start voltage. When solving a non-equilibrium problem, it is sometimes difficult to get convergence if the drain voltage is ramped to a high value. Specifying **vd_initial**, less than **vd**, enables the simulator establish a low Vd solution which can be used as an initial guess to speed up convergence for high drain voltages.

vgstep: Step size for the gate voltage

ngstep: Number of gate voltage steps

vdstep: Step size for the drain voltage

ndstep: Number of drain voltage steps

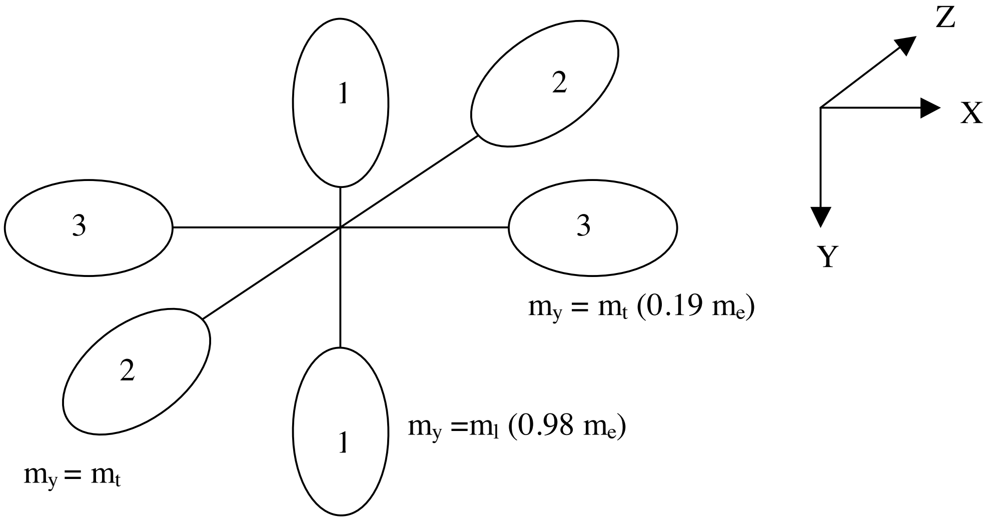


Fig. C2. Valley information

Solve

dvmax: For any choice of transport model, the Poisson equation is solved self-consistently with the corresponding transport equation. The self-consistent solution of a transport equation with the Poisson equation, is said to have converged if the maximum change in potential between two self-consistent iterations is less than **dvmax** (eV).

Material parameters

The user has the flexibility of altering the following material parameters:

wfunc_top: Top gate contact work function (eV)

wfunc_bot: Bottom gate contact work function (eV)

mlong: Longitudinal relative electron mass ratio

mtrans: Transverse relative electron mass ratio

eps_top: Top insulator relative dielectric constant

kox_top: Top insulator relative dielectric constant

kox_bot: Bottom insulator relative dielectric constant

dec_top: Conduction band offset between the substrate and the top gate insulator (eV)

dec_bot: Conduction band offset between the substrate and the bottom gate insulator (eV)

3) Output:

At the end of each simulation, a number of postscript files and ".dat" files will be written to the "output/" directory. These data contained in these files is described at the end of this section. Errors can be reported to celab@ecn.purdue.edu by forwarding the Matlab error message and the input deck to the aforementioned email address. The ".dat" files contain ascii data that has been plotted in the corresponding postscript file with the same name but with a ".ps" extension. The data files generated are:

Output.dat: Containing the entire simulation convergence information.

Esub_X.dat: The subband energy vs x for the lowest subband, for the entire bias range.

Fermi_X.dat: The calculated Fermi energy of each Büttiker proble, for the entire bias range.

ID_VG.dat/ID_VD.dat: The I-V characteristics for the bias range specified.

N2D_X.dat: Integrated 2D electron density (cm^{-2}) along the channel.

Ec_X_Y.dat: The full 2D potential profile for the last bias point.

Ne_X_Y.dat: The electron concentration (cm^{-2}) for the last bias point.