Stony Brook University

Theoretical Analysis of Prospective Nanoelectronic Devices

A Dissertation Presented

by

Thomas John Walls

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Physics

Stony Brook University

August 2008

Stony Brook University

The Graduate School

Thomas John Walls

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Konstantin K. Likharev – Dissertation Advisor
Distinguished Professor, Department of Physics and Astronomy

Harold J. Metcalf – Chairperson of Defense
Distinguished Teaching Professor, Department of Physics and Astronomy

Dmitri V. Averin
Professor, Department of Physics and Astronomy

Serge Luryi
Distinguished Professor and Chair
Department of Electrical and Computer Engineering

This dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Theoretical Analysis of Prospective Nanoelectronic Devices

by

Thomas John Walls

Doctor of Philosophy

in

Physics

Stony Brook University

2008

The goal of this work is theoretical analyses of two prospective nanoelectronic devices. The first of them, the metal-oxide-semiconductor field-effect transistor (MOSFET) is the cornerstone of present day integrated circuit technology. We explore the ultimate size scaling limits of MOSFETs as their critical feature dimensions are scaled down below 10 nm. At that size, the physics of electron transport in these devices radically changes from quasi-equilibrium drift-diffusion to ballistic propagation. A proper description of such a regime requires a quantitative account of two-dimensional electrostatics and quantum mechanical effects such as direct source-to-drain tunneling. We have carried out extensive numerical simulations of nanoscale transistors, including these effects, using a self-consistent solution of the Poisson and Schrödinger equations. The results show that advanced silicon transistors can provide voltage gain at gate lengths as small as 4 nm. However, the device sensitivity to unavoidable variations of the dimensions during fabrication, and power consumption grow exponentially in this regime.

The second device under study, the superconductor balanced comparator, is based on the quantum mechanical quantization of flux through superconducting loops. This device is a key component of Rapid Single-Flux-Quantum (RSFQ) circuits which can be used for digital signal processing at sub-THz frequencies, with extremely small power consumption, albeit at deep refrigeration. Alternatively, the comparator may be used for measurement of current (or magnetic flux) with sub-picosecond time resolution. We show that the signal detection sensitivity of the balanced comparator, with realistic parameters, may be limited only by the fundamental quantum fluctuations.

Contents

List of Figures

List of Tables

Nomenclature

Physical Constants and Global Functions

e	Unit of electron charge: 1.602×10^{-19} C
$\hbar$	Reduced Plank's constant: 1.054×10^{-34} J
i	$\sqrt{-1}$
Φ_0	Magnetic flux quantum: 2.067×10^{15} Wb
m_0	Free electron mass: 9.109×10^{-31} kg
$\Theta(x)$	Heaviside step function

Ballistic NanoFET

L_c	Transistor channel length
L_g	Transistor gate length
L_{ext}	Length of doped thin extensions
L_{BB}	Source to drain "bulk-to-bulk" length
t_B	Bulk electrode thickness
t_c	Transistor channel length
t_{ox}	Insulator oxide thickness
W	Transistor width
E_{F}	Electrode Fermi energy
μ_{F}	Electrode chemical potential
$\Phi(x, z)$	Electrostatic potential

T	System temperature
$E_{z,n}$	Channel confinement energy
V_d	Transistor source-drain voltage
V_g	Transistor source-gate voltage
V_s	Source electrode potential, taken to be ground
V_t	Transistor threshold voltage
V_{DD}	Circuit drive voltage
α	Ratio of silicon and insulator dielectric constants $\epsilon_{si}/\epsilon_{ox}$
ϵ_{ox}	Dielectric constant of the insulator oxide
ϵ_{si}	Dielectric constant of silicon
ρ	Transistor electron density
G_v	Voltage gain
$g_{s,v}$	Electron spin and valley degeneracies
J	Transistor current density
J_{OFF}	Off-state current density
J_{ON}	On-state current density
n_D	Electrode doping density
P	Transistor power consumption
$\bar{A}$	Channel coupling strength of sub-band modes
$\bar{m}$	Effective 3-D electron mass
$\varphi_n(z)$	Wavevector perpendicular to the channel in the thin channel
$\varphi_w(z)$	Confined state $\hat{z}$ wavefunction
$\xi_w(z)$	Wavevector perpendicular to the channel in the bulk electrode
a_{nw}	Overlap of electrode, channel wavefunctions

m_h Heavy electron mass

m_l Light electron mass

$m_{x,y,z}$ Effective electron mass in $\hat{x}$, $\hat{y}$, $\hat{z}$

$q_n(z)$ Wavevector perpendicular to the channel

Λ Transistor characteristic scaling length

λ Effective "switching activity" factor

μ Electron mobility

θ Electron bath incident angle

l Electron mean free path

$\mathcal{F}_j(x)$ Fermi-Dirac integral of order j, 2.21

$\mathcal{D}(E)$ Probability of electron transmission through the channel

$f(E)$ Fermi distribution function, 2.16

$\Delta_{x,z}$ FDE mesh points spacings in $\hat{x}$, $\hat{z}$

Josephson Junction Comparator

β_c Stewart-McCumber damping parameter, 5.85

λ Effective inductance parameter, 6.20

φ Josephson phase difference

σ_i Pauli matrices

γ Comparator damping parameter, 5.51

γ_Q Comparator scale of quantum fluctuations, 5.41

γ_T Comparator scale of thermal fluctuations, 5.41

$\mu(t)$ Comparator potential inversion function, 5.43

I_x Comparator signal current

i_x Normalized signal current, 5.44

ω_c	Characteristic frequency, 5.22
ω_J	Angular Josephson frequency, 5.11
ω_p	Plasma frequency, 5.29
E_c	Josephson junction charging energy
E_J	Characteristic Josephson junction energy, 5.15
I_c	Josephson junction critical current
ρ	Comparator density matrix
J	System propagator
k	Comparator-qubit mutual inductance parameter, 6.7
κ	Comparator-qubit coupling coefficient, 6.10
L	Inductance
Γ	Dephasing of measurement
C	Junction capacitance
I	Fidelity of information in measurement
K	System kinetic energy
M_J	Comparator equivalent mass, 5.36
Q	Charge
$U(\varphi)$	System potential energy

Acknowledgements

The author would first and foremost like to thank his adviser, professor Konstantin Likharev, whose guidance and tutelage have been a source of constant inspiration. I owe him no small debt of gratitude as his invaluable insights, dedication and vision have been the driving force of this work. Any results contained herein are simply a testament to his strength as a scientist and a mentor. It has truly been an honor to know him both personally and professionally.

I would also like to thank my committee members, professors Harold Metcalf, Dmitri Averin and Serge Luryi, who have kindly offered time from their extremely hectic schedules to review the dissertation and sit on the oral exam board. I thank professor Metcalf personally for offering timely advice and support throughout my graduate career. I also thank professor Averin for the fantastic opportunity to collaborate on the study of qubit measurements. His development of the "information/backaction" measure of information fidelity was the foundation of those results.

The base of this dissertation is built in large part on direct collaboration with Timur Filippov and Viktor Sverdlov, post-doctoral researchers with professor Likharev. They were instrumental to my understanding of the dynamics of Josephson junctions and ballistic MOSFETs respectively. I thank Jingbin Li, a fellow graduate student, who calculated and analyzed the single-gate transistor results presented below.

I sincerely appreciate the entirety of the faculty and staff at Stony Brook University but would especially like to mention Pernille Jensen, Pat Peiliker, Diane Siegel, Sara Lutterbie, and Bob Segnini for their personal support and often underappreciated hard work.

I acknowledge financial support at different stages by the Office of Naval Research and the Semiconductor Research Corporation. I also acknowledge the generous donation from Harold Weinstock at the Air Force Office of Scientific Research of computing time at the Department of Defense's High Performance Computing Center. Additional computing resources were provided by the "Seawulf" cluster at Stony Brook University and "New York Blue" cluster at Brookhaven National Lab.

Finally, I would like to especially thank my fiancee, Deborah Swantek. Without her love, support and seemingly unending patience, this work would not have been possible.

Chapter 1

Introduction

The technological capacity for processing, storage and communication of information has progressed at an astonishing, exponential rate for the past half-century. The electronic integrated circuit (IC) has been the workhorse for information processing which has certainly seen a golden age over the past fifty years. The performance improvements of processing ability have come largely from the continued shrinking of circuit components allowing for lower capacitive recharging time (*i. e.* faster clocking frequency) and greater component packing density per chip.

The key point of the era was the introduction of complementary metal-oxide-semiconductor (CMOS) technology. CMOS builds logic circuits from metal-oxide-semiconductor field-effect transistors (MOSFETs) which have identical, symmetric current characteristics as a function of applied gate and drain bias. The MOSFET is essentially two semiconductor electrodes doped with atoms which donate either electrons to the conduction band (n doping) or holes to the valence band (p doping). The electrodes are connected through a semiconductor of the opposite doping. The channel is covered by a gate electrode whose field penetrates the middle region. A n-type MOSFET consists of n-doped electrodes and the basic principle of operation is that a voltage applied to the gate depletes the hole concentration in the p-doped region creating a conduction channel for the flow of electrons. A reverse bias increases the hole concentration further restricting the electron flow, so the device acts as a three terminal switch. The compliment to the n-type transistor is a p-n-p device; a positive gate bias turns on a n-type MOSFET while shutting off a p-type MOSFET.

As an example, Fig. 1.1 shows a basic building block of integrated circuitry, the CMOS inverter. Two oppositely doped MOSFETs are tied together at their gates with input voltage V_{in}, the source node of the n-FET is connected the source line (taken here as ground) and the source node of p-FET is con-

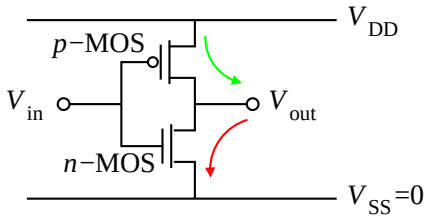

Figure 1.1: CMOS inverter circuit.

nected to circuit drive voltage V_{DD}. The transistors are connected at their drain nodes as output voltage V_{out}. An input voltage $V_{in} = V_{DD}$ turns off the p-FET while turning on the n-FET. This allows the flow of current (shown by the red line) through the n-FET bringing the V_{out} to the source line level $V_{out} = 0$. On the other hand, zero voltage turns off the n-FET and opens the p-FET allowing current shown by the green line and brings the output voltage $V_{out} = V_{DD}$.

CMOS technology has two large advantages of over other kinds of electronic circuits. First, the complementary nature of the transistors mean that noise components tend to cancel leaving CMOS highly immune to noise fluctuations. Second, the inverter example demonstrates that current only flows through the device during the transient of switching from the on/off states. Hence, CMOS circuits have a very low static power consumption.

Since its inception, the number of transistors in CMOS circuits has doubled roughly every 24 months, an observation first made famous by G. Moore [1] (his original assessment was doubling every year [2]). The "International Technology Roadmap for Semiconductors" (ITRS), a consortium of industry microchip manufacturers, projects this scaling will continue for at least another decade [3]. The increase in transistor count comes in large part from shrinking the device dimensions, and while impressive, this fantastic trend cannot continue indefinitely as sizes reach near atomic scales. The days of exponentially increasing processing power from technology generation to generation by traditional device scaling techniques are drawing to a close.

New and novel architectures will be required to continue to push our capacity for processing, transmitting and storing information much beyond present day limits. One such idea is the use of quantum entanglement for massively parallel computation in an ever increasing number of physical systems [4]. The range of problems where such an architecture will be advantageous is still limited however, and a feasible implementation of such a system still seems a long way off. Another proposed technique is the use of quantized magnetic flux in superconducting circuitry. Termed "rapid single-flux quantum" (RSFQ) logic [5, 6], quantized flux units may be manipulated orders of magnitude faster

than today's electronics with extremely low power requirements. Unfortunately, the deep refrigeration required for the superconducting circuitry makes this architecture impractical for wide spread adoption. There are numerous other proposed solutions for next generation computation (*e. g.* single-electron transistor, single-atom transistor, hybrid CMOL technology, etc). All have evident advantages and drawbacks and all are being actively pursued as unique paradigms.

In fact, even pushing traditional silicon structures to their ultimate size limits will require a fundamental shift in their operational physics. In present day devices, current carrying electrons can be accurately described by a particle theory. As the devices are pushed to their nanoscale limits, the quantum mechanical wave properties of the current carrying particles will have an increasing impact on an accurate description of the device properties.

This dissertation is divided in to two main parts. Part I develops theory and numerical algorithms to analyze the performance characteristics, and find optimal operating conditions of traditional silicon devices when scaled down to their ultimate nanoscale limits. Part II is aimed to develop an understanding of the noise characteristics of an analog (superconducting) signal sampling circuit used in existing RSFQ electronic devices and its potential application for future quantum computing implementations.

Part I

Ballistic Thin-Channel MOSFETs

Introduce knowledge gradually, avoiding bloodshed if possible.
M. E. Saltykov-Shchedrin, taken from "Dynamics of Josephson Junctions and Circuits" by K. Likharev.

Chapter 2

Theory of Ballistic NanoFETs

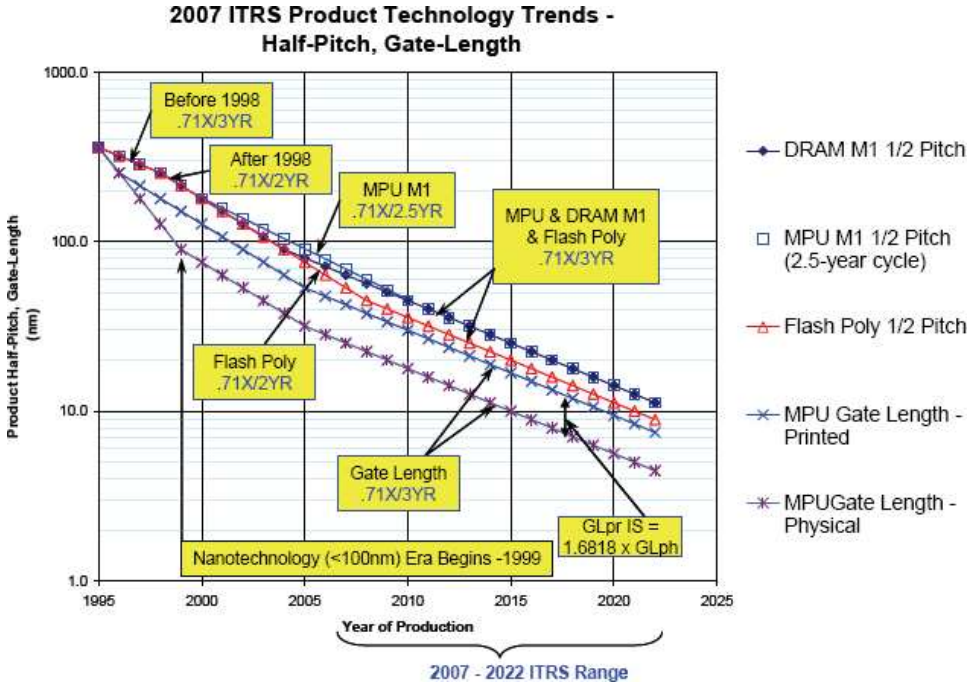

Figure 2.1: Projections of transistor critical length scaling [7].

The metal-oxide-semiconductor field-effect-transistor (MOSFET) is the corner stone of integrated circuit technology used for the vast majority of present day information processing. The continued shrinking of the MOSFET components increases the overall speed at which information may be processed in two primary ways. First, increased circuit packing density allows more transistors; providing more computing power in a single chip. Second, lower device and circuit capacitance allows the system to be clocked at a higher speed,

and smaller device dimensions reduce the power consumed per clocking cycle. Figure 2.1 shows the historical trend and projections of scaling down the physical device lengths and device half-pitch (the length between wires in adjacent components) for present day MOSFETs. Namely, the projections show that the physical length of the transistor gate will fall below 10 nm by 2015.

A model of the traditional n-type "bulk" channel MOSFET is shown in Fig. 2.2. It consists of two n-doped (phosphorus) silicon source / drain regions separated by a p-doped (boron) channel region. The channel is covered by a gate electrode insulated from the channel traditionally by silicon dioxide due to its ease of fabrication and perfect match with the silicon lattice. The transistor acts as a three terminal switch and a voltage amplifier, controlling very large current with relatively small signals. The gate electrode was historically a metallic conductor, most commonly aluminum. Modern implementations use highly doped polysilicon electrodes which have better fabrication properties and ease tuning of the gate workfunction which affects the device threshold voltage.

Formally, we should analyze both n-type and p-type MOSFETs used in CMOS logic, as the different hole mass slightly changes the behavior of the p-type device. However, much modern circuitry uses so-called dynamic logic which requires a factor of 2 fewer transistors over static implementations. Figure 2.3 shows a typical precharge-evaluate two footed NAND gate [8]. Initially, the clock pulse is low, and the p-FET (shown in red) is precharged so $V_{out} = V_{DD}$. When the clocking pulse goes high, the evaluation n-FET (shown in green) is turned on, a discharge path to ground is created and V_{out} may be pulled to ground depending on input nodes A_1, A_2. In a typical case, the p-FET may shut before the evaluation pulse. The voltage V_{DD} remains stored for a short time in the line capacitance near V_{out}. The speed of the gate is then entirely defined by the switching dynamics of the n-FETs. Hence, most of the work in this field and all the work in this dissertation is focused on evaluation of the n-type device.

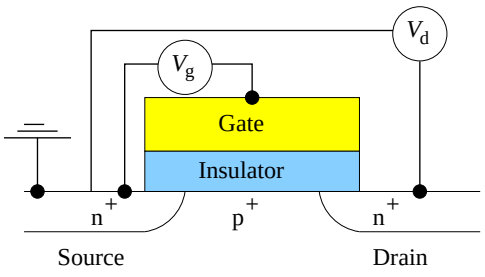

Figure 2.2: Single gate transistor.

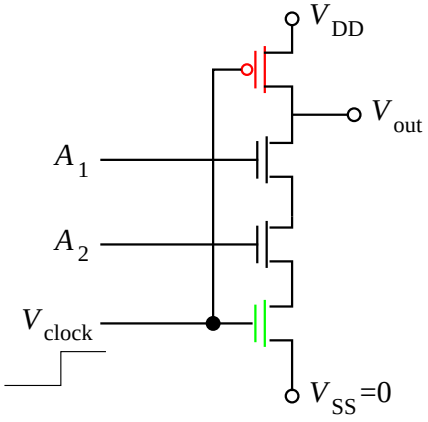

Figure 2.3: Two terminal Dynamic Logic NAND gate.

A full discussion of the physics of this device is out of the scope of this work, for an in-depth treatment see the monograph by Sze [9]. The important point to note is that as the length of the gate is reduced, electrostatic control of the channel region may be degraded, hindering device performance. Additionally, the field produced by the drain electrode may begin to affect the electrostatics inside the depletion region. To mitigate these so-called "short channel" effects, all device lengths and the supply voltage should be scaled down similarly, maintaining a constant electric field in the device [10]. To reduce the channel thickness, the doping density of the electrodes is also increased by the same proportion, reducing the effective depletion width at a given voltage.

As the physical gate length of the transistor approaches 10 nm however, increasing the channel doping by the proper scaling ratio may not produce a depletion width scaled by the same factor. 2-D electrostatic effects, such as "drain-induced barrier lowering" (DIBL) may begin to affect the channel potential.

At such small gate lengths, constant field scaling rules dictate extremely thin channels. The channel thickness reaches a level where quantum electron states becomes strongly confined in the z-direction. The confined nature of the electric field across the depletion region requires an exponential dependence in the transport direction. To first order the potential in the center of the channel is proportional to $\propto \exp(-\pi L_c/2\Lambda)$, where L_c is the device channel length, and Λ the characteristic scaling length. Using standard separation of variables [11] and matching approximate boundary conditions, Frank *et al.* [12] showed that length Λ must satisfy the following estimate:

$$\epsilon_{si} \tan(\pi t_{ox}/\Lambda) + \epsilon_{ox} \tan(\pi t_c/\Lambda) = 0, \tag{2.1}$$

where ϵ_{si}, ϵ_{ox} are the dielectric constants of the channel and insulator oxide, t_{ox} is the oxide thickness and t_c is the channel thickness or depletion width. Short-channel and DIBL effects will begin to onset when $L_c/\Lambda < 2$, or for practical devices when the gate length L_g reaches approximately $10 \sim 20$ nm [13]. Additionally, for sub-20 nm devices, the drift-diffusion model, which has had so much success describing device physics, may begin to fail due to the limited number of particle interactions through the transport channel. Coupled with the lowered particle mobility from such high doping levels, a completely new model of transistor dynamics needs to be explored to push device scaling beyond 10 nm gate lengths.

The goal of this work is to characterize and optimize MOSFET structures beyond the 10 nm frontier. We explore the performance of these nanometer scale devices and examine the ultimate scaling limits of silicon transistors including consideration for fabrication tolerances and power consumption limitations. We have developed a novel and flexible device simulator as a self-consistent solution of the Poisson and Schrödinger equations which calculates dynamics for a vast range of device parameters and material properties. The simulator is capable of calculating several treatments of channel electrons in increasing levels of complexity including classical [14], WKB approximation [15], 1-D [16], and a full 2-D quantum solution. In addition it is written to run on either a single processor machine or distributed parallel architecture (see appendix B.2), and will configure itself for either architecture with a simple compile time flag. The simulator was written (and re-written multiple times) entirely within our group, using standard algorithms where appropriate.

2.1 Ballistic Model

Proper scaling techniques will control unwanted short channel effects up to a point. Ultimately however, more advanced device structures will likely be required to reach sub-10 nm device lengths. In particular, dual gate (DG) structures have become an attractive option for reaching ultimate size scaling limits because the electrostatic potential closely follows that of the gate, minimizing the short-channel effects over the single gate counterparts. A comparison of the properties of single and double gate transistors is discussed in chapter 3. Specifically, DG structures have twice better control of the channel electrostatics over their single gate counterparts while lowering the device power consumption almost an order of magnitude.

Fabrication of double-gate structures however, requires a move to more complex silicon-on-insulator (SOI) technology where layered silicon-insulator structures are employed. Because the silicon channel is surrounded on both

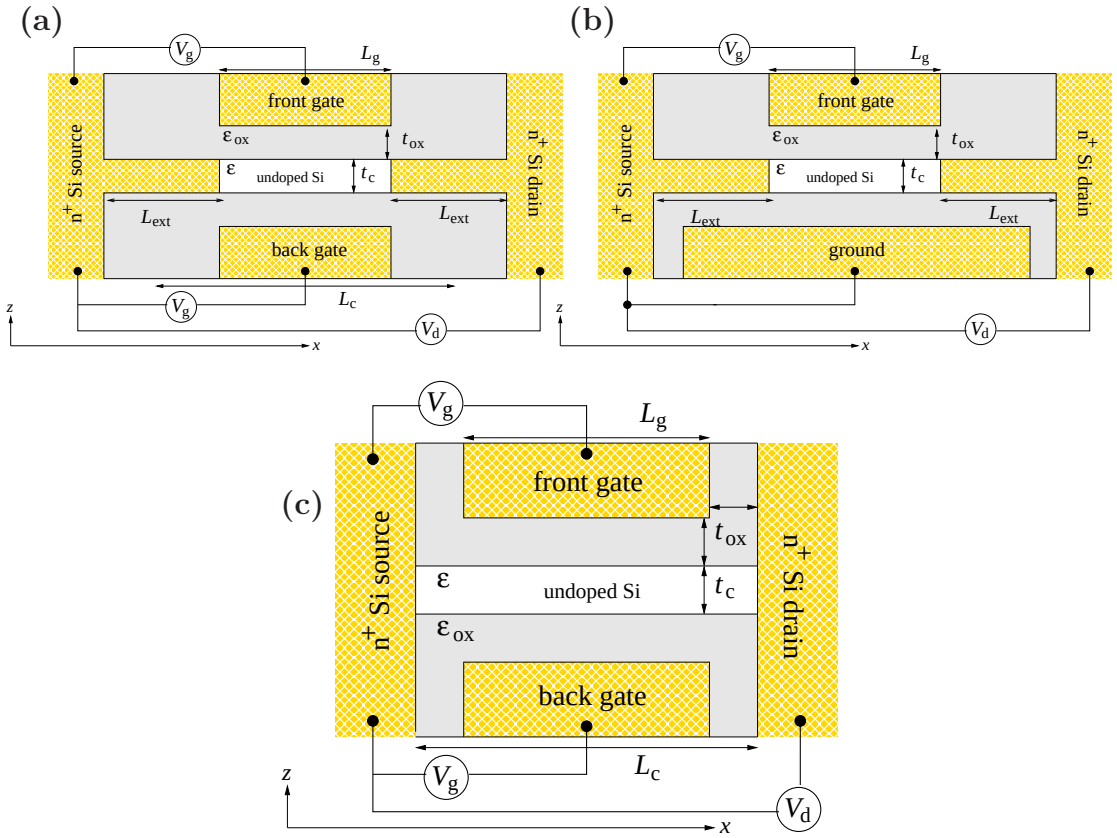

Figure 2.4: Model of ballistic MOSFET structures for device with (a) thin, doped channel extensions, (b) single-gate, grounded back-plane, and (c) bulk electrodes.

sides by an insulator, the particles are confined by the abrupt Si-SiO_2 interface at the channel edges. The silicon layer creating the channel may be made nearly arbitrarily thin with the electrons perfectly confined. Advanced fabrication processes are being actively being explored. Popular options are the self-aligned planar structures, shown in Figs. 2.4(a), 2.4(c), so-called "finFET" structures, shown in Fig. 2.5 [17], nearly identical "vertical"-fin structures, and gate-all-around structures [18]. While the gate electrode capping the silicon channel makes the finFET device slightly different than the planar structures, the fringing fields near these corners are not too important. The key point is that in all these structures, the gate has optimal control over the channel electrostatics and they will all scale similarly. The planar DG MOSFET may be seen as a close approximation to the "ultimate" MOSFET.

Figure 2.4 shows the three model FET structures under consideration. One with thin doped channel extension regions (panel (a)) and the other with the conducting channel connecting two wide bulk electrodes (panel (c)). Panel (b) shows the single gate transistor similar to the model with doped extensions with a long, grounded back-plane.

The electrodes are connected to a silicon channel separated from two similarly biased doped silicon gate electrodes by a silicon-dioxide insulator. In these systems, the particles are confined by the discontinuity at the Si-SiO_2 interface, so the channel doping becomes unnecessary. In fact, any doping of the channel produces randomly located single scatters degrading the electron mobility and causes unpredictability in device fabrication. So doping of the channel becomes completely undesirable and we consider an intrinsic channel without any fixed scatters.

The inelastic, electron-phonon scattering time in silicon, at room temperature, is about 10 fs [19], which yields an energy relaxation time for high energy carriers near $\tau \approx 200$ fs [20, 21]. The kinetic energy of electrons in the channel is near their thermal injection energy with average velocity

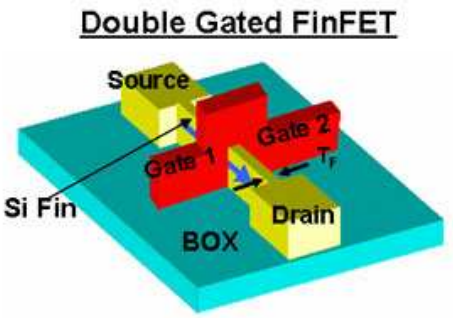

Figure 2.5: Schematic of a double-gate finFET [17].

$\langle v \rangle = \sqrt{2T/m_x} \approx 2 \times 10^7$ cm/s, with the effective electron mass in the transport direction $m_x = 0.19 m_0$, T the system temperature (in energy units), and room temperature thermal value $T = 26$ meV. This gives a mean free path for inelastic scattering $l = \tau \langle v \rangle \approx 40$ nm, *i. e.* much larger than any device dimensions under consideration. Hence, we are left with only elastic surface roughness scattering.

The best experimental results [22–25], and models [26], indicate that the electron mobility may be quite high indeed, at least as high as $\mu = e\tau/m_x \approx 250$ cm^2/Vs for a channel thickness of 2.5 nm. This leads to an elastic scattering time $\tau = \mu m_x/e \approx 27$ fs with a elastic mean free path $l \approx 5.5$ nm. While this may still be shorter than the total channel length, the transport is effectively regulated in a fairly narrow section of the channel near the top of the potential barrier. In this case, the electrons may be modeled as traversing the channel without encountering any scattering events and the transport considered ballistic. These are the best mobilities for now, future investigation may further reduce the surface roughness, increasing l even higher. But even at present mobilities, the ballistic assumption may be quite accurate. In any case, ballistic transport represents the limit of an ideal device when scaled down to ultimate lengths.

Ballistic transport implies that the particle wavefunctions will be coherent along the entire length of the channel and a full quantum treatment to account for their wave nature is required. In this respect, ballistic devices may seen as one implementation of so-called "electron optics" systems [27, 28].

For the electrodes, analysis of dopant fluctuations yields the opposite conclusion regarding doping levels. Any number of dopants N_D will have natural fluctuations with standard deviation $\sigma_{N_D} = \langle (\Delta N_D)^2 \rangle = N_D^{1/2}$. In order to maintain device reproducibility above 90%, the doping should be high enough to keep $N_D/N_D^{1/2} < 10\%$. Since the volume of the electrode regions is on the order 100 nm^3, the electrode doping should be at least as high as 0.1 nm^{-3}, well past the degeneracy threshold for silicon, but still below the solid solubility limit. In most of this work we accept a doping density $n_D = 3 \times 10^{20}$ cm^{-3} $= 0.3$ nm^{-3}, and all donors are assumed activated. Deviations from this doping value will be clearly noted.

Due to its close approximation to the "ultimate" MOSFET, it is not surprising that these devices have been aggressively studied in recent years. The first analytical description a ballistic MOSFET was given by K. Natori [29], based on the 1-D capacitive response of the channel electrons to the gate potential (also see the pioneering work of Frank *et al* [30]). A similar approach was taken by Lundstrom [31] to describe ballistic FETs in terms of a somewhat phenomenological back-scattering probability. The pioneering results of Natori

are outlined in section 2.2, but his 1-D theory is inadequate to describe the two dimensional electrostatic effects. These effects were subject of an early analytical study [32] based on a parabolic approximation for the channel potential [33, 34]. Unfortunately, the field confinement in thin channel devices gives an exponential behavior of the potential [12] making the parabolic approximation only asymptotically valid for long channels. This exponential model of Frank has been used to develop a compact model of the threshold voltage for a surrounding gate configuration [35] and to derive closed-form expressions for the current density [36–39] in long channel devices.

While none of the analytic theories above can fully account for all the effects contributing to performance degradation as the transistor is scaled down, the reduced system size makes it an attractive option for numeric simulation. Other groups have developed models of transport which are capable of including scattering in the channel region [40]. In particular, an approach based on calculation of non-equilibrium (Keldysh) Green's functions [41, 42] has received a lot of attention. The "non-equilibrium Green's function" (NEGF) method has been used in its one dimensional form to analyze the back-scattering coefficient proposed by Lundstrom [43] and to analyze the effect of source-to-drain tunneling through the potential barrier [44, 45]. This model was later expanded to a two dimensional solution of the Green's function in both real [46] and k-space [47]. Most recently a full three dimensional model was proposed [48]. Despite being more general, inclusion of scattering effects is extremely cumbersome so most results presented in these works neglect all scattering in the undoped part of the channel as well. In this case the NEGF method essentially reduces to the direct solution of the Schrödinger equation developed in this dissertation. For example, Fig. 2.6 shows the calculated source-drain current in the NEGF formalism and the results of our work for two different devices [44, 49]. The simple, rapid solution of the Schrödinger equation indeed reproduces the results of the more complex NEGF method.

Another group has developed a self-consistent numeric solution of the Schrödinger-Poisson equations based on the quantum-transmitting boundary method (QTBM) [50, 51]. However, the main analyses of these works was for the so-called done-bone structure and the results presented not extensive enough for a clear picture of ultimate scaling to emerge.

The simulator developed within our own group to analyze ultra-small MOSFET devices originally began as a solution of the Schrödinger equation in the WKB approximation [14, 15] and was later expanded with the results presented in this dissertation [16, 49, 52, 53].

In all cases, the solution of the Schrödinger equation is based on the Hartree approximation, ignoring the exchange interaction term resulting for the overlap

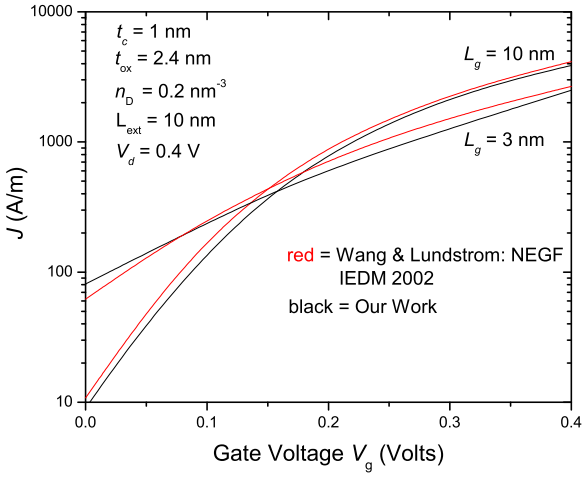

Figure 2.6: Comparison of numerical results calculated with the NEGF to the direct solution of the Schrödinger equation.

of electron wavefunctions. In the bulk electrode regions, this effect amounts to a universal shift in the level of the conduction band. It is accounted for in the bulk regions phenomenologically by taking the bottom of the conduction band as zero as using experimentally obtained values for the band gap. In addition, the Fermi energy is calculated as integral over momentum states, independent of the exchange-correlation terms, so these effects are encompassed through the use of experimental values for the effective masses in each conduction valley. In the channel region, there is no positive charge background, so the net Coulomb effect is not screened by a background charge. As a result, the net Coulomb potential felt by an electron in the channel far eclipses the exchange term and the exchange correlation effect may be ignored here to a good approximation. As a crude model we may write the Coulomb potential between two particles a distance L apart (assuming a cylindrical shape in the device width) as

$$U_c \approx e^2 n_D t_c L \ln\left(\frac{L}{t_c}\right). \tag{2.2}$$

The exchange correlation term

$$U_{xc} = -\frac{e^2}{2}\int \frac{\psi_i^*(\boldsymbol{r})\psi_j(\boldsymbol{r})\psi_j^*(\boldsymbol{r}')\psi_i(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\, d\tau d\tau', \tag{2.3}$$

has integrand proportional to inverse radial distance $1/r$, hence the exchange

potential is roughly proportional to the cube root of the charge density [54]

$$U_{xc} \propto \frac{e^2}{2} n_D^{1/3}. \tag{2.4}$$

This yields ratio

$$U_c/U_{xc} \propto 2n_D^{2/3} t_c L \ln\left(\frac{L}{t_c}\right) \gtrsim 10, \tag{2.5}$$

for practical cases. In the absence of the positive background, the exchange correlation effect is largely overshadowed by the Coulomb potential.

Several experimental devices approaching the nanoscale limits have also been realized [55–60] including finFET structures [61–63], which have scaled the physical gate length to the 10 nm threshold. Some of the results have been quite dramatic, with evidence that the devices were sufficiently short to demonstrate evidence for the observation of direct source-to-drain tunneling of the electrons. For reviews see [64–67].

2.2 Main Analytical Relations

The dominant feature of the DG devices shown in Fig. 2.4 is the ultra-thin body of the channel. For devices of practical interest, the channel thickness $t_c < 3$ nm. In this case the particle wavefunction is strongly confined in the z-direction with energy spectrum

$$E_{z,n} = \frac{\hbar^2 n^2 \pi^2}{2 m_z t_c^2}, \tag{2.6}$$

with Plank's constant $\hbar = 1.054 \times 10^{-34}$ Js, effective mass m_z perpendicular to the channel. For $t_c = 2$ nm and free electron mass $m_z \approx m_0$, the confinement energy of the lowest state is $E_{z,1} \approx 100$ meV.

The constant energy surface [9] for the band structure of silicon is shown in Fig. 2.7. Each doubly degenerate valley is composed of a light electron effective mass ($m_l = 0.19 m_0$) and one heavy effective mass ($m_h = 0.98 m_0$). The confinement energy for valleys with the heavy mass oriented in $\hat{x}$, $\hat{y}$ is roughly 5 times larger than the $\hat{z}$ valleys and may be ignored as a first approximation.

The large confinement energy creates effective sub-bands for transport in the channel with only the lowest level or levels contributing. Considering only the lowest sub-band, and a sufficiently wide device, the electron wavefunction

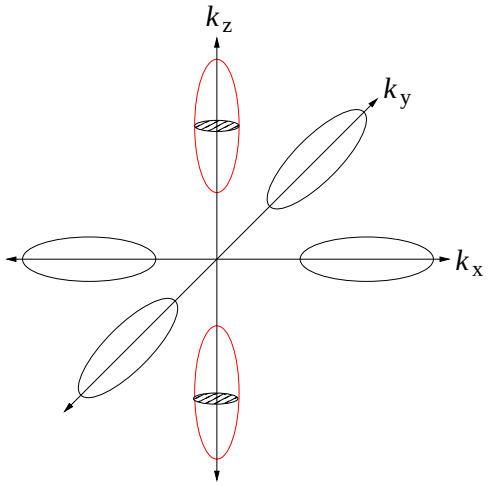

Figure 2.7: Constant energy surfaces for Si [100]. The red ellipses indicated electron heavy mass oriented in $\hat{z}$.

may be partitioned

$$\Psi(x,y,z) = \frac{2}{t_c}\cos^2\left(\frac{\pi z}{t_c}\right)\psi(x)e^{ik_y y}, \tag{2.7}$$

and the electron "sees" effective potential energy in the transport direction

$$U(x) = \bar{\Phi}(x) + E_{z,1}, \tag{2.8}$$

with average potential (see section 3.1)

$$\bar{\Phi}(x) = \frac{2}{t_c}\int_{-t_c/2}^{t_c/2}\Phi(x,z)\cos^2\left(\frac{\pi z}{t_c}\right)dz. \tag{2.9}$$

The total effective potential energy is simply increased by confinement energy $E_{z,1}$. Note that this also increases the particle band-gap by $E_{z,1} + E^h_{z,1}$ with hole confinement energy $E^h_{z,1}$. All devices considered will assume the silicon channel is oriented in the [001] direction.

2.2.1 Fermi Energy

Deep inside the doped electrodes, all electric fields are screened and the potential becomes constant. We define the Fermi level for the electrode in this region relative to the bottom of the conduction band. An ellipsoid with prin-

cipal axis lengths a,b,c occupies a volume $V = (4\pi/3)abc$. Hence the volume of phase space with energy $E < E_{\mathrm{F}}$ is

$$V_p = \frac{4\pi}{3}(2m_x E_{\mathrm{F}})^{1/2}(2m_y E_{\mathrm{F}})^{1/2}(2m_z E_{\mathrm{F}})^{1/2}, \tag{2.10}$$

with total number of occupied states

$$N = g_s g_v \frac{V(2E_{\mathrm{F}})^{3/2}}{6\pi^2\hbar^3}\sqrt{m_x m_y m_z}, \tag{2.11}$$

accounting for spin and valley degeneracies g_s, g_v. For a given density n_D of activated donors, the Fermi energy for the electrode is then found as

$$E_{\mathrm{F}} = \frac{\hbar^2}{2\bar{m}}\left(\frac{\pi^2 n_D}{2}\right)^{2/3}, \tag{2.12}$$

where we have used $g_s = 2$, $g_v = 6$, and

$$\bar{m} \equiv (m_x m_y m_z)^{1/3} = \left(m_l^2 m_h\right)^{1/3}. \tag{2.13}$$

2.2.2 Bulk Electrode Density

The density of electrons with energy

$$E = \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y} + \frac{k_z^2}{m_z}\right), \tag{2.14}$$

in the bulk electrodes may be found by integration

$$n_{3D} = \frac{g_s g_v}{(2\pi)^3}\iiint f(E)\, dk_x\, dk_y\, dk_z, \tag{2.15}$$

where $f(E)$ is the Fermi distribution function

$$f(E) \equiv \left\{1 + \exp\left[(E - E_{\mathrm{F}})/T\right]\right\}^{-1}, \tag{2.16}$$

and μ_{F} the electrode chemical potential.

Replacement of the terms k_{i}, $\mathrm{i} \in \{x, y, z\}$ in ellipsoid Eq. (2.14) with

$$k_{\mathrm{i}} = \sqrt{2m_{\mathrm{i}}}k'_{\mathrm{i}}, \tag{2.17}$$

yields constant energy surface $r^2/\hbar^2 = k'_x + k'_y + k'_z = [E - \Phi(x,z)]/\hbar^2$ with differential volume element for incident angles θ, ϕ,

$$dk_x\, dk_y\, dk_z = \frac{(2\bar{m})^{3/2}}{\hbar^3} r^2 \sin(\theta)\, dr\, d\theta\, d\phi, \tag{2.18}$$

and $\bar{m}$ defined by Eq. (2.13). In the two dimensional case, the constant energy surface gives differential area element

$$dk_x\, dk_y = \frac{\sqrt{m_x m_y}}{\hbar^2}\, dE\, d\theta. \tag{2.19}$$

Again using $g_s = 2$, $g_v = 6$, Eq. (2.15) may be rewritten

$$n_{3D}(x,z) = \frac{3(2\bar{m}T)^{3/2}}{\pi^2\hbar^3} \mathcal{F}_{1/2}\left\{[\mu_{\mathrm{F}} - \Phi(x,z)]\,/T\right\}. \tag{2.20}$$

The Fermi-Dirac integral used in Eq. (2.20) of general order j is defined as

$$\mathcal{F}_j(x) \equiv \int_0^\infty \frac{y^n}{1+\exp(y-x)}\, dy. \tag{2.21}$$

Note that this definition lacks a more standard normalization by gamma function $1/\Gamma(n+1)$ [68]. The chemical potential μ_{F} for each electrode is found numerically at the beginning of program execution by enforcing electro-neutrality between the background dopant charge and the free particle charge

$$\frac{3(2\bar{m}T)^{3/2}}{\pi^2\hbar^3} \mathcal{F}_{1/2}\left\{\mu_{\mathrm{F}}/T\right\} = n_D. \tag{2.22}$$

2.2.3 Doped Extension Density

For devices with doping in the ultra thin body region, wavefunction partition (2.7) implies that the full free electron density in the thin doped region may be written as

$$n_{3D}^{\mathrm{ext}} = \frac{2}{t_c} \cos^2\left(\frac{\pi z}{t_c}\right) n_{2D}^{\mathrm{ext}}(x), \tag{2.23}$$

for two dimensional sheet density $n_{2D}^{\rm ext}(x)$. The sheet density may again be found from Fermi distribution (2.16)

$$\begin{aligned} n_{2D}^{\rm ext}(x) &= \frac{g_s g_v}{(2\pi)^2} \iint f(E) k_x \, k_y, \\ &= \frac{g_s g_v \sqrt{m_x m_y}}{(2\pi\hbar)^2} \int_0^\infty \int_0^{2\pi} f(E) \, dE \, d\theta, \end{aligned} \tag{2.24}$$

where we have used area element (2.19). This integral is evaluated as

$$n_{2D}^{\rm ext}(x) = \frac{\sqrt{m_x m_y}\, T}{\pi\hbar^2} \ln\left[1 + \exp\left((\mu_{\rm F} - \bar{\Phi}(x))/T\right)\right]. \tag{2.25}$$

As a side note, this result can also be derived from the 2-D density of states. The number of points in k-space area $\mathcal{A} = \pi k^2$ is given by

$$N = g_s g_v \mathcal{U} \mathcal{A}, \tag{2.26}$$

with unit cell area

$$\mathcal{U} = \left(\frac{L}{2\pi}\right)^2. \tag{2.27}$$

The density of states is given by

$$G_2(E) = \frac{dN}{dk}\frac{dk}{dE}, \tag{2.28}$$

with density per unit area

$$g_2(E) = g_s g_v \frac{\sqrt{m_x m_y}}{2\pi\hbar^2}. \tag{2.29}$$

The total particle density is then

$$n_{2D} = \int g_2(E) f(E) \, dE, \tag{2.30}$$

which reproduces Eq. (2.25).

2.2.4 Charge Density

Given the donor density n_D and electrostatic potential $\Phi(x,z)$, the total charge density in the electrode regions is then calculated by Eqs. (2.20), (2.23) as

$$\rho = -e \times \begin{cases} \left(n_D - \dfrac{3\left(2\bar{m}T\right)^{3/2}}{\pi^2\hbar^3}\mathcal{F}_{1/2}\left[(\mu_{\mathrm{F}} - \Phi(x,z))/T\right] \right), \\ 2\cos^2\left(\dfrac{\pi z}{t_c}\right)\left(n_D - \dfrac{2\sqrt{m_x m_y}\,T}{\pi\hbar^2 t_c}\ln\left[1+\exp(\mu_{\mathrm{F}} - \bar{\Phi}(x))/T\right]\right), \end{cases} \tag{2.31}$$

in the bulk and doped extension regions respectively.

A detailed description of the calculation of the electron density in the channel will be deferred to later chapters. Chapter 3 discusses the 1-D Schrödinger approximation relevant for device with thin electrode extensions and chapter 4 discusses the full 2-D solution required for devices with bulk electrodes.

2.2.5 Device Current

Once a self-consistent solution for the device potential is known, we may calculate the one-direction current density as a summation over incident states. The current of a single quantum (plane wave) state in the bath is

$$\begin{aligned} I_1 &= e\tfrac{\mathrm{i}\hbar}{2m_x}\left(\Psi\partial_x\Psi^* - \Psi^*\partial_x\Psi\right), \\ &= e\frac{\hbar k_x}{m_x}\left|\Psi\right|^2, \\ &= e\frac{\hbar k_x}{m_x L_B}, \end{aligned} \tag{2.32}$$

for a particle with wavevector $k_x = \sqrt{2m_x E}/\hbar$ and wavefunction Ψ normalized to length L_B. The current of the single state moving one direction through the device channel (*i. e.* left to right) is

$$I_k = e\frac{g_s g_v \hbar k_x}{m_x L_B}\mathcal{D}(\boldsymbol{k})f(E), \tag{2.33}$$

where $\mathcal{D}(\boldsymbol{k})$ is the probability that a particle in state $\boldsymbol{k}$ transmits through the channel. The total one direction device current may then calculated by the

usual summation over states

$$I = \sum_{\boldsymbol{k}} I_k = e\frac{g_s g_v \hbar}{m_x L_B} \sum_{\boldsymbol{k}} k_x \mathcal{D}\left(\boldsymbol{k}\right) f(E). \tag{2.34}$$

2.2.6 Landauer Conductance

For a 2-D system, and assuming a large number of incident modes, Eq. (2.34) may be written

$$I = e\frac{g_s g_v \hbar t_B}{m_x (2\pi)^2} \int\limits_{k_x>0} k_x \mathcal{D}(\boldsymbol{k}) f(E)\, d^2k, \tag{2.35}$$

for bulk electrode thickness t_B. Introducing "incident angle" θ such that $k_x = |\boldsymbol{k}|\cos(\theta)$ and using ellipsoidal differential area (2.19) the current may be expressed as

$$I = e\frac{g_s g_v}{2\pi\hbar} \int\limits_0^\infty dE\, f(E) \left[\frac{t_B\sqrt{2m_z E}}{\pi\hbar} \int\limits_0^1 d(\sin(\theta)) \mathcal{D}(E,\theta) \right]. \tag{2.36}$$

The term in square brackets is nothing more that the total transmission probability $\mathcal{D}(E)$ (see section 4.1.2) and the total current is found from the balance of left and right moving states

$$I = e\frac{g_s g_v}{2\pi\hbar} \int\limits_0^\infty dE\, \mathcal{D}(E) \left[f(E) - f(E - eV_d) \right]. \tag{2.37}$$

For small applied drain voltage, in the limit $T \to 0$, $f(E) - f(E - eV_d) \approx eV_d$, and assuming perfect transmission through the channel $\mathcal{D}(E) \to 1$. Equation (2.37) reduces to the Landauer formula for conductance through a single mode quantum point contact

$$I = GV_d = e^2\frac{g_s g_v}{2\pi\hbar}. \tag{2.38}$$

This is the natural result of a transmission probability based formalism [69].

2.2.7 Aperture Limit

In the limit that the channel length $L_c \to 0$, the channel represents a small opening in a thin insulating diaphragm between the conducting electrodes. The conductance of this aperture may be found from the extension of Sharvin's

conductance to the diffusive limit [70]. The device conductance, per unit width is found to be

$$G/W = G_S \left(1 + \frac{3\pi}{8}\frac{t_c}{l}\right), \tag{2.39}$$

which is Sharvin's conductance [71]

$$G_S/W = \frac{2e^2}{h}\frac{k_F^2 t_c}{4\pi}, \tag{2.40}$$

corrected for the diffusive environment with mean free path l.

2.2.8 Capacitance Model

When the channel length is much larger than the both the channel and oxide thicknesses $L_c \gg t_c, t_{ox}$, the potential along the center of the channel develops a long, constant in $\hat{x}$, plateau region. In this case the potential only depends on variation in the z-direction and the device may be modeled simply as two capacitors connected in series along $\hat{z}$ [65]

$$C_{\text{eff}}^{-1} = C_g^{-1} + C_q^{-1}. \tag{2.41}$$

The first term represents the regular geometric capacitance [65]

$$C_g^{-1} = \frac{t_{ox}}{2\epsilon_{ox}} + \left(\frac{1}{12} + \frac{5}{8\pi^2}\right)\frac{t_c}{\epsilon_{si}}, \tag{2.42}$$

where the numerical factors are modified from the single gate result [29]

$$C_g^{-1} = \frac{t_{ox}}{\epsilon_{ox}} + \left(\frac{1}{3} + \frac{5}{8\pi^2}\right)\frac{t_c}{\epsilon_{si}}, \tag{2.43}$$

to account for the double gate structure. The second term in Eq. (2.41) is the "quantum" capacitance whose equilibrium value is simply proportional to the two dimensional density of states [72]

$$C_q = e^2 \frac{2\sqrt{m_x m_y}}{\pi\hbar^2}, \tag{2.44}$$

with electron charge e. For most cases of practical interest $C_q \gg C_g$, so

$$C_{\text{eff}} \approx \frac{t_{ox}}{2\epsilon_{ox}}, \tag{2.45}$$

and the charge in the channel may be written

$$Q_{\text{chan}} = C_{\text{eff}}(V_g - V_t), \tag{2.46}$$

with gate voltage V_g and threshold voltage shift V_t. This shift may seen as simply the difference between confinement energy $E_{z,1}$ and the Fermi energy in the gate electrode [65]. Note that the charge in the channel, and thus the channel potential depend upon the gate bias and threshold voltage, which changes proportional to the Fermi energy in the gate electrode, but is independent of the source and drain voltages. Thus, shifting the threshold voltage may be achieved through engineering of the gate workfunction.

The capacitive model leads to a simple, yet somewhat bulky expression for the current density [29]

$$J \equiv I/W, \tag{2.47}$$

$$J = e\frac{4\sqrt{2m_y}T^{3/2}}{\pi^2\hbar^2}\left[\mathcal{F}_{1/2}(u) - \mathcal{F}_{1/2}(u-\nu_d)\right], \tag{2.48}$$

and in terms of normalized drain voltage $\nu_d = eV_d/T$,

$$\begin{aligned} u &= \ln\left[\sqrt{(1+e^{\nu_d})^2 + 4e^{\nu_d}(e^{\rho}-1)} - (1+e^{\nu_d})\right] - \ln(2), \\ \rho &= \frac{\pi\hbar^2 C_{\text{eff}}(V_g - V_t)}{2eTm_y}. \end{aligned} \tag{2.49}$$

Equation (2.48) may be reduced to more useful expressions in the most interesting cases of subthreshold current $V_g < V_t$, $eV_d \gg T$ [65]

$$J_{\text{T}} = e\left(\frac{2\sqrt{m_y}T^{3/2}}{\pi^{3/2}\hbar^2}\right)\exp\left[e(V_g - V_t)/T\right], \tag{2.50}$$

and saturation current $\mathcal{F}_{1/2}(u) \approx (2/3)u^{3/2}$ [29, 65],

$$J_{\text{S}} = \frac{4\sqrt{2}\hbar}{3\sqrt{e\pi}m_y}\left[C_{\text{eff}}(V_g - V_t)\right]^{3/2}. \tag{2.51}$$

While this model misses important factors required to analyze the device scaling to short channel lengths, they provide intuitive insight into the dynamics of ballistic devices. Namely, Eqs. (2.50), (2.51) show that a long channel ballistic transistor would have ideal, subthreshold slope $\propto 1/T$ and perfect current saturation. These are the most important transistor characteristics: the

subthreshold slope determines how much current leaks through the transistor in the off-state and the saturation current specifies how well the transistor may be engineered into an integrated circuit. The mechanism of current saturation for the ballistic transistor is different than a traditional MOSFET device. It is clear from Eq. (2.48) that the net device current can be viewed as a balance between left and right flowing current states. For small applied drain voltage $eV_d < T$, the difference between the Fermi-Dirac integrals is approximately linear and the device expresses a clear linear transport region. When $eV_d \gg T$, the second term in square brackets in Eq. (2.48) becomes negligible and current saturation is the result of exhaustion of available electrons between the potential plateau and the source Fermi level.

While the above results indicate that the ballistic transistor may be a viable option for ultimately scaled silicon devices, they cannot describe the 2-D electrostatic effects important for short channel devices or direct source-to-drain tunneling of electrons through the potential barrier. Additionally, even in the limit $L_c \to \infty$, the device current may not be regulated by the long plateau region, but rather accumulated electrons near the source electrode (see section 4.2.1). For an accurate account of these effects and hence proper description of scaling dynamics we implement a numeric solution of the two dimensional Poisson equation and Schrödinger equation for a description of channel electrons.

In contrast with long devices, the channel potential for short channel devices in the off-state is roughly quadratic (*e. g.* see section 4.2.4). The transmission probability through a quadratic barrier, which determines the transistor current (see below) is exactly solvable in the WKB approximation by the well-known Kemble formula (Ch. 50 of Ref. [73]). The transmission probability is given by

$$\mathcal{D}(E_x) = \left(1 + \exp\left[2\pi\left(\Phi_0 - E_x\right)/\hbar\omega\right]\right)^{-1}, \tag{2.52}$$

with inverted oscillator frequency

$$\omega^2 = \frac{8\Phi_0}{m_x L_c^2}, \tag{2.53}$$

and potential maximum

$$\Phi_0 = \Phi_{\max}(x) + E_{z,1}. \tag{2.54}$$

By coincidence, Eq. (2.52) takes the exact same form as the Fermi distribution and the tunneling current may be evaluated in terms of an "inversion

temperature" [65]

$$T_{\text{inv}} \equiv \hbar\omega/2\pi. \tag{2.55}$$

Direct comparison the T_{inv} with the lattice temperature implies that tunneling will begin to dominate when $L_c < L_Q$, where

$$L_Q = \frac{2\hbar}{\pi T}\left(\frac{\Phi_0}{m_x}\right)^{1/2}. \tag{2.56}$$

For barrier height $\Phi_0 = 0.2$ eV and $T = 26$ meV, the tunnel length $L_Q \approx 7$ nm, in good general agreement with the results of chapter 4.

2.2.9 Voltage Gain

For an ultimate analysis of device performance we are looking to find acceptable trades-offs between the device size and the quality of control over the current including saturation properties and sub-threshold characteristics. All of these properties may be encapsulated by a single figure-of-merit at fixed current density (per unit device width), the voltage gain. Defined as

$$G_v \equiv \left.\frac{\partial V_d}{\partial V_g}\right|_{J=\text{const}}, \tag{2.57}$$

the voltage gain may also be seen as the ratio of the device transconductance ($\partial J/\partial V_g$) to the differential resistance ($\partial J/\partial V_d$). In the subthreshold domain it is a measure of inverse DIBL effects and in the on-state, a measure of the quality of the saturation. The voltage gain is not necessarily a popular engineering figure since for an ideal device in saturation $G_v \to \infty$. However, it is a powerful way to visualize the response of the effective channel potential to applied gate bias and is useful for placing a lower bound on device performance requirements. Fundamentally, integrated circuits require $G_v > 1$, but a more practical measure may be $G_v > 2$ to consider the device a useful candidate.

We may find an analytic result for the voltage gain in the subthreshold regime where all electrons have been effectively expelled from the channel. We assume a simple three point finite-difference representation for the channel potential, which essentially represents a quadratic solution between the node points. The node points considered are shown in Fig. 2.8, and the electron density taken to be strictly zero everywhere. We consider the finite difference solution for points ϕ_c, ϕ_{ox} in the middle of the channel and oxide and connect

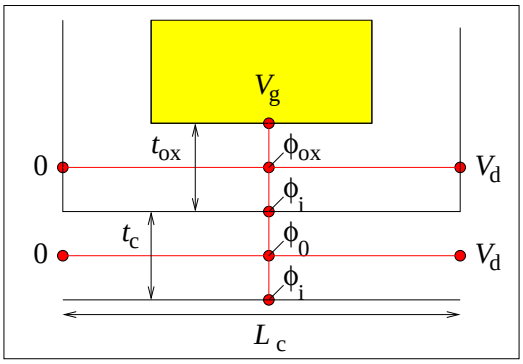

Figure 2.8: Quadratic model for the channel potential.

the solutions at the channel-oxide interface point ϕ_i. This yields relations

$$\begin{aligned} \frac{V_d - 2\phi_{ox}}{L_c^2} + \frac{V_g + \phi_i - 2\phi_{ox}}{t_{ox}^2} &= 0, \\ \frac{V_d - 2\phi_c}{L_c^2} + \frac{2\phi_i - 2\phi_c}{t_c^2} &= 0, \\ \frac{\phi_i - \phi_{ox}}{t_{ox}} = \alpha \frac{\phi_c - \phi_i}{t_c}, & \end{aligned} \tag{2.58}$$

where $\alpha \equiv \epsilon_{si}/\epsilon_{ox}$ is the ratio of the dielectric constants in the channel and oxide. Solution of these equations yields the potential in the center of the channel

$$\phi_c = \frac{2L_c^4 V_g + \left(2\alpha t_{ox}^3 t_c + 2t_{ox}^2 t_c^2 + 2\alpha L_c^2 t_{ox} t_c + L_c^2 t_c^2 + 2L_c^2 t_{ox}^2\right) V_d}{2\left(2t_{ox}^2 L_c^2 + L_c^4 + 2\alpha t_{ox}^3 t_c + 2t_{ox}^2 t_c^2 + 2\alpha L_c^2 t_{ox} t_c + L_c^2 t_c^2\right)}. \tag{2.59}$$

The result is not terribly intuitive in this form, but for subthreshold current

$$J_T \propto \exp\left(-e\phi_c/T\right), \tag{2.60}$$

the voltage gain may be calculated

$$G_v = \frac{2r_c^2}{1 + 2\alpha\beta_{ox}\left(1 + r_{ox}^{-2}\right) + 2r_{ox}^{-2} + 2\beta_{ox}^2}, \tag{2.61}$$

in term of length ratios

$$\begin{aligned} r_c &\equiv L_c/t_c, \\ r_{ox} &\equiv L_c/t_{ox}, \\ \beta_{ox} &\equiv t_{ox}/t_c. \end{aligned} \tag{2.62}$$

A comparison of this model to the full numeric simulations are shown in section

4.2.5.

2.2.10 Power

A major problem encountered by circuit designers is chip power consumption. Circuit power consumption is already taxing today's VLSI designs, a problem which may only become worse as components begin to operate in the ballistic regime [74]. The total IC power requirements depend a lot on circuit architecture and design. However, we may gain insight into the issue with a crude model which captures the two most important aspects of circuit power, dynamic power used during the switching transients and static power leaked in the off-state. As a first approximation, the total circuit power may be represented as a sum [75, 76]

$$P = P_{\text{stat}} + P_{\text{dyn}} \tag{2.63}$$

of static,

$$P_{\text{stat}} = I_{\text{OFF}} V_{DD}, \tag{2.64}$$

and dynamic,

$$P_{\text{dyn}} = \sum_i \alpha_i C_i V_{DD}^2 f, \tag{2.65}$$

contributions with drive voltage V_{DD}. Here C_i is the total capacitance of circuit block i, f the clocking frequency and α_i the circuit "activity factor" of the i_{th} circuit block. For a given circuit block with capacitive recharging time

$$\tau_i = \frac{C_i V_{DD}}{I_{\text{ON},i}}, \tag{2.66}$$

no more than fraction

$$p = f\tau, \tag{2.67}$$

of the clock frequency should be taken for capacitive recharging. In general, $p \ll 1$ and for a constant circuit speed requirement, assuming constant on-state current density in each circuit block, equation (2.63) may be conveniently rewritten in terms of the power density

$$P/W = (\lambda J_{\text{ON}} + J_{\text{OFF}})\, V_{DD}, \tag{2.68}$$

in terms of on and off state current densities

$$\begin{aligned} J_{OFF} &\equiv I_{OFF}/W, \\ J_{ON} &\equiv I_{ON}/W. \end{aligned} \tag{2.69}$$

The "effective activity factor" is given by

$$\lambda \equiv p \sum_i \alpha_i W_i / W, \tag{2.70}$$

where W_i is the width of the i_{th} circuit block, and W the total circuit width.

The strength of Eq. (2.68) is its simplicity. It clearly shows the trade-off between static and dynamic power requirements and a minimum power operating point may be found in terms of just two parameters λ, J_{ON}. The static and dynamic components may be calculated readily from the transistor characteristics $J(V_d, V_g)$. For a fixed V_{DD} and J_{ON}, the gate voltage is found such that

$$J_{\text{ON}} - J(V_{DD}, V_{ON}) = 0. \tag{2.71}$$

The off-state current is then calculated

$$J_{\text{OFF}} = J(V_{DD}, V_{ON} - V_{DD}). \tag{2.72}$$

The first step may be interpreted as optimization of the gate workfunction for the transistor to provide some specified operating current for the circuit J_{ON}. The second step is then a measure of how well the circuit drive voltage V_{DD} shuts the current in the transistor off-state.

A typical plot of total power density P/W versus drive voltage V_{DD} is shown in Fig. 2.9 for decreasing gate lengths. The dashed line shows the linearly increasing contribution of dynamic consumption to the total power. The colored dotted lines show the fall-off of static leakage for a given gate length. The trade-off between static and dynamic components is clear; for small V_{DD}, the transistor is not well shut, while the total power at high V_{DD} is dominated by large switching transient currents. For each device, a clear minimum total power develops, which may be accurately found numerically with Brent's method [77].

2.3 Numeric Methods

The numerical Poisson equation solver is designed specifically for the problem of different semiconducting materials with abrupt interfaces. It is intended to be as general as possible and the properties of each material may be tuned independently for a general calculation scheme. It accounts for the field discontinuities at the material interfaces and fully describes the relevant two dimensional electrostatic effects. It also automatically accounts for band-bending near the electrode interfaces and field penetration in the electrodes.

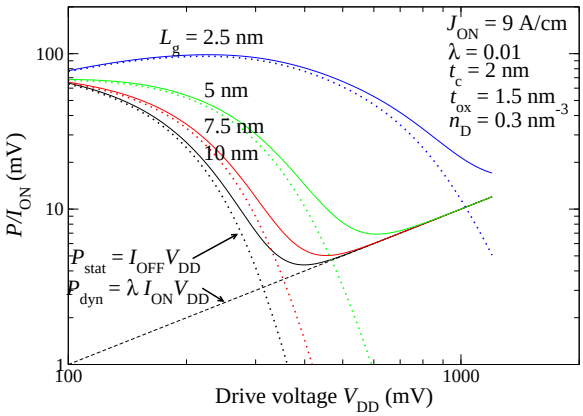

Figure 2.9: Total device power versus supply voltage V_{DD} for decreasing gate lengths. Dashed and dotted lines represent that dynamic and static components respectively.

2.3.1 Poisson Solver

We begin with Maxwell's equations for the electric field in a medium [11],

$$\boldsymbol{\nabla} \times \boldsymbol{E} = 0, \tag{2.73}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{D} = 4\pi\rho(\boldsymbol{x}), \tag{2.74}$$

where $\rho(\boldsymbol{x})$ is the electron charge and the displacement vector $\boldsymbol{D}$ is related to the electric field $\boldsymbol{E}$ through the permittivity tensor $\bar{\epsilon}$

$$\boldsymbol{D} = \bar{\epsilon} \cdot \boldsymbol{E}. \tag{2.75}$$

In the usual way, Eq. (2.73) is automatically satisfied by the introduction of the scalar potential

$$-\boldsymbol{\nabla}\Phi(\boldsymbol{x}) = \boldsymbol{E}. \tag{2.76}$$

Plugging the relation (2.75) into Eq. (2.74) we find the standard result

$$\boldsymbol{\nabla} \cdot [\bar{\epsilon}(\boldsymbol{x}) \cdot \boldsymbol{\nabla}\Phi(\boldsymbol{x})] = -4\pi\rho(\boldsymbol{x}). \tag{2.77}$$

For all devices considered, we will assume an isotropic medium, so equation (2.77) reduces to Poisson's equation

$$\nabla^2\Phi(\boldsymbol{x}) = -\frac{4\pi}{\epsilon}\rho(\boldsymbol{x}). \tag{2.78}$$

There are a variety of methods available to numerically approximate a

solution to Eqs. (2.77), (2.78). Popular methods include finite-elements [78] and the so-called "method of lines" [79, 80]. Since the potential does not vary wildly in our domain of interest, we choose to solve equation (2.78) using a finite difference equation (FDE) scheme for its computational simplicity and speed. We map the potential on a rectilinear grid with node spacings Δ_x and Δ_z respectively. While the basic expressions for the five-point finite difference scheme are widely available [81, 82], it is nevertheless instructive to look at their derivation as special care should be taken at the interface between different materials. The derivation of the basic FDE equations will also illuminate higher order nine-point approximations.

2.3.1.1 1-D Finite Differences

We begin by deriving the fundamental finite difference relations. The goal of a general finite difference relation is to approximate some continuous derivative of a function $\Phi(x)$ at point x_* using discrete points of known values

$$\left.\frac{\partial^n \Phi(x)}{\partial x^n}\right|_{x_*} \approx \sum_j \gamma_j \Phi_j, \tag{2.79}$$

where $\Phi_j \equiv \Phi(x_j)$ is the function value at fixed (arbitrarily spaced) points x_j and γ_j are undetermined weighting values. Evaluation point x_* does not necessarily coincide with a grid point x_j. To calculate the weighting factors γ_j we write the function at each x_j as a Taylor expansion around x_*,

$$\Phi_j \approx \Phi(x_*) + \Delta_j \partial_x \Phi(x_*) + \frac{\Delta_j^2}{2}\partial_x^2 \Phi(x_*) + \ldots + \frac{\Delta_j^k}{k!}\partial_x^k \Phi(x_*), \tag{2.80}$$

where $\Delta_j \equiv (x_j - x_*)$. Thus if we have q node points

$$\sum_{j=1}^{q} \gamma_j \Phi_j = \sum_{k=0} B_k (\gamma_1 + \cdots + \gamma_q) \partial_x^k \Phi(x_*), \tag{2.81}$$

where

$$\begin{aligned} B_0 &= \gamma_1 + \cdots + \gamma_q, \\ B_1 &= \Delta_1\gamma_1 + \cdots + \Delta_q\gamma_q, \\ &\vdots \\ B_n &= \frac{\Delta_1^n}{n!}\gamma_1 + \cdots + \frac{\Delta_q^n}{n!}\gamma_q. \end{aligned}$$

In order for Eq. (2.79) to remain a consistent approximation of the true derivative we must require [81]

$$\begin{aligned} B_k &= 0, \quad k = 0, \ldots, \neq n, \\ B_n &= 1. \end{aligned} \tag{2.82}$$

The general solution for the weighting factors γ_i is then given by system of equations

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \Delta_1 & \Delta_2 & \cdots & \Delta_q \\ \vdots & & & \\ \Delta_1^n & \Delta_2^n & \cdots & \Delta_q^n \\ \Delta_1^{n+1} & \Delta_2^{n+1} & \cdots & \Delta_q^{n+1} \\ \vdots & & & \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_q \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ n! \\ 0 \\ \vdots \end{bmatrix}. \tag{2.83}$$

This implies that in general, we need at least $q = n + 1$ node points to approximate a derivative of order n. If $q > n + 1$, there will be $q - n - 1$ free parameters which in general should be solved with constraints (2.82) [81].

For the special case of $\partial_x^2 \Phi(x_j)$, using node points γ_{j-1},γ_j,γ_{j+1} with fixed node spacing Δ_x, this yields system of equations

$$\begin{bmatrix} 1 & 1 & 1 \\ -\Delta_x & 0 & \Delta_x \\ \Delta_x^2 & 0 & \Delta_x^2 \end{bmatrix} \begin{bmatrix} \gamma_{j-1} \\ \gamma_j \\ \gamma_{j+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}, \tag{2.84}$$

and solving for each γ_j,

$$\partial_x^2 \Phi(x_j) \approx \left(\Phi_{i+j} + \Phi_{i-j} - 2\Phi_j\right) \Delta_x^{-2} + \mathcal{O}(\Delta_x^2). \tag{2.85}$$

It should be noted that the approximation to the derivative is $\mathcal{O}(\Delta_x^2)$ only when $x_* = x_j$. If $x_* \neq x_j$, this approximation is $\mathcal{O}(\Delta_x)$. In general, the error associated with a one dimensional finite difference approximation is $\mathcal{O}(\Delta_x^{q-n-1})$ [81]. Plugging into the one dimensional form of the Poisson equation (2.78) we find the standard finite difference relation

$$\Phi_{j+1} + \Phi_{j-1} - 2\Phi_j = -4\pi \Delta_x^2 \rho_j / \epsilon_j. \tag{2.86}$$

This expression may also be derived by defining the centered difference operator

$$\begin{aligned} \bar{\delta}_c \Phi_j &\equiv \Phi_{j+1/2} - \Phi_{j-1/2}, \\ \bar{\delta}_c x_j &\equiv x_{j+1/2} - x_{j-1/2} = \Delta_x. \end{aligned} \tag{2.87}$$

The continuous derivative can them be approximated $\partial_x^2 \Phi(x_j) \approx \bar{\delta}_c^2 \Phi(x_j) / \bar{\delta}_c^2 x_j$.

At the interface between different device regions, the finite difference relations require a slightly special treatment. Because of the discontinuity of the dielectric constant, the Taylor expansion at points across the interface is not strictly valid. Furthermore, if the node point falls exactly on the interface, there is an ambiguity in the value of ϵ in Eq. (2.77). In our model, all interface fall in FDE node points. To derive the expression for node points that fall on the interface, we examine the Poisson equation at (non-node) points a small distance δ to the left and right of the interface (see Fig. 2.10), and writing the Taylor expansions as

$$\begin{aligned}
\Phi(x - \Delta_x - \delta) &\approx \Phi_*^- - \Delta_x \partial_x \Phi_*^- + \frac{\Delta_x^2}{2} \partial_x^2 \Phi_*^- + \cdots , \\
\Phi(x + \Delta_x + \delta) &\approx \Phi_*^+ + \Delta_x \partial_x \Phi_*^+ + \frac{\Delta_x^2}{2} \partial_x^2 \Phi_*^+ + \cdots ,
\end{aligned}$$

Across the boundary, the displacement vector $\boldsymbol{D}$ must obey (Ch. 4, Ref. [11]), $\boldsymbol{D}^- - \boldsymbol{D}^+ = \sigma$, where σ is the free charge on the surface, in our case $\sigma = 0$. Since the surface charge is zero, by Eq. (2.78) we have in the limit that $\delta \to 0$, $\epsilon_1 \partial_x^2 \Phi_*^- = 0 = \epsilon_2 \partial_x^2 \Phi_*^+$, or we can write

$$\lim_{\delta \to 0} \partial_x^2 \Phi_*^- = \lim_{\delta \to 0} \partial_x^2 \Phi_*^+ = 0. \tag{2.88}$$

The potential across the boundary is continuous so we also have condition

$$\lim_{\delta \to 0} \Phi_*^- = \lim_{\delta \to 0} \Phi_*^+ = \Phi_j. \tag{2.89}$$

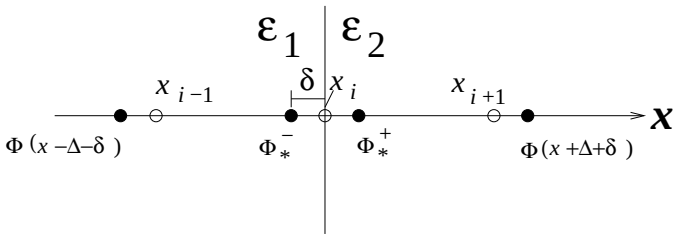

Figure 2.10: Interface between difference transistor materials.

Plugging these conditions into the Taylor expansions we see

$$\sum_j \gamma_j \Phi_j =$$

$$(\gamma_{j+1} + \gamma_j + \gamma_{j-1})\,\Phi_j + (-\frac{\epsilon_2}{\epsilon_1}\gamma_{j-1} + \gamma_{j+1})\Delta_x \partial_x \Phi_j + (\gamma_{j+1} + \gamma_{j-1})\frac{\Delta_x^2}{2}\partial_x^2 \Phi_j. \tag{2.90}$$

Consistency of the summation over node points as a valid approximation to the continuous derivative (2.79) requires the system of equations

$$\begin{bmatrix} 1 & 1 & 1 \\ -\frac{\epsilon_2}{\epsilon_1}\Delta_x & 0 & \Delta_x \\ \Delta_x^2 & 0 & \Delta_x^2 \end{bmatrix} \begin{bmatrix} \gamma_{j-1} \\ \gamma_j \\ \gamma_{j+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}. \tag{2.91}$$

Solution of this system yields approximation for the second derivative on the boundary

$$\partial_x^2 \Phi(x_j) \approx \left[\epsilon_2 \Phi_{j+1} + \epsilon_1 \Phi_{j-1} - (\epsilon_1 + \epsilon_2)\Phi_j\right] \Delta_x^{-2} = 0, \tag{2.92}$$

or

$$\epsilon_2 \Phi_{j+1} + \epsilon_1 \Phi_{j-1} - (\epsilon_1 + \epsilon_2)\Phi_j = 0. \tag{2.93}$$

Comparing Eqs. (2.86), (2.93), the term multiplying Φ_j can be seen as the average value of the dielectric between the two regions. Note that in the limit $\epsilon_1 \to \epsilon_2$, the two results are identical. Equation (2.93) is the same that one would get by simply writing Eq. (2.93) as the backward and forward differences at the boundary $\epsilon_1(\Phi_j - \Phi_{j-1})/\Delta_x = \epsilon_2(\Phi_{j+1} - \Phi_j)/\Delta_x$.

Once the electron density is calculated everywhere in the device (through Eqs. 2.31, and the channel density calculated below), the result is used as the right-hand part of the finite difference representation of the Poisson equation. The simulator is written to allow for use of either a standard five-point finite difference scheme or a slightly more accurate nine-point finite difference scheme.

Given N mesh points, system of equations (2.86) can be written as a $N \times N$

tri-diagonal matrix equation

$$\begin{bmatrix} b_1 & c_1 & 0 & \cdots & & & \\ a_2 & b_2 & c_2 & \cdots & & & \\ & & & \vdots & & & \\ & & & \cdots a_{N-1} & b_{N-1} & c_{N-1} & \\ & & & \cdots & 0 & a_N & b_N \end{bmatrix} \cdot \begin{bmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_{N-1} \\ \Phi_N \end{bmatrix} = \begin{bmatrix} \rho_1 - a_1 \\ \rho_2 \\ \vdots \\ \rho_{N-1} \\ \rho_N - c_N \end{bmatrix}, \tag{2.94}$$

where we have assumed Dirichlet (fixed) boundary conditions at mesh points a_1 and c_N. The tri-diagonal matrix equation can be solved in $O(N)$ operations using back-substitution [77]. The back-substitution method is actually more general and may be used to solve an arbitrary ordinary differential equation

$$\partial_x^2 q(x) + \partial_x p(x) + r(x) = 0, \tag{2.95}$$

given Dirichlet boundary conditions [83].

2.3.1.2 2-D Finite Differences

In the same manner as section 2.3.1.1, we want to approximate a general derivative of function $\Phi(z_*, x_*)$ using a rectilinear grid of known values $\Phi_{i,j} \equiv \Phi(z_i, x_j)$,

$$\left.\frac{\partial^n \Phi(z,x)}{\partial_x^m \partial_z^{n-m}}\right|_{z_*,x_*} \approx \sum_{i,j} \gamma_{i,j} \Phi_{i,j}. \tag{2.96}$$

This notation for i,j and $\Phi(z,x)$ is adopted because of the programming convention implemented. The $\hat{x}$ direction is defined parallel to the device channel and positive in the source-to-drain direction. $\hat{z}$ is defined as perpendicular to the channel and positive in the gate-to-channel direction.

2.3.1.3 5-point FDE

For node points which do not fall on the interface between device regions, we are looking for an approximation to the derivative

$$(\partial_x^2 + \partial_z^2)\Phi_{z_i,x_j} \approx \gamma_{i-1,j}\Phi_{i-1,j} + \gamma_{i+1,j}\Phi_{i+1,j} + \gamma_{i,j-1}\Phi_{i,j-1} + \gamma_{i,j+1}\Phi_{i,j+1} + \gamma_{i,j}\Phi_{i,j}. \tag{2.97}$$

To determine the weighting factors $\gamma_{i,j}$, we expand the potential at each $\Phi_{i,j}$ around the central point Φ_{i_0,j_0},

$$\begin{aligned} \Phi_{i,j} \approx\ & \Phi_{i_0,j_0} + \Delta_x \partial_x \Phi_{i_0,j_0} + \Delta_z \partial_z \Phi_{i_0,j_0} + \\ & \tfrac{1}{2!}\left[\Delta_x^2 \partial_x^2 \Phi_{i_0,j_0} + 2\Delta_x \Delta_z \partial_x \partial_z \Phi_{i_0,j_0} + \Delta_z^2 \partial_z^2 \Phi_{i_0,j_0}\right] + \cdots . \end{aligned} \tag{2.98}$$

Collecting terms by derivative order, we find

$$\sum_{i,j} \gamma_{i,j}\Phi_{i,j} = B_{0,0}\Phi_{i_0,j_0} + B_{1,0}\partial_z\Phi_{i_0,j_0} + B_{0,1}\partial_x\Phi_{i_0,j_0} + B_{2,0}\partial_z^2\Phi_{i_0,j_0} + B_{0,2}\partial_x^2\Phi_{i_0,j_0},$$

where each $B_{m,n}(\gamma_{i,j})$ is a function of the weighting factors of the derivative of order $\partial_z^i\partial_x^j$ in the expansion. Again, consistency arguments (see section 2.3.1.1) require $B_{2,0} = B_{0,2} = 1$, else $B_{i,j} = 0$. In the 5-point case, this gives system of equations

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ \Delta_z & -\Delta_z & 0 & 0 & 0 \\ 0 & 0 & \Delta_x & -\Delta_x & 0 \\ \frac{\Delta_z^2}{2} & \frac{\Delta_z^2}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{\Delta_x^2}{2} & \frac{\Delta_x^2}{2} & 0 \end{bmatrix} \begin{bmatrix} \gamma_{i+1,j} \\ \gamma_{i-1,j} \\ \gamma_{i,j+1} \\ \gamma_{i,j-1} \\ \gamma_{i,j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}. \tag{2.99}$$

The solution of this system is unique and gives $\gamma_{i+1,j} = \gamma_{i-1,j} = \Delta_z^{-2}$, $\gamma_{i,j+1} = \gamma_{i,j-1} = \Delta_x^{-2}$, $\gamma_{i,j} - 2\left(\Delta_z^{-2} + \Delta_x^{-2}\right)$, which gives approximation for the operator

$$\begin{aligned} \left(\partial_z^2 + \partial_x^2\right)\Phi_{i,j} \quad \approx \quad & \Delta_z^{-2}\left(\Phi_{i+1,j} + \Phi_{i-1,j}\right) + \Delta_x^{-2}\left(\Phi_{i,j+1} + \Phi_{i,j-1}\right) \\ & -2\left(\Delta_x^{-2} + \Delta_z^{-2}\right)\Phi_{i,j} + \mathcal{O}(\Delta_z^2, \Delta_x^2). \end{aligned} \tag{2.100}$$

Plugging into the Poisson equation (2.78) yields the standard 5-point finite difference equation

$$\begin{aligned} \Delta_z^{-2}\left(\Phi_{i+1,j} + \Phi_{i-1,j}\right) + \Delta_x^{-2}\left(\Phi_{i,j+1} + \Phi_{i,j-1}\right) - 2\left(\Delta_x^{-2} + \Delta_z^{-2}\right)\Phi_{i,j} = \\ -4\pi\rho_{i,j}/\epsilon_{i,j}, \end{aligned} \tag{2.101}$$

which has an associated error term $\mathcal{O}(\Delta_z^2, \Delta_x^2)$. It is also easily derived by direct application of the centered difference operator (2.87) $\left(\partial_x^2 + \partial_z^2\right) \approx \left(\bar{\delta}_{c,x}^2/\Delta_x^2 + \bar{\delta}_{c,z}^2/\Delta_z^2\right)$.

Node points which lay on the interface between device regions again require careful attention. Figure 2.11 shows a surface interface between 2 transistor materials with different dielectric constants ϵ_1, ϵ_2 respectively. Again the dielectric is assumed to jump instantaneously at the interface. At the boundary, the solution must obey boundary conditions (Ch. 4, Ref. [11])

$$\left(\boldsymbol{D}^- - \boldsymbol{D}^+\right)\cdot\hat{x} = \sigma = 0, \tag{2.102}$$

$$\left(\boldsymbol{E}^- - \boldsymbol{E}^+\right)\cdot\hat{z} = 0. \tag{2.103}$$

As in section 2.3.1.1 we again only write the Taylor series for points in the

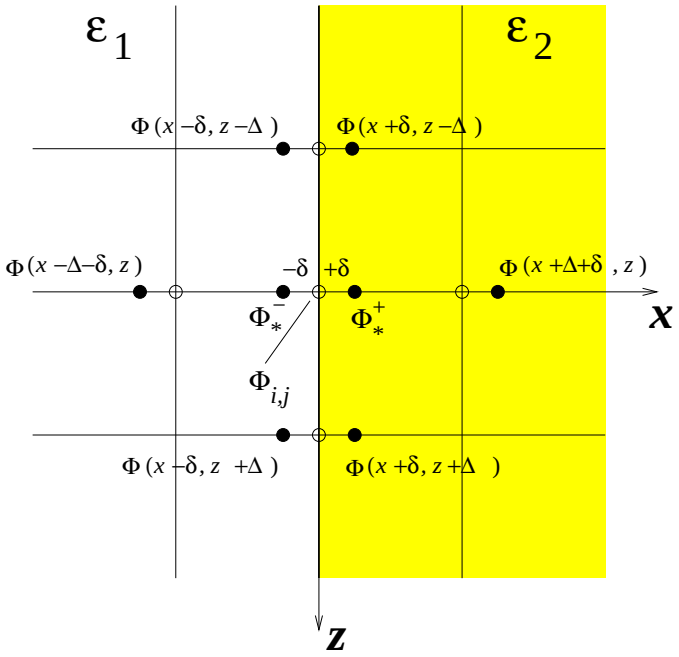

Figure 2.11: Plane interface between difference transistor materials. Open circles represent FDE points and closed circles effective evaluation points.

same device region.

$$\begin{aligned}
\Phi(z, x+\Delta_x+\delta) &\approx \Phi^+_* + \Delta_x\partial_x\Phi^+_* + \tfrac{\Delta_x^2}{2}\partial_x^2\Phi^+_* + \cdots, \\
\Phi(z, x-\Delta_x-\delta) &\approx \Phi^-_* - \Delta_x\partial_x\Phi^-_* + \tfrac{\Delta_x^2}{2}\partial_x^2\Phi^-_* + \cdots, \\
\Phi(z+\Delta_z, x+\delta) &\approx \Phi^+_* + \Delta_z\partial_z\Phi^+_* + \tfrac{\Delta_z^2}{2}\partial_z^2\Phi^+_* + \cdots, \\
\Phi(z-\Delta_z, x+\delta) &\approx \Phi^+_* - \Delta_z\partial_z\Phi^+_* + \tfrac{\Delta_z^2}{2}\partial_z^2\Phi^+_* + \cdots, \\
\Phi(z+\Delta_z, x-\delta) &\approx \Phi^-_* + \Delta_z\partial_z\Phi^-_* + \tfrac{\Delta_z^2}{2}\partial_z^2\Phi^-_* + \cdots, \\
\Phi(z-\Delta_z, x-\delta) &\approx \Phi^-_* - \Delta_z\partial_z\Phi^-_* + \tfrac{\Delta_z^2}{2}\partial_z^2\Phi^-_* + \cdots.
\end{aligned} \tag{2.104}$$

The continuity of $\Phi(x,z)$ requires $\lim\delta \to 0 : \Phi^-_* = \Phi^+_*$. Using boundary conditions (2.102), (2.103), and zero surface charge we have relations

$$\lim_{\delta\to 0}\begin{cases}
\epsilon_1\partial_x\Phi^-_* &= \epsilon_2\partial_x\Phi^+_*, \\
\partial_z\Phi^-_* &= \partial_z\Phi^+_*, \\
\Phi^-_* &= \Phi^+_* = \Phi_{i,j}, \\
\partial_x^2\Phi^-_* &= \partial_x^2\Phi^+_* (=0), \\
\partial_z^2\Phi^-_* &= \partial_z^2\Phi^+_* (=0),
\end{cases}$$

where we have multiplied ∂_x terms by their respective dielectric constants.

These boundary conditions exactly cancel all the first order terms in the sum over (2.104) and we find

$$\begin{aligned}\partial_z^2\Phi_{i,j} &= \Delta_z^{-2}\left[\Phi_{i+1,j}+\Phi_{i-1,j}-2\Phi_{i,j}\right], \\ \partial_x^2\Phi_{i,j} &= 2\Delta_x^{-2}\left[\frac{\epsilon_1}{\epsilon_1+\epsilon_2}\Phi_{i,j-1}+\frac{\epsilon_2}{\epsilon_1+\epsilon_2}\Phi_{i,j+1}-\Phi_{i,j}\right].\end{aligned} \tag{2.105}$$

Thus we have approximation for the differential operator

$$\begin{aligned}(\partial_z^2+\partial_x^2)\Phi_{z_i,x_j} \approx\ & \frac{2}{\Delta_x^2(\epsilon_1+\epsilon_2)}(\epsilon_1\Phi_{i,j-1}+\epsilon_2\Phi_{i,j+1})+\Delta_z^{-2}(\Phi_{i+1,j}+\Phi_{i-1,j}) \\ & -2(\Delta_x^{-2}+\Delta_z^{-2})\Phi_{i,j}.\end{aligned}$$

The finite difference expression for points on a surface interface is then $\nabla^2\Phi(x,z) = (\partial_x^2+\partial_z^2)\Phi(x,z)=0$ written as

$$\begin{aligned}2\Delta_x^{-2}\left(\epsilon_2\Phi_{i,j+1}+\epsilon_1\Phi_{i,j-1}\right)+(\epsilon_1+\epsilon_2)\Delta_z^{-2}\left(\Phi_{i+1,j}+\Phi_{i-1,j}\right) \\ -2\left(\epsilon_1+\epsilon_2\right)\left(\Delta_x^{-2}+\Delta_z^{-2}\right)\Phi_{i,j}=0.\end{aligned} \tag{2.106}$$

Equation (2.106) reduces to the Laplace form of Eq. (2.101) in the limit that $\epsilon_1 \to \epsilon_2$.

At the device corners, the finite difference equations are derived the same way as above (Fig. 2.12). Writing the Taylor expansion for points only in the same quadrant, we find

$$\begin{aligned}
\Phi(z+\Delta_z+\delta,x+\delta) &\approx \Phi_*^{++}+\Delta_z\partial_z\Phi_*^{++}+\tfrac{\Delta_z^2}{2}\partial_z^2\Phi_*^{++}+\mathcal{O}(\Delta_z^2), \\
\Phi(z+\Delta_z+\delta,x-\delta) &\approx \Phi_*^{+-}+\Delta_z\partial_z\Phi_*^{+-}+\tfrac{\Delta_z^2}{2}\partial_z^2\Phi_*^{+-}+\mathcal{O}(\Delta_z^2), \\
\Phi(z-\Delta_z+\delta,x+\delta) &\approx \Phi_*^{-+}-\Delta_z\partial_z\Phi_*^{-+}+\tfrac{\Delta_z^2}{2}\partial_z^2\Phi_*^{-+}+\mathcal{O}(\Delta_z^2), \\
\Phi(z-\Delta_z+\delta,x-\delta) &\approx \Phi_*^{--}-\Delta_z\partial_z\Phi_*^{--}+\tfrac{\Delta_z^2}{2}\partial_z^2\Phi_*^{--}+\mathcal{O}(\Delta_z^2), \\
\Phi(z+\delta,x+\Delta_x+\delta) &\approx \Phi_*^{++}+\Delta_x\partial_x\Phi_*^{++}+\tfrac{\Delta_x^2}{2}\partial_x^2\Phi_*^{++}+\mathcal{O}(\Delta_x^2), \\
\Phi(z-\delta,x+\Delta_x+\delta) &\approx \Phi_*^{-+}+\Delta_x\partial_x\Phi_*^{-+}+\tfrac{\Delta_x^2}{2}\partial_x^2\Phi_*^{-+}+\mathcal{O}(\Delta_x^2), \\
\Phi(z+\delta,x-\Delta_x-\delta) &\approx \Phi_*^{+-}-\Delta_x\partial_x\Phi_*^{+-}+\tfrac{\Delta_x^2}{2}\partial_x^2\Phi_*^{+-}+\mathcal{O}(\Delta_x^2), \\
\Phi(z-\delta,x-\Delta_x-\delta) &\approx \Phi_*^{--}-\Delta_x\partial_x\Phi_*^{--}+\tfrac{\Delta_x^2}{2}\partial_x^2\Phi_*^{--}+\mathcal{O}(\Delta_x^2).
\end{aligned} \tag{2.107}$$

Again we multiply each expansion by its respective ϵ_i and using boundary conditions (2.102),(2.103) the first order terms cancel. Since $\Phi(z,x)$ is continuous

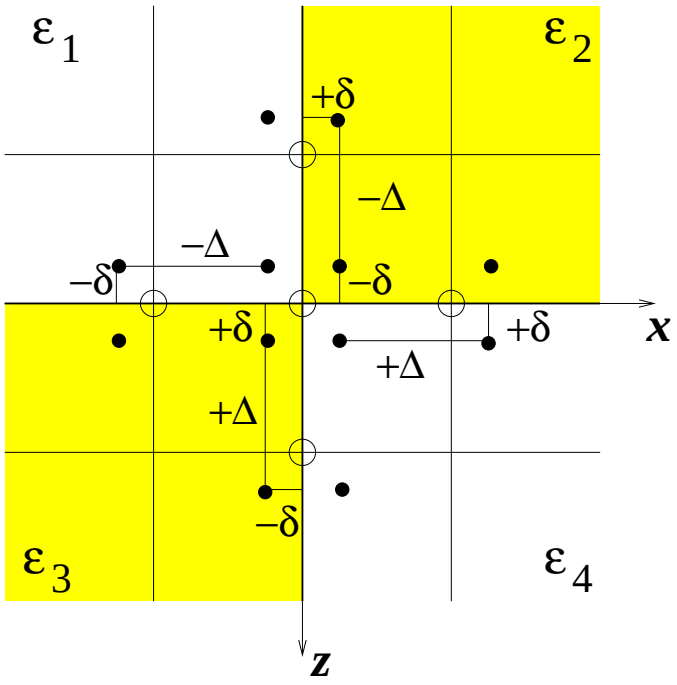

Figure 2.12: Corner interface between difference transistor materials. Open circles indicate FDE node points. Closed circles indicate effective evaluation points.

we find approximations for the operators

$$\begin{aligned} \partial_z^2\Phi_{i,j} &\approx \frac{2}{\Delta_z^2(\Sigma_i\epsilon_i)}\left[(\epsilon_3+\epsilon_4)\Phi_{i+1,j}+(\epsilon_1+\epsilon_2)\Phi_{i-1,j}\right]-2\Delta_z^{-2}\Phi_{i,j}, \\ \partial_x^2\Phi_{i,j} &\approx \frac{2}{\Delta_x^2(\Sigma_i\epsilon_i)}\left[(\epsilon_2+\epsilon_4)\Phi_{i,j+1}+(\epsilon_1+\epsilon_3)\Phi_{i,j-1}\right]-2\Delta_x^{-2}\Phi_{i,j}. \end{aligned} \tag{2.108}$$

Plugging into $\nabla^2\Phi(z,x)=0$ we find the finite difference expression at corner to be

$$\begin{aligned} \Delta_z^{-2}\left[(\epsilon_3+\epsilon_4)\Phi_{i+1,j}+(\epsilon_1+\epsilon_2)\Phi_{i-1,j}\right]+\Delta_x^{-2}\left[(\epsilon_2+\epsilon_4)\Phi_{i,j+1}+(\epsilon_1+\epsilon_3)\Phi_{i,j-1}\right] \\ -(\epsilon_1+\epsilon_2+\epsilon_3+\epsilon_4)\left(\Delta_x^{-2}+\Delta_z^{-2}\right)\Phi_{i,j}=0. \end{aligned} \tag{2.109}$$

Equation (2.109) reduces to the expression at a surface (2.106) in the appropriate limit and may be used to describe an arbitrary device corner. For example, at the bottom left corner of the gate $\epsilon_1=\epsilon_3=\epsilon_4=\epsilon_{ox}$, $\epsilon_2=\epsilon_{si}$. These results are also in agreement with early authors [84, 85].

We are now left with the boundary conditions at the device edges. The left and right domain boundaries have fixed values given by the specified electrode

potentials $V_s = 0, V_d$. For the top boundary, we solve the 1-D Poisson equation via the tri-diagonal system of equations 2.94 and fix the value of the potential along the top border.

For a symmetric dual gate device, we only need to solve the Poisson equation in the top half of the device as the potential will be symmetric in $\hat{z}$. Thus, the middle of the channel will be the bottom boundary of the domain and the boundary condition is a von Neumann type, where the normal derivative of the potential is zero $\partial\Phi(x,z)/\partial_z = 0$. Hence

$$\frac{\Phi_{M+1,j} - \Phi_{M-1,j}}{2\Delta_z} = 0,$$

or $\Phi_{M-1,j} = \Phi_{M+1,j}$, giving discrete finite difference equation

$$2\Delta_z^{-2}\Phi_{i-1,j} + \Delta_x^{-2}(\Phi_{i,j+1} + \Phi_{i,j-1}) - 2\left(\Delta_x^{-2} + \Delta_z^{-2}\right)\Phi_{i,j} = -4\pi\rho_{i,j}/\epsilon_{i,j}. \quad (2.110)$$

Equations (2.101), (2.106), (2.109), and (2.110) constitute the complete set of 5-point finite difference equations. Given N mesh points in $\hat{x}$ and M mesh points in $\hat{z}$, equation (2.101) can be written as an $MN \times MN$ block diagonal matrix equation of the form

$$\begin{bmatrix} \bar{T}_1 & \bar{D}_1^d & 0 & 0 & 0 \\ \bar{D}_2^u & \bar{T}_2 & \bar{D}_2^d & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \bar{D}_{M-1}^u & \bar{T}_{M-1} & \bar{D}_{M-1}^d \\ 0 & 0 & 0 & \bar{T}_M^u & \bar{D}_M \end{bmatrix} \cdot \begin{bmatrix} \bar{\Phi}_{i=1} \\ \bar{\Phi}_{i=2} \\ \cdots \\ \bar{\Phi}_{i=M-1} \\ \bar{\Phi}_{i=M} \end{bmatrix} = \begin{bmatrix} \bar{\rho}_{i=1} \\ \bar{\rho}_{i=2} \\ \cdots \\ \bar{\rho}_{i=M-1} \\ \bar{\rho}_{i=M} \end{bmatrix}, \quad (2.111)$$

where $\bar{T}_i$ is a $N \times N$ tri-diagonal matrix with non-zero elements corresponding to the weights of terms $\Phi_{i,j-1}$, $\Phi_{i,j}$, and $\Phi_{i,j+1}$. $\bar{D}_i^{u,d}$ is $N \times N$ diagonal matrix with diagonal elements corresponding to the weights of term $\Phi_{i-1,j}$ and $\Phi_{i+1,j}$. $\bar{\Phi}_i$ and $\bar{\rho}_i$ are N element arrays with values $\rho_{i,j}$ corresponding to the electron density at mesh point (i,j). It would be grossly inefficient to try to store the entire $MN \times MN$ array, so we use the sparse matrix storage format and solve the matrix equation using the Linear Bi-congruential Gradient Method algorithm [77].

2.3.1.4 9-point FDE

For the 9-point finite difference scheme, the derivation proceeds in the same way as section 2.3.1.3. We want to approximate the continuous operator $\nabla^2 = \partial_x^2 + \partial_z^2 \approx \sum_{i,j} \gamma_{i,j}\Phi_{i,j}$ as a sum over discrete known points $\Phi_{i,j}$ and $\gamma_{i,j}$ are yet unknown weighting coefficients. Now however, we add points $\Phi_{i+1,j+1}$,

$\Phi_{i+1,j-1}$, $\Phi_{i-1,j+1}$, $\Phi_{i-1,j-1}$. As before we write the Taylor expansions (to third order this time) for each point $\Phi_{i,j}$ around Φ_{i_0,j_0}, plug into $\sum_{i,j}\gamma_{i,j}\Phi_{i,j}$ and collect terms of similar derivative order to find

$$\sum_{i,j}\gamma_{i,j}\Phi_{i,j} = \sum_{m=0,n=0}^{3,3} B_{m,n}(\gamma_{i,j})\partial_z^m\partial_x^n\Phi_{i_0,j_0}.$$

Again, consistency arguments require (see section 2.3.1.1) $B_{2,0} = B_{0,2} = 1$, else $B_{m,n} = 0$ which gives system of equations

$$\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\Delta_z & -\Delta_z & 0 & 0 & \Delta_z & \Delta_z & -\Delta_z & -\Delta_z & 0 \\
0 & 0 & \Delta_x & -\Delta_x & \Delta_x & -\Delta_x & \Delta_x & -\Delta_x & 0 \\
0 & 0 & 0 & 0 & \Delta_z\Delta_x & -\Delta_z\Delta_x & -\Delta_z\Delta_x & \Delta_z\Delta_x & 0 \\
\frac{\Delta_z^2}{2} & \frac{\Delta_z^2}{2} & 0 & 0 & \frac{\Delta_z^2}{2} & \frac{\Delta_z^2}{2} & \frac{\Delta_z^2}{2} & \frac{\Delta_z^2}{2} & 0 \\
0 & 0 & \frac{\Delta_x^2}{2} & \frac{\Delta_x^2}{2} & \frac{\Delta_x^2}{2} & \frac{\Delta_x^2}{2} & \frac{\Delta_x^2}{2} & \frac{\Delta_x^2}{2} & 0 \\
0 & 0 & 0 & 0 & \frac{\Delta_z^2\Delta_x}{2} & -\frac{\Delta_z^2\Delta_x}{2} & \frac{\Delta_z^2\Delta_x}{2} & -\frac{\Delta_z^2\Delta_x}{2} & 0 \\
0 & 0 & 0 & 0 & \frac{\Delta_z\Delta_x^2}{2} & \frac{\Delta_z\Delta_x^2}{2} & -\frac{\Delta_z\Delta_x^2}{2} & -\frac{\Delta_z\Delta_x^2}{2} & 0 \\
\frac{\Delta_z^3}{6} & -\frac{\Delta_z^3}{6} & 0 & 0 & \frac{\Delta_z^3}{6} & \frac{\Delta_z^3}{6} & -\frac{\Delta_z^3}{6} & -\frac{\Delta_z^3}{6} & 0 \\
0 & 0 & \frac{\Delta_x^3}{6} & -\frac{\Delta_x^3}{6} & \frac{\Delta_x^3}{6} & -\frac{\Delta_x^3}{6} & \frac{\Delta_x^3}{6} & -\frac{\Delta_x^3}{6} & 0
\end{bmatrix} \times$$

$$\begin{bmatrix} \gamma_{i+1,j} \\ \gamma_{i-1,j} \\ \gamma_{i,j+1} \\ \gamma_{i,j-1} \\ \gamma_{i+1,j+1} \\ \gamma_{i+1,j-1} \\ \gamma_{i-1,j+1} \\ \gamma_{i-1,j-1} \\ \gamma_{i,j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{2.112}$$

Because not all of the equations are linearly independent, the solution has one free parameter which, without loss of generality, we choose to be $\gamma_{i+1,j+1}$. The

solution of equation (2.112) is then written

$$\begin{aligned}
\gamma_{i+1,j} = \gamma_{i-1,j} &= \frac{1-2\Delta_z^2\gamma_{i+1,j+1}}{\Delta_z^2}, \\
\gamma_{i,j+1} = \gamma_{i,j-1} &= \frac{1-2\Delta_x^2\gamma_{i+1,j+1}}{\Delta_x^2}, \\
\gamma_{i+1,j-1} = \gamma_{i-1,j+1} = \gamma_{i-1,j-1} &= \gamma_{i+1,j+1}, \\
\gamma_{i,j} &= -\frac{2\left(\Delta_z^2+\Delta_x^2-2\Delta_z^2\Delta_x^2\gamma_{i+1,j+1}\right)}{\Delta_z^2\Delta_x^2}.
\end{aligned} \tag{2.113}$$

The choice $\gamma_{i+1,j+1} = 0$ returns us the 5-point equation of section 2.3.1.3. The choice $\gamma_{i+1,j+1} = 1/6\Delta_z\Delta_x$ corresponds to the appropriate weight in the 4th order expansion of the difference operator (2.87)

$$\nabla^2 \approx \frac{\bar{\delta}_z^2}{\Delta_z^2} + \frac{\bar{\delta}_x^2}{\Delta_x^2} + \left(\frac{1}{6\Delta_z\Delta_x}\right)\bar{\delta}_z^2\bar{\delta}_x^2, \tag{2.114}$$

which has associated error term $\mathcal{O}(\Delta_z^4, \Delta_x^4)$. The operator ∇^2 may then be then expressed as

$$\begin{aligned}
(\partial_x^2+\partial_z^2)\Phi_{i,j} \approx & \Delta_z^{-2}\Delta_x^{-2}\left\{\left(\Delta_x^2-\frac{\Delta_x\Delta_z}{3}\right)(\Phi_{i+1,j}+\Phi_{i-1,j})\right. \\
& + \left(\Delta_z^2-\frac{\Delta_x\Delta_z}{3}\right)(\Phi_{i,j+1}+\Phi_{i,j-1}) \\
& + \frac{\Delta_z\Delta_x}{6}(\Phi_{i+1,j+1}+\Phi_{i+1,j-1}+\Phi_{i-1,j+1}+\Phi_{i-1,j-1}) \\
& \left. - 2\left(\Delta_x^2+\Delta_z^2-\frac{\Delta_x\Delta_z}{3}\right)\Phi_{i,j}\right\}.
\end{aligned} \tag{2.115}$$

Plugging into equation (2.78) we find the general 9-point finite difference expression

$$\begin{aligned}
(\partial_x^2+\partial_z^2)\Phi_{i,j} \approx & \left(\Delta_x^2-\frac{\Delta_x\Delta_z}{3}\right)(\Phi_{i+1,j}+\Phi_{i-1,j}) + \left(\Delta_z^2-\frac{\Delta_x\Delta_z}{3}\right)(\Phi_{i,j+1}+\Phi_{i,j-1}) \\
& + \frac{\Delta_z\Delta_x}{6}(\Phi_{i+1,j+1}+\Phi_{i+1,j-1}+\Phi_{i-1,j+1}+\Phi_{i-1,j-1}) \\
& - 2\left(\Delta_x^2+\Delta_z^2-\frac{\Delta_x\Delta_z}{3}\right)\Phi_{i,j} = -4\pi\Delta_z^2\Delta_x^2\rho_{i,j}/\epsilon_{i,j}.
\end{aligned} \tag{2.116}$$

An alternative (but equivalent) solution may be found by direct summation of the Taylor expansion terms for each node point

$$\begin{aligned}
\Phi(z \pm \Delta_z, x) \approx & \Phi_{i,j} \pm \Delta_z \partial_z \Phi_{i,j} + \frac{\Delta_z}{2} \partial_z^2 \Phi_{i,j} \pm \frac{\Delta_z^3}{6} \partial_z^3 \Phi_{i,j}, \\
\Phi(z, x \pm \Delta_x) \approx & \Phi_{i,j} \pm \Delta_x \partial_x \Phi_{i,j} + \frac{\Delta_x}{2} \partial_x^2 \Phi_{i,j} \pm \frac{\Delta_x^3}{6} \partial_x^3 \Phi_{i,j}, \\
\Phi(z \pm \Delta_z, x + \Delta_x) \approx & \Phi_{i,j} \pm \Delta_z \partial_z \Phi_{i,j} + \Delta_x \partial_x \Phi_{i,j} + \frac{\Delta_z^2}{2} \partial_z^2 \Phi_{i,j} + \frac{\Delta_x^2}{2} \partial_x^2 \Phi_{i,j} \\
& \pm \Delta_z \Delta_x \partial_z \partial_x \Phi_{i,j} \pm \frac{\Delta_z^3}{6} \partial_z^3 + \Delta_z^2 \Delta_x \partial_z^2 \partial_x \pm \Delta_z \Delta_x^2 \partial_z \partial_x^2 \\
& + \frac{\Delta_x^3}{6} \partial_x^3 \Phi_{i,j}, \\
\Phi(z \pm \Delta_z, x - \Delta_x) \approx & \Phi_{i,j} \pm \Delta_z \partial_z \Phi_{i,j} - \Delta_x \partial_x \Phi_{i,j} + \frac{\Delta_z^2}{2} \partial_z^2 \Phi_{i,j} + \frac{\Delta_x^2}{2} \partial_x^2 \Phi_{i,j} \\
& \mp \Delta_z \Delta_x \partial_z \partial_x \Phi_{i,j} \pm \frac{\Delta_z^3}{6} \partial_z^3 + \Delta_z^2 \Delta_x \partial_z^2 \partial_x \pm \Delta_z \Delta_x^2 \partial_z \partial_x^2 \\
& - \frac{\Delta_x^3}{6} \partial_x^3 \Phi_{i,j}.
\end{aligned} \tag{2.117}$$

The sum of the first two expansions (the node points in line with $\Phi_{i,j}$) gives the same result as section 2.3.1.3

$$(\partial_x^2 + \partial_z^2)\Phi_{i,j} \approx \Delta_z^{-2}(\Phi_{i+1}, \Phi_{i,j}) + \Delta_x^{-2}(\Phi_{i,j+1} + \Phi_{i,j-1}) - 2\left(\Delta_z^{-2} + \Delta_x^{-2}\right)\Phi_{i,j}.$$

Hence, the sum of the remaining points must contribute nothing on the right-hand side. It is easily seen that

$$\begin{aligned}
\Delta_z^{-2}\Delta_x^{-2}\beta \Big[& \Phi_{i+1,j+1} + \Phi_{i+1,j-1}\Phi_{i-1,j+1} + \Phi_{i-1,j-1} \\
& - 2(\Phi_{i+1,j} + \Phi_{i-1,j}\Phi_{i,j+1} + \Phi_{i,j-1}) + 4\Phi_{i,j} \Big] = 0,
\end{aligned} \tag{2.118}$$

where we have multiplied by free parameter β. The choice $\beta = \Delta_z\Delta_x/6$ corresponds the fourth order centered difference approximation $\nabla^2 \approx \bar{\delta}_z^2 + \bar{\delta}_x^2 + (1/6\Delta_z\Delta_x)\bar{\delta}_z^2\bar{\delta}_x^2$. When the off-line node points are added, we again arrive at Eq. (2.116).

At device interfaces, we write the expansions only for node points in the

same device region (Fig. 2.11)

$$\begin{aligned}
\Phi(z \pm \Delta_z, x \pm \delta) &\approx \Phi_*^{\pm} \pm \Delta_z \partial_z \Phi_*^{\pm} + \frac{\Delta_z^2}{2}\partial_z^2 \Phi_*^{\pm} \pm \frac{\Delta_z^3}{6}\partial_z^3 \Phi_*^{-}, \\
\Phi(z, x \pm \Delta_x \pm \delta) &\approx \Phi_*^{\pm} \pm \Delta_x \partial_x \Phi_*^{\pm} + \frac{\Delta_x^2}{2}\partial_x^2 \Phi_*^{\pm} \pm \frac{\Delta_x^3}{6}\partial_x^3 \Phi_*^{\pm}, \\
\Phi(z \pm \Delta_z, x + \Delta_x + \delta) &\approx \Phi_*^{+} \pm \Delta_z \partial_z \Phi_*^{+} + \Delta_x \partial_x \Phi_*^{+} \pm \Delta_z \Delta_x \partial_z \partial_x \Phi_*^{+} \\
&\quad + \frac{\Delta_z^2}{2}\partial_z^2 \Phi_*^{+} + \frac{\Delta_x^2}{2}\partial_x^2 \Phi_*^{+} + \frac{\Delta_z^2 \Delta_x}{3}\partial_z^2 \partial_x \Phi_*^{+} \\
&\quad \pm \frac{\Delta_z \Delta_x^2}{3}\partial_z \partial_x^2 \Phi_*^{+} \pm \frac{\Delta_z^3}{6}\partial_z^2 \Phi_*^{+} + \frac{\Delta_x^3}{6}\partial_x^3 \Phi_*^{+}, \\
\Phi(z \pm \Delta_z, x - \Delta_x - \delta) &\approx \Phi_*^{-} \pm \Delta_z \partial_z \Phi_*^{-} - \Delta_x \partial_x \Phi_*^{-} \mp \Delta_z \Delta_x \partial_z \partial_x \Phi_*^{-} \\
&\quad + \frac{\Delta_z^2}{2}\partial_z^2 \Phi_*^{-} + \frac{\Delta_x^2}{2}\partial_x^2 \Phi_*^{-} - \frac{\Delta_z^2 \Delta_x}{3}\partial_z^2 \partial_x \Phi_*^{-} \\
&\quad \pm \frac{\Delta_z \Delta_x^2}{3}\partial_z \partial_x^2 \Phi_*^{-} \pm \frac{\Delta_z^3}{6}\partial_z^2 \Phi_*^{-} - \frac{\Delta_x^3}{6}\partial_x^3 \Phi_*^{-}.
\end{aligned} \tag{2.119}$$

The cancellation of the first order derivatives is again done through boundary conditions (2.102),(2.103) and the continuity condition $\lim \delta \rightarrow : 0\Phi_*^{+} = \lim \delta \rightarrow 0\Phi_*^{-}$. As above, the differential operator may be expressed entirely in terms of the node points in-line with $\Phi_{i,j}$,

$$\begin{aligned}
(\partial_z^2 + \partial_x^2)\Phi_{i,j} &\approx \Delta_z^{-2}(\Phi_{i+1,j} + \Phi_{i-1,j}) + \Delta_x^{-2}\left(\frac{2\epsilon_2}{\epsilon_1 + \epsilon_2}\Phi_{i,j+1} + \frac{2\epsilon_1}{\epsilon_1 + \epsilon_2}\Phi_{i,j-1}\right) \\
&\quad -2(\Delta_z^{-2} + \Delta_x^{-2})\Phi_{i,j}.
\end{aligned} \tag{2.120}$$

So inclusion of any additional node points must sum to zero over their expansions. The sum of the diagonal node points is

$$\begin{aligned}
\Delta_x^{-2}\Delta_z^{-2}\left[\frac{\epsilon_2}{\epsilon_1 + \epsilon_2}(\Phi_{i+1,j+1} + \Phi_{i-1,j+1}) + \frac{\epsilon_1}{\epsilon_1 + \epsilon_2}(\Phi_{i+1,j-1} + \Phi_{i-1,j-1})\right] = \\
2\Delta_x^{-2}\Delta_z^{-2}\Phi_{i,j} + (\Delta_x^{-2}\partial_z^2 + \Delta_z^{-2}\partial_x^2)\Phi_{i,j},
\end{aligned} \tag{2.121}$$

or

$$2\beta\Delta_x^{-2}\Delta_z^{-2}\left[\frac{\epsilon_2}{\epsilon_1+\epsilon_2}(\Phi_{i+1,j+1}+\Phi_{i-1,j+1}+\frac{\epsilon_1}{\epsilon_1+\epsilon_2}(\Phi_{i+1,j-1}+\Phi_{i-1,j-1})\right.$$
$$\left.-\Phi_{i+1,j}-\Phi_{i-1,j}-\frac{2\epsilon_2}{\epsilon_1+\epsilon_2}\Phi_{i,j+1}-\frac{2\epsilon_1}{\epsilon_1+\epsilon_2}\Phi_{i,j-1}+2\Phi_{i,j}\right]=0, \tag{2.122}$$

where we have multiplied by free weighting term 2β. Adding these terms to the previous expression, we find the somewhat unwieldy

$$\begin{aligned}(\partial_z^2+\partial_x^2)\Phi_{i,j} \approx\ & \Delta_z^{-2}\Delta_x^{-2}\left\{(\Delta_x^2-2\beta)(\Phi_{i+1,j}+\Phi_{i-1,j})\right.\\ & +(\epsilon_1+\epsilon_2)^{-1}\left[2\epsilon_2(\Delta_z^2-2\beta)\Phi_{i,j+1}+2\epsilon_1(\Delta_z^2-2\beta)\Phi_{i,j-1}\right.\\ & \left.+2\beta\left(\epsilon_2(\Phi_{i+1,j+1}+\Phi_{i-1,j+1})+\epsilon_1(\Phi_{i+1,j-1}+\Phi_{i-1,j-1})\right)\right]\\ & \left.-2(\Delta_x^2+\Delta_z^2+2\beta)\Phi_{i,j}\right\},\end{aligned} \tag{2.123}$$

where $\beta=\Delta_z\Delta_x/6$ is again the appropriate weighting factor.

At a device corner (see figure 2.12), we proceed in exactly the same manner, with $\beta=\Delta_z\Delta_x/6$,

$$\begin{aligned}(\partial_z^2+\partial_x^2)\Phi_{i,j} \approx\ & \Delta_z^{-2}\Delta_z^{-2}\times\\ & \left\{(\sum_{i=1}^{4}\epsilon_i)^{-1}\left[(\Delta_x^2-4\beta)\left[(\epsilon_3+\epsilon_4)\Phi_{i+1,j}+(\epsilon_1+\epsilon_2)\Phi_{i-1,j}\right]\right.\right.\\ & +(\Delta_z^2-4\beta)\left[(\epsilon_2+\epsilon_4)\Phi_{i,j+1}+(\epsilon_1+\epsilon_3)\Phi_{i,j-1}\right]\\ & \left.+4\beta(\epsilon_4\Phi_{i+1,j+1}+\epsilon_3\Phi_{i+1,j-1}+\epsilon_2\Phi_{i-1,j+1}+\epsilon_1\Phi_{i-1,j-1})\right]\\ & \left.-(\Delta_z^2+\Delta_x^2-4\beta)\Phi_{i,j}\right\}.\end{aligned} \tag{2.124}$$

2.3.2 Mixing Methods

For a given electron density $p(x,z)$, the electrostatic potential $\Phi(x,z)$,is solved. This potential yields a new electron density and the process continues iteratively until a full self-consistent solution is found. The simplest approach of iteration, using the calculated electron density for a given potential directly as input to the next iteration step has notoriously poor convergence properties. In fact, even for relatively simple systems, this procedure can easily iterate forever or even diverge. More advanced iterative procedures are required to

make the solution of the self-consistent potential numerically feasible.

The problem can be seen conceptually as minimization over a large dimensional vector space. Hence we may use the progress made in multidimensional minimization algorithms to advance our solution. We have implemented three such procedures. One based on a modified secant approach which aims to minimize a general residual vector, and two based on Johnson's formulation [86, 87] of Broyden's method [77, 88], which is a quasi-Newton-Raphson approach. Intuitively one might expect Broyden's update, which incorporates knowledge of the function derivatives, should be vastly superior to the secant method. In these systems however, we do not have knowledge of the full Jacobian and an approximation of this matrix is required. This estimation reduces Broyden's method to be numerically identical to the secant approach [87].

All three implementations may be used in the simulator and are left as configuration options. While one group [51] noted modest improvement in convergence time using a further modified version of the Broyden's update, and even more advanced schemes have been proposed (see *e. g.* [89, 90]) all the results in this thesis have used the secant method (also known as Anderson's method). Anderson's method is described in the next section, and is preferred due to its conceptual simplicity, ease of parameter adjustment, and low storage requirements as compared to more complex methods [87].

2.3.2.1 Anderson's Method

A generalized secant method for multivariate minimization was first proposed by Wolfe [91] and later greatly improved by Anderson [87, 92]. We write the electron density at the node points defined by the Poisson finite difference mesh as a vector $|n\rangle$. Given input vector $|n_{in}\rangle$ we want to find the potential which tends difference vector

$$|F\rangle \equiv |n_{out}\rangle - |n_{in}\rangle, \tag{2.125}$$

to zero for resulting output density $|n_{out}\rangle$. To accelerate the convergence process, we may use the information obtained in the previous M iterations to minimize this difference. We define a general input vector as a linear combination of the subspace spanned by the previous M vectors at iteration step i as

$$|u^i\rangle = |n^i_{in}\rangle + \sum_{j=1}^{M} \Theta^i_j \left(|n^{i-j}_{in}\rangle - |n^i_{in}\rangle \right), \tag{2.126}$$

and general residual vector

$$|R^i\rangle = |F^i\rangle + \sum_{j=1}^{M} \Theta_j^i \left(|F^{i-j}\rangle - |F^i\rangle \right). \tag{2.127}$$

Coefficients Θ_j^i are yet undetermined and are selected to minimize the norm of the residual vector $\langle R^i|R^i\rangle$. This leads to a numeric solution of the system of equations

$$\sum_{j=1}^{M} \langle F^i - F^{i-m}|F^i - F^{i-j}\rangle \Theta_j^i = \langle F^i - F^{i-m}|F^i\rangle, \quad \forall_{k=1\ldots M}. \tag{2.128}$$

Once the optimal coefficients Θ_j^i are found, we define the input density for the next iteration step

$$|n_{in}^{i+1}\rangle = |u^i\rangle + \beta|R^i\rangle, \tag{2.129}$$

where parameter β represents the strength of which the optimal residual vector is mixed in to the new iteration. While increasing this parameter $\beta \to 1$ increases the rate of convergence, it also becomes unstable for some systems because linear dependencies develop in system of equations (2.128) near convergence [92]. Empirically we find that $\beta = 0.4$ produces a universally stable algorithm while maintaining an acceptable rate of convergence.

When $M \to 0$, we recover the simple mixing algorithm

$$|n_{in}^{i+1}\rangle = n_{in}^i\rangle + \beta|F^i\rangle. \tag{2.130}$$

As noted by the original author, the performance gains induced by increasing M larger than inclusion of more than a few previous iterations is reduced because we mix in information from the poor guesses of early iterations. For most simulation results, we have accepted a value $M = 5$. In contrast to results demonstrated previously [86], Anderson's method for density in ballistic MOSFETs is quite stable for the majority of systems and produces dramatically increased performance over the simple mixing algorithm.

2.3.2.2 Broyden's Update

Broyden's second method [88] may be derived by expanding difference vector (2.125) to first order

$$|F\rangle \approx |F^i\rangle + \mathcal{J}^i \left(|n_{in}\rangle - |n_{in}^i\rangle \right), \tag{2.131}$$

where matrix $\mathcal{J}^i$ is some approximation to the actual Jacobian

$$J^i_{lm} \equiv \frac{\partial |F\rangle_l}{\partial |n\rangle_m}. \tag{2.132}$$

The requirement that $|F\rangle$ vanish yields update formula for the input vector

$$|n^{i+1}_{in}\rangle = |n^i_{in}\rangle - \left[\mathcal{J}^i\right]^{-1} |F^i\rangle, \tag{2.133}$$

and minimization of

$$\left|\mathcal{J}^{i+1} - \mathcal{J}^i\right|^2, \tag{2.134}$$

subject to the constraint

$$|F^{i+1}\rangle - |F^i\rangle - \mathcal{J}^{i+1}\left(|n^{i+1}_{in}\rangle - |n^i_{in}\rangle\right) = 0, \tag{2.135}$$

yields update formula for the approximate Jacobian [93]

$$\mathcal{J}^{i+1} = \mathcal{J}^i + \frac{\left[|\Delta F^i\rangle - \mathcal{J}^i |\Delta n^i_{in}\rangle\right] \langle \Delta n^i_{in}|}{\langle \Delta n^i_{in} | \Delta n^i_{in}\rangle}. \tag{2.136}$$

Here the normalized difference vectors are given by

$$\begin{array}{rcl} |\Delta F^i\rangle & = & \left(|F^{i+1}\rangle - |F^i\rangle\right) / \left|\langle F^{i+1} - F^i | F^{i+1} - F^i\rangle\right|, \\ |\Delta n^i_{in}\rangle & = & \left(|n^{i+1}_{in}\rangle - |n^i_{in}\rangle\right) / \left|\langle F^{i+1} - F^i | F^{i+1} - F^i\rangle\right|. \end{array} \tag{2.137}$$

This scheme has been implemented directly for some physical systems [94], however it requires storage and inversion of the full $N \times N$ Jacobian matrix which is infeasible for even the smallest FET systems.

Srivastava derived a rank-1 update scheme [93] for the inverse Jacobian directly which requires storage of difference vectors of size N for each previous $i-1$ iterations. The initial approximation of the Jacobian is written in terms of the identity matrix $\hat{I}$

$$\left(\mathcal{J}^1\right)^{-1} = -\beta \hat{I}. \tag{2.138}$$

The update formulas may be written

$$|n^{i+1}_{in}\rangle = |n^i_{in}\rangle - \left(\mathcal{J}^i\right)^{-1} |F^i\rangle, \tag{2.139}$$

for in input density, and in terms of the difference vectors (2.137) as

$$\left(\mathcal{J}^{i+1}\right)^{-1} = \left(\mathcal{J}^i\right)^{-1} - \left[|\Delta n^i_{in}\rangle - \left(\mathcal{J}^i\right)^{-1} |\Delta F^i\rangle\right] \langle \Delta F^i|. \tag{2.140}$$

For the Jacobian which again minimizes $\left|\left(\mathcal{J}^{i+1}\right)^{-1} - \left(\mathcal{J}^{i}\right)^{-1}\right|$ subject to the constraint [86]

$$|\Delta n_{in}^{i}\rangle - \left(\mathcal{J}^{i}\right)^{-1}|\Delta F^{i}\rangle = 0. \tag{2.141}$$

This formulation of Broyden's second method eliminates the requirement of numerically inverting the Jacobian matrix and requires only storage i vectors. However, only information from the previous iteration is included in the update formula. Based on the work of Vanderbilt and Louie [95], Johnson derived an update formula using the information obtained in all previous $i-1$ iterations [86]. In this scheme, difference $\left|\left(\mathcal{J}^{i+1}\right)^{-1} - \left(\mathcal{J}^{1}\right)^{-1}\right|$ is minimized and the update formula may be written

$$|n_{in}^{i+1}\rangle = |n_{in}^{i}\rangle + \beta|F^{i}\rangle - \sum_{j=1}^{i-1}\omega_j\gamma_{ij}\left[\beta|\Delta F^{i}\rangle + |\Delta n_{in}^{i}\rangle\right], \tag{2.142}$$

where

$$\gamma_{ij} = \sum_{l=1}^{i-1}\omega_l\langle\Delta F^{l}|F^{i}\rangle\left[\hat{A}^{-1}\right]_{lj}. \tag{2.143}$$

Matrix $\hat{A}$ is given by

$$\hat{A}_{lj} = \omega_l\omega_j\langle\Delta F^{l}|\Delta F^{j}\rangle, \tag{2.144}$$

and following the recommendation by Johnson, the weighting coefficients are selected

$$\omega_i = \min\begin{cases} 1, \\ \langle F^{i}|F^{i}\rangle^{-1/2}. \end{cases} \tag{2.145}$$

A comparison of the methods is shown in Fig. 2.13 for a typical system $L_g = 10$ nm, $V_d = V_g = 0$ V and classical channel electrons (see appendix A.3). The black line shows Anderson's method, preferred in this work and the red line is Johnson's formulation of Broyden's update. In both cases, the mixing strength $\beta = 0.4$. The green line shows the results of simple mixing with mixing strength $\beta = 0.05$. While this comparison may seem a little unfair, raising the strength for simple mixing much higher than this value causes the scheme to diverge.

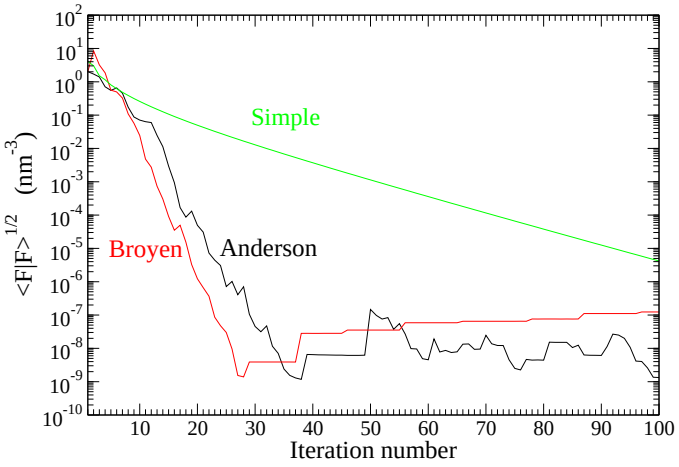

Figure 2.13: RMS error term $\langle F^i|F^i\rangle^{1/2}$ versus iteration number for simple mixing method, Anderson's method and Johnson's formulation of Broyden's method.

Chapter 3

Transistors with Thin Extensions

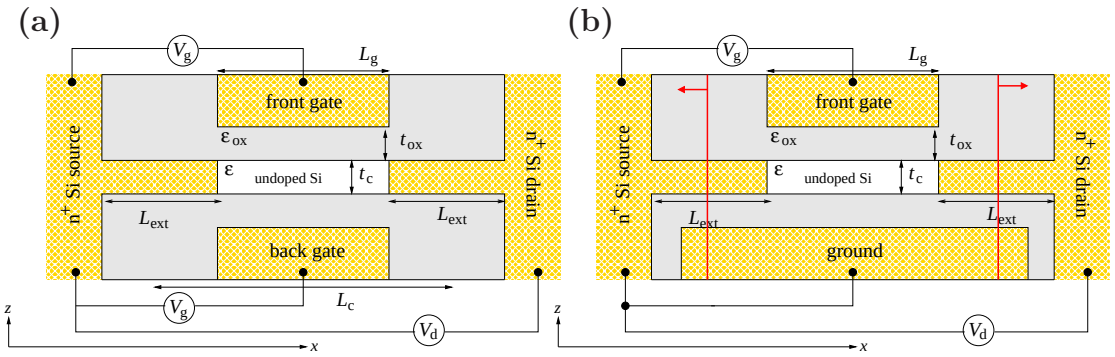

Figure 3.1: Model ballistic FET with thin electrode extensions for (a) double gate and (b) single gate structures. Red lines indicate cutoff where the transistor dynamics become independent of the model.

The double and single gate models of the ballistic FET structure with thin electrode extensions are shown in Fig. 3.1. Due to their closer relation to current fabrication techniques, these models are widely discussed in the literature (see *e. g.* [64] and references therein). The thin extension device has advantages and drawbacks over its bulk counterpart (Fig. 2.4(b)). The thin channel extensions increase the overall "bulk-to-bulk" device length reducing the maximum packing density. Additionally, the thin electrodes are less able to dissipate the energy of impinging ballistic electrons, so these devices may have unacceptable heating and higher thermal scaling limits [96]. This has motivated discussion of so-called "flared" or "dog-bone" structures in the literature [18]. Additionally, the shorter channel length increases direct source-to-drain

tunneling current. On the other hand, since the gate completely encompasses the channel region, the thin extensions provide better electrostatic control over the entire channel.

When a channel electron propagates into the doped extension region, elastic scattering with the background dopants creates a non-zero probability that the particle will be back-scattered into the channel region. In the solution to the Schrödinger equation derived below a "completely absorbing" boundary condition for the wavefunction is assumed so this model does not account for back-scattered electrons. The reflection probability for high doping levels was studied numerically in both a Monte Carlo model and through solution of the Boltzmann equation [97]. The results indicate the reflection probability for high energy electrons takes a analytical form [97],

$$R_f = \frac{1}{1 + \left[\frac{2}{\pi}\left(\frac{L_{\text{ext}}}{l} + \beta\right)\sin(\theta)\right]^{-1}}, \tag{3.1}$$

for mean free path l, doped extension length L_{ext} and incident phase space angle $\theta = \cos^{-1}(\Phi_0/E)^{1/2}$. Here β is a small correction term caused by the non-uniform distribution of "hot" carriers in the $[p_x, p_y]$ phase plane. When the length of the doped extension is on the order of, or larger than the mean free path, $L_{\text{ext}}/l \gg \beta$, so we may re-write Eq. (3.1) in terms of transmission probability $\mathcal{D}$ as

$$\left[\frac{2}{\pi}\frac{L_{\text{ext}}}{l}\sin(\theta)\right] = \frac{1-\mathcal{D}}{\mathcal{D}}. \tag{3.2}$$

The $\sin(\theta)$ term may be interpreted as the fraction of channel transverse modes used for transport, so Eq. (3.1) represents nothing more than the ballistic Landauer resistance $G_L^{-1} = h/2e^2$ connected in series with the normal Ohmic resistance of the thin extension

$$G^{-1} = G_L^{-1} + G_O^{-1}. \tag{3.3}$$

Hence, the effect of back-scattering from the doped extension region can modeled by addition of the potential drop across the two doped extensions with resistance (per unit width) $R = 2\sigma^{-1}L_{\text{ext}}/t_c$ while maintaining the assumption of the completely absorbing boundary conditions. For our standard doping, $n_D = 0.3$ nm^{-3} we may use the bulk resistivity (see p. 31 of Ref. [9]) $\sigma^{-1} = 40$ Ωcm.

3.1 1-D Schrödinger Approximation

We begin with the time independent Schrödinger equation

$$\begin{aligned} \mathcal{H}\Psi(x,y,z) &= E\Psi(x,y,z), & (3.4)\\ &= -\frac{\hbar^2}{2}\left(\frac{1}{m_x}\partial_x^2+\frac{1}{m_y}\partial_y^2+\frac{1}{m_z}\partial_z^2\right)+\Phi(x,z). & (3.5)\end{aligned}$$

Assuming that the device is sufficiently wide, wavefunction $\Psi(x,y,z)$ may be partitioned

$$\Psi(x,y,z)=\Psi(x,z)e^{\mathrm{i}k_y y}, \tag{3.6}$$

and through standard separation of variables, Eq. (3.5) may be written

$$\frac{1}{\Psi(x,z)}\left(\frac{1}{m_x}\partial_x^2+\frac{1}{m_z}\partial_z^2\right)\Psi(x,z)+\frac{1}{\psi(y)}\partial_z^2\psi(y)-\frac{2}{\hbar^2}\left(\Phi(x,z)-E\right)=0.$$

Noting that only the third term contains any y dependence, $\frac{1}{\psi(y)}\partial_y^2\psi(y)$ yields a constant $-(2m_y/\hbar^2)E_y$ and the result is re-written

$$\frac{1}{\Psi(x,z)}\left(\frac{1}{m_x}\partial_x^2+\frac{1}{m_z}\partial_z^2\right)\Psi(x,z)+\frac{2}{\hbar^2}\left(E-E_y-\Phi(x,z)\right)=0. \tag{3.7}$$

We look for a series solution to equation (3.7)

$$\Psi(x,z)=\sum_n \varphi_n(z)\psi_n(x), \tag{3.8}$$

and choose $\varphi_n(z)$ to be the solution to the equation

$$-\frac{\hbar^2}{2m_z}\frac{\partial^2}{\partial z^2}\varphi_n(z)=E_{z,n}\varphi_n(z), \tag{3.9}$$

These infinite well basis states closely approximate the full solution and we have made no constraints on function $\psi_n(x)$. Plugging into equation (3.7) yields

$$\begin{aligned}\sum_{n'}\Bigg[\left(-\frac{\hbar^2}{2m_x}\frac{\partial^2}{\partial x^2}\psi_{n'}(x)-(E-E_{z,n'}-E_y)\psi_{n'}(x)\right)\varphi_{n'}(z)\\ +\Phi(x,z)\psi_{n'}(x)\varphi_{n'}(z)\Bigg]=0.\end{aligned} \tag{3.10}$$

Because $\varphi_n(z)$ is orthonormal, multiplying both sides by $\varphi_n^*(z)$ and integrating over the channel thickness we find

$$\left[-\frac{\hbar^2}{2m_x}\frac{\partial^2}{\partial x^2} - (E - E_{z,n} - E_y) + \sum_{n'}\left(\int_{-t_c/2}^{t_c/2} \varphi_{n'}(z)\Phi(x,z)\varphi_n^*(z)dz\right)\right]\psi_n(x) = 0. \tag{3.11}$$

The $\Phi(x,z)$ term in parenthesis can be seen as the matrix element of the electrostatic potential

$$\bar{\Phi}_n(x) = \sum_{n'}\int_{-t_c/2}^{t_c/2} \varphi_{n'}(z)\Phi(x,z)\varphi_n^*(z)dz, \tag{3.12}$$

and equation (3.11) reduces to the effective one dimensional form

$$\left[\frac{\partial^2}{\partial x^2} + \frac{2m_x}{\hbar^2}\left(E - E_{z,n} - E_y - \bar{\Phi}_n(x)\right)\right]\psi_n(x) = 0, \tag{3.13}$$

where

$$E_{z,n} = \frac{\hbar^2 n^2 \pi^2}{2m_z t_c^2}. \tag{3.14}$$

Rapid evaluation of $\bar{\Phi}_n(x)$ for arbitrary node n may be done by recognition that the sum over n' is a Fourier transform of $\Phi(x,z)$ (see appendix A.2). Equation (3.13) may further be simplified by the observation that for electrons injected from thin source / drain electrodes, they already have confinement energy $E_{z,n}$. Hence this term falls out altogether and the effective 1-D Schrödinger equation becomes

$$\left[\frac{\partial^2}{\partial x^2} + \frac{2m_x}{\hbar^2}\left(E - E_y - \bar{\Phi}_n(x)\right)\right]\psi_n(x) = 0. \tag{3.15}$$

Equations (3.13), (3.15) are the effective 1-D Schrödinger equation for particles in channel sub-band n. When the channel thickness t_c is small, only the lowest sub-band contributes significantly and the average potential is reduced to

$$\bar{\Phi}(x) = \frac{2}{t_c}\int_{-t_c/2}^{t_c/2} \Phi(x,z)\cos^2\left(\frac{\pi z}{t_c}\right)dz. \tag{3.16}$$

3.1.1 Channel Density

The number of electrons in the channel is given by the summation over states

$$N(x,y,z) = g_s g_v \sum_{\boldsymbol{k}} |\Psi(x,y,z)|^2 \, f(E). \tag{3.17}$$

Using wavefunction partition (3.6) and assuming only contributions from lowest sub-band, the electron density in the channel may be calculated as

$$n_{3D}(x,z) = \left(\frac{2}{t_c}\right) \cos^2\left(\frac{\pi z}{t_c}\right) n_{2D}(x). \tag{3.18}$$

The channel sheet density is found from the single state wavefunctions

$$n_{2D}(x) = \frac{g_s g_v}{4\pi^2} \int_{-\infty}^{\infty} dky \int_{0}^{\infty} |\psi(x)|^2 \, f(E_x + E_y), \tag{3.19}$$

where $\psi(x)$ is the solution to (3.15) at energy E_x inside the channel region. Using the dispersion relation $k = \sqrt{2mE}/\hbar$, the sheet density may be expressed

$$n_{2D}(x) = \frac{g_s g_v \sqrt{m_x m_y}}{4\pi^2} \int_{0}^{\infty} dE_x \, E_x^{-1/2} \, |\psi(x)|^2 \int_{0}^{\infty} dE_y \, E_y^{-1/2} f(E_x + E_y). \tag{3.20}$$

Converting to dimensionless energy variables $\epsilon \equiv E/T, \epsilon_F \equiv \mu_{\mathrm{F}}/T$,

$$n_{2D}(x) = \frac{g_s g_v \sqrt{m_x m_y} T}{4\pi^2 \hbar^2} \int_{0}^{\infty} \frac{d\epsilon_x}{\sqrt{\epsilon_x}} |\psi(x)|^2 \int_{0}^{\infty} \frac{d\epsilon_y}{\sqrt{\epsilon_y}} \frac{1}{1 + \exp(\epsilon_x + \epsilon_y - \mu_{\mathrm{F}})}, \tag{3.21}$$

and using degeneracy values $g_s = 2$, $g_v = 2$, the sheet density is expressed

$$n_{2D}(x) = \frac{\sqrt{m_x m_y} T}{\pi^2 \hbar^2} \int_{0}^{\infty} |\psi(x)|^2 \mathcal{F}_{-1/2}(\epsilon_F - \epsilon_x) \epsilon_x^{-1/2} d\epsilon_x, \tag{3.22}$$

in terms of the Fermi-Dirac integral, Eq. (2.21). This sheet density should be calculated twice. Once each for electrons injected from the source and drain electrodes. For a discussion of the numeric solution of this integral, see B.1.

3.1.1.1 Quantum Limit

In the limit that temperature $T \to 0$, we replace the Fermi distribution with Heaviside step function $f(E_x + E_y) = \Theta(E_\mathrm{F} - E_x - E_y)$, in equation (3.20). Evaluating

$$\int_0^\infty \Theta(E_\mathrm{F} - E_x - E_y) E_y^{-1/2}\, dE_y = 2\left(E_\mathrm{F} - E_x\right)^{1/2}, \tag{3.23}$$

the sheet density is written

$$n_{2D}(x) = \frac{g_s g_v \sqrt{m_x m_y}}{2\pi^2\hbar^2} \int_0^{E_\mathrm{F}} dE_x\, E_x^{-1/2} \left(E_\mathrm{F} - E_x\right)^{1/2} |\psi(x)|^2 . \tag{3.24}$$

3.1.2 Current Density

The assumption of single sub-band transport also allows simplification of the expression for the device current. Since the particles are injected from a thin region matching the transport channel, they already have energy $E_{z,1}$ required to overcome the confinement potential. The summation (2.34) then becomes two dimensional integral

$$\sum_{\boldsymbol{k}} \to \frac{W L_B}{(2\pi)^2} \int d^2k, \tag{3.25}$$

and the current density is expressed

$$J = I/W = e\frac{g_s g_v \hbar}{m_x (2\pi)^2} \int d^2k k_x \mathcal{D}(E_x) f(E_x + E_y). \tag{3.26}$$

The transmission probability is only a function of the particles wavevector in the $\hat{x}$ direction. Using relations

$$\begin{aligned} k_x dk_x &= \frac{m_x}{\hbar^2} dE_x, \\ dk_y &= \frac{\sqrt{2m_y}}{2\hbar} E_y^{-1/2} dE_y, \end{aligned}$$

the current density is written for one direction (*i. e.* left to right) as

$$J = e\frac{g_s g_v\sqrt{2m_y}}{(2\pi)^2\hbar^2}\int_0^\infty dE_x\,\mathcal{D}(E_x)\int_0^\infty dE_y\,E_y^{-1/2}f(E_x+E_y). \tag{3.27}$$

Or again in terms of the Fermi-Dirac integral, the total current density is

$$J = \frac{J_0}{\pi}\int_0^\infty d\epsilon_x\mathcal{D}(\epsilon_x T)\left[\mathcal{F}_{-1/2}(\epsilon_F-\epsilon_x)-\mathcal{F}_{-1/2}(\epsilon_F-\nu_D-\epsilon_x)\right], \tag{3.28}$$

where $\nu_d \equiv eV_d/T$ and

$$J_0 \equiv e\frac{\sqrt{2m_y}T^{3/2}}{\pi\hbar^2}. \tag{3.29}$$

3.1.2.1 "Simple Transistor" Limit

In the simple transistor limit, all electrons with energies below the potential maximum in the channel are reflected and all the electrons above are transmitted as $T \to 0$. The transmission coefficient is then

$$\mathcal{D}(E_x) = \Theta(E_x-\Phi_0), \tag{3.30}$$

where Φ_0 is the maximum of the potential in the channel

$$\Phi_0 = \Phi_{\max}(x), \tag{3.31}$$

and using $g_s = 2$ and $g_v = 2$, the current density reduces to

$$J = e\frac{2(2m_y)^{1/2}}{\pi^2\hbar^2}\int_{\Phi_0}^{E_\mathrm{F}} dE_x\Theta(E_x-\Phi_0)(E_\mathrm{F}-E_x)^{1/2}. \tag{3.32}$$

Evaluating the integral,

$$J = e\frac{2(2m_y)^{1/2}}{\pi^2\hbar^2}\left[-\frac{2}{3}(E_\mathrm{F}-E_x)^{3/2}\right]_{E_{z,1}+\Phi_0}^{E_\mathrm{F}}, \tag{3.33}$$

or

$$J = e\frac{4(2m_y)^{1/2}}{3\pi^2\hbar^2}(E_\mathrm{F}-\Phi_0)^{3/2}. \tag{3.34}$$

3.1.3 Evaluation of the Wavefunction

To evaluate the wavefunction in the channel, we divide the device into three regions: source, drain and channel (Fig. 3.2). Using plane wave solutions for the wavefunctions in the electrode regions, the solutions to Eq. (3.15) may be written

$$\begin{aligned} \psi_s(x) &= e^{\mathrm{i}k_w x} + Be^{-\mathrm{i}k_w x}, \\ \psi(x) &= Cf(x) + Dg(x), \\ \psi_d(x) &= Fe^{\mathrm{i}k_\nu (x-L_c)}. \end{aligned} \tag{3.35}$$

The two terms for $\psi_s(x)$ represent the incident and reflected states and the undetermined constants B, C, D, and F are assumed normalized to the amplitude of the incident wave.

The functions $f(x)$ and $g(x)$ are linearly independent functions to be found numerically. The appropriate selection of the boundary conditions for the numeric solutions to $f(x)$, $g(x)$ is key to the rapid solution of (3.15). The speed of this solution underlies the speed of overall execution time because the integration of the wavefunctions over energy (3.22) at each channel mesh point becomes the limiting factor for execution time. We choose the boundary conditions to satisfy

$$\begin{array}{c|c} f(0) = 1 & g(0) = 0, \\ f(L_c) = 0 & g(L_c) = 1. \end{array} \tag{3.36}$$

For numerical calculations, such presentation has a distinct advantage over the usual Cauchy approach, in frequent cases when the $\psi(L_c)/\psi(0)$ ratio is very small, for example as a result of the exponential tunneling of $\psi(x)$ through the potential barrier formed in the channel by negative gate voltage. In this case, the smallness may be expressed by a small D/C ratio and does not require a very precise calculation of functions $f(x)$ and $g(x)$.

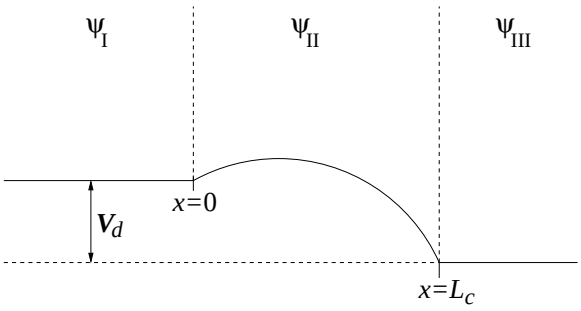

Figure 3.2: Arbitrary 1-D potential profile.

We may easily express equation (3.15) as finite difference equation

$$\psi_{i+1} + \psi_{i-1} + \left[\frac{2m_x\Delta_s^2}{\hbar^2}\left(E - E_y - \bar{\Phi}(x_i)\right) - 2\right]\psi_i = 0, \tag{3.37}$$

with mesh spacing Δ_s. With the choice of boundary conditions (3.36), this FDE is simply a tri-diagonal system of equations which may be rapidly solved in $\mathcal{O}(N)$ time (see section 2.3.1.1).

Because of the rapid oscillation of higher energy wavefunctions, the mesh spacing used for the Poisson solver is far too crude. In order to maintain overlap between the wavefunction mesh and the Poisson mesh (and hence points where knowledge of the electron density is desired) we divide the Poisson spacing by a factor 2^{M_m} to solve the wavefunction FDE. $M_m = 8$ by default as a starting value but may be increased dynamically in cases where higher accuracy is required. We then calculate integral (3.22) at node points corresponding to the FDE for the Poisson equation.

We require at the interface between device regions

$$\begin{array}{c|c} \psi_s(0) = \psi(0) & \partial_x\psi_s(0) = \partial_x\psi(0), \\ \psi(L_c) = \psi_d(L_c) & \partial_x\psi(L_c) = \partial_x\psi_d(L_c). \end{array} \tag{3.38}$$

Enforcement of boundary conditions (3.38) yields expressions for the unknown constants

$$\begin{aligned} B &= \frac{\mathrm{i}k_w\left(\mathrm{i}k_\nu - g'_{L_c}\right) - f'_0\left(\mathrm{i}k_\nu - g'_{L_c}\right) - g'_0 f'_{L_c}}{\left(\mathrm{i}k_w + f'_0\right)\left(\mathrm{i}k_\nu - g'_{L_c}\right) + g'_0 f'_{L_c}}, \\ C &= \frac{2\mathrm{i}k_w\left(\mathrm{i}k_\nu - g'_{L_c}\right)}{\left(\mathrm{i}k_w + f'_0\right)\left(\mathrm{i}k_\nu - g'_{L_c}\right) + g'_0 f'_{L_c}}, \\ D &= \frac{2\mathrm{i}k_w f'_{L_c}}{\left(\mathrm{i}k_w + f'_0\right)\left(\mathrm{i}k_\nu - g'_{L_c}\right) + g'_0 f'_{L_c}}, \\ F &= \frac{2\mathrm{i}k_w f'_{L_c}}{\left(\mathrm{i}k_w + f'_0\right)\left(\mathrm{i}k_\nu - g'_{L_c}\right) + g'_0 f'_{L_c}}, \end{aligned} \tag{3.39}$$

where $f'_0 \equiv \partial_x f(0)$, $f'_{L_c} \equiv \partial_x f(L_c)$, $g'_0 \equiv \partial_x g(0)$, $g'_{L_c} \equiv \partial_x g(L_c)$, are the derivatives of the numeric solutions at the boundaries.

With constants determined in terms of the numerical solutions, the transmission probability $\mathcal{D}(E) = (k_\nu/k_w)|F|^2$ is readily found to be

$$\mathcal{D}(E) = \frac{4k_\nu k_w f'^2_{L_c}}{\left(g'_0 f'_{L_c} - g'_{L_c} f'_0 - k_\nu k_w\right)^2 + \left(k_\nu f'_0 - k_w g'_{L_c}\right)^2}. \tag{3.40}$$

We also need to calculate the square modulus of the wavefunction in the channel, but for completeness show the results for all three device regions. We consider two cases independently.

1. k_ν real:

   In this case, the particle's energy is greater than the constant potential in region III, and is relevant for all electrons originating in the source and for drain electrons with $E > eV_d$. For the wavefunction in the source, we can take advantage of the unit boundary conditions, recognize that $B = C - 1$ and immediately write

$$\psi_s(x) = 2\mathrm{i}\left[\sin(k_w x) + \frac{k_w(\mathrm{i}k_\nu - g'_{L_c})}{(\mathrm{i}k_w + f'_0)(\mathrm{i}k_\nu - g'_{L_c}) + g'_0 f'_{L_c}} e^{-\mathrm{i}k_w x}\right]. \quad (3.41)$$

   The wavefunction modulus in the three regions is easily calculated as

$$|\psi_s(x)|^2 = 4\left[\sin^2(k_w x) + \frac{k_w^2(k_\nu^2 + g'^2_{L_c})}{\left(g'_0 f'_{L_c} - g'_{L_c} f'_0 - k_\nu k_w\right)^2 + \left(k_\nu f'_0 - k_w g'_{L_c}\right)^2} + 2\sin(k_w x) k_w \Re\left(\frac{\mathrm{i}k_\nu - g'_{L_c}}{(\mathrm{i}k_w + f'_0)(\mathrm{i}k_\nu - g'_{L_c}) + g'_0 f'_{L_c}} e^{-\mathrm{i}k_w x}\right)\right], \quad (3.42)$$

$$|\psi(x)|^2 = \frac{4k_w^2\left[k_\nu^2 f^2(x) + \left(g'_{L_c} f(x) - f'_{L_c} g(x)\right)^2\right]}{\left(g'_0 f'_{L_c} - g'_{L_c} f'_0 - k_\nu k_w\right)^2 + \left(k_\nu f'_0 - k_w g'_{L_c}\right)^2}, \quad (3.43)$$

$$|\psi_d(x)|^2 = \frac{4k_w^2 f'^2_{L_c}}{\left(g'_0 f'_{L_c} - g'_{L_c} f'_0 - k_\nu k_w\right)^2 + \left(k_\nu f'_0 - k_w g'_{L_c}\right)^2}. \quad (3.44)$$

2. k_ν imaginary: We will also need to consider the contribution of electrons originating in the drain regions with $E < eV_d$. In this case, k_ν in the source electrode becomes imaginary. Again using the relation $B = C - 1$, we may write

$$\psi_s(x) = 2\mathrm{i}\left[\sin(k_w x) + \frac{k_w(|k_\nu| + g'_{L_c})}{(\mathrm{i}k_w + f'_0)(|k_\nu| + g'_{L_c}) - g'_0 f'_{L_c}} e^{-\mathrm{i}k_w x}\right]. \quad (3.45)$$

The wavefunction modulus in the three regions is then found

$$|\psi_s(x)|^2 = 4\left[\sin^2(k_w x) + \frac{k_w^2(|k_\nu| + g'_{L_c})^2}{k_w^2\left(|k_\nu| + g'_{L_c}\right)^2 + \left(f'_0(|k_\nu| + g'_{L_c}) - g'_0 f'_{L_c}\right)^2} + 2\sin(k_w x)k_w(|k_\nu| + g'_{L_c})\Re\left(\frac{e^{-\mathrm{i}k_w x}}{(\mathrm{i}k_w + f'_0)(|k_\nu| + g'_{L_c}) - g'_0 f'_{L_c}}\right)\right], \tag{3.46}$$

$$|\psi(x)|^2 = \frac{4k_w^2\left[\left(|k_\nu| + g'_{L_c}\right) f(x) - f'_{L_c} g(x)\right]^2}{k_w^2\left(|k_\nu| + g'_{L_c}\right)^2 + \left(f'_0(|k_\nu| + g'_{L_c}) - g'_0 f'_{L_c}\right)^2}, \tag{3.47}$$

$$|\psi_d(x)|^2 = \frac{4k_w^2 f'^2_{L_c} e^{-2|k_\nu|(x-L_c)}}{k_w^2\left(|k_\nu| + g'_{L_c}\right)^2 + \left(f'_0(|k_\nu| + g'_{L_c}) - g'_0 f'_{L_c}\right)^2}. \tag{3.48}$$

3.1.3.1 Linear Potential

Many practical potential profiles may be well approximated by a linear potential drop between the source and drain electrodes. While this is only an approximation to an actual device profile, convenient expressions may be found which enable rapid estimation of device dynamics.

The effective potential

$$U(x) = V_0 + V_d\left(1 - \frac{x}{L_c}\right), \tag{3.49}$$

shown in Fig. 3.3, with arbitrary shift V_0, has general solution in the channel

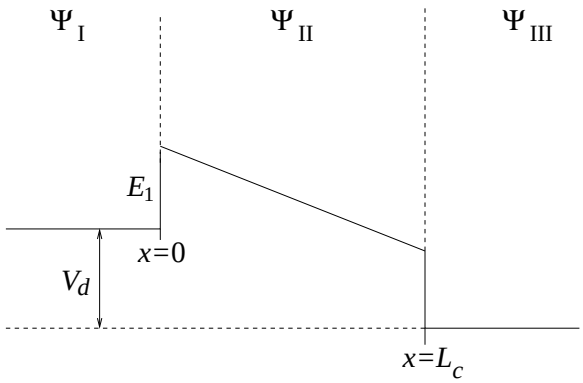

Figure 3.3: Linear 1-D potential profile.

region

$$\psi(x) = C\mathrm{Ai}\left(\alpha(x)\right) + D\mathrm{Bi}\left(\alpha(x)\right), \tag{3.50}$$

where

$$\alpha(x) \equiv \left(\frac{2mL_c^2}{\hbar^2 V_d^2}\right)^{1/3}(V_0 - E - V_d\frac{x}{L_c}), \tag{3.51}$$

$$\frac{d}{dx}\alpha = -\left(\frac{2mV_d}{\hbar^2 L_c}\right)^{1/3}. \tag{3.52}$$

For most linear problems, the raw use of Eq. (3.50) becomes numerically unstable. A better solution is to use the unit boundary conditions (3.36). This yields expressions

$$\begin{aligned} f(x) &= \frac{\mathrm{Bi}(\alpha(L_c))\mathrm{Ai}(\alpha(x)) - \mathrm{Ai}(\alpha(L_c))\mathrm{Bi}(\alpha(x))}{\mathrm{Ai}(\alpha(0))\mathrm{Bi}(\alpha(L_c)) - \mathrm{Bi}(\alpha(0))\mathrm{Ai}(\alpha(L_c))}, \\ g(x) &= \frac{\mathrm{Ai}(\alpha(0))\mathrm{Bi}(\alpha(x)) - \mathrm{Bi}(\alpha(0))\mathrm{Ai}(\alpha(x))}{\mathrm{Ai}(\alpha(0))\mathrm{Bi}(\alpha(L_c)) - \mathrm{Bi}(\alpha(0))\mathrm{Ai}(\alpha(L_c))}. \end{aligned} \tag{3.53}$$

When $\alpha(x) \gg 1$, calculation of the individual Airy functions is numerically unfeasible; to first order, the Airy functions are approximated by [98]

$$\begin{aligned} \mathrm{Ai}(x \to \infty) &\approx \frac{e^{-\gamma}}{2\sqrt{\pi}x^{1/4}}, \\ \mathrm{Ai}'(x \to \infty) &\approx -\frac{x^{1/4}e^{-\gamma}}{2\sqrt{\pi}}, \\ \mathrm{Bi}(x \to \infty) &\approx \frac{e^{\gamma}}{\sqrt{\pi}x^{1/4}}, \\ \mathrm{Bi}'(x \to \infty) &\approx \frac{x^{1/4}e^{\gamma}}{\sqrt{\pi}}, \\ \mathrm{Ai}(x \to -\infty) &\approx \frac{\cos(\gamma - \pi/4)}{\sqrt{\pi}x^{1/4}}, \\ \mathrm{Ai}'(x \to -\infty) &\approx -\frac{4\cos(\gamma + \pi/4)x^{3/2} - \sin(\gamma + \pi/4)}{4\sqrt{\pi}x^{5/4}}, \\ \mathrm{Bi}(x \to -\infty) &\approx -\frac{\sin(\gamma - \pi/4)}{\sqrt{\pi}x^{1/4}}, \\ \mathrm{Bi}'(x \to \infty) &\approx \frac{4\sin(\gamma + \pi/4)x^{3/2} + \cos(\gamma + \pi/4)}{4\sqrt{\pi}x^{5/4}}. \end{aligned} \tag{3.54}$$

where

$$\gamma \equiv \frac{2}{3}x^{3/2}.$$

With these approximations functions $f(x)$, $g(x)$ are approximated

$$\begin{aligned} f(x) &\approx \frac{\alpha^{1/4}(0)\left(\exp\left[2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(x))\right]-\exp\left[-2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(x))\right]\right)}{\alpha^{1/4}(x)\left(\exp\left[2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right]-\exp\left[-2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right]\right)}, \\ g(x) &\approx \frac{\alpha^{1/4}(0)\left(\exp\left[2/3(\alpha^{3/2}(x)-\alpha^{3/2}(0))\right]-\exp\left[-2/3(\alpha^{3/2}(x)-\alpha^{3/2}(0))\right]\right)}{\alpha^{1/4}(x)\left(\exp\left[2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right]-\exp\left[-2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right]\right)}, \end{aligned} \tag{3.55}$$

with derivatives

$$\begin{aligned} f'(x) &= \left(\frac{d\alpha}{dx}\right)\frac{(\mathrm{Bi}(\alpha(L_c))\mathrm{Ai}'(\alpha)-\mathrm{Ai}(\alpha(L_c))\mathrm{Bi}'(\alpha))}{\mathrm{Ai}(\alpha(0))\mathrm{Bi}(\alpha(L_c))-\mathrm{Bi}(\alpha(0))\mathrm{Ai}(\alpha(L_c))}, \\ g'(x) &= \left(\frac{d\alpha}{dx}\right)\frac{\mathrm{Ai}(\alpha(0))\mathrm{Bi}'(\alpha)-\mathrm{Bi}(\alpha(0))\mathrm{Ai}'(\alpha)}{\mathrm{Ai}(\alpha(0))\mathrm{Bi}(\alpha(L_c))-\mathrm{Bi}(\alpha(0))\mathrm{Ai}(\alpha(L_c))}. \end{aligned} \tag{3.56}$$

The derivative $(d\alpha/dx)$ is calculated as

$$\frac{d\alpha}{dx} = -\left(\frac{2m_xV_d}{\hbar^2L_c}\right)^{1/3}, \tag{3.57}$$

yielding approximations

$$\begin{aligned} f'(x) &\approx \left(\frac{2m_xV_d}{\hbar^2L_c}\right)^{1/3}\frac{(\alpha(0)\alpha(x))^{1/4}\left(\exp\left[2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(x))\right]+\exp\left[-2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(x))\right]\right)}{\left(\exp\left[2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right]-\exp\left[-2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right]\right)}, \\ g'(x) &\approx -\left(\frac{2m_xV_d}{\hbar^2L_c}\right)^{1/3}\frac{(\alpha(0)\alpha(x))^{1/4}\left(\exp\left[2/3(\alpha^{3/2}(x)-\alpha^{3/2}(0))\right]+\exp\left[-2/3(\alpha^{3/2}(x)-\alpha^{3/2}(0))\right]\right)}{\left(\exp\left[2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right]-\exp\left[-2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right]\right)}. \end{aligned} \tag{3.58}$$

At the boundaries, these approximations reduce to

$$\begin{aligned} f'(0) &\approx \left(\frac{2m_xV_d}{\hbar^2L_c}\right)^{1/3}\frac{(\alpha(0))^{1/2}}{\tanh\left(2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right)}, \\ f'(L_c) &\approx \left(\frac{2m_xV_d}{\hbar^2L_c}\right)^{1/3}\frac{(\alpha(0)\alpha(L_c))^{1/4}}{\sinh\left(2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right)}, \\ g'(0) &\approx -\left(\frac{2m_xV_d}{\hbar^2L_c}\right)^{1/3}\frac{(\alpha(0))^{1/2}}{\sinh\left(2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right)}, \\ g'(L_c) &\approx -\left(\frac{2m_xV_d}{\hbar^2L_c}\right)^{1/3}\frac{(\alpha(0)\alpha(L_c))^{1/4}}{\tanh\left(2/3(\alpha^{3/2}(L_c)-\alpha^{3/2}(0))\right)}. \end{aligned} \tag{3.59}$$

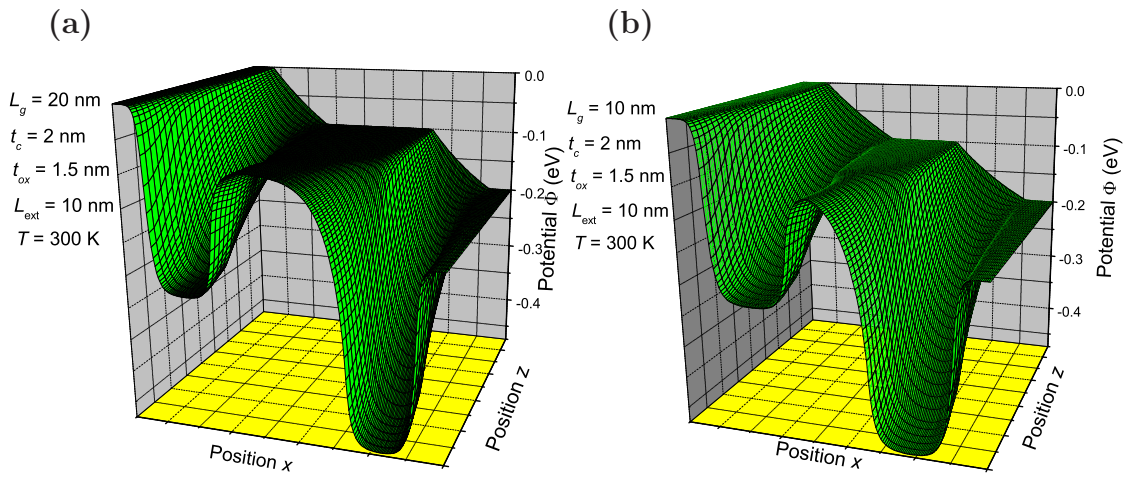

Figure 3.4: Sample 2D potential profiles for device with thin extensions with (a) $L_g = 20$ nm and (b) $L_g = 10$ nm. For each plot $V_g = 0.1$ V and $V_d = 0.2$ V.

3.2 Double-Gate Transistor

3.2.1 Potential

Typical results for the potential distribution are shown in Fig. 3.4 for two long channel devices. Only the upper half of the transistor is shown due to device symmetry. In both cases we use parameters $t_c = 2$ nm, $t_{ox} = 1.5$ nm, $L_{\text{ext}} = 10$ nm. Panel (a) shows the results for a 20 nm gate length device. While assuming wavefunction coherence over this length may be optimistic for present devices see section 2.1, the only impedance to the ballistic assumption is the surface roughness. If the surfaces can made clean enough, ballistic transport may be assumed over any device length [99]. Panel (b) shows the potential distribution for a more realistic 10 nm gate length device. All devices are calculated at room temperature using a standard doping density $n_D = 0.3$ nm^{-3}.

The transistor dynamics are dominated by the potential in the middle of the channel, shown in Fig. 3.5 for same parameters and three different gate lengths. The dotted lines mark the division between the bulk, thin extension and channel regions. For the longest gate device, the plateau region assumed in the analytical models of section 2.2.8 is clearly visible and the field produced by the drain bias is effectively shielded by the gate electrode. As the gate length is reduced to 10 nm, the onset of DIBL effects can clearly be seen. As the length of the gate is further reduced to 5 nm, the drain field not only reduces the barrier, but a significant reduction in the barrier thickness can

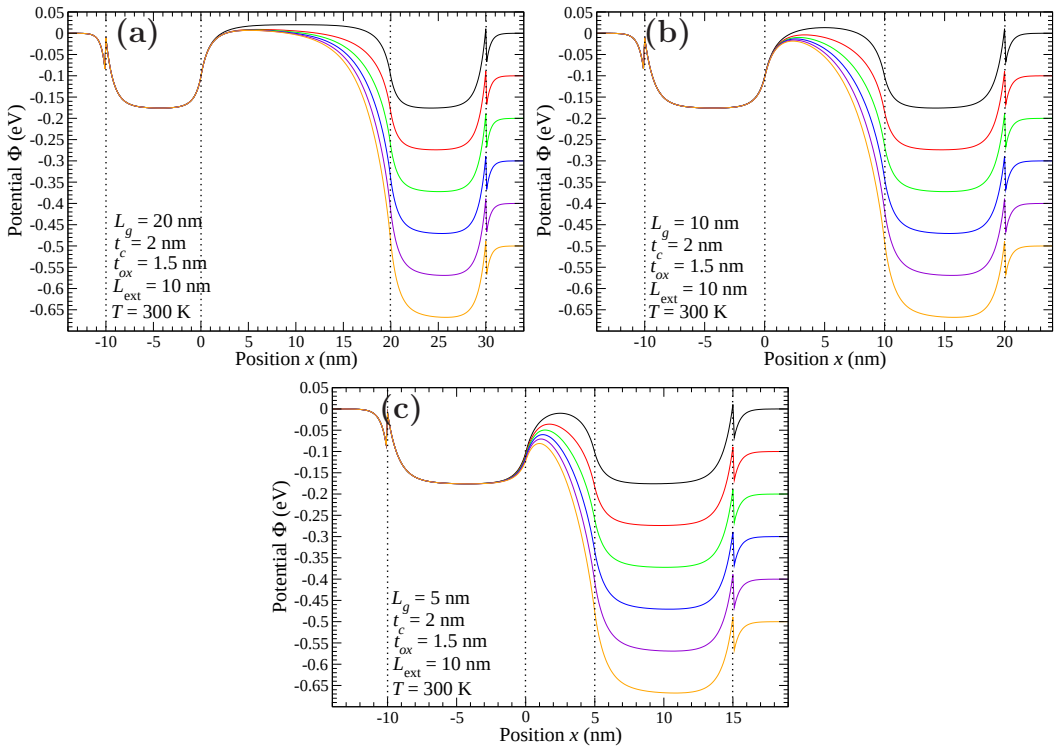

Figure 3.5: Potential through the middle of the channel for device with thin extensions for $V_g = 0.1$ V in V_d steps of 100 mV. The gate length for each panel is (a) $L_g = 20$ nm, (b) $L_g = 10$ nm and (c) $L_g = 5$ nm.

also be noted. This reduction of the barrier width may be sufficiently reduced that the onset of direct source-to-drain tunneling current begins.

3.2.2 $I - V_d$ Families

Figure 3.6 shows $I - V_d$ families in gate voltage steps of 50 meV for decreasing gate length from 10 nm to 2.5 nm. The solid lines are the current assuming no voltage drop across the thin electrodes (Fig. 3.5). The dashed lines are the current results scaled to account for voltage dropped in the doped extensions using resistivity $\rho = 3 \times 10^{-4}$ Ωcm. All the devices demonstrate high current densities for modest values of the gate voltage. In fact, this device may be seen as a "normally on" transistor ($J_{\text{sat}}(V_g = 0) \approx 7$ A/cm for $L_g = 10$ nm). While not ideal, the $J(V_g = 0)$ current may be a lowered by a shifted threshold voltage through gate work-function engineering (see section 2.2.8). The long channel device shows excellent current saturation, but noticeable degradation may be seen for the 5 nm gate length. The shortest channel considered demonstrates DIBL has completely ruined the saturation characteristics. Admittedly however, this configuration is a bit pathological as the conducting channel is nearly as thick as it is long.

3.2.3 Subthreshold Current

The subthreshold properties for the same set of devices is shown in Fig. 3.7. This oxide thickness has been increased to $t_{ox} = 2.5$ nm to minimize the gate leakage current. The near vertical dotted line shows the ideal thermal subthreshold slope

$$\Delta V_g = \ln(10)T = 60 \text{ mV/dec.} \tag{3.60}$$

The near horizontal line is an estimation of the gate leakage current, based on a WKB solution of tunneling through a trapezoidal barrier assuming a gate overlap of 2 nm [14]. For further discussion of the gate leakage estimation, see appendix A.1. The long channel transistor shows near perfect subthreshold slope and very little DIBL. As the gate length is scaled to 5 nm, the drain effects are clearly seen and the slope is severely degraded. DIBL effects also begin to affect the device, expressed as the spread of the slopes for different V_d. The onset of direct tunneling may also be seen by the (very slight) upward curvature near the bottom of the subthreshold region. However, the 5 nm device still shows nearly 6 orders of magnitude difference between on the and off states, which is close to what one would require for use in a memory application. As the gate is further scaled to the impracticable 2.5 nm, the gate clearly loses all control over channel electrostatics, and the device impractical

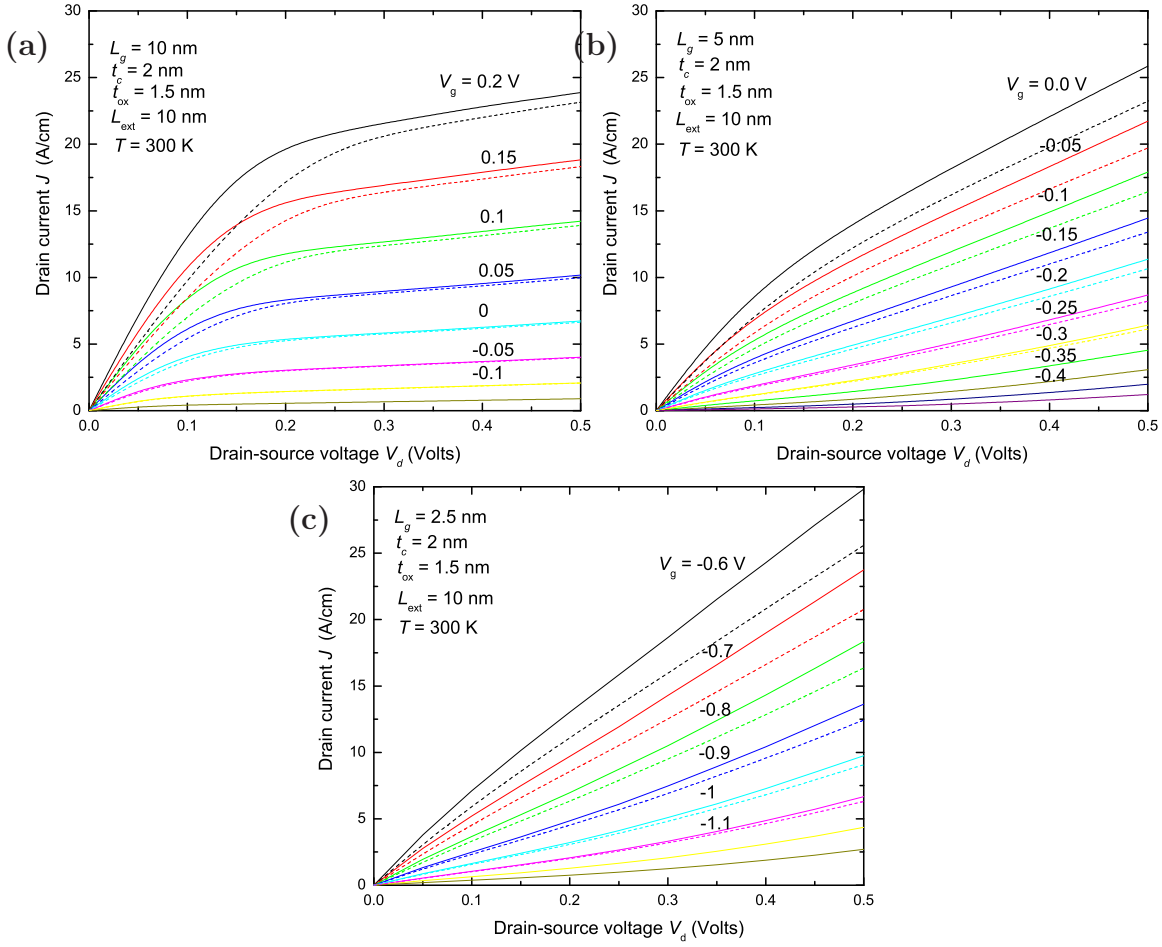

Figure 3.6: Source-Drain $I - V_d$ curves for DG MOSFET with thin extensions for (a) $L_g = 10$ nm, (b) $L_g = 5$ nm, (c) $L_g = 2.5$ nm. Solid lines show results for total voltage drop across the intrinsic channel. The dashed lines are the current scaled to account for the voltage spread over the entire thin region.

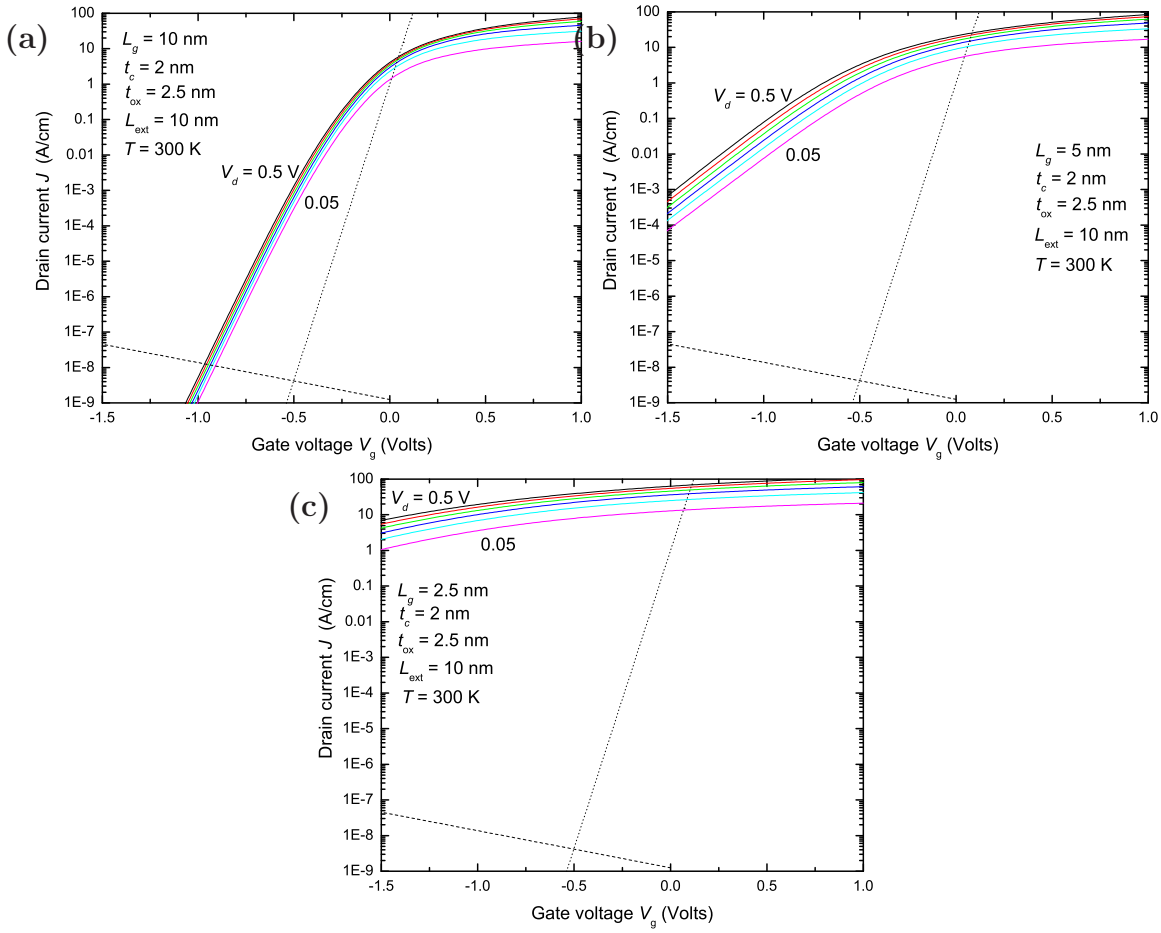

Figure 3.7: Subthreshold curves for the same gate lengths as Fig. 3.6. The oxide thickness has been increased $t_{ox} = 2.5$ nm to minimize gate-oxide leakage. Nearly vertical dotted lines represent ideal slope. Dashed lines are an estimate of gate-oxide leakage current.

for integrated circuits.

3.2.4 Voltage Gain

Numerically, calculation of G_v (2.57) may be considered a root finding problem. For given device voltages V_d, V_g, the current density is calculated. The gate voltage is then shifted twice by fixed, small amounts $\Delta_{vg}/2$ and a V_d is found such that $J - J(V_d, V_g \pm \Delta_{vg}/2) = 0$. This condition is rapidly found using Brent's method [77]. Using three different values for Δ_{vg} we may establish average and standard deviation values for G_v.

Figure 3.8 shows G_v for the range of gate voltage from subthreshold to saturation. The results indicate that a relatively long device with $L_g = 7.5$ nm will have sufficient signal gain over the entire range of operating voltages to be a functional transistor. However, G_v falls off rapidly with decreasing gate length, and can be seen in a simple view as the fractional response of the maximum of the potential barrier in the channel to the gate voltage $G_v \propto \alpha\Phi_0$. However, these results indicate that even for an ultra-short total transport length $L_c = L_g = 5$ nm the ballistic FET may still be an acceptable candidate, and the fundamental limit on scaling from a performance perspective may end for the thin extension device near $L_g \approx 5$ nm.

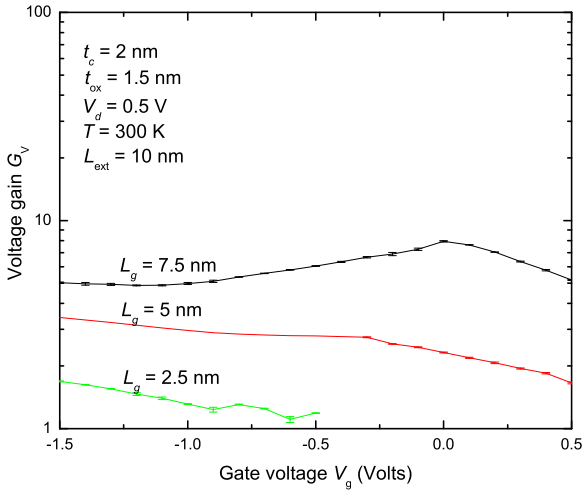

Figure 3.8: Figure-of-merit voltage gain at $V_d = 0.5$ V versus gate voltage for decreasing gate lengths.

3.2.5 Threshold Voltage Rolloff

While the results above are very encouraging from a pure performance perspective, they do not account for more practical circuit design considerations. Minute fluctuations in the fabrication process may lead to unacceptably large deviations in transistor properties within the same circuit. We define the "rolloff" in the transistor as the deviation of the threshold voltage at gate length L_g from an equivalent device at $L_g \to \infty$. Here we define the threshold voltage as the required gate voltage such that $J[V_d, V_t(L_g)] = J_{\text{thresh}} \equiv 1 \times 10^{-4}$ A/cm. For the device with thin extensions and infinite gate length, we may assume only classically transmitted electrons with potential barrier height $\Phi_0 = V_g + E_{z,1}$. The threshold voltage is then found numerically

$$J_{\text{cl}}(V_d, V_g) - J_{\text{thresh}} = 0, \tag{3.61}$$

using Brent's method for finding roots [77]. J_{cl} is given as

$$J_{\text{cl}} = \frac{J_0}{\pi} \left[\mathcal{J}(\nu_g - \epsilon_F) - \mathcal{J}(\nu_g - \epsilon_F - \nu_d) \right], \tag{3.62}$$

where J_0 is given by Eq. (3.29), $\nu_{d,g} = V_{d,g}/T$, $\epsilon_F = \mu_{\text{F}}/T$, and

$$\mathcal{J}(\eta) \equiv \int_0^\infty \epsilon_y^{-1/2} \ln \left[1 + \exp\left(-\eta - \epsilon_y \right) \right] d\epsilon_y, \tag{3.63}$$

(see appendix A.3).

The rolloff in the threshold voltage is shown in Fig. 3.9 for the standard doping $n_D = 0.3$ nm^{-3} for decreasing gate length L_g over a range of channel and oxide thicknesses. In all cases, for sub-10 nm devices, the change in threshold from the ideal value begins to grow exponentially. This exponential growth is a reflection of the loss of electrostatic control of the gate over the channel potential. In the subthreshold region $J_T \propto \exp(G_v)$. As G_v is reduced the device must supply exponentially more voltage to shut the current.

Moreover, ultra-tight fabrication control cannot be achieved at the expense of the other device critical dimensions. The change in the threshold voltage from the "standard" device versus channel thickness and oxide thickness is shown in Fig. 3.10. For ultra-short gate length devices, the swing in the threshold voltage may be even more sensitive to variations in the channel in oxide thicknesses. All of the results may be seen intuitively from the electrostatic picture. In this picture, the devices with the thickest channels and oxide layers (shown by the green lines in both panels) allow the weakest field penetra-

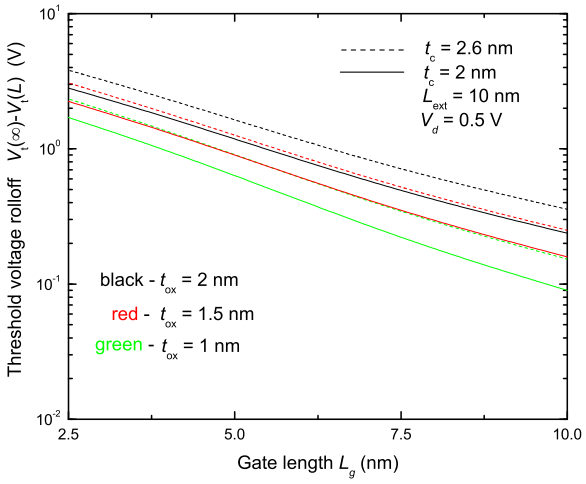

Figure 3.9: "Rolloff" of the threshold voltage of the device with thin extensions for two different channel thickness t_c and three values of the oxide thickness t_{ox}.

tion into the channel, and thus require the higher applied bias to compensate for the change in device dimension.

The exponentially growing sensitivity of V_t to natural geometrical deviations may lead an unacceptably high cost of fabrication. Unless radically new methods for large scale creation of very-large-scale-integrated (VLSI) circuits are developed, economic limitations may end device scaling long before the physical limits are reached.

3.2.6 Power

In addition to concerns over fabrication tolerances, the circuit power consumption is already a major problem for today's VLSI circuitry. The minimum power operating point (2.68) is plotted in Fig. 3.11 versus the device gate length for two different values of effective activity "switching parameter" λ. In stark contrast with the expectations of scaling for traditional devices, the transistor operating power increases with shrinking gate lengths. The power minimum not only increases, but grows exponentially in all devices considered. This is a direct result of transition from diffusion physics, where current saturation is provided by channel scattering, to the ballistic regime where current saturation is the result of exhaustion of supply electrons [15]. As L_g is decreased in the ballistic regime device performance degrades, either through loss of electrostatic gate control or tunneling current, and larger V_{DD} is required to shunt the leakage current. The larger V_{DD} in turn requires larger overall minimum power characteristics.

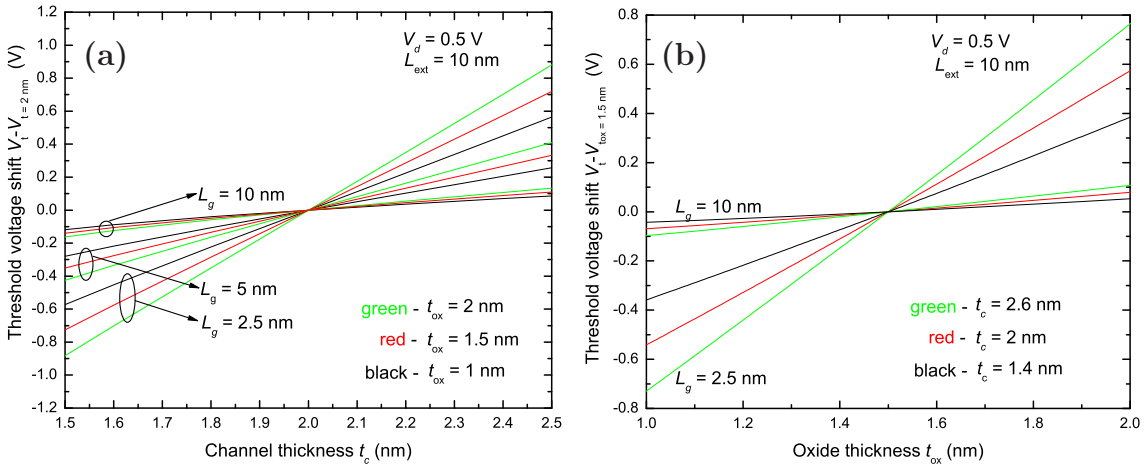

Figure 3.10: Change in the threshold voltage V_t versus (a) channel thickness and (b) oxide thickness for a range of gate lengths.

3.3 Single-Gate Transistor

The primary difference between the single gate structure (Fig. 3.1(b)) and the double-gate configuration is that the back gate is no longer biased at V_g, but maintained at the source level (ground) [100]. To model the ground plane below the buried-oxide (BOX) layer in typical single-gate SOI devices we have extended the back gate to encompass (nearly) the entire doped extension region. We restrict ourselves here to the case of very long extension regions. In this case, the device dynamics are independent of the length of the doped extensions as the density of injected particles is determined by the equilibrium plateau that develops inside the doped thin region (see *e. g.* 3.5). Hence, the device properties are independent of everything to the left of the middle of the source extension and to the right of the drain extension, shown by the red lines in Fig. 3.1(b). This model is an accurate representation of typical single-gate SOI MOSFET structures.

The main effect of the ground plane is that the effective channel capacitance (2.41) is cut roughly in half as the gate field now drops over the channel region. All other aspects of the calculation proceed in the same manner as described above. Namely, we calculate the charge density in the doped regions via equilibrium Eqs. 2.31 and the channel electron density and current density with the one dimensional approximation to the Schrödinger equation.

The device is no longer symmetric along the center of the channel however. To simulate this device we remove the Cauchy boundary condition in the center of the channel and solve the Poisson equation over the entire region with fixed

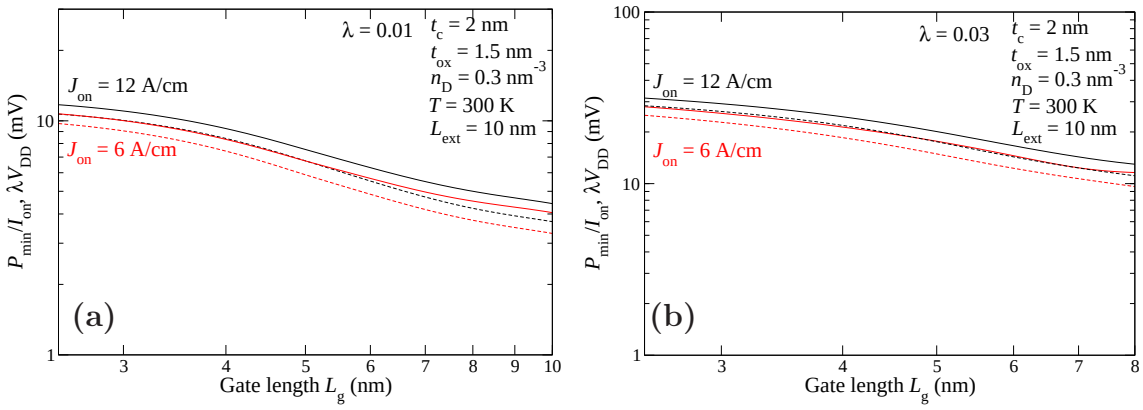

Figure 3.11: Minimum power (solid lines) and minimum drive voltage V_{DD} (dashed lines) versus gate length for two values of effective activity factor: (a) $\lambda = 0.01$,) (b) $\lambda = 0.03$.

boundary conditions at the domain edges.

3.3.1 Device Potential

Typical mid-channel potential profiles are shown in the left column of Fig. 3.12 compared with the profiles of the double gate structure (Fig. 3.1) which provides the same source-drain current J. Panels (a), (b) show constant gate voltage V_g and panels (c), (d) show constant source-drain voltage V_d. In all cases the potential profiles for the single and double gate structures are nearly identical except that the sensitivity of the potential maximum to the gate bias is twice stronger for the DG structure. This result is intuitively expected for long channel devices in the subthreshold region. In this case, the two dimensional field effects may be ignored. The channel potential for the double gate structure remains fixed at V_g while in the single gate structure it drops linearly across the channel with mid-channel value $V_g/2$. It is somewhat surprising however that this relation still holds at higher current densities where channel electrons may screen the gate field.

3.3.2 Performance

The families of $I - V_d$ and subthreshold curves for the single and double gate devices are shown in Fig. 3.13. In the saturation region (3.13(a), 3.13(b)) the current density of the single gate is nearly half that of its double gate counterpart at high gate voltages V_g. This factor of two comes from the fact that it takes nearly double the applied gate bias for the single gate device to

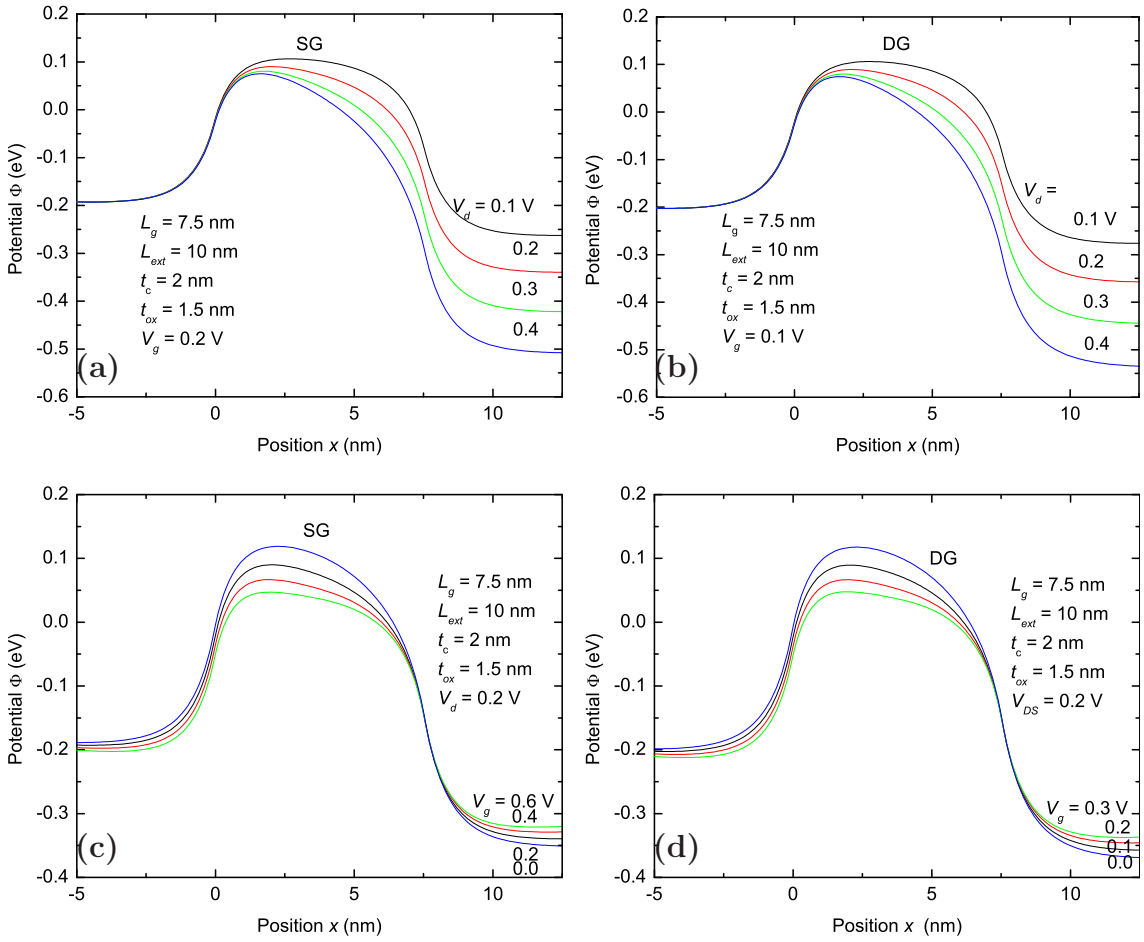

Figure 3.12: Effective potential for single-gate (panels (a),(c)) and double-gate (panels (b),(d)) MOSFETs with $L_g = 7.5$ nm and identical source-drain currents J.

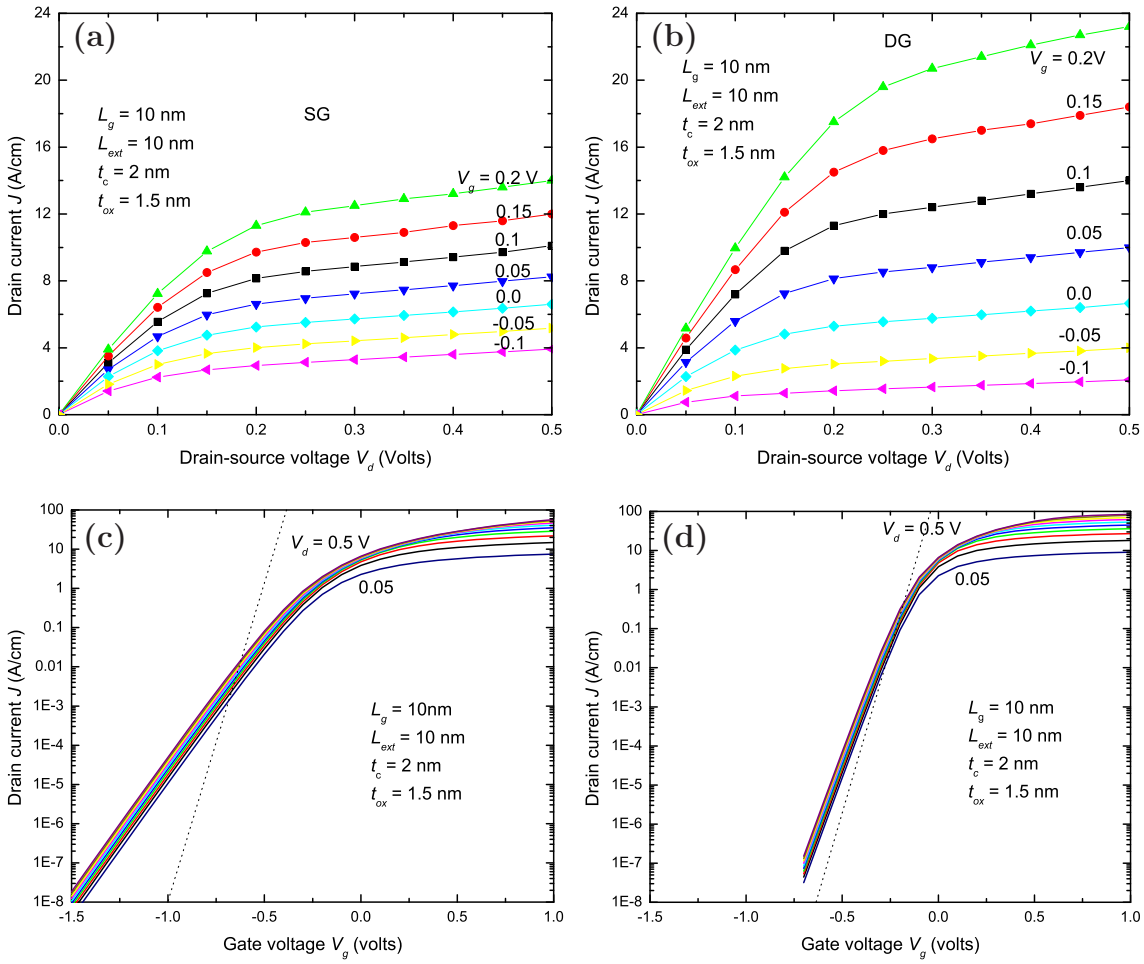

Figure 3.13: Source-drain $I-V_d$ curves of single-gate and double-gate (panels(a), (b) respectively) MOSFETs with channel length $L_g = 10$ nm. Panels (c), (d) show the subthreshold curves for the same devices. The dashed line shows the ideal thermal subthreshold slope.

suppress the potential barrier in the channel. Examination of the subthreshold slope (3.13(c), 3.13(d)) shows that even for a relatively long device, the slope is roughly half the ideal thermal slope of 60 mV/dec.

These results are again encapsulated in the voltage gain G_v. For a fair comparison of the two devices, Fig. 3.14(a) shows the voltage gain for the single and double gate devices at equivalent source-drain current J. The factor of two relationship holds over the entire range of currents. So the double gate transistor may scaled down approximately twice further than the single gate configuration. For example, enforcing $G_v > 1$ shows that the single gate transistor crosses this threshold at $L_g = 5$ nm while, as shown in Fig. 3.8, the double gate device may still provide gain down to $L_g \approx 2.5$ nm. Equivalently, the double gate device will have twice better performance at fixed gate length.

The minimum power operating point versus (top) gate length L_g is shown in Fig. 3.14(a) for both the single and double gate devices. At long gate lengths we again see the power requirements for the double gate structure are twice less than that for the single gate. As the gate length is reduced, the drive voltage V_{DD} required to reach the minimum power point for the single gate device pulls the potential sufficiently high that the hole concentration in the channel from Zener tunneling becomes quite substantial (see section 4.2.4). The net effect of the holes would be a positive background charge decreasing the potential maximum forcing an even higher V_{DD} to close the transistor.

The effect is not included numerically in our calculations but we simulate it by placing a bound at $V_{DD} = 1.4$ V, roughly the point where the band edges begin to cross. The minimum power is evidently at higher V_{DD} than this limit (see *e. g.* Fig. 2.9) and the result is to increase the effective power of the device. This calculation is given by the solid lines in Fig. 3.14 while the dashed lines are the result without consideration for Zener tunneling. The double gate structure does not reach this limit, but the effect on the single gate device is substantial for $L_g \leq 6$ nm. As the gate length approaches $L_g = 3$ nm, the double gate device has nearly an order of magnitude lower operating power and twice the voltage gain over the single gate, providing evidence that the more complex double gate structure (or something similar such as surrounding gates) will be required to reach these ultra-short gate lengths.

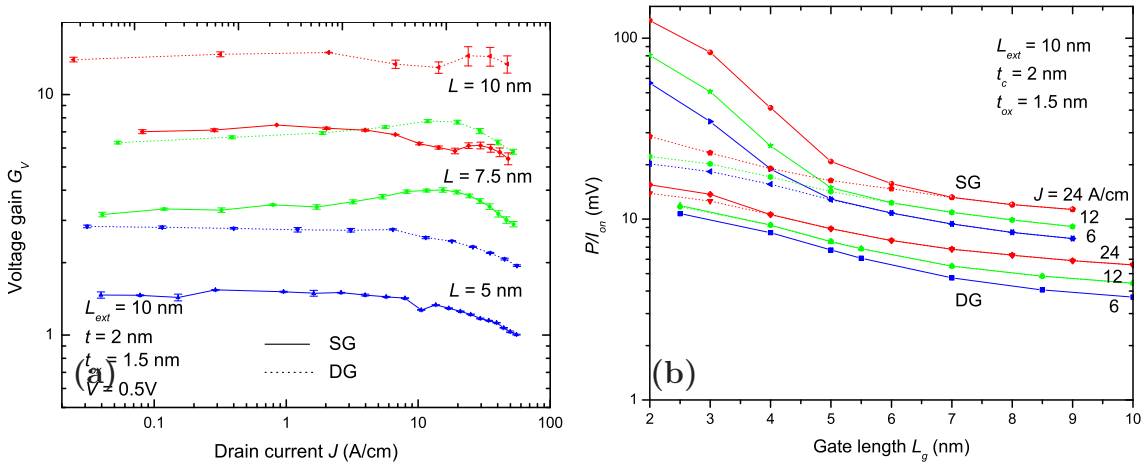

Figure 3.14: Voltage gain G_v for single and double gate devices as a function of drain current density (panel (a)) and minimum power for single and double gate models versus gate length L_g (panel (b)). Dashed lines represent power minimum without effect of Zener tunneling.

Chapter 4

Double-Gate Device with Bulk Electrodes

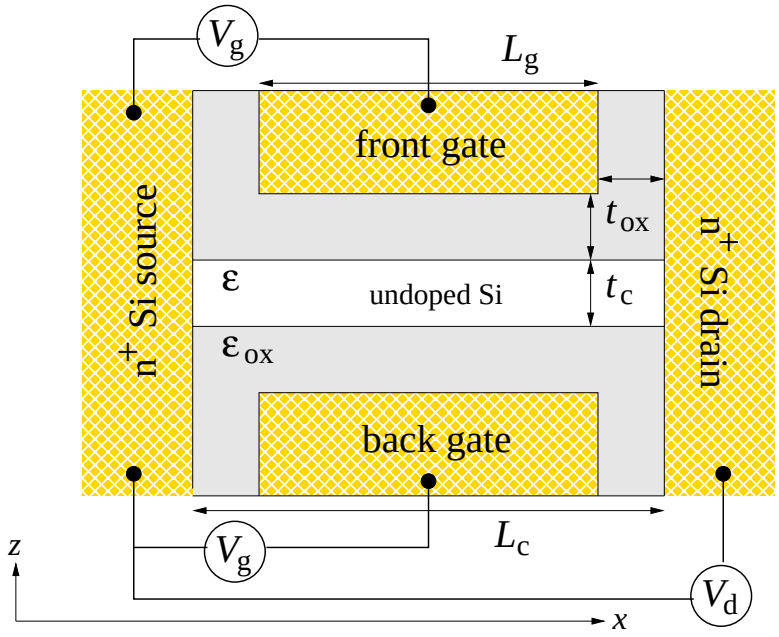

Figure 4.1: Model ballistic FET with bulk electrodes.

The results of chapter 3 show that the DG FET structure is a promising candidate for reaching nanometer scale lengths. The structure may be improved from a circuitry standpoint by removal of the thin extension regions and connecting the channel directly to the bulk regions (Fig. 4.1). Removal of the doped extensions allows for lower overall "bulk-to-bulk" device length

and better thermal scaling. As a simple example, the thin extensions are able to dissipate power

$$P/W = \Delta T \frac{t_c}{L_{\text{ext}}} \kappa_{si}, \tag{4.1}$$

where $\kappa_{si} = 1.5$ W/cm^{o}C is the thermal conductivity of silicon. For a 50 A/cm current density at $V_d = 0.5$ V, this leads to a rise in system temperature

$$\Delta T \approx \frac{L_{\text{ext}}}{t_c} \frac{J_{ON} V_d}{\kappa_{si}} \approx 356 \text{ K}. \tag{4.2}$$

The bulk configuration is much less susceptible to these thermal problems.

In the bulk-electrode model, the channel is longer than the gate electrode $L_c = L_g + 2t_{ox}$. As a result, the gate is not absolutely effective in screening charge near the electrode regions, and charge accumulation can occur. While not ideal, this accumulation does not hinder the performance characteristics and the longer channel delays the onset of direct tunneling allowing for a smaller gate electrode for an equivalent channel length as compared to the devices with extensions.

As a first approach, we may assume that the sharp transition could be approximated by a smooth transition and thus the particle enter the channel adiabatically [101]. In the adiabatic limit, the electron obtains the confinement energy $E_{z,1}$ without scattering and we may use the results of sections 3.1.1 and 3.1.2 within the raised channel potential

$$\bar{\Phi}(x) \rightarrow \bar{\Phi}(x) + E_{z,1}. \tag{4.3}$$

Many results were obtained with this model [16, 49, 53], but the 1-D Schrödinger equation is not strictly valid for this device. The phase-space mismatch between the channel and drain make back-scattering from this interface negligible [97], but back-scattering at the source interface may be substantial, an effect ignored by the adiabatic assumption. Additionally, near the electrode interfaces, the higher order wave function modes will have their greatest impact, exactly where the charge accumulation which dominates transport develops. So an accurate description of the device requires a full 2-D solution of the Schrödinger equation in the channel. Additionally, to enable discussion of wider channel devices, we include the effect of all valley contributions (see Fig. 2.7) to the electron and current densities.

4.1 2-D Schrödinger Solution

While a numerical solution of the Schrödinger equation (with the appropriate "open" boundary conditions deep inside the electrodes, and the first-type conditions at the channel/oxide interface) on a 2D direct-space grid is conceptually the easiest approach, it requires the inversion of a sparse matrix of the size $N_{tot} = N_x \times N_z$, where N_x, N_z are the grid sizes in the $\hat{x}$ and $\hat{z}$ directions (Fig. 4.1), respectively. Such a solution requires as many as roughly $\mathcal{O}(N_{tot}^2)$ operations [77]. The necessity of wavefunction integration over energies of all incident electron waves, within a substantial energy interval, to obtain the full current and charge density distribution further exasperates the necessary computing resources. On the other hand, the full Fourier transform to the 2D momentum space would be highly unnatural, due to the hump-like nature of the effective potential profile in the $\hat{x}$ direction.

This is why, inspired by the renowned analysis of the transversal quantization effects on transport, carried out by Szafer and Stone [102], we have opted for a mixed approach in which the electron wavefunctions in the channel are Fourier-expanded in the $\hat{z}$ direction, while the coefficients of this expansion are computed as explicit functions of x. This approach results in a tri-diagonal matrix which can be solved with $\mathcal{O}(N_{tot})$ operations, for practicable accuracy giving a factor of 100 or so advantage in computation resources over a 2D real-space method. In fact, a direct comparison of two slightly different methods based on the real and mode space solutions of the "Non-equilibrium Green's function" (NEGF) equations for a DG-FET with thin doped extensions and a channel body thickness of 1.5 nm [47] demonstrated the a mode space approach within the NEGF was over 135 times faster than the comparable real-space implementation at a single bias point. However, the NEGF mode space approach did not include the coupling between channel sub-bands modes and thus was only an approximate solution to the full wavefunction. This decoupling of the modes is equivalent to the δ-$\bar{A}$ approximation discussed in appendix A.4.2.

We derive here a full 2-D mixed mode-space solution, including sub-band coupling,to describe coherent transport through the narrow ballistic FET channel.

To solve the full two dimensional Schrodinger equation,

$$\left[-\frac{\hbar^2}{2m_x}\frac{\partial}{\partial x} - \frac{\hbar^2}{2m_z}\frac{\partial}{\partial z} + \Phi(x,z)\right]\Psi(x,z) = (E - E_y)\Psi(x,z), \tag{4.4}$$

we make the usual assumption that the electrons scattered from the transistor back into bulk electrodes dephase and thermalize before their possible return

to the channel. This assumption is fully justified if the mean free path of electrons in the electrodes l is much longer than the channel thickness t_c, while the electron-phonon scattering length $l_{\rm ph}$ is not too much longer than l. For the electrons making the largest contribution into the net current, and the parameters used in our work, $l \approx 10$ nm, and $l_{\rm ph} \approx 50$ nm [97, 103], so that these conditions are relatively well satisfied. Due to this fact, the net current and electric charge densities may be calculated as sums of incoherent contributions by wavefunctions corresponding to plane waves incident on the channel, at various angles, from the drain and source.

In the same manner as Eq. (3.6) we may factor out the y component of the wavefunction in the (uniform) device width. The solution to Eq. (4.4) for the energy eigenfunction in the source, drain and channel regions may be presented as

$$\Psi(x,y,z) = e^{{\rm i}k_y y} \times \begin{cases} \Psi_s(x,z), & x \le 0, \\ \Psi(x,z), & 0 \le x \le L_c, \\ \Psi_d(x,z), & x \ge L_c. \end{cases} \tag{4.5}$$

Again, k_y is conserved throughout the device due to uniformity and assuming a quadratic dispersion relation related to the energy in that direction by

$$E_y = \frac{\hbar^2 k_y^2}{2m_y}. \tag{4.6}$$

Following the mixed-momentum space approach of Szafer and Stone [102], the solution for the energy eigenfunctions in the electrodes may be presented as a series of plane-wave solutions

$$\begin{aligned} \Psi_s(x,z) &= \chi_w(z) e^{{\rm i}k_w x} + \sum_{w'} r_{ww'} \chi_{w'}(z) e^{-{\rm i}k_{w'} x}, \\ \Psi_d(x,z) &= \sum_{\nu} F_{\nu w} e^{{\rm i}k_\nu (x - L_c)} \chi_\nu(z), \end{aligned} \tag{4.7}$$

where

$$\chi_w(z) = \left(\frac{2}{t_B}\right)^{1/2} \times \begin{cases} \cos(q_w z), & w = 1,3,5,\ldots, \\ \sin(q_w z), & w = 2,4,6,\ldots, \end{cases} \tag{4.8}$$

are the standing wave eigenfunctions in the z direction with wavevector

$$q_w \equiv \frac{w\pi}{t_B}. \tag{4.9}$$

The scattering amplitudes $r_{w,w'}$ and $F_{w\nu}$ describe the plane wave compo-

nents, respectively, reflected from the channel back to the source and transmitted though the channel into the drain, normalized to the amplitude of the incident wave. For a region with a constant potential, such a plane-wave expansion is the exact general solution to the Schrödinger equation, and is thus more exact than the 1D approximation (which essentially corresponds to taking into account just one reflected and one transferred wave). In our case this is still a certain approximation, because of some penetration of the electric field into source and drain - see Fig. 4.3 below. However, as these figures show, due to high electrode doping the field is screened very fast in the electrodes, so that the maximal electrostatic potential of the penetrated field is just a few percent of the voltages applied to the device. We believe that the effect of this approximation on the final results are virtually negligible.

Within the quadratic dispersion law approximation, and with the electric potential of the source taken for the reference, the total energy of eigenfunction (4.7) is

$$E = \frac{\hbar^2 k_w^2}{2m_x} + E_y + \frac{\hbar^2 q_w^2}{2m_z}. \tag{4.10}$$

For the waves incident from the drain (which may give a substantial contribution to the net current and charge, especially at low source-drain voltages V_d), we use the expression similar to Eq. (4.7), with the characteristic equation

$$E = \frac{\hbar^2 k_\nu^2}{2m_x} + E_y + \frac{\hbar^2 q_\nu^2}{2m_z} - eV_d. \tag{4.11}$$

The eigenfunctions in the channel may be also expressed through series expansions

$$\Psi(x,z) = \sum_n \psi_n(x)\varphi_n(z), \text{ for } 0 \le x \le L_c, \tag{4.12}$$

where eigenfunctions $\varphi_n(z)$ are defined similarly to Eq. (4.8),

$$\varphi_n(z) = \left(\frac{2}{t_c}\right)^{1/2} \times \begin{cases} \cos\left(n\pi z/t_c\right), & n = 1,3,5,\ldots, \\ \sin\left(n\pi z/t_c\right), & n = 2,4,6,\ldots, \end{cases} \tag{4.13}$$

but with the replacement of the electrode thickness t_B for the much smaller thickness of the channel $t_c \ll t_B$. Plugging (4.5), (4.12) into the 3D Schröinger equation, and using the orthonormality of functions $\varphi_n(z)$, we readily arrive at an effective 1D Schrödinger equation for electrons in the n-th subband:

$$\left[-\frac{\hbar^2}{2m_x}\frac{d^2}{dx^2} - E_x - e\bar{\Phi}_n(x)\right]\psi_n(x) = 0. \tag{4.14}$$

Here the effective 1D potential

$$\bar{\Phi}_n(x) \equiv \sum_{n'} \int \varphi_n(z)\Phi(x,z)\varphi_{n'}\, dz, \tag{4.15}$$

while the effective 1D energy E_x is related to the total energy E as

$$E = E_x + E_y + E_{z,n}, \tag{4.16}$$

where the last term presents the lateral quantum confinement,

$$E_{z,n} = \frac{\hbar^2\pi^2 n^2}{2m_z t_c^2}. \tag{4.17}$$

For numerical solution, it is convenient to present the general solution of Eq. (4.14) in the following form:

$$\psi_n(x) = C_n f_n(x) + D_n g_n(x), \tag{4.18}$$

where C_n, D_n are undetermined weights, while $f_n(x)$, $g_n(x)$ are the linearly independent, particular solutions of the same equation. At the source-to-channel interface ($x = 0$), we enforce the continuity of the wavefunction. According to the first of Eqs. (4.7) and Eq. (4.12), the condition has the form

$$\chi_w(z) + \sum_{w'} r_{ww'}\chi_{w'}(z) = \sum_n \left(C_n f_n(0) + D_n g_n(0)\right)\varphi_n(z). \tag{4.19}$$

Multiplying both sides of that equation by $\chi_{w''}(z)$ and integrating the result over the channel thickness, we get a set of linear equations

$$r_{w,w'} = \sum_n \left[C_n f_n(0) + D_n g_n(0)\right] a_{nw'} - \delta_{w,w'}, \tag{4.20}$$

where coefficients

$$a_{nw} \equiv \int_{-t_c/2}^{t_c/2} \chi_w(z)\varphi_n(z)\, dz, \tag{4.21}$$

represent the strength of the overlap of transverse wavefunctions in the channel and electrode. Using the confined state wavefunctions (4.8) and (4.13), the

integral is calculated explicitly as

$$a_{nw} = \begin{cases} (-1)^{n/2} \dfrac{4nt_B^{3/2}t_c^{1/2}}{\pi(w^2t_c^2 - n^2t_B^2)} \sin\left(\dfrac{w\pi t_c}{2t_B}\right) & w, n \text{ even}, \\ (-1)^{(n+1)/2} \dfrac{4nt_B^{3/2}t_c^{1/2}}{\pi(w^2t_c^2 - n^2t_B^2)} \cos\left(\dfrac{w\pi t_c}{2t_B}\right) & w, n \text{ odd}, \\ \sqrt{\dfrac{t_c}{t_B}} & w^2t_c^2 = n^2t_B^2, \\ & (w \mod 2) = (n \mod 2), \\ 0 & \text{else}. \end{cases} \tag{4.22}$$

When n, w are odd, a_{nw}^2 has the following limiting expressions

$$a_{nw}^2 \approx \begin{cases} \dfrac{16}{(n\pi)^2}\left(\dfrac{t_c}{t_B}\right) & : w\frac{t_c}{t_B} \ll 1, \\ \dfrac{16n^2\left(\frac{t_c}{t_B}\right)}{\pi^2(w^2(t_c/t_B)^2 - n^2)^2} & : w\frac{t_c}{t_B} \gg 1, \end{cases} \tag{4.23}$$

and when n, w are even,

$$a_{nw}^2 \approx \begin{cases} \dfrac{4w^2}{n^2}\left(\dfrac{t_c}{t_B}\right) & : w\frac{t_c}{t_B} \ll 1, \\ \dfrac{16}{(n\pi)^2}\left(\dfrac{t_c}{t_B}\right) & : w\frac{t_c}{t_B} \approx \pi/2, \\ \dfrac{16n^2\left(\frac{t_c}{t_B}\right)}{\pi^2(w^2(t_c/t_B)^2 - n^2)^2} & : w\frac{t_c}{t_B} \gg 1. \end{cases} \tag{4.24}$$

From the continuity of the wavefunction derivative we get another relation:

$$\mathrm{i}k_w\chi_w(z) - \mathrm{i}\sum_{w'} k_{w'}r_{ww'}\chi_{w'}(z) = \sum_n \frac{\partial}{\partial x}\psi_n(0)\varphi_n(z). \tag{4.25}$$

Multiplying by $\varphi_{n'}(z)$ and integrating the result, and combining it with Eq.

(4.20), we get

$$2\mathrm{i}k_w a_{nw} - \mathrm{i}\sum_m \bar{A}^w_{nm}\left[C_m f_m(0) + D_m g_m(0)\right] = C_n f'_n(0) + D_n g'_n(0), \qquad (4.26)$$

where the prime represents differentiation over x, and kernel elements

$$\bar{A}^w_{nm} \equiv \sum_{w=1}^{\infty} a_{nw} k_w a_{mw}, \qquad (4.27)$$

characterize the coupling between sub-band modes.

Absolutely similar calculations at the channel-drain interface yield conditions

$$F_{\nu w} = \sum_m \left[C_m f_m(L_c) + D_m g_m(L_c)\right] a_{m\nu}, \qquad (4.28)$$

$$C_n f'_n(L_c) + D_n g'_n(L_c) = \mathrm{i}\sum_m \bar{A}^\nu_{nm}\left[C_m f_m(L_c) + D_m g_m(L_c)\right]. \qquad (4.29)$$

Here, coefficients $\bar{A}^\nu$ are defined similarly to Eq. (4.27), with the replacement of the electron wavevector k_w in the source by k_ν in the drain, and the sum taken over the drain modes numbered with index ν.

Relations (4.20), (4.26), (4.28), (4.29) form a full set of linear relations for the Fourier expansion amplitudes, which completely determine coefficients $r_{ww'}$, $F_{\nu w}$, C_n, and D_n, and, together with functions $f_n(x)$, $g_n(x)$, determine all eigenfunctions $\Psi(x,z)$.

4.1.1 Channel Electron Density

The total number of electrons in the channel, arriving from the source, may be found by the summation over all incident waves:

$$N(x,y,z) = g_s g_v \sum_{\boldsymbol{k}} |\Psi(x,z)|^2 f(E), \qquad (4.30)$$

where $\Psi(x,z)$ is assumed normalized to the amplitude of the incident state wavefunction. For a large number of incident states, the summation becomes integral

$$N(x,y,z) = \frac{g_s g_v L_B t_B W}{(2\pi)^3} \int\limits_{k_x>0} d^3k\, f(E)\, |\Psi(x,z)|^2 . \qquad (4.31)$$

The three dimensional electron density is given by

$$n_{3D}(x,y,z) = \frac{g_s g_v t_B}{(2\pi)^3} \int\limits_{k_x>0} d^3k\, f(E)\, |\Psi(x,z)|^2 \,. \tag{4.32}$$

Using relation

$$dk_y = \frac{1}{2\hbar}\left(\frac{2m_y}{E_y}\right)^{1/2} dE_y, \tag{4.33}$$

and Fermi-Dirac integral (2.21), the integral becomes

$$n_{3D}(x,y,z) = \frac{g_s g_v t_B \sqrt{2m_y}}{(2\pi)^3\hbar} \int\limits_{k_x>0} d^2k\ |\Psi(x,z)|^2\, \mathcal{F}_{-1/2}\left((\mu_{\rm F} - E_{x,z})/T\right), \tag{4.34}$$

where

$$E_{x,z} \equiv E_x + E_{z,n}. \tag{4.35}$$

When the bulk electrode thickness t_B is large we can introduce angle θ such that

$$\begin{aligned} k_w &= \left(\sqrt{2m_x E_{x,z}}/\hbar\right)\cos\theta, \\ q_w &= \left(\sqrt{2m_z E_{x,z}}/\hbar\right)\sin\theta. \end{aligned}$$

For arbitrary effective masses m_x, m_y, the differential area is given by (*c. f.* 2.19)

$$d^2k = \frac{\sqrt{m_x m_z}}{\hbar^2} dE_{x,z}\, d\theta, \tag{4.36}$$

and using dimensionless energy variables $\epsilon \equiv E/T$, the expression for the three dimensional electron density is written

$$\begin{aligned} n_{3D}(x,y,z) = \frac{g_s g_v (2m_x m_y m_z)^{1/2} T^{3/2}}{(2\pi\hbar)^3} &\int\limits_0^\infty d\epsilon_{x,z} \mathcal{F}_{-1/2}(\epsilon_F - \epsilon_{x,z}) \\ &\times \int\limits_{-\pi/2}^{\pi/2} d\theta\ |\Psi(x,z)|^2 \,. \end{aligned} \tag{4.37}$$

In our case, however, it is more convenient to express the integral over θ as a direct summation over modes w. Such a summation is convenient because it allows the charge and current densities to be expressed explicitly in terms

of coupling strength $\bar{A}$, which only needs to be calculated once, before any self-consistent iteration begins. To calculate the integration over the incident angle of the wavefunction, we use the relations

$$\int_{-\pi/2}^{\pi/2} d\theta = 2 \int_{0}^{\pi/2} d\left[\sin(\theta)\right] \left[\cos(\theta)\right]^{-1}, \tag{4.38}$$

and

$$\begin{aligned} \sin(\theta) &= \frac{\hbar\pi}{t_B\sqrt{2m_z E_{x,z}}} w, \\ \cos(\theta) &= \frac{\hbar}{k_w}\sqrt{2m_x E_{x,z}}, \\ d\left(\sin(\theta)\right) &= \frac{\pi\hbar}{t_B\sqrt{2m_z E_{x,z}}} dw, \end{aligned} \tag{4.39}$$

the integral over θ of the wavefunction may be expressed as a summation over incident states

$$\int_{-\pi/2}^{\pi/2} \left|\Psi(x,z)\right|^2 d\theta = \frac{2\pi}{t_B}\sqrt{\frac{m_x}{m_z}} \sum_w k_w^{-1} \left|\Psi(x,z)\right|^2. \tag{4.40}$$

Plugging this relation into (4.37), we arrive at the expression for the electron density

$$n_{3D}(x,y,z) = \frac{\left(2m_x^2 m_y\right)^{1/2} T^{3/2}}{\pi^2\hbar^3} \int_0^\infty d\epsilon_{x,z} \mathcal{F}_{-1/2}(\epsilon_F - \epsilon_{x,z}) \left[\sum_{w=1}^{\infty} k_w^{-1} \left|\Psi(x,z)\right|^2\right], \tag{4.41}$$

where we have used degeneracies $g_s = 2$, $g_v = 2$.

The total electron density in the channel is then calculated as a sum of Eq. (4.41) and a similar expression for electrons incident from the drain with the shift energy of reference by eV_d.

4.1.2 Current Density

Taking into account the full two dimensional solution for the wavefunction, the summation for the device current (2.34) may be written

$$I = \sum_{\boldsymbol{k}} I_k = \frac{W t_B L_B}{(2\pi)^3} \int d^3k I_k. \tag{4.42}$$

Noting that the transmission coefficient is independent of the wavevector component k_y, and evaluating the current density we write

$$J = I/W = e\frac{g_s g_v \hbar t_B}{m_x (2\pi)^3} \int\limits_{k_x>0} d^2k k_x \mathcal{D}(\boldsymbol{k}_{x,z}) \int\limits_{-\infty}^{\infty} dk_y f(E). \tag{4.43}$$

Converting dk_y to energy units yields

$$J = e\frac{g_s g_v t_B (2m_y)^{1/2}}{m_x (2\pi)^3} \int\limits_{k_x>0} d^2k k_x \mathcal{D}(\boldsymbol{k}_{x,z}) \int\limits_{0}^{\infty} dE_y E_y^{-1/2} f(E). \tag{4.44}$$

Again using the differential area (2.19), the current density may be written

$$\begin{aligned} J = e\frac{g_s g_v (2m_y)^{1/2}}{4\pi^2\hbar^2} \int\limits_0^\infty dE_{x,z} \mathcal{F}_{-1/2}\left[(\mu_{\mathrm{F}} - E_{x,z})/T\right] \\ \times \left[\frac{t_B\sqrt{2m_z E_{x,z}}}{\pi\hbar} \int\limits_0^1 d(\sin\theta)\mathcal{D}(E_{x,z},\theta)\right]. \end{aligned} \tag{4.45}$$

Again, it will be more convenient to represent the integration in square brackets as a sum over incident modes. The transmission probability for incident mode w is the sum over all possible transmitted modes ν,

$$\mathcal{D}_w(E_{x,z}) = \sum_\nu \left|\frac{k_\nu}{k_w}\right| |F_{w,\nu}|^2, \tag{4.46}$$

and using relation (4.39) the total probability of transmission for a particle of energy $E_{x,z}$ is found to be

$$\sum_w \mathcal{D}_w(E_{x,z}) = \frac{t_B\sqrt{2m_z E}}{\pi\hbar} \int\limits_0^1 d(\sin\theta)\mathcal{D}(E_{x,z}). \tag{4.47}$$

Hence, the term in square brackets of Eq. (4.45) is simply the total transmission coefficient (4.47) and the total current density may be written

$$J = \frac{J_0}{\pi} \int_0^\infty d\epsilon_{x,z} \mathcal{D}(\epsilon_{x,z} T) \left[\mathcal{F}_{-1/2}(\epsilon_F - \epsilon_{x,z}) - \mathcal{F}_{-1/2}(\epsilon_F - \nu_d - \epsilon_{x,z}) \right], \quad (4.48)$$

where $\nu_d \equiv eV_d/T$, we have used degeneracy values $g_s = 2$, $g_v = 2$ and

$$J_0 \equiv e \frac{\sqrt{2m_y} T^{3/2}}{\pi \hbar^2}. \quad (4.49)$$

The first term in the square brackets represents the total left-to-right moving current for particles incident from the source. The net source-to-drain current is balanced by the motion of particles moving right-to-left from drain, represented by the second term.

4.1.3 Evaluation of the Wavefunction

As in section 3.1, $f_n(x)$ and $g_n(x)$ are any two linearly independent solutions of (4.14). For notational simplicity we will use the conventions

$$\begin{array}{rclcrcl} f_m(0) & = & f_{m0} & & f_m(L_c) & = & f_{mL}, \\ f'_m(0) & = & f'_{m0} & & f'_m(L_c) & = & f'_{mL}, \\ g_m(0) & = & g_{m0} & & g_m(L_c) & = & g_{mL} \\ g'_m(0) & = & g'_{m0} & & g'_m(L_c) & = & g'_{mL}, \end{array} \quad (4.50)$$

It is most convenient to present the solution of linear system (4.20), (4.26), (4.28), (4.29) in matrix notation. Defining

$$\begin{array}{rcl} \bar{M}_1 & \equiv & \mathrm{i}\bar{A}^w_{nm} g_{m0} + \delta_{nm} g'_{m0}, \\ \bar{M}_2 & \equiv & \mathrm{i}\bar{A}^w_{nm} f_{m0} + \delta_{nm} f'_{m0}, \end{array} \quad (4.51)$$

the solution to Eqs. (4.20), (4.26) may be rewritten compactly as

$$|a\rangle = \bar{M}_2 |C\rangle + \bar{M}_1 |D\rangle, \quad (4.52)$$

where $|C\rangle$ and $|D\rangle$ are vectors over unknown channel weights C_n, D_n at fixed w, and $|a\rangle$ is a vector over the same n states with elements

$$|a\rangle_n \equiv 2\mathrm{i}k_w a_{nw}. \quad (4.53)$$

The solution of equations (4.28), (4.29) yield relations

$$\bar{M}_3 \left|C\right\rangle = \bar{M}_4 \left|D\right\rangle, \tag{4.54}$$

where we have defined matrices

$$\begin{array}{rcl} \bar{M}_3^{nm} & \equiv & \mathrm{i}\bar{A}_{nm}^{\nu} f_{mL} - \delta_{nm} f'_{mL}, \\ \bar{M}_4^{nm} & \equiv & -\mathrm{i}\bar{A}_{nm}^{\nu} g_{mL} + \delta_{nm} g'_{mL}. \end{array} \tag{4.55}$$

Results (4.52) and (4.54) solved together yield expressions for wavefunction coefficient vectors $\left|C\right\rangle$, $\left|D\right\rangle$

$$\begin{array}{rcl} \left|C\right\rangle & = & \left[\bar{M}_2 + \bar{M}_1 \bar{M}_4^{-1} \bar{M}_3\right]^{-1} \cdot \left|a\right\rangle, \\ \left|D\right\rangle & = & \bar{M}_4^{-1} \bar{M}_3 \left[\bar{M}_2 + \bar{M}_1 \bar{M}_4^{-1} \bar{M}_3\right]^{-1} \cdot \left|a\right\rangle. \end{array} \tag{4.56}$$

Because matrix $\bar{M}_3$ has diagonal elements shifted by $f'_{mL} \ll 1$, the calculation becomes numerically unstable for the alternate solution using $\bar{M}_3^{-1}\bar{M}_4$. For computational convenience we define coefficient matrices

$$\begin{array}{rcl} \bar{M}_C & \equiv & \left[\bar{M}_2 + \bar{M}_1 \bar{M}_4^{-1} \bar{M}_3\right]^{-1}, \\ \bar{M}_D & \equiv & \bar{M}_4^{-1} \bar{M}_3 \bar{M}_C, \end{array} \tag{4.57}$$

then the weights $\left|C\right\rangle$ and $\left|D\right\rangle$ may be expressed as simply

$$\begin{array}{rcl} \left|C\right\rangle & = & \bar{M}_C \left|a\right\rangle, \\ \left|D\right\rangle & = & \bar{M}_D \left|a\right\rangle. \end{array} \tag{4.58}$$

Notice that $\bar{M}_C$ and $\bar{M}_D$ are functions in terms of t_c, t_B, E, L_c, and applied potential V_d but **not** the specific incoming or outgoing states w, ν. Using result (4.58) and relation (4.28), we find the transmitted amplitude to be

$$F_{\nu w} = \left[\left\langle a_f\right| \bar{M}_C + \left\langle a_g\right| \bar{M}_D\right] \cdot \left|a\right\rangle, \tag{4.59}$$

where $\left\langle a_f\right|$, $\left\langle a_g\right|$ are row vectors over m with elements definition

$$\begin{array}{rcl} \left|a_f\right\rangle_m & \equiv & a_{m\nu} f_{mL}, \\ \left|a_g\right\rangle_m & \equiv & a_{m\nu} g_{mL}. \end{array} \tag{4.60}$$

4.1.3.1 Unit Boundary Conditions

As discussed in section 3.1.3, the most appropriate choice of boundary conditions for the numerical solution of $f_n(x)$, $g_n(x)$ is

$$\begin{array}{rclcrcl} f_n(0) & = & 1 & \Big| & f_n(L_c) & = & 0, \\ g_n(0) & = & 0 & \Big| & g_n(L_c) & = & 1. \end{array} \tag{4.61}$$

The matrix expressions (4.51), (4.55) are then simplified

$$\begin{array}{rcl} \bar{M}_1 & = & \delta_{nm} g'_{m0}, \\ \bar{M}_2 & = & \mathrm{i}\bar{A}^w_{nm} + \delta_{nm} f'_{m0}, \\ \bar{M}_3 & = & -\delta_{nm} f'_{mL}, \\ \bar{M}_4 & = & -\mathrm{i}\bar{A}^\nu_{nm} + \delta_{nm} g'_{mL}. \end{array} \tag{4.62}$$

Since $\bar{M}_1$ and $\bar{M}_3$ are now diagonal matrices, the amount of matrix manipulation we have to do in the numerical calculation can be reduced. The expressions for coefficient vectors $|C\rangle$, $|D\rangle$ remains unchanged, but it becomes most convenient to work with matrix $\bar{M}_{43} \equiv \bar{M}_4^{-1}\bar{M}_3$ whose elements are given by

$$\left[\bar{M}_{43}\right]_{nm} = -\left[\bar{M}_4^{-1}\right]_{nm} f'_{mL}, \tag{4.63}$$

and matrix $\bar{M}_{143} \equiv \bar{M}_1\bar{M}_4^{-1}\bar{M}_3$ whose elements are given by

$$\left[\bar{M}_{143}\right]_{nm} = -\left[\bar{M}_4^{-1}\right]_{nm} g'_{n0} f'_{mL}. \tag{4.64}$$

Hence, we may write the coefficient matrices as

$$\begin{array}{rcl} \bar{M}_C & = & \left[\bar{M}_2 + \bar{M}_{143}\right]^{-1}, \\ \bar{M}_D & = & \bar{M}_{43}\bar{M}_C. \end{array} \tag{4.65}$$

with vector solutions (4.58).

As mentioned previously, a major benefit of this method is that both the square modulus of the eigenfunction $\Psi(x,z)$ and transmission probability may be written explicitly in terms of $\bar{A}$, which only needs to be tabulated once. Since $f_{mL} = 0$ by definition, the expression for the transmitted amplitude reduces to

$$F_{\nu w} = 2\mathrm{i}k_w \sum_{nm} \left[\bar{M}_D\right]_{nm} a_{n\nu} a_{mw}, \tag{4.66}$$

and the total transmission co-efficient, Eq. (4.47), becomes

$$\mathcal{D}(E_{x,z}) = 4 \sum_{n,m,n',m'} \bar{\mathcal{K}}^\nu_{nn'} \bar{\mathcal{K}}^w_{mm'} \left[\bar{M}_D\right]_{nm} \left[\bar{M}^*_D\right]_{n'm'}, \tag{4.67}$$

where $\bar{\mathcal{K}}_{nm} \equiv \Re(\bar{A}_{nm})$ is the average incident wavevector over modes w which couples to states n,m [102].

The wavefunction inside the channel is written as

$$\Psi(x,z) = \sum_n \Big(|C\rangle_n f_n(x) + |D\rangle_n g_n(x) \Big) \varphi_n(z), \tag{4.68}$$

with square modulus

$$|\Psi(x,z)|^2 = \sum_{nm} \Big[|C\rangle_n |C\rangle_m^* f_n(x) f_m(x) + |D\rangle_n |D\rangle_m^* g_n(x) g_m(x) + |C\rangle_n |D\rangle_m^* f_n(x) g_m(x) + |D\rangle_n |C\rangle_m^* g_n(x) f_m(x) \Big] \varphi_n(z) \varphi_m(z), \tag{4.69}$$

and the star represents conjugation of the complex variable. For the channel electron density, the relevant summation is the square modulus of the wavefunction normalized to the incident wavevector $\sum_w k_w^{-1} |\Psi(x,z)|^2$. Since $f_n(x)$, $g_n(x)$, and $\varphi_n(z)$ are independent of w, we can calculate the sum *apriori*. The summation over states may then be expressed as

$$\sum_w k_w^{-1} |\Psi(x,z)|^2 = \sum_{nm} \Big[P(C_n, C_m) f_n(x) f_m(x) + P(D_n, D_m) g_n(x) g_m(x) + P(D_n, C_m) f_m(x) g_n(x) + P(C_n, D_m) f_n(x) g_m(x) \Big] \varphi_n(z) \varphi_m(z), \tag{4.70}$$

where $C_n(D_n)$ is the n^{th} element of the $|C\rangle$ ($|D\rangle$) vector. The useful term $P(V_n, V_m)$represents the general summed product of the n-th and (conjugated) m-th elements of vectors $|C\rangle$, $|D\rangle$,

$$P(V_n, V_m) \equiv \sum_w k_w^{-1} V_n V_m^*. \tag{4.71}$$

For arbitrary elements V_n, V_m, this result may be simplified by writing $|C\rangle$ and $|D\rangle$ in terms of their respective matrices (4.65)

$$V_n = 2\mathrm{i} k_w \sum_m \left[\bar{M}_V\right]_{nm} a_{mw}. \tag{4.72}$$

Thus,

$$V_n V_m^* = 4k_w^2 \sum_{ll'} \left[\bar{M}_V\right]_{nl} \left[\bar{M}_V^*\right]_{ml'} a_{lw} a_{l'w}, \tag{4.73}$$

or explicitly in terms of $\bar{A}$,

$$P(V_n, V_m) = 4 \sum_{ll'} \left[\bar{M}_V\right]_{nl} \left[\bar{M}_V^*\right]_{ml'} \bar{\mathcal{K}}_{ll'}^w. \tag{4.74}$$

The reflection term for the incident wavefunction is then found to be

$$r_{ww'} = 2\mathrm{i}k_w \sum_{nm} \bar{M}_C a_{nw'} a_{mw} - \delta_{ww'}, \tag{4.75}$$

and the wavefunctions in the bulk regions are given as

$$\Psi_s(x,z) = 2\mathrm{i}\left[\chi_w(z)\sin(k_w x) + k_w \sum_{n,m} \bar{M}_C a_{mw} \bar{\Lambda}_n^w\right], \tag{4.76}$$

$$\Psi_d(x,z) = 2\mathrm{i}k_w \sum_{nm} \bar{M}_D a_{mw} \bar{\Lambda}_n^\nu, \tag{4.77}$$

where

$$\begin{aligned} \bar{\Lambda}_n^w &\equiv \sum_{w'} \chi_{w'}(z) e^{(-\mathrm{i}k_{w'}x)} a_{nw'}, \\ \bar{\Lambda}_n^\nu &\equiv \sum_{\nu} \chi_\nu(z) e^{[\mathrm{i}k_\nu(x-L_c)]} a_{n\nu}. \end{aligned} \tag{4.78}$$

The normalized sums of the bulk electrode wavefunctions is found to be

$$\begin{aligned} &\sum_w k_w^{-1} \left|\Psi_s(x,z)\right|^2 = \\ &4\left\{\sum_w \left[k_w^{-1}\chi_w^2(z)\sin^2(k_w x) + 2\chi_w(z)\sin(k_w x)\sum_{nm} \Re\left(\left[\bar{M}_C\right]_{nm} a_{mw}\bar{\Lambda}_n^w\right)\right]\right. \\ &\qquad \left. + \sum_{n,m,n',m'} \left[\bar{M}_C\right]_{nm}\left[\bar{M}_C^*\right]_{n'm'} \bar{\mathcal{K}}_{mm'}^w \bar{\Lambda}_n^w \bar{\Lambda}_{n'}^{*,w}\right\}, \end{aligned} \tag{4.79}$$

$$\sum_w k_w^{-1} \left|\Psi_d(x,z)\right|^2 = 4 \sum_{n,m,n',m'} \left[\bar{M}_D\right]_{nm}\left[\bar{M}_D^*\right]_{n'm'} \bar{\mathcal{K}}_{m,m'}^w \bar{\Lambda}_n^\nu \bar{\Lambda}_{n'}^{*,\nu}. \tag{4.80}$$

The matrix $\bar{A}$ only depends on physical parameters t_c, t_B, which can not change through the self-consistent iteration, and electron energy E. Hence, we may tabulate the matrix values over E before the calculation begins and the full two dimensional wavefunction solution can be calculated very rapidly.

Unfortunately, direct calculation of (4.27) converges very slowly at high w. However, the wavevector may be expanded exactly in this region to provide a rapidly converging expression.

4.1.4 Numerical Evaluation of $\bar{A}$

Because summation (4.27) tends to converge very slowly at high w terms, we need a more efficient method to evaluate it. This first step is to recognize that since $a_{nw} = 0$ if $(n \mod 2) \neq (w \mod 2)$, then

$$\bar{A}_{nm} = 0 \text{ if } (n \mod 2) \neq (m \mod 2). \tag{4.81}$$

Hence, plugging in the explicit expression for a_{nw} we find in terms of ratio $r \equiv t_B/t_c$,

$$\bar{A}_{nm} = \begin{cases} 0 & (n \mod 2) \neq (m \mod 2), \\ \mathrm{sgn}(n)\mathrm{sgn}(m) \sum\limits_{w=1,3\ldots} k_w c_w \cos^2\left(\dfrac{w\pi}{2r}\right) & (n,m) = \text{odd}, \\ \mathrm{sgn}(n)\mathrm{sgn}(m) \sum\limits_{w=2,4\ldots} k_w c_w \sin^2\left(\dfrac{w\pi}{2r}\right) & (n,m) = \text{even}, \end{cases} \tag{4.82}$$

where

$$\mathrm{sgn}(n) = \begin{cases} (-1)^{n/2} & n = \text{even}, \\ (-1)^{(n+1)/2} & n = \text{odd}, \end{cases} \tag{4.83}$$

and

$$c_w = \begin{cases} 1/r & w = nr \text{ and } w = mr, \\ \dfrac{4mr}{\pi(w^2 - m^2r^2)} & w = nr, \\ \dfrac{4nr}{\pi(w^2 - n^2r^2)} & w = mr, \\ \dfrac{16nmr^3}{\pi^2(w^2 - m^2r^2)(w^2 - n^2r^2)} & \text{else}. \end{cases} \tag{4.84}$$

The wavevector k_w can be expressed as

$$k_w = \sqrt{\zeta - \zeta_w}, \tag{4.85}$$

where $\zeta = 2m_x E/\hbar^2$ and $\zeta_w = w^2\pi^2 m_x/m_z t_B^2$. When $\zeta_w > \zeta$ then the wavevec-

tor can be expressed

$$k_w = \mathrm{i}\zeta_w^{1/2}\sqrt{1-\beta_w}, \tag{4.86}$$

with terms

$$\begin{aligned} \beta_w &\equiv \alpha E w^{-2}, \\ \alpha &\equiv \frac{2m_z t_B^2}{\pi^2\hbar^2}. \end{aligned} \tag{4.87}$$

When $\zeta_w \gg \zeta$, the wavevector can be expanded in the Taylor series

$$k_w = \mathrm{i}\zeta_w^{1/2}\sum_j c_j \alpha^j E^j w^{-2j}, \tag{4.88}$$

where

$$c_j \equiv \frac{\left(\frac{1}{2}-j-1\right)!}{j!}. \tag{4.89}$$

Then the high w terms of the summation can be expressed

$$\sum_{w=\text{high cutoff}}^{\infty} a_{nw}k_w a_{mw} = \mathrm{i}\sum_j c_j\alpha^j E^j \sum_{w=\text{high cutoff}} \zeta_w^{1/2} a_{nw}a_{mw}w^{-2j}. \tag{4.90}$$

The utility of this expression is that it converges relatively rapidly in j terms and the second part is completely independent of E. Hence, this part only needs to be tabulated once and plugged into the summation over j at a given energy E. Other useful approximations for kernel $\bar{A}$ including those of the original authors, are provided in appendix A.4.

4.2 Device Characteristics

4.2.1 Potential

Two typical potential distributions for the bulk electrode device are shown in Fig. 4.2 for two values of the applied gate voltage V_g. In the on-state ($V_g = 0.1$ V), the potential "hump" near the source electrode, a defining characteristic for this device, is clearly visible. This hump is the result of screening of the gate field by the source and drain electrodes allowing for the accumulation of uncompensated charge.

The potential along the middle of the device is shown by the black lines in Fig. 4.3 for a long gate, $L_g = 10$ nm (panel (a)) and short gate, $L_g = 2.5$ nm (panel (b)) device. The solid lines are the results calculated with the full 2-D solution and the dashed lines are the result calculated with the 1-D

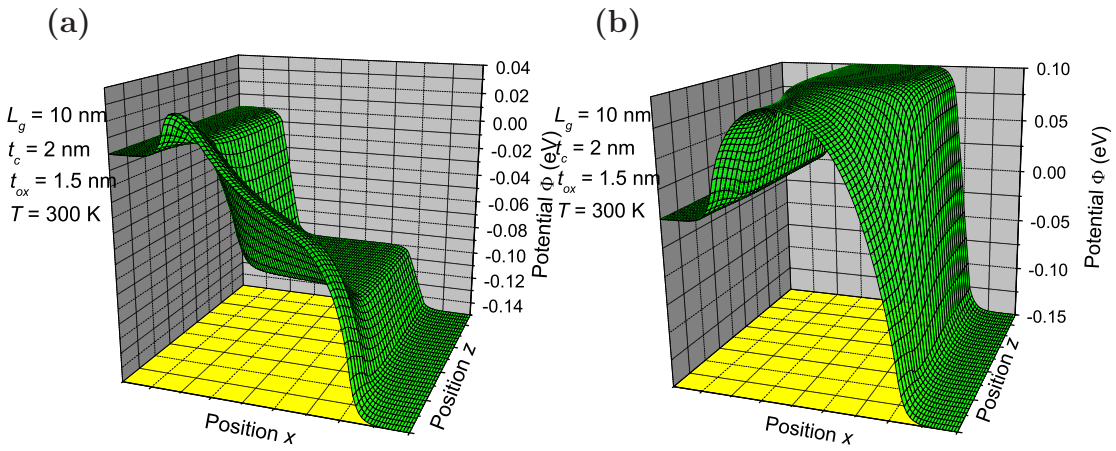

Figure 4.2: Sample 2D potential profiles for $L_g = 10$ nm, $t_{ox} = 1.5$ nm, $t_c = 2$ nm, $V_d = 0.2$ V device with (a) $V_g = 0.1$ V, (b) $V_g = -0.1$ V.

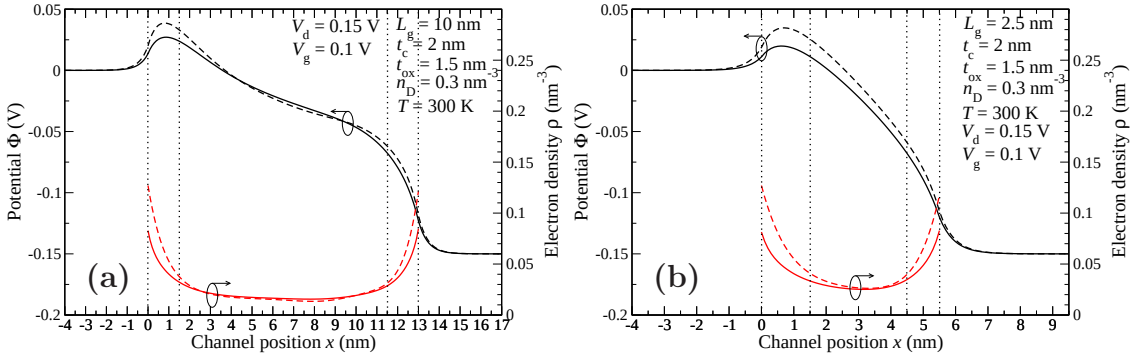

Figure 4.3: Mid-channel potential (black lines) and electron density (red lines) for (a) the same device as in Fig. 4.2 a (b) device with $L_g = 2.5$ nm. Dashed lines are the results in the 1-D approximation.

solution of the Schrödinger equation presented in section 3.1. Because the 1-D approximation neglects the back-scattering of the wavefunction at the channel entrance (*i. e.* source electrode for left-to-right moving electrons and drain electrode for right-to-left), it overestimates the wavefunction penetration into the channel. The red lines show the electron density through the middle of the device. The results show that the adiabatic (1D) approximation overestimates the electron density especially near the electrodes where the tunneling states with energy $E < E_{z,1}$ contribute. It is exactly this region where the potential hump develops hence the 1-D approximation tends to exaggerate its size. The hump near the source is the bottleneck which effectively regulates transport through the device, so small deviations in this region can lead to yield large deviations in the calculation of the current density.

4.2.1.1 Valley Contributions

Figure 4.4 shows the contribution of each doubly degenerate valley for the silicon band structure. The dashed line shows the calculated potential through the middle of the device channel. The X, Y, and Z curves are the contributions when the electron heavy mass is oriented in the $\hat{x}$, $\hat{y}$, and $\hat{z}$ directions respectively (see Fig. 2.7). For the thin channel, $t_c = 2$ nm, the one valley assumption taken for the 1-D Schrödinger calculations is validated. For thicker channels however, all transport valleys need to be considered for accurate device evaluation. As an interesting side note, the curvature of the potential near the drain tends to create a confinement well, producing the small oscillations of the density near the drain. Close examination of these oscillations reveals the difference between the de-Broglie wavelengths for each effective mass.

4.2.2 Potential Pockets

When the applied gate voltage is positive and very high, the bottom of the potential profile may fall below the level of the conduction band in the drain region. Panel (a) of Fig. 4.5 shows one such case calculated in the 1-D approximation. The red line with the dashed section shows the calculated electron density including only electrons from the source and drain electrodes with energies $E > 0$, with zero defined from the bottom of their respective conduction bands.

In equilibrium, the potential "pocket" that forms below the drain level would be filled with electrodes through inelastic relaxation processes (see Fig. 4.6). The electron-phonon interaction time is on the order of 100 fs (see section 2.1), so the pocket will be filled with electrons in the on-state (when the channel density is high) on the order of nanoseconds so this pocket is somewhat

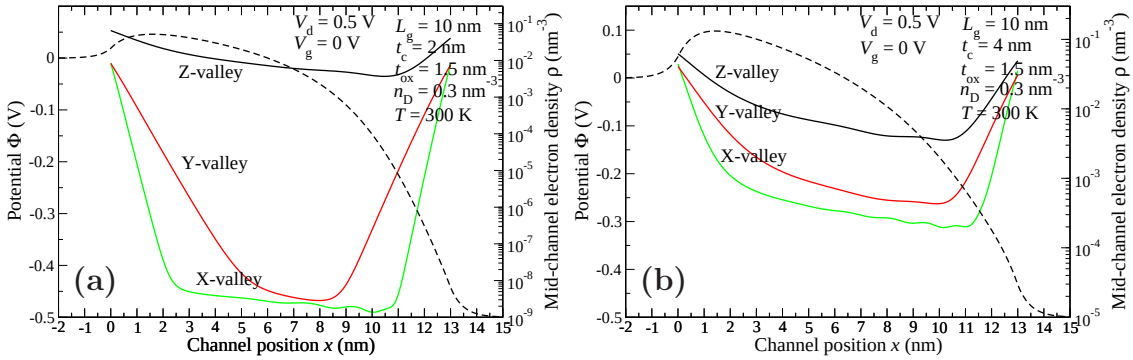

Figure 4.4: Contribution of each silicon valley to the total channel electron density for $t_c = 2$ nm and $t_c = 4$ nm. The dashed line represent the mid-channel potential profile.

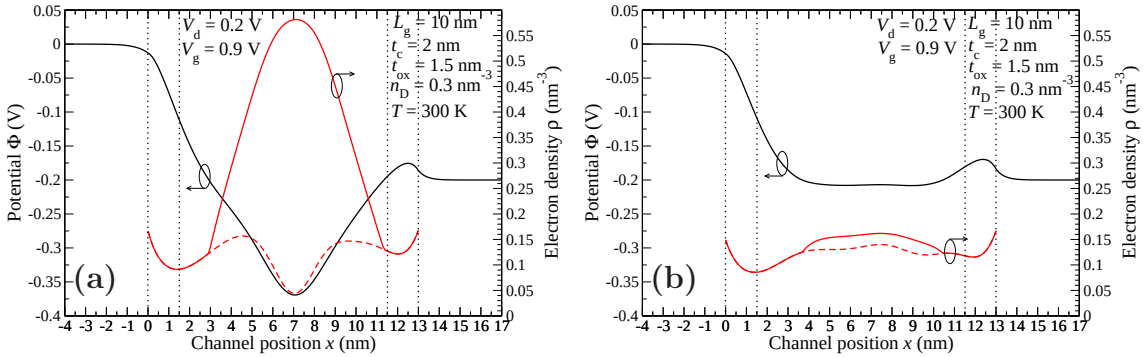

Figure 4.5: Example potential profiles (black lines) converged (a) without and (b) with inclusion of sub-V_d energy electrons. The dashed red line indicates the calculated electron density without the pocket electrons.

unrealistic in a steady-state device. This effect has not been accounted for in the results of chapter 3. To account for the sub-V_d electrons, we add to the electron density a term proportional to the two dimensional density of states

$$n_{\text{pocket}}(x,z) = e\frac{2}{t_c}\cos^2\left(\frac{\pi z}{t_c}\right)\frac{4m_x}{\pi\hbar^2}\left(\Phi(x) - V_d\right), \tag{4.91}$$

where we have included a factor 2 for the spin degeneracy and 2 for the valley degeneracy. The continuous red line in panel (a) shows what the total electron density would be added at this iteration step. Panel (b) shows the self-consistent solution same device including the sub-V_d equilibrium electrons, where again the dashed red-line is the density considering only the source /

drain electrons. The net effect of the inclusion of the sub-V_d electrons is to push the bottom of the conduction band in the channel to not exceed the drain level.

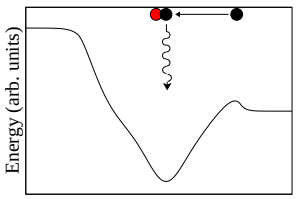

Figure 4.6: Schematic of electron (black circle) and phonon (red circle) relaxation.

4.2.3 $I - V_d$ Families

The $I - V_d$ families for thin, $t_c = 2$ nm, and thick $t_c = 4$ nm channel devices are shown in Fig. 4.7 for gate voltage steps of 100 meV. The dashed lines are the results calculated in the 1-D approximation. For small applied gate bias, $V_g < 0.3$ V, the overestimation of the potential bottleneck yields an underestimation of the current density. For larger applied bias, the opposite effect occurs. The potential hump is eliminated and the potential becomes pinned at the source level. The overestimate of the channel transparency then produces a severely higher calculation of the current density.

The overall performance for the thin channel device are encouraging. The device demonstrates excellent current saturation and very high current densities for gate lengths as small as $L_g = 5$ nm and may even show suitable saturation properties down to $L_g = 2.5$, although a direct comparison of the these results with 3.2.2 is somewhat unfair because of the longer channel $L_c = 5.5$ nm which reduces the impact of direct source-to-drain tunneling. For thicker channel devices, the results are less impressive. Even at the longest gate length $L_g = 10$ nm, DIBL effects severely degrade the current saturation. The prospect of scaling devices with wide channels below 10 nm will be limited. Also, the present day rule of thumb $t_c \approx L_g/2$ [13] for present day MOSFETs will clearly need to be more aggressive for ballistic devices.

The $I-V_d$ families for devices with ultra-thin $t_c = 1$ nm channels are shown in Fig. 4.8. While electron mobilities acceptable for the ballistic assumption have not yet been demonstrated at this ultra thin level, we study the device here as an ideal case device to reach the ultimate scaling limits. For our standard doping level $n_D = 0.3$ nm^{-3}, shown in panel (a), the device again shows near perfect current saturation. The current density is however, unacceptably

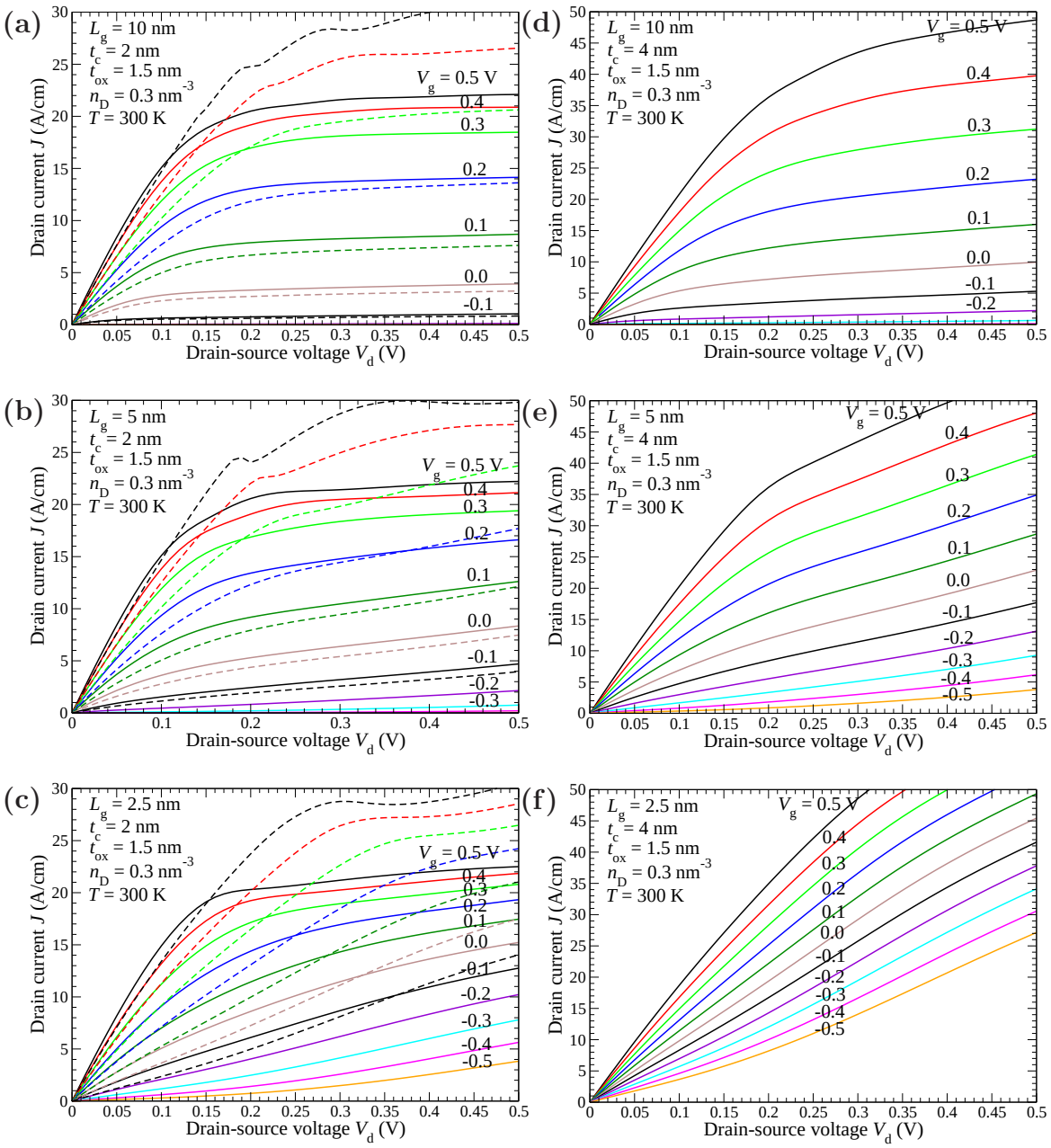

Figure 4.7: Source-Drain $I - V_d$ curves for DG MOSFET with bulk electrodes for $L_g = 10$ nm, $L_g = 5$ nm, $L_g = 2.5$ nm ((a)-(c) and (d)-(f)). The left column shows the $t_c = 2$ nm device and $t_c = 4$ nm in the right column. Dashed lines represent show the results calculated in the 1-D approximation.

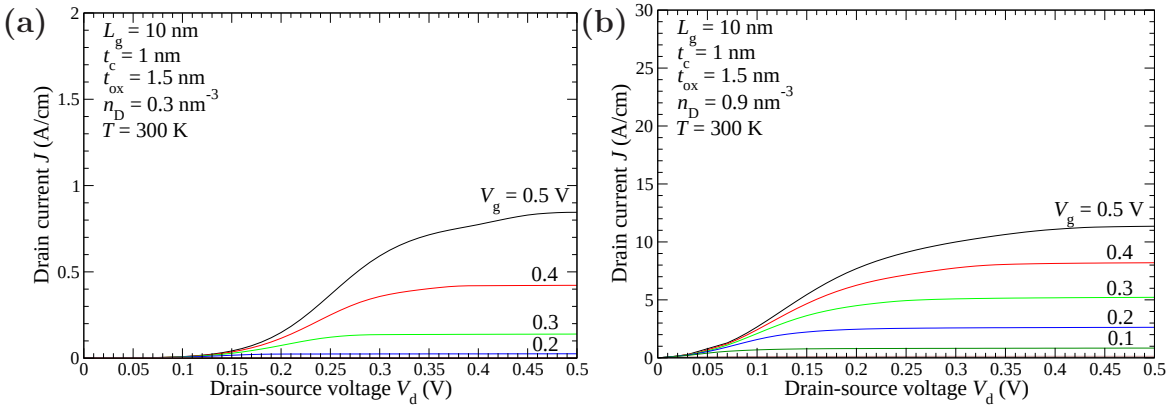

Figure 4.8: Source-Drain $I - V_d$ curves for ultra thin channel $t_c = 1$ nm for (a) standard doping density $n_D = 0.3$ nm^{-3} and (b) high doping $n_D = 0.9$ nm^{-3}.

low to be considered for future circuit applications [3]. To overcome this difficulty, panel (b) shows the results with the doping density increased to near the solid-solubility limit for silicon $n_D = 0.9$ nm^{-3}. The achieved current densities are more acceptable and the utlra-thin channel MOSFET may be considered the "ideal" candidate for ultimate scaling.

4.2.4 Subthreshold Current

The subthreshold characteristics for the same set of devices as figure 4.7 are shown in figure 4.9, where again the oxide thickness has been increased $t_{ox} = 2.5$ nm to minimize the gate leakage current. The gate leakage is represented by the near horizontal dashed lines and is a numerical calculation of the leakage current using the actual potential profile (see appendix A.1). The dotted lines represent the approximate point at which the peak of the valence band reaches the minimum in the conduction band and inter-band tunneling will begin to cross. At this point, inter-band Zener tunneling from the channel to the drain may begin. The tunneling of electrons from the channel valence band to the drain conduction band will result of an accumulation of holes at the point of the potential maximum. This loss of electrons may be compensated by electron-hole recombination, but estimates show that this process will likely be insufficient to counterbalance the effect. The holes produce positive charge exactly at the point of potential maximum, which lowers the potential and reduces the transistors ability to close the current. This effect is not accounted for in our simulation so results below this point may be somewhat approximate. The results of the 1-D calculation are shown as the colored dashed lines and are nearly identical to the 2-D simulation. The high potential barrier means

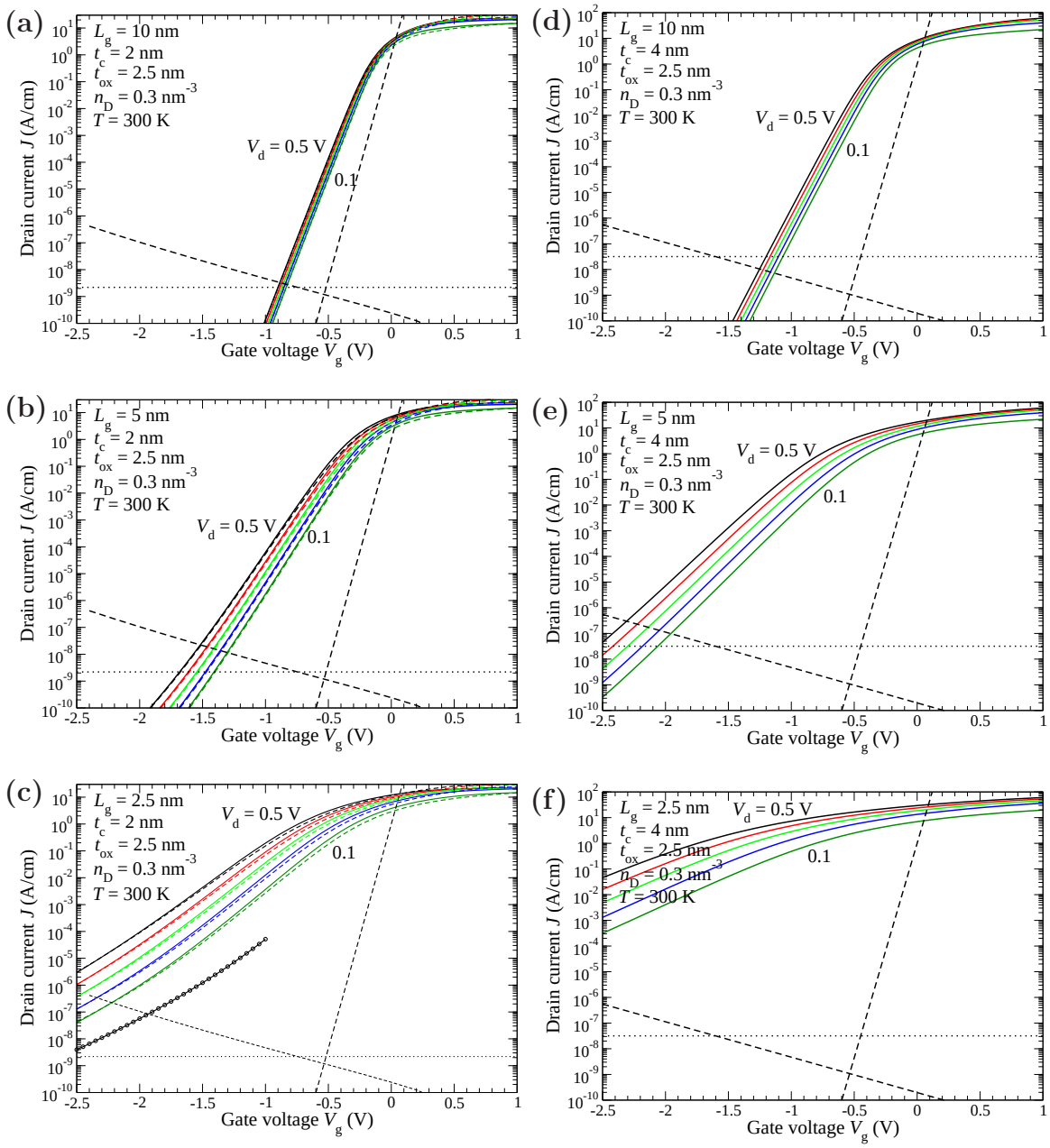

Figure 4.9: Subthreshold curves for the same gate lengths as Fig. 4.7 with increased oxide $t_{ox} = 2.5$ nm. Colored dashed lines represent 1-D calculation. Dotted lines represent the onset of inter-band tunneling.

that the current is dominated by very high energy electrons which are less susceptible to back-scattering at the source interface which makes the two simulations more compatible.

The results again show that for the standard thickness $t_c = 2$ nm (left column of Fig. 4.9), the device demonstrates near perfect, thermal subthreshold slope for the longest gate length $L_g = 10$ nm. In contrast to the device with thin extensions, as the gate length is scaled to $L_g = 5$ nm, the slope is still very acceptable due to the longer channel length. In the 5 nm case, the onset on tunneling current is more clear as the bending of the curves near the bottom of the curves below $V_g = 0.1$ V. Even with the gate length scaled to $L_g = 2.5$ nm, the device still demonstrates around seven orders of magnitude between the on-state and off-state which approaches what is fundamentally required for application in memory circuits.

The thick channel, $t_c = 4$ nm are shown in the right column of Fig. 4.9. Again, even at the longest gate length, the subthreshold slope is severely degraded due to the poor electrostatic control of the gate over the potential inside the channel.

The subthreshold curves for the ultra-thin channel device are shown in Fig. 4.10. Panel (a) shows the full subthreshold profile. As hinted at in section 4.2.3, the device can be essentially seen as a "tunnel transistor", very similar to those with Schottky-barrier junctions (see. *e. g.*, Ref. [104] and references therein), because the confinement energy for the thin channel surpasses the Fermi energy $E_{z,1} \gg E_{\mathrm{F}}$. The crossing point of the confinement energy and the Fermi energy occurs at doping level

$$n_D t_c = 2\pi \left(\frac{\sqrt{m_x m_y}}{m_z} \right). \tag{4.92}$$

In order to provide acceptable current densities, the doping should be as close to (or high than) this threshold as possible. For a 1 nm channel thickness, this corresponds to doping density $n_D \approx 1.2$ nm^{-3}, slightly higher than the doping value accepted in Fig. 4.8.

The subthreshold plot also demonstrates an exponential slope different than the thermal slope in the region $0.1 < V_g < 0.7$. Two sample potential profiles in the region are shown in panel (b). The dashed lines show the potential plus the confinement energy $E_{z,1}$. This different exponential slope can be seen as Fowler-Nordheim tunneling as the channel potential becomes pinned at the source level. Further increase of the gate bias then has the effect of shrinking the barrier width. The line with circles in panel (a) shows the current density calculated in the limit of transmission through a trapezoidal barrier where the length is scaled proportionally with the applied gate voltage.

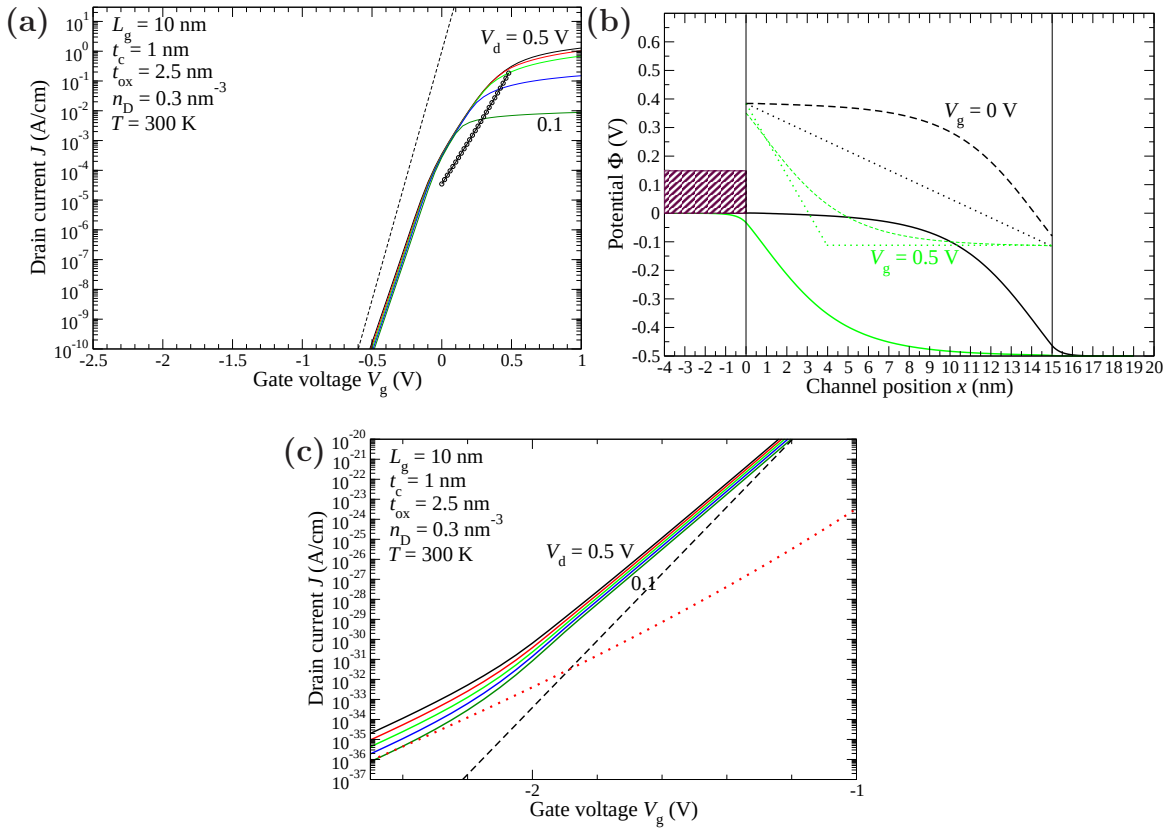

Figure 4.10: Subthreshold curves for ultra thin channel $t_c = 1$ nm for (panel (a)). Panel (b) shows the potential profiles (and potential including confinement as dashed lines) for the Fowler-Nordheim tunneling region, demonstrated as the line with circles in panel a. Panel (c) shows the onset of tunneling current. The dotted red line shows the current density calculated in an exactly solvable WKB model.

Panel (c) shows the current density at the onset of tunneling. While this result is much lower than the inter-band tunneling point, it is illustrative of tunneling effects in all devices considered. For high negative gate voltage, the potential barrier is well approximated as quadratic (see panel b of Fig. 4.2), and the solution may be found analytically in the WKB approximation (2.52). The tunneling current, when calculated in the quadratic limit is found to be

$$J_Q = e\frac{\sqrt{m_y}}{2\pi^3\hbar^{1/2}}\omega^{3/2}e^{(2\pi/\hbar\omega)(E_F-\Phi_0)}, \tag{4.93}$$

with oscillator frequency 2.53. The dotted red line shows the current density (4.93) and comparison of the inversion potential with thermal subthreshold current shows that the onset of tunneling current occurs at

$$V_g \approx 1.4 \text{ eV}, \tag{4.94}$$

in excellent agreement with the full 2-D simulation.

4.2.5 Device Performance

The voltage gain G_v for the standard device thickness $t_c = 2$ nm versus applied gate voltage from off-state to on-state is shown in Fig. 4.11 for three different values of the doping density n_D. While it was previously expected that doping densities greater than $n_D = 0.3$ nm^{-3} would produce uncontrollable currents, these results show that the overall device performance is insensitive to the electrode doping density. The dashed lines in panel (a) show the voltage gain calculated in the 1-D approximation. Because G_v is effectively a measure of electrostatic response to changes in the gate voltage, it is not surprising that the two simulations give very similar results. The results hint that even gate lengths scaled to $L_g = 2.5$ nm may demonstrate fundamental performance characteristics sufficient for engineering into some integrated circuits.

For devices with very long channels, the quadratic model is not valid as the potential takes exponential form

$$\Phi(x,z) \approx \phi(z)\sinh(x/\Lambda), \tag{4.95}$$

where characteristic length Λ may be found as a solution to equation [12]

$$\alpha\tan\left(\frac{t_c}{2\Lambda}\right)\tan\left(\frac{t_{ox}}{\Lambda}\right) = 1. \tag{4.96}$$

The voltage gain calculated in the exponential model, appropriate for very

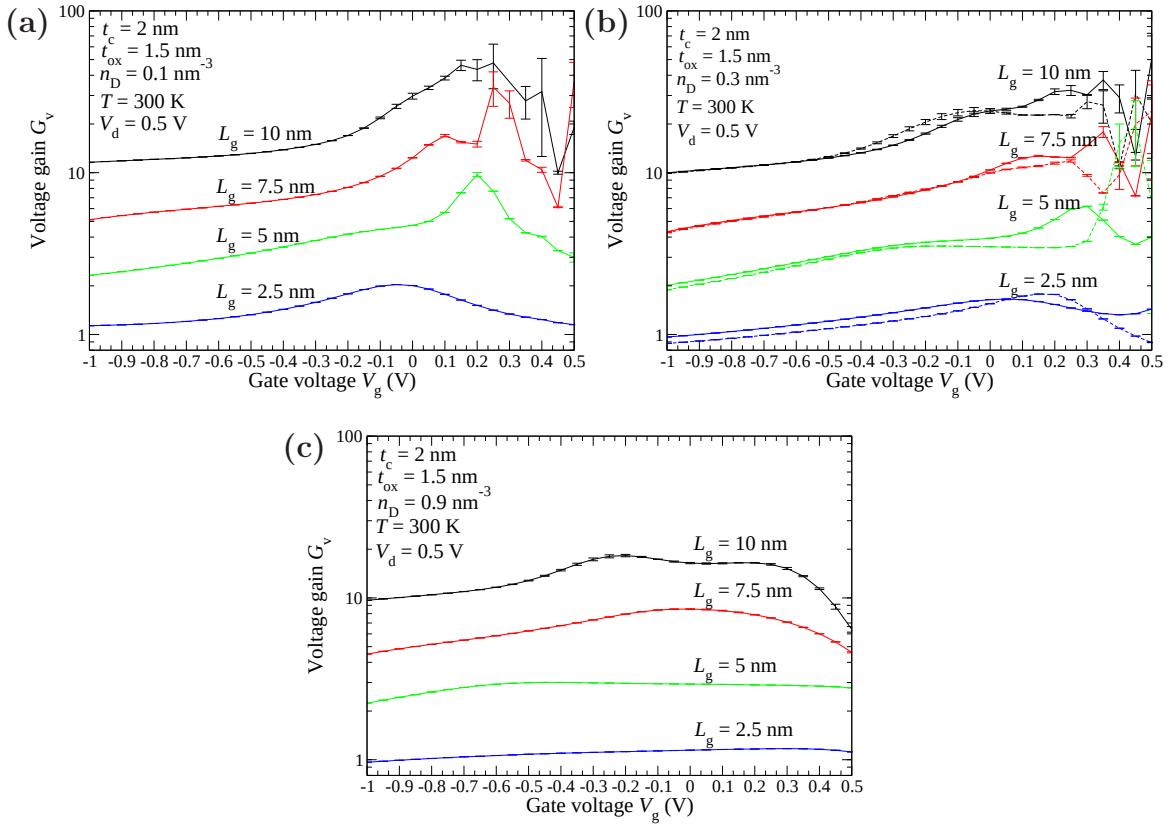

Figure 4.11: Voltage gain G_v for decreasing gate length at three different doping levels: (a) $n_D = 0.1$ nm^{-3}, b) $n_D = 0.3$ nm^{-3}, c) $n_D = 0.9$ nm^{-3}. Dashed lines in panel (b) shows the G_v calculated in the 1-D approximation.

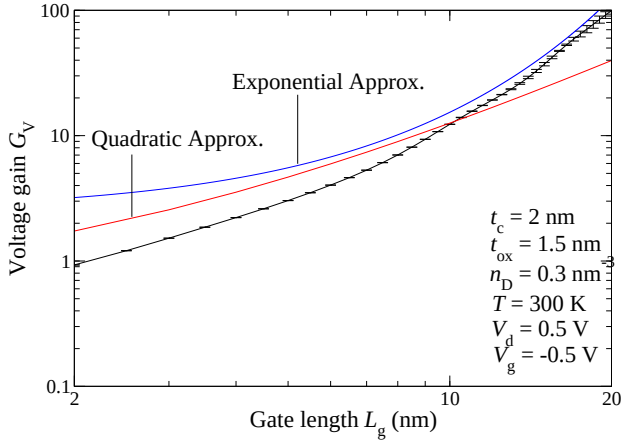

Figure 4.12: Voltage gain G_v for bulk model versus L_g. The red and blue lines show the results of the quadratic and exponential models respectively.

long devices is shown as the blue line in Fig. 4.12. The red line shows the result of the quadratic model (2.61). The difference between the results of the quadratic model and the numerical results for very short gates lengths is caused by quantum tunneling current, an effect which cannot be described by a simple electrostatics model. According to the full numerical simulations, the scaling requirement $G_v \sim> 2$ means that the gate length for bulk devices may safely be scaled to $L_g \approx 4$ nm, and scaling to $L_g = 3$ nm may still be possible depending on engineering requirements.

In order to reach ultimately small gate lengths, the restrictions on the possible width of the channel become quite severe. The device voltage gain versus channel thickness is shown in Fig. 4.13 for decreasing gate length. For example, in order to maintain viability in a $L_g = 5$ nm device the channel thickness should be constrained to no larger than 3 nm, more ideally approaching the 2 nm value assumed in this work.

The dashed line in Fig. 4.13 is again the result of model (2.61). It is clear that the channel thickness requirements for even the longest gate devices are the direct consequence of simple electrostatic limitations.

4.2.5.1 Gate Capacitance

The total gate charge and large signal capacitance are shown in Fig. 4.14. These results are generally on scale with the ITRS 2007 projections [3] for DG structures with 10 nm physical gate lengths predicted for 2015, $C_{g,ideal} = 3.68 \times 10^{-12}$ F/cm. However, the ITRS results are based on the assumption of higher drive supply voltage $V_{dd,ITRS} = 0.8$ V and lower saturation current $J_{sat,ITRS} = 7.02$ A/cm than the ballistic transistor shown here. The DG-FET demonstrates near linear capacitance scaling down to even the smallest gate lengths. The gate thickness has been taken to be $t_g = 4$ nm. While this choice is a bit arbitrary, the capacitance also scales linearly with the choice of this parameter.

The gate capacitance remains the limiting parameter for circuit clocking in the absence of circuit power and thermal requirements. For example, Fig. 4.14(b) shows that our standard $t_c = 2$ nm, $n_D = 0.3$ nm^{-3} device with a gate length $L_g = 5$ nm has a total gate capacitance (per unit width) around 5×10^{-12} F/cm. Fig. 4.7(b) shows an on-state current density $J_{ON} = 27$ A/cm at drive voltage $V_{DD} = 0.5$ V. This yeilds a capactive recharging time $\tau = C_g V_{DD}/J_{ON} \approx 93$ ps. Assuming fraction of clock pulse used for recharging $p = 1/16$ this leads to an absolute ceiling for the clocking frequency $f \approx 11$ GHz. Of course, even in present day circuits, the clock frequency is limited by the thermal dissipation of chip and not the gate capacitance.

Using the velocity estimate from section 2.1, $\langle v \rangle \approx 2 \times 10^7$ cm/s the device

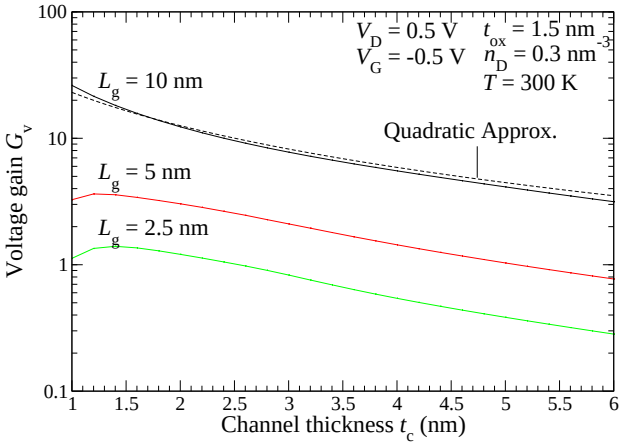

Figure 4.13: Voltage gain G_v for bulk model versus t_c for decreasing gate length. The dashed black line is the result of the quadratic model.

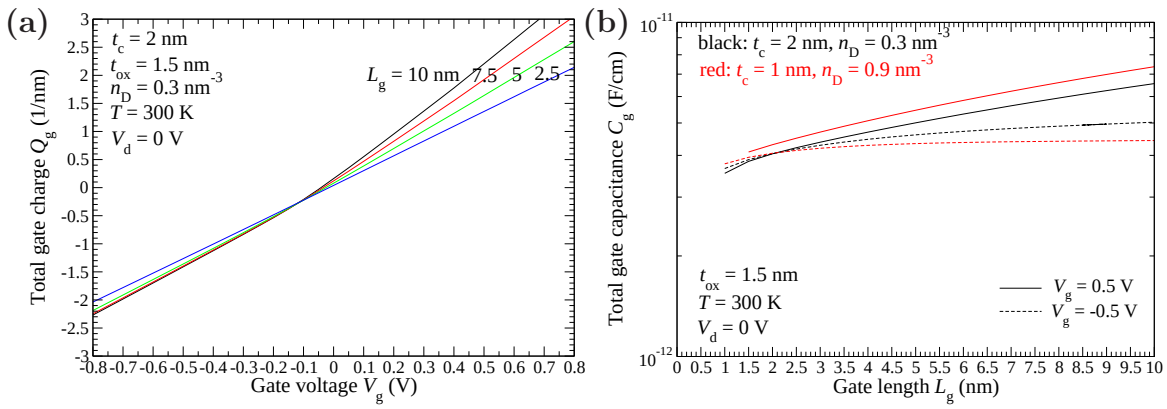

Figure 4.14: Total electron charge (panel (a)) and capacitance (panel (b)) for devices with decreasing gate lengths.

has an intrinsic time delay $t_i \approx L_c/\langle v \rangle \approx 40$ ps, over a twice as small as the capacitive recharging time. The estimate for the intrinsic delay is actually un upper limit as the particle dynamics are essential determined in a small region around the potential maximum, near the source electrode in the on-state.

4.2.5.2 Advanced Materials

High-k dielectric materials are actively being sought as a replacement for the standard silicon dioxide. Hafnium oxide, with $\epsilon \approx 30\epsilon_0$ has seen spectacular recent success due to it's close compatibility with the silicon lattice and thermal stability. So much so that it has even been incorporated into some of the most recent production devices [105]. The primary benefit of these HfO_2 materials is the reduction of gate leakage currents [57] due to their higher band-gap energy ($E_{\rm gap} = 5.65$ eV) [106] and the ability to use a thicker dielectric for the same silicon "equivalent oxide thickness"

$$t_{EOT} = t_{\rm phys}\frac{\epsilon_{\rm SiO_2}}{\epsilon_{ox}}. \tag{4.97}$$

However, the question of how these advanced materials will impact the ultimate scaling limits has not been fully addressed.

The device voltage gain at constant physical oxide thickness t_{ox} versus the oxide dielectric constant, calculated within our model, is shown in panel (a) of Fig. 4.15. The dashed line shows the result of the simple quadratic model (2.61). The results indicate that the performance benefits from even the very highest permittivity oxides will be limited. As the permittivity of the insulator is raised much higher than that for SiO_2, $\epsilon_{ox} \approx 3.9\epsilon_0$, the device voltage becomes flat. Although the gate field may have increased penetration into the channel, the response of the potential barrier saturates.

A similar result is obtained for low-k channel materials (*i. e.* graphene), shown in panel (b). This second result is more artificial however because the calculation assumes a silicon band structure not relevant for graphene channels. The limited impact of scaling the dielectric constant may seen intuitively from electrostatic arguments. For all practical device $r_{ox} \gg 1$, so the two dominant terms in the quadratic model (2.61) for the reduction of G_v are $2\alpha\beta_{ox}$, and $2\beta_{ox}^2$. The second term in the denominator may be presented as

$$2\alpha\beta_{ox} = 2\frac{t_{EOT}}{t_{ECT}}, \tag{4.98}$$

where t_{ECT} is the "effective channel thickness" defined similarly to Eq. (4.97). Even in the case when either dielectric is scaled to the ideal limit $t_{EOT}/t_{ECT} \rightarrow$

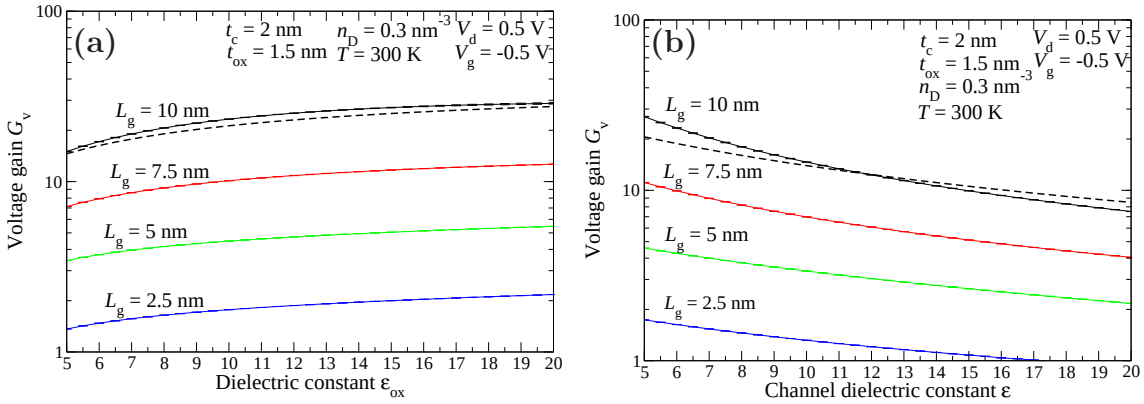

Figure 4.15: Voltage gain G_v versus dielectric constant in (a) oxide and (b) channel for decreasing gate lengths. The dashed line in both figures is the result of the simple quadratic model.

0, the electrostatic response is still hindered by a term proportional to the physical oxide and channel thicknesses. For our standard configuration, $t_o x = 1.5$ nm, $t_c = 2$ nm, this yields $2\beta_{ox}^2 = 1.125$.

In the limit that $\alpha \ll 1$, and $t_{ox} \ll t_c, L_c$ the device will demonstrate near perfect electrostatic response and the subthreshold voltage gain reduces

$$G_v \to 2\left(\frac{L_c}{t_c}\right)^2. \tag{4.99}$$

Hence, the device performance in the ideal limit is determined simply by the ratio of channel dimensions where the factor of two represents the doubled gate capacitance. This may also been seen in terms of the present rule of thumb for scaling, the device lengths should be scaled proportionally to maintain constant field inside the channel to scale down the device while maintaining constant performance. The requirement to maintain device voltage gain above $G_v \approx 2$ leads to ultimate scaling limit

$$L_c^{\lim} = t_c. \tag{4.100}$$

Fundamentally, the channel length may be scaled to near the channel thickness if the physical oxide thickness may be made much smaller than the channel while allowing acceptably low leakage currents and maintaining reproducibility. A heavy requirement to say the least. Slightly more realistically, for ultimately scaled devices $t_{ox} \approx t_c$ and extremely high oxide permittivity

$$L_c^{\lim} = \sqrt{3}\, t_c, \tag{4.101}$$

or slightly better than the present day rule of thumb for scaling.

4.2.6 Threshold Voltage Rolloff

Similar to the device with thin doped channel extensions, the performance characteristics of the device connected to bulk electrodes are very promising to maintain scaling to ultra small gate lengths. The natural question is whether the scaling of this device is more sensitive to deviations in the fabrication parameters. The development of the potential "hump" near the source electrode means that the 1-D theory used in section 3.2.5 is not necessarily valid for the bulk device. For this configuration, the threshold voltage for the infinite length device is calculated at $L_g = 30$ nm.

The rolloff in the threshold voltage is shown in Fig. 4.16 for bulk electrodes with standard doping $n_D = 0.3$ nm^{-3} (panel (a)), high doping $n_D = 0.9$ nm^{-3} (panel (b)) for a range of channel and oxide thicknesses. We find that sensitivity to fluctuations in the fabrication scale similarly to the results presented in section 3.2.5, *i. e.* the sensitivity to changes in the gate length grows exponentially as the device is scaled down.

The bulk devices demonstrate a crossover point at around $L_g \approx 4$ nm where the dependence of ΔV_t on the oxide thickness changes sign. This crossover is stable over a surprisingly large range of channel thicknesses. For large gate devices, the main impact of changing the oxide thickness is a change in device electrostatics, and the resulting change in channel length $L_c = L_g + 2t_{ox}$ is relatively unimportant. For short gate devices however, direct source-to-drain tunneling begins to dominate device characteristics and the resulting change in the width of the tunnel barrier becomes the dominant effect. Smaller oxides have better electrostatic properties, but also smaller L_c tunnel barriers and hence larger ΔV_t beyond $L_g \approx 4$. Note that this result is obtained assuming the oxide scaled similarly in all device directions. This is mainly the product of conceptual convenience. The crossover would disappear for a configuration with a thin oxide layer between the gate and channel and a constant thick layer between the gate and source / drain electrodes.

In practical integrated circuits, fluctuations in the threshold voltage should be held to be much smaller than the circuit supply voltage V_{DD}. For example, panel (a) shows that in order to control threshold voltage swing to within 50 meV in a standard bulk electrode device(a) with gate length $L_g = 5$ nm, the gate dimension of each circuit element should not vary more than 0.2 nm. In other words, the fabrication process should be held within 4% at the 3σ level, *i. e.* much tighter than the 12% projected in the latest ITRS [3]. The situation becomes twice worse as the gate length is thinned to 1 nm as the fabrication process should be controlled to within 2% deviation of the critical

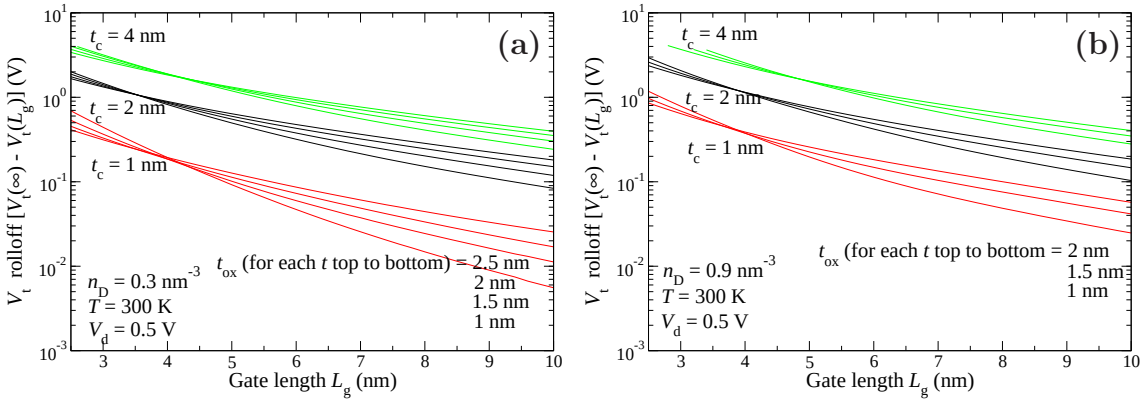

Figure 4.16: "Rolloff" of the threshold voltage V_t for bulk device in (a) standard and (b) ultra-high doped bulk device for three channel thickness.

dimension.

Again, close attention must be paid to fabrication control over all the device dimensions. The change in the threshold voltage from the "standard" device versus channel thickness and oxide thickness is shown in Fig. 4.17 for the standard doping level. For ultra-short gate length devices, the swing in the threshold voltage may be even more sensitive to variations in the channel in oxide thicknesses. The crossover in the relation between tunneling and electrostatics is also expressed in Fig. 4.17. For example, the sensitivity to a change in channel thickness shown in panel 4.17a is higher for the ultra-thin $t_c = 1$ nm device as the gate length reaches $L_g = 2.5$ nm.

4.2.7 Power

The minimum power operating point is plotted in Fig. 4.18 versus the device gate length for our standard device ($t_c = 2$ nm, $n_D = 0.3$ nm^{-3}, panels (a), (c)) and an "ideal" device ($t_c = 1$ nm, $n_D = 0.9$ nm^{-3}, panels (b), (d)) for two different values of λ. The exponential growth of the power minimum operating point is evident in all devices considered, even for "ideal" devices where the minimum power is lower, but only slightly from the "standard" device configuration. This is the expected result as the mechanism for saturation of the current is the same in both devices, *i. e.* exhaustion of the supply of source electrons (see section 3.2.6).

To illustrate this point, the projected physical gate will reach $L_g = 13$ nm by year 2013 [3], similar to our bulk device with $L_g = 10$ nm, $L_c = 13$ nm. The maximum acceptable chip power for performance computing logic applications is stated to be 198 W with a chip size of 140 mm^2 and 0.62×10^9 transistors

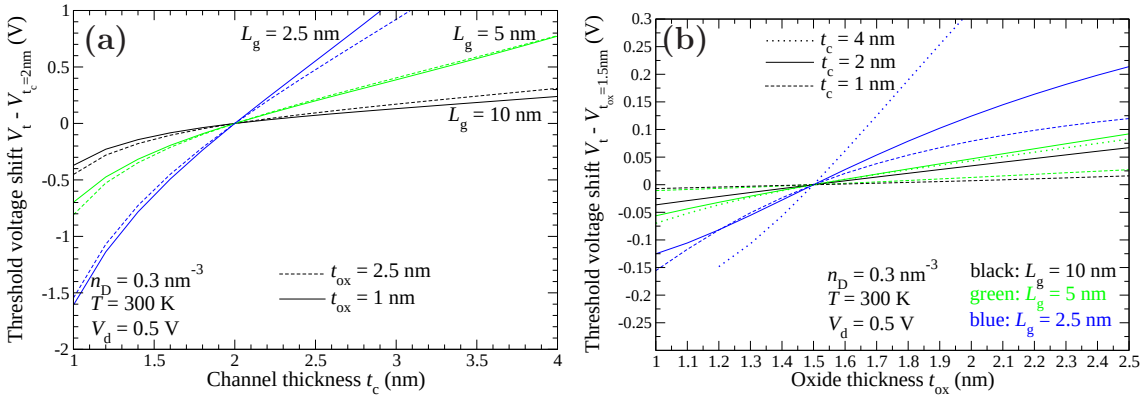

Figure 4.17: Change in the threshold voltage V_t versus (a) channel thickness and (b) oxide thickness for a range of gate lengths.

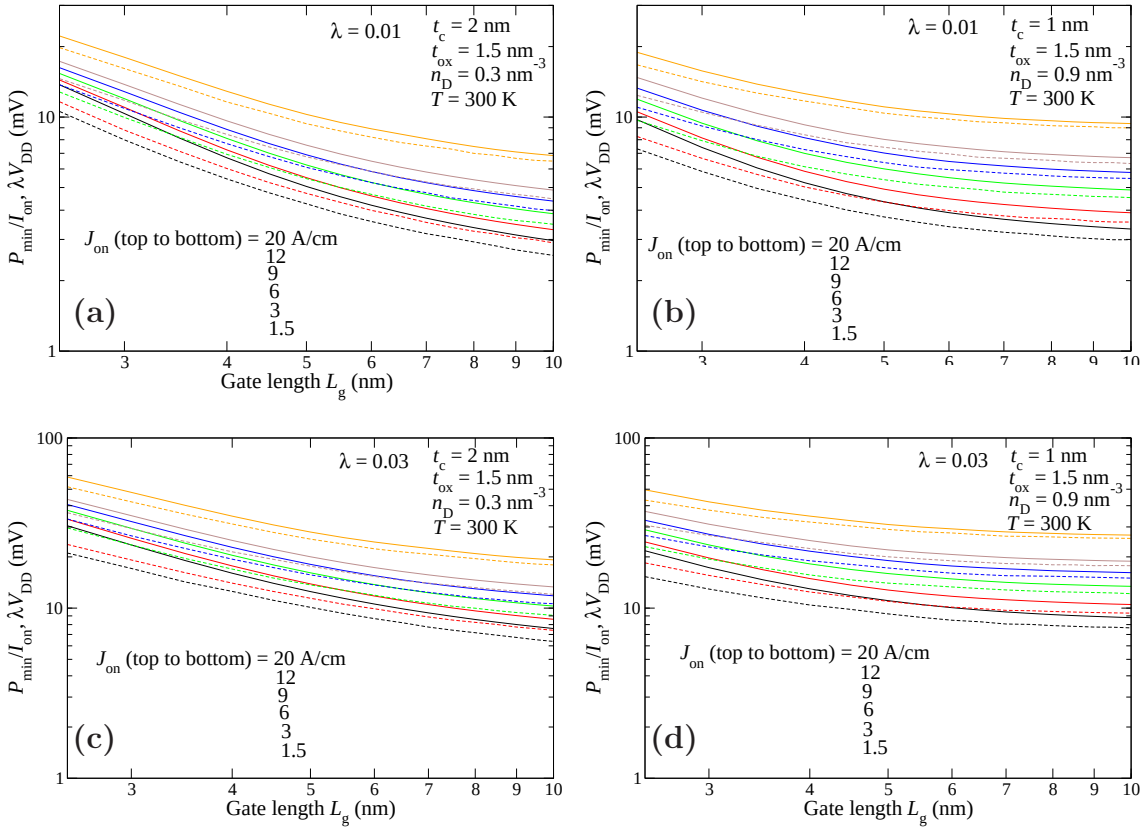

Figure 4.18: Minimum power (solid lines) and minimum drive voltage V_{DD} (dashed lines) versus gate length for standard device $t_c = 2$ nm,$n_D = 0.3$ nm^{-3} (panels (a),(c)) and "ideal" device $t_c = 1$ nm,$n_D = 0.9$ nm^{-3} (panels (b),(d)).

per square centimeter (see tables 6a and 1i of [7]). Assuming standard ratio $W/L_g = 10$, this leads to maximum power per device width for $L_g = 10$ nm

$$P_{\max}/W = 0.0225 \text{ W/cm}. \tag{4.102}$$

These values correspond to a switching activity $\alpha = 0.15$ at $V_{DD} = 0.3$ mV and $J_{ON} = 6$ A/cm, very near today's value [65]. As the transistor gate length is scaled down, the activity factor λ will also need to scale to keep the total power below the required limit. Hence, only a small fraction of transistors will be available at any one time. Radically new device architectures may be required for low power requirements to accommodate these ultra small transistors.

For the same parameters as Fig. 4.18, the same consequence of power scaling is shown in Fig. 4.19 as leakage current versus specified on-state current against ITRS stated requirements from the 2003 and 2007 editions. Table 4.1 shows these values explicitly from Tab. 47 of Ref. [107] and Tab. PIDS2 of Ref. [3].

Clearly, the analytical and numeric scientific results had an impact on the engineering predictions as the ITRS requirements were significantly loosened between 2003 and 2007. However, even the latest projections seem overly optimistic for even the longest gate devices. The projected on-state current densities also seem overly aggressive as reaching current densities higher than around $J_{\text{ON}} \approx 22$ A/cm requires a drive voltage $V_{DD} > 1.2$ V. Reaching the projected level of leakage current may be possible, but it would require operating the device away from the minimum power point further increasing overall circuit power consumption.

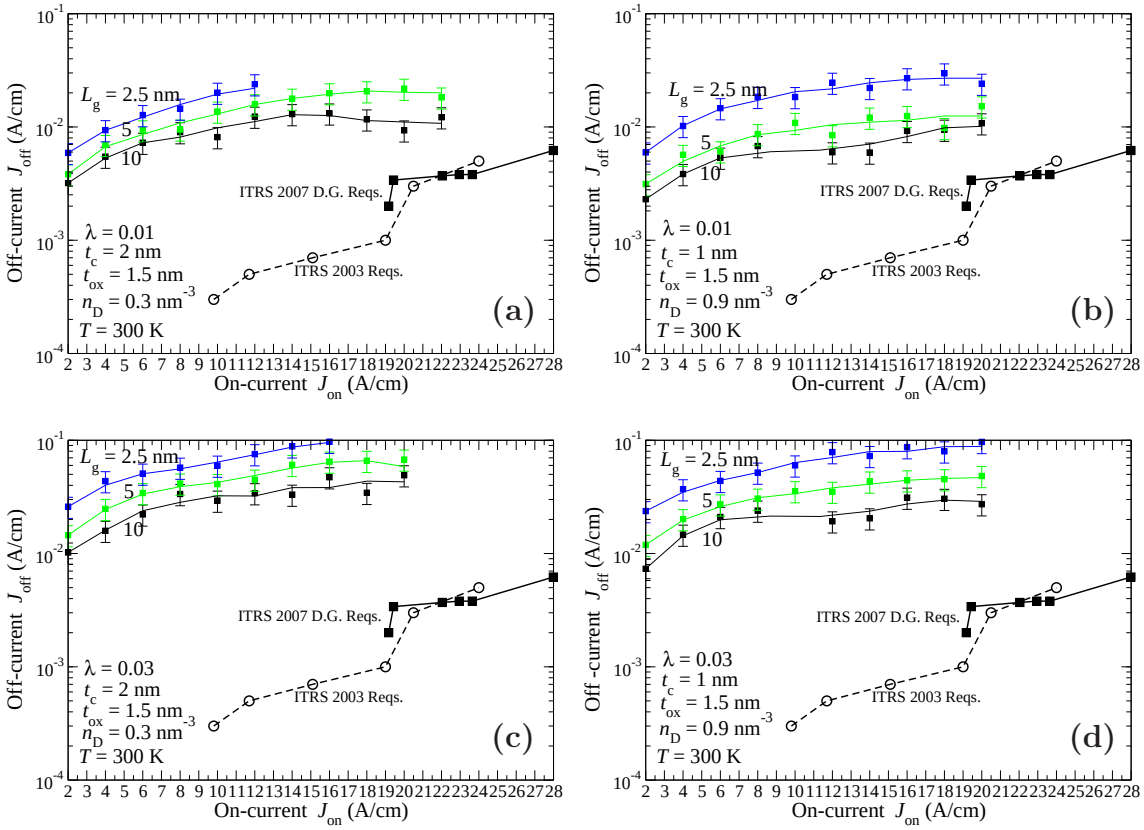

Figure 4.19: Off-state current versus specified on-state current J_{ON} calculated in the point of power minimum. Panels (a), (c) show the standard device parameters and panels (b), (d) show the "ideal" device. The dashed and solid line represent the ITRS requirements as stated in 2003 and 2007.

Year	**Phys. gate len. (nm)**	V_{DD} **(V)**	J_{sat} **(A/cm)**	J_{leak} $\mathbf{10^{-3}}$**(A/cm)**
ITRS 2003				
2003	45	1.2	9.80	0.3
2006	28	1.1	11.7	0.5
2007	25	1.1	15.1	0.7
2010	18	1.0	19.0	1.0
2013	13	0.9	20.5	3.0
2016	9	0.8	24.0	5
ITRS 2007				
2011	16	1.0	19.17	2.0
2012	14	0.9	19.43	3.4
2013	13	0.9	22.04	3.7
2014	11	0.9	23.65	3.8
2015	10	0.8	22.95	3.8
2021	5	0.65	27.99	6.2

Table 4.1: ITRS predicted values for saturation and maximum drain-source subthreshold leakage current densities [107], [3]

4.3 Comparison of Thin-extension and Bulk-electrode Devices

Figure 4.20 shows the source-drain $I - V_d$ families of section 4.2.3 and chapter 3 side by side for similar gate lengths. Figure 4.21 shows this same comparison for the subthreshold slopes. When the gate length L_g is equal in the two models, the bulk device displays better saturation and subthreshold performance over its thin-extension counterpart due to the longer channel length L_c which lower electron tunneling through the potential barrier. On the other hand when the channel lengths are the same, the thin extension model performs better because the gate has electrostatic control over the entire channel region. The trade off between these two effects means that the bulk electrode and thin-extensions transistors will scale similarly when the thin extension gate length is between L_g and $L_g + 2t_{ox}$ for the bulk model.

All transistor properties only get worse in both models as the gate is scaled to the sub-10 nm regime. Hence, a primary benefit of continued scaling is the increased chip packing density and the relevant length scale is not L_g, but the total "bulk-to-bulk" length $L_{BB} = L_g + 2L_{\text{ext}}$.

The minimum power results for both models is shown in Fig. 4.22 versus the total "bulk-to-bulk" length L_{BB}. Lines of similar color represent fixed L_{ext} and dashed lines fixed L_g. In the limit $L_{BB} \to \infty$, all the numerical results approach the value found from the 1-D theory 2.2.

The same results as 4.22 for the voltage gain G_v are shown in 4.23. Both in terms of minimum power and voltage gain, the bulk electrode device outperforms the model with thin extensions at equivalent L_{BB}. From a performance point of view, the bulk device is the preferable candidate for ultimate scaling.

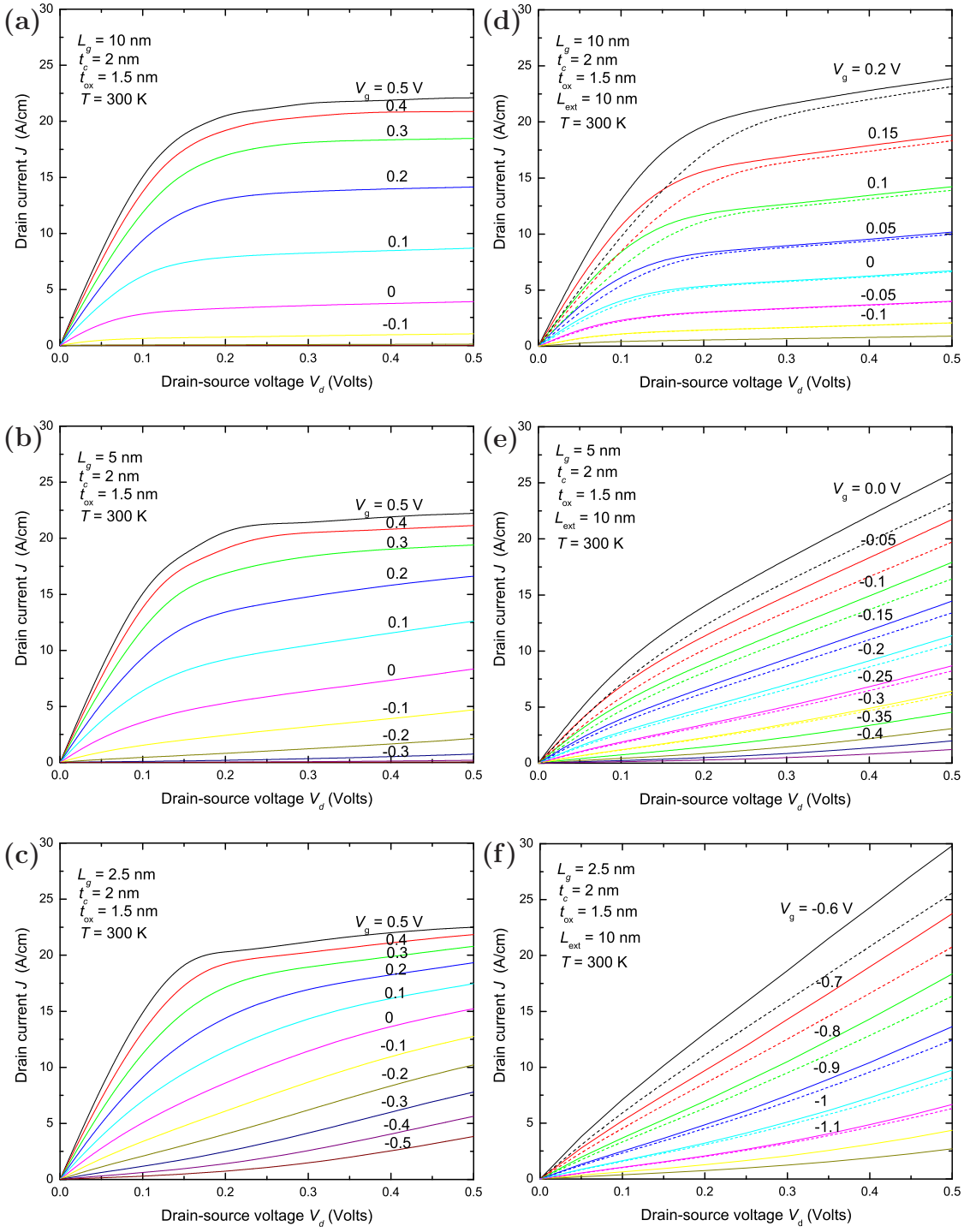

Figure 4.20: Source-Drain $I - V_d$ curves for DG MOSFET for $L_g = 10$ nm, $L_g = 5$ nm, $L_g = 2.5$ nm ((a)-(c) and (d)-(f)). The bulk electrode and thin-extension devices are shown in the left and right columns respectively.

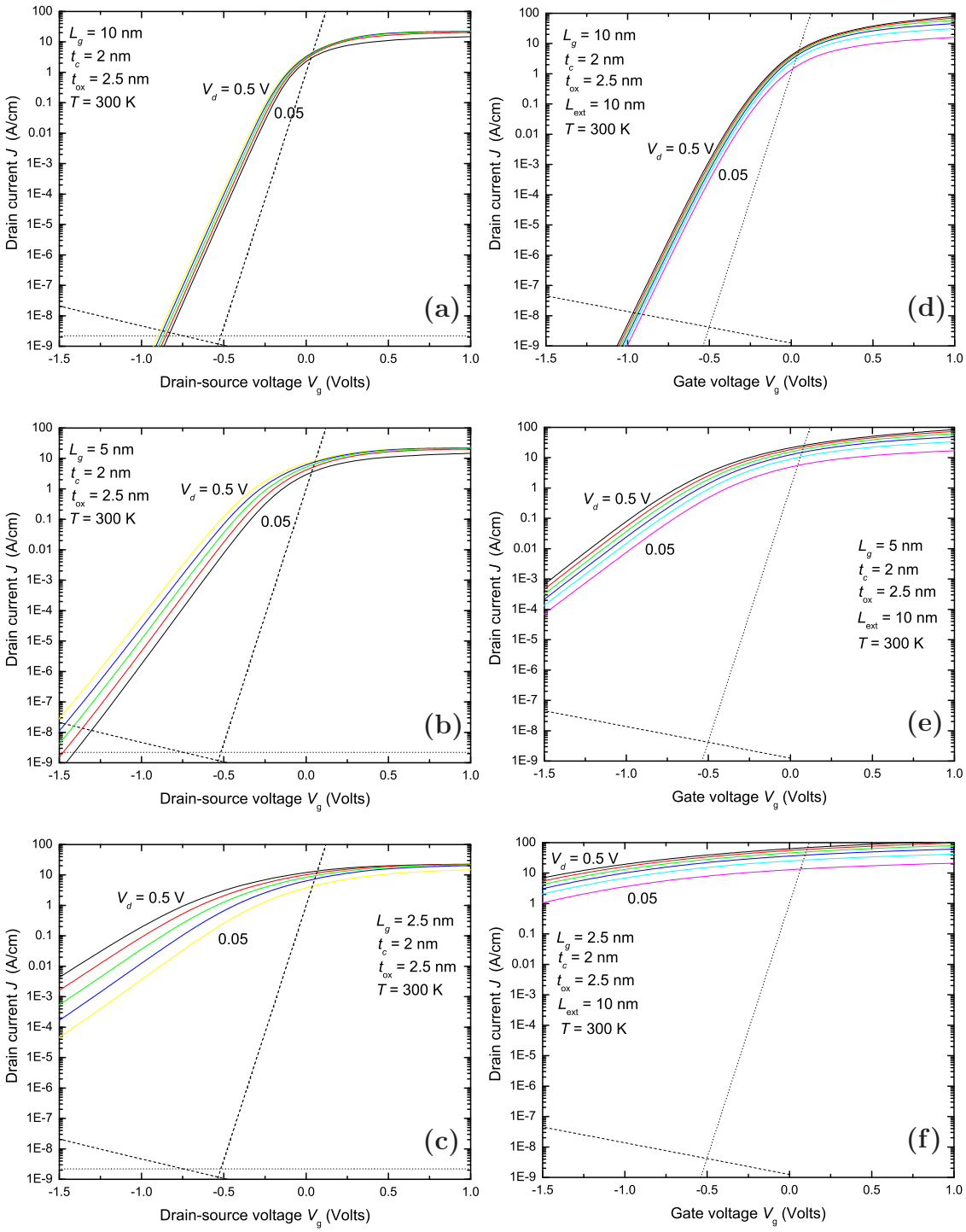

Figure 4.21: Source-Drain $I - V_d$ curves for DG MOSFET for $L_g = 10$ nm, $L_g = 5$ nm, $L_g = 2.5$ nm ((a)-(c) and (d)-(f)). The bulk electrode and thin-extension devices are shown in the left and right columns respectively.

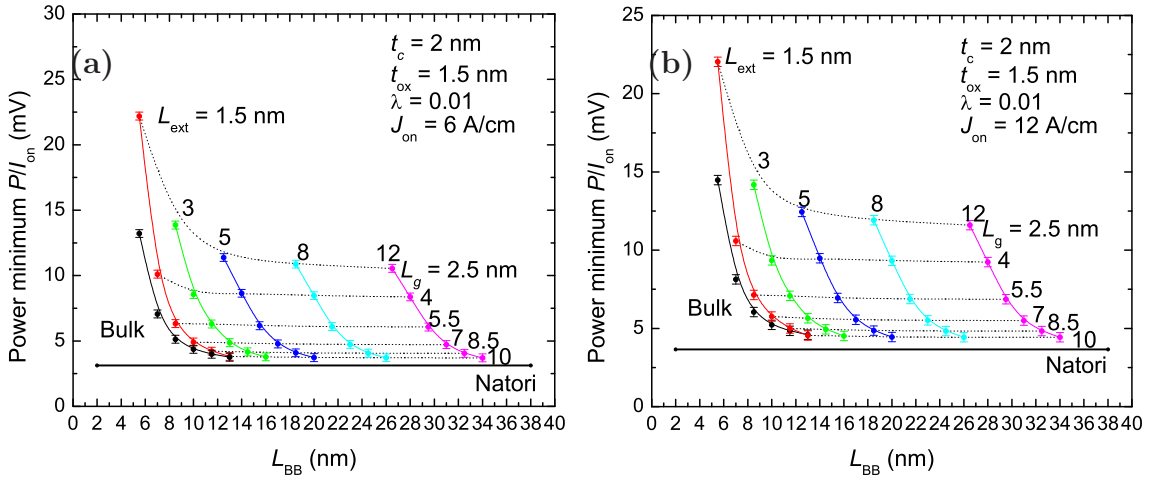

Figure 4.22: Minimum power supply voltage versus total "bulk-to-bulk" length L_{BB} for two specified on-state currents (panel (a) $J_{\rm ON} = 6$ A/cm, (b) $J_{\rm ON} = 12$ A/cm). The colored lines show a DG MOSFET for increasing electrode extension lengths. Dashed lines represent equivalent gate lengths and the horizontal line the result of 1-D theory of Sec. 2.2.8

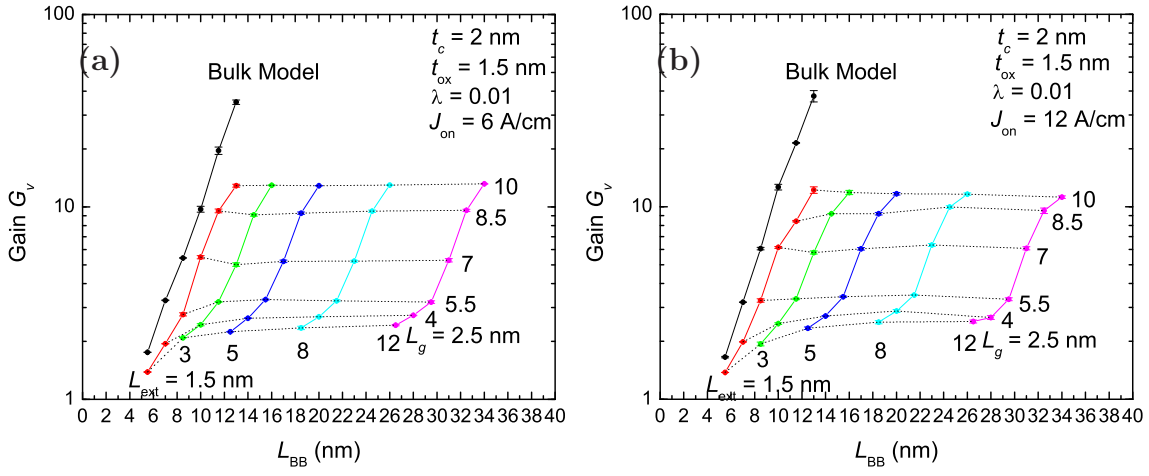

Figure 4.23: Same plot as Fig. 4.22 but for the voltage gain calculated in the point of power minimum.

4.4 Conclusions

We have developed theoretical and numerical models for MOSFET characterization in the ultra-small gate length ballistic limit. The simulator is capable of many approximations for channel electrons ranging from a purely classical description to a full solution of the two dimensional Schrödinger equation. Extensive device simulation shows that the double gate FET structure is an ideal candidate to reach the ultimate scaling limits. Fundamentally, Moore's exponential law may continue into the ballistic regime and gate lengths may shrink to at least 5 nm for devices connected directly to bulk electrodes. Beyond this point, direct source-to-drain tunneling of electrons begins to seriously impact device characteristics even in ideal cases. Sub-5 nm may still be useful for integrated circuits however the engineering requirements to compensate for the degraded performance would be large.Unfortunately, the results are not as positive when the economics of fabrication cost and power consumption are considered. Without major advances in lithographic techniques and chip cooling, which is the limiting factor for circuit power limits, Moore's law may come to end far before the physical limits have been reached. There are other proposals to continue the trend for circuit density, such as three dimensional transistor stacking techniques. But it is clear that device scaling, which has driven the exponential growth to this point, is fast coming to an end.

Ultimately, if the challenges facing fabrication are overcome, there will likely not be one answer to the minimum MOSFET size. Devices may be engineered for specific chip requirements as such high performance circuits for computing and low power circuits for mobile devices. The trade-offs of feature size versus performance may be tailored to meet the needs of each designer.

Part II

Josephson Junction Comparator as a Quantum Limited Detector

I am a strange loop.
D. R. Hofstadter.

Chapter 5

Comparator for High-Impedance Signal Readout

Traditional electronics are just now beginning to reach the quantum mechanical threshold. On the other hand, the application of coherent wavefunction manipulation for digital logic has existed for over three decades in superconducting circuits [5]. Superconducting digital logic is a very attractive replacement for present day electronics because the limiting frequencies of operation can approach 1 THz with nearly negligible power cost for gate operations. Specifically, devices based on the quantization of magnetic flux have found a wide range of applications from a recent dramatic demonstration of quantum mechanical coherence effects (avoided crossing) at a macroscopic level [108], to defining the international standard volt unit [109], to commercially viable, highly accurate magnetic field sensors and analog-to-digital (A/D) converters [110]. However, the ultimate sensitivity limits of the later set of devices is still not well understood. In this chapter, we will develop a model based on the Caldeira-Leggett formalism [111] for the signal resolution of direct measurement of current from a high impedance source for one important component of superconducting circuits, the balanced Josephson junction comparator. Chapter 6 is dedicated to development of a less cumbersome Heisenberg-Langevin-Lax model [112] and the possible application for rapid single-shot measurements via inductive coupling.

5.1 Single Josephson Junction

In the simplest view of superconductivity, lattice phonons may screen the electrostatic repulsion between electrons near the Fermi surface creating a net attractive force between two particles. This net attraction allows conduction

band electrons of opposite spins and wavevectors to bind together into singlet state pairs, known as Cooper pairs [113], which resemble bosons with an effective wavefunction that describes the center of mass of the bound "particle". At negligible temperatures, all of the conduction band electrons are collected into pairs and collapse into a single quantum state

$$\boldsymbol{\Psi} = \boldsymbol{\Psi}_0 e^{\mathrm{i}\theta}, \tag{5.1}$$

with constant magnitude $\boldsymbol{\Psi}_0$ and phase θ. This theory is applicable in cases where the spatial variation of the potential is slow compared with the ground state wavevector.

The spatially invariant density of Cooper pairs is then

$$n_S = |\boldsymbol{\Psi}_0|^2, \tag{5.2}$$

whose phase can vary as $\theta = \boldsymbol{p}\cdot\boldsymbol{r}/\hbar + \omega t$ with the ground state energy $E = \hbar\omega$.

If we consider a loop of superconducting material, continuity of the wavefunction (5.1) requires that the phase must change by exactly $2\pi n$, where n is an integer, for one complete path around the loop. Magnetic fields cannot penetrate the superconducting material itself [113], and the requirement that the phase changes by integer values (modulo 2π) leads to the conclusion that any magnetic flux $\boldsymbol{\Phi}$ enclosed by the loop must also be quantized

$$|\boldsymbol{\Phi}| = n\Phi_0, \tag{5.3}$$

in integer values of flux quantum

$$\Phi_0 \equiv h/2e = 2.067 \times 10^{-15} \text{ Wb}. \tag{5.4}$$

Any voltage applied to a broken superconducting loop (Fig. 5.1) will change the ground state energy $\Delta E = 2eV$, rotating the phase with angular frequency $\omega = 2eV/\hbar$. Application of a voltage pulse such that $\int V(t)\,dt = \Phi_0$,

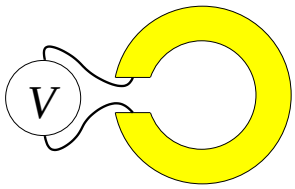

Figure 5.1: Schematic broken superconducting loop with applied voltage V.

causes phase change

$$\Delta\theta = \int 2eV(t)/\hbar \, dt = 2\pi, \tag{5.5}$$

and is equivalent to insertion of one flux quantum Φ_0 into the loop. This is the so-called Single-Flux-Quantum (SFQ) pulse and is the basis for Rapid Single-Flux-Quantum (RSFQ) digital circuitry [6, 110].

We consider the case of two identical superconducting slabs of area A separated by an insulating material of thickness d, weak enough to allow overlap of the superconducting wavefunctions but strong enough for the ground-state wavefunctions in each material to maintain independent phases, shown schematically in Fig. 5.2. We may write the wavefunction in the insulator material as a superposition $\Psi = \Psi_1 + \Psi_2$ of wavefunctions

$$\begin{aligned} \Psi_1 &= \Psi_0 \exp[-\alpha(x + d/2)] \exp(\mathrm{i}\theta_1), \\ \Psi_2 &= \Psi_0 \exp[-\alpha(d/2 - x)] \exp(\mathrm{i}\theta_2), \end{aligned} \tag{5.6}$$

with decay constant α determined by the insulator material. The probability current through the insulator is then calculated as

$$\begin{aligned} I_s &= e\frac{\mathrm{i}A\hbar}{m} \left(\Psi \nabla \Psi^* - \Psi^* \nabla \Psi\right), \\ &= I_c \sin\varphi, \end{aligned} \tag{5.7}$$

with junction critical current

$$I_c = e\frac{2A\hbar\alpha|\Psi_0|^2}{m} e^{-\alpha d}, \tag{5.8}$$

and phase difference

$$\varphi \equiv \theta_1 - \theta_2. \tag{5.9}$$

Named the Josephson effect [114], Eq. (5.7) shows that a supercurrent flows through the insulating junction even in the absence of an applied voltage. Supplying a voltage to the superconducting slabs creates a split in the ground state energies $\Delta E = E_1 - E_2 = 2eV$. The supercurrent then oscillates in time

$$I_s = I_c \sin\left(\omega_J t\right), \tag{5.10}$$

at angular Josephson frequency

$$\omega_J = \frac{d\varphi}{dt} = \frac{2\pi}{\Phi_0} V. \tag{5.11}$$

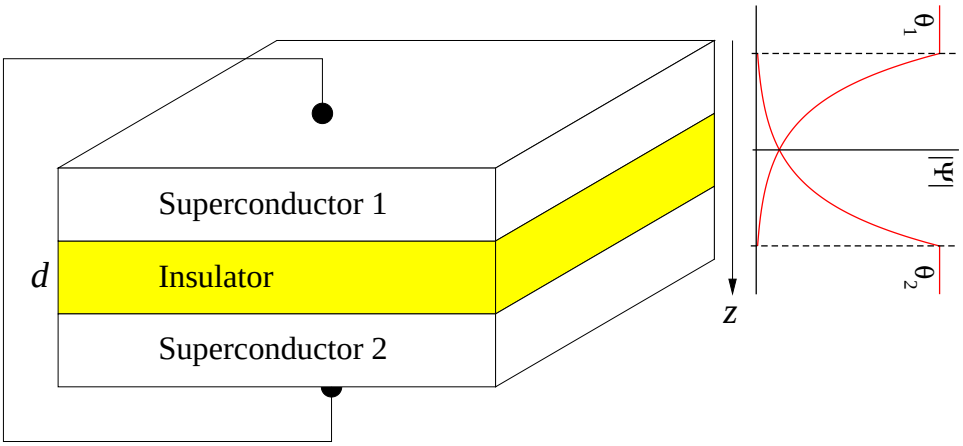

Figure 5.2: Model SIS junction. The wavefunction overlap is schematically shown by the red lines with the dash lines representing the superconductor-insulator interface.

The Josephson effect is a direct consequence of the overlap of the decaying ground state wavefunctions in the region between the superconductors, and independent of the material providing the separation. The effect may also be observed by replacing the insulating material with a non-superconducting metal or even a simple thin constriction between the slabs.

As a device, a single junction may be seen as a non-linear inductor with inductance

$$L_J = \frac{L_0}{\cos(\varphi)}, \tag{5.12}$$

with inductive amplitude

$$L_0 = \frac{\Phi_0}{2\pi I_c} = \frac{\hbar}{2eI_c}. \tag{5.13}$$

The potential energy stored in the junction is given by

$$U(\varphi) = \int I_s V dt = E_J \left[1 - \cos(\varphi)\right], \tag{5.14}$$

with characteristic Josephson energy

$$E_J = \frac{\hbar I_c}{2e} = \frac{\Phi_0 I_c}{2\pi}. \tag{5.15}$$

At non-zero voltages, we can write the current through the junction as a

sum of current components

$$I_t = I_s(\varphi) + I_n(V), \tag{5.16}$$

where I_n is a component of the current in addition to the supercurrent carried by non-bound electrons ("quasiparticles") created by thermal excitation of the Cooper pairs. Because it is carried by quasiparticles, $I_n(V)$ is referred to as the "normal" current and the junction acts as a nearly ideal Ohmic resistor for this current channel given by

$$I_n = G_n V, \tag{5.17}$$

with conductance $G_n = 1/R_n$ the inverse of the junction resistance. The normal current also becomes the dominant current component for applied voltages larger than the superconducting energy gap $eV > \Delta(T)$, independent of the system temperature. Combing the normal resistance and the natural current scale for the junction, we find the characteristic voltage scale

$$V_c = I_c R_n. \tag{5.18}$$

This voltage scale specifies a maximum rate of phase change through relation (5.11), or a characteristic frequency

$$\omega_c = \frac{2\pi}{\Phi_0} I_c R_n, \tag{5.19}$$

above which the junction's response begins to degrade. The resistance may be seen as a damping parameter, and the characteristic frequency

$$\omega_c L_0 = R_n, \tag{5.20}$$

may be viewed as just the inverse relaxation time. In the work below, we will consider Josephson junctions which are coupled in parallel with a shunting resistor R such that

$$R_n \gg R, \tag{5.21}$$

so the characteristic frequency is written only in terms of the junction resistance as

$$\omega_c = \frac{2\pi}{\Phi_0} I_c R. \tag{5.22}$$

The device also has a dynamic capacitance C, adding "displacement" cur-

rent component

$$I_d = C\frac{dV}{dt}, \tag{5.23}$$

for total junction current

$$I_t = I_s(\varphi) + I_n(V) + I_d(\dot{V}). \tag{5.24}$$

The classical energy of the electric field adds *kinetic* energy component of the Josephson charging energy

$$E_c = \frac{(2e)^2}{2C}, \tag{5.25}$$

$$K = \frac{C}{2}V^2 = \frac{Q^2}{2C}, \tag{5.26}$$

with total charge

$$Q = \int I_t\, dt = CV = \frac{2\pi C}{\Phi_0}\dot{\varphi}. \tag{5.27}$$

The device capacitance contributes kinetic term

$$K = \frac{1}{2}E_J\omega_p^{-2}\dot{\varphi}^2 \tag{5.28}$$

with the device plasma frequency

$$\omega_p^{-2} = L_0 C = \frac{\Phi_0 C}{2\pi I_c}. \tag{5.29}$$

So the total energy $E = K + U(\varphi)$ of a single Josephson junction, in terms of the phase as the sole principle variable, including damping is given

$$E = E_J\left[\frac{1}{2}\omega_p^{-2}\dot{\varphi}^2 + \omega_c^{-1}\dot{\varphi} + 1 - \cos(\varphi)\right]. \tag{5.30}$$

5.1.1 Mechanical Analogs

The dynamics of the shunted Josephson junction are easily conceptualized in terms of two simple mechanical analogs. The torque on a simple pendulum of mass m and length l (Fig. 5.3), in a uniform gravitational field is

$$Q = \mu\frac{d^2\varphi}{dt^2} = mgl\sin(\varphi) - \eta\frac{d\varphi}{dt}. \tag{5.31}$$

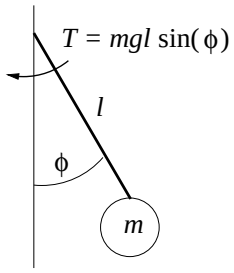

Figure 5.3: Pendulum Mechanical analog of a single Josephson junction.

This is exactly the same as Eq. (5.30) where the displacement from vertical corresponds to the phase difference, the angular velocity takes the place of the voltage and pendulum frequency

$$\omega_0 = \sqrt{\frac{g}{l}} \tag{5.32}$$

corresponds to the plasma frequency.

A second, perhaps more convenient, analogy is that of a classical particle with effective mass

$$M_0 = E_J \omega_p^{-2}, \tag{5.33}$$

moving in "washboard" potential

$$U(x) = M_0(1 - \cos(\varphi)). \tag{5.34}$$

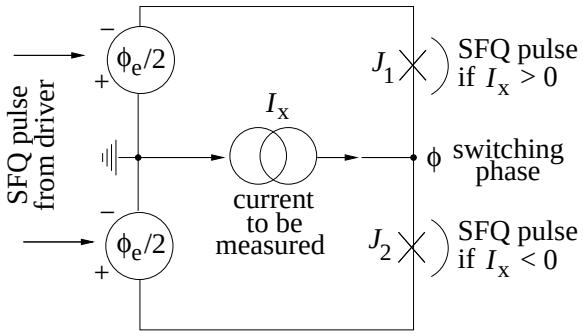

Figure 5.4: Balanced comparator circuit setup to measure high impedance current source.

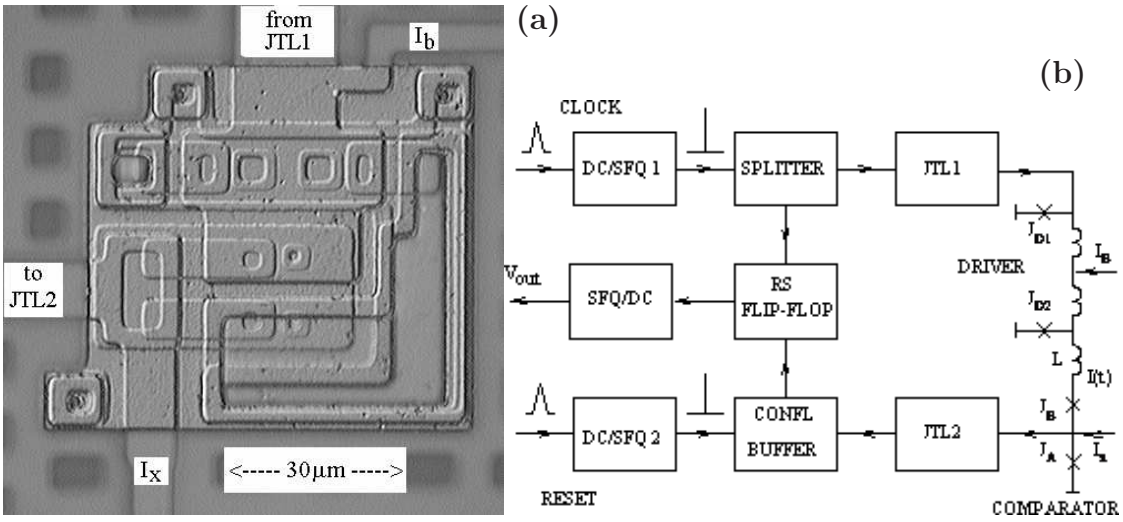

Figure 5.5: Experimental balanced Josephson junction comparator [115]. (a) STM image of the device and (b) its circuit schematic.

5.2 Comparator Circuit

A "balanced comparator" is composed of two identical shunted Josephson junctions biased in series by an external driver providing a source of phase difference $\varphi_e(t)$ (shown in Fig. 5.4). The junctions are connected through low inductance lines, and biased in parallel by the current I_x whose sign is to be measured. An experimental implementation and circuit schematic are shown in Fig. 5.5.

The device (essentially a dc SQUID [65]) has total potential energy

$$U(\varphi) = -2E_J \cos\left[\frac{\varphi_e(t)}{2}\right] - \frac{\hbar}{2e} I_x \varphi. \tag{5.35}$$

The net effect of the signal current I_x is to tilt the equivalent washboard potential, shown for a small value of I_x in Fig. 5.6. We assume the system has sufficient time to settle into an equilibrium state $\varphi = \varphi_i$. In the ideal case, nether I_x nor φ_e depend on the state of the comparator and the external driver then injects a single flux quantum ($\Delta\varphi_e = 2\pi$) into the superconducting loop created by the comparator and the driver's output stage. This pulse inverts the potential (5.35) creating a local instability for the phase. The phase must then settle into one of two adjacent states $\varphi_f = \varphi_i \pm \pi$. Because the loop is non-quantizing, this transient process creates a discrete SFQ (5.5) pulse at one of the junctions depending on the sign of I_x. This pulse is relatively easily detected [110], so the accuracy of the measurement is entirely determined by

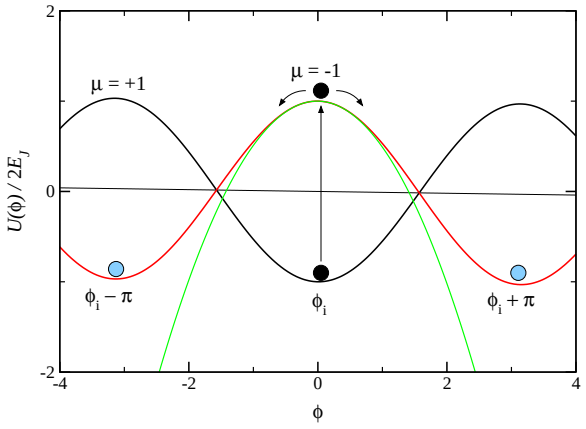

Figure 5.6: $U(\varphi)$ for $I_x = 0.01I_c$ and two values of injected phase difference $\mu = \cos(\varphi_e/2)$. The green line shows the linearized potential near $\varphi = \varphi_i$.

the evolution of the comparator. The dynamics of the phase may be seen the motion of a particle with effective mass

$$M_J = 2E_J\omega_p^{-2}, \tag{5.36}$$

moving in the tilted washboard board potential (5.35). In the absence of any fluctuations within the comparator, the particle will move either left or right depending entirely upon the sign of I_x. Fluctuations in the particles motion however create a non-zero probability that the particle will shift to the opposite minimum.

Let P_2 be the probability that the SFQ pulse is detected at junction J_2 (*i. e.* $\varphi_f = \varphi_i - \pi$). In the idealized case, the probability of switching would be given by the Heaviside step function

$$P_2 = \Theta(-I_x), \tag{5.37}$$

shown by the red line in Fig. 5.7. Fluctuations in the particle motion create a so called "gray zone", typically defined as

$$\Delta I_x \equiv \left| \frac{dP_2}{dI_x} \right|_{I_x=0}, \tag{5.38}$$

which quantifies the accuracy of the measurement. The width of this gray zone was characterized in experimental devices for single shot measurements [115], and later expanded for repeated measurements [116]. More recently the comparator gray zone was measured for devices with additional cooling of the

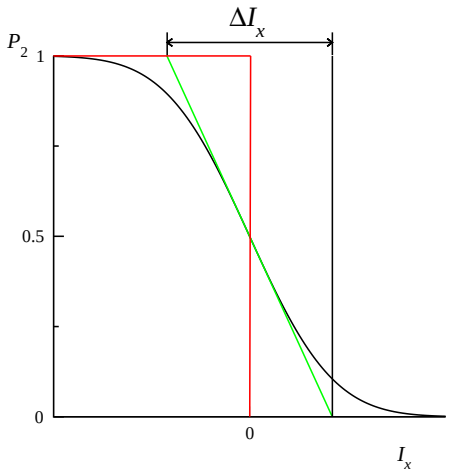

Figure 5.7: Switching probability P_2 versus current I_x. Red lined indicates perfect sign measurement.

shunt resistors [117]. A theoretical model [118, 119] was developed for an analytical form of the driver pulse $\varphi_e(t) \propto \exp(\kappa t)$ which is not quite relevant for practical devices.

In the thermal and quantum limits, the fluctuations of the current may characterized on the natural scales [118, 119]

$$\begin{array}{rcl} I_Q & \equiv & (2\pi/\Phi_0)\hbar\omega_p/2 = (2e/\hbar)\hbar\omega_p/2 = e\omega_p, \\ I_T & \equiv & (2\pi/\Phi_0)T = (2e/\hbar)T, \end{array} \tag{5.39}$$

with transition temperature [119]

$$T_{\text{trans}} = \hbar\omega_p/2. \tag{5.40}$$

We consider the case when these natural scales are small

$$\begin{array}{rcl} \gamma_Q \equiv I_Q/I_c = \hbar\omega_p/2E_J & \ll & 1, \\ \gamma_T \equiv I_T/I_c = T/E_J & \ll & 1, \end{array} \tag{5.41}$$

and the potential inversion is faster than the scale of system dynamics, which can readily be done with RSFQ circuitry [110]. Then the choice of the final state may be well described by the dynamics of the potential near $\varphi = \varphi_i$ and we may expand the potential in a Taylor series, keeping only the two leading terms

$$U(\varphi) = E_J\left(\mu(t)\varphi^2 - i_x\varphi + \text{ const}\right), \tag{5.42}$$

where

$$\mu(t) \equiv \cos\left[\frac{\varphi_e(t)}{2}\right], \tag{5.43}$$

$$i_x \equiv I_x/2I_c. \tag{5.44}$$

This expansion is shown by the green line in Fig. 5.6. The system may then be interpreted as motion of a damped time-dependent harmonic oscillator with frequency

$$\omega^2(t) = \omega_p^2\mu(t), \tag{5.45}$$

where $\mu(t)$ is switched rapidly from initial state $\mu_i \approx 1$ to inverted final state $-\mu_f = \mu_i$.

5.3 System Propagator

The probability of switching to state $\varphi_f = \varphi_i - \pi$, may be found as

$$P_2 = \lim_{t\to\infty} \int_{-\infty}^{\varphi_{\max}(t)} \rho(\varphi, \varphi, t)\, d\varphi, \tag{5.46}$$

where $\rho(\varphi, \varphi, t)$ are the diagonal elements of the system's density matrix [120] $\rho(\varphi, \varphi', t)$, and $\varphi_{\max}(t)$ is the coordinate of the potential maximum (5.42) after inversion. The density matrix at arbitrary time t may be expressed in terms of centralized coordinates

$$\begin{aligned} \eta &\equiv \varphi + \varphi', \\ \xi &\equiv \varphi - \varphi', \end{aligned} \tag{5.47}$$

$\rho(\varphi, \varphi, t) = (1/2)\rho(\eta, 0, t)$ from the initial matrix at time $t = 0$, traced over degrees of freedom of the environment, via the system propagator

$$\rho(\eta, 0, t) = \iint_{-\infty}^{\infty} J(\eta, 0, t|\eta_i, \xi_i, 0)\rho(\eta_i, \xi_i, 0)\, d\eta_i\, d\xi_i. \tag{5.48}$$

To find the system propagator, we use the approach of Caldeira and Leggett [121, 122] where the phase is coupled to an environment of linearly distributed oscillators. This theory, which provides a correct description for systems with externally shunted junctions yields expression for the propagator (see Eqs.

6.1-6.4 of [121]):

$$J(\varphi,\varphi',t)=\oint\oint\exp\left\{\frac{1}{\hbar}\left[\mathrm{i}S(\varphi,\varphi')-\Theta_T(\varphi,\varphi')\right]\right\}D\varphi\,D\varphi', \tag{5.49}$$

with action

$$S(\varphi,\varphi')=\int_0^t\mathcal{L}d\tau-\int_0^t M_J\gamma\varphi\dot{\varphi}d\tau+\int_0^t M_J\gamma\varphi'\dot{\varphi}'d\tau \tag{5.50}$$

expressed in terms of equivalent mass 5.36, damping parameter

$$\gamma=\frac{\omega_p^2}{2\omega_c}, \tag{5.51}$$

and thermal contribution given by

$$\begin{aligned}\Theta_T(\varphi,\varphi')=&\frac{2M_J\gamma}{\pi}\int_0^\Omega\nu\coth\left(\frac{\hbar\nu}{2T}\right)\\&\times\int_0^t\int_0^\tau(\varphi(\tau)-\varphi'(\tau))\cos(\nu(\tau-s))(\varphi(s)-\varphi'(s))\,dsd\tau d\nu.\end{aligned} \tag{5.52}$$

Here $\Omega\gg\omega_p$ is the maximum frequency of the bath oscillators and the dot represent differentiation over τ. The system Lagrangian may be written as a function of phase

$$\mathcal{L}=\frac{M_J}{2}\left[\omega_p^{-2}(\dot{\varphi}^2-\dot{\varphi}'^2)-2\omega_c^{-1}(\varphi\dot{\varphi}'-\dot{\varphi}\varphi')-\mu(\tau)(\varphi^2-\varphi'^2)+2i_x(\varphi-\varphi')\right]. \tag{5.53}$$

Finding the paths which minimize the action

$$\delta_x S=\frac{d}{dt}\frac{\partial\mathcal{L}}{\partial\dot{x}}-\frac{\partial\mathcal{L}}{\partial x}=0, \tag{5.54}$$

with terms

$$\begin{array}{rcl|rcl} \dfrac{\partial \mathcal{L}}{\partial \dot{\varphi}} & = & M_J\left(\omega_p^{-2}\dot{\varphi}+\omega_c^{-1}\varphi'\right) & \dfrac{\partial \mathcal{L}}{\partial \varphi} & = & -M_J\left(\mu(\tau)\varphi+\omega_c^{-1}\dot{\varphi}'-i_x\right) \\ \dfrac{\partial \mathcal{L}}{\partial \dot{\varphi}'} & = & -M_J\left(\omega_p^{-2}\dot{\varphi}'+\omega_c^{-1}\varphi\right) & \dfrac{\partial \mathcal{L}}{\partial \varphi'} & = & M_J\left(\mu(\tau)\varphi'+\omega_c^{-1}\dot{\varphi}\right), \end{array} \tag{5.55}$$

yields equations of motion

$$\begin{array}{rcl} \omega_p^{-2}\ddot{\varphi}+\omega_c^{-1}\dot{\varphi}+\mu(\tau)\varphi & = & 2i_x, \\ \omega_p^{-2}\ddot{\varphi}'+\omega_c^{-1}\dot{\varphi}+\mu(\tau)\varphi & = & 0. \end{array} \tag{5.56}$$

Changing to centralized coordinates (5.47), we find fundamental equations of motion

$$\omega_p^{-2}\ddot{\eta}+\omega_c^{-1}\dot{\eta}+\mu(\tau)\eta = 2i_x, \tag{5.57}$$

$$\omega_p^{-2}\ddot{\xi}-\omega_c^{-1}\dot{\xi}+\mu(\tau)\xi = 0. \tag{5.58}$$

Noting that

$$\varphi\dot{\varphi}=\frac{\partial}{\partial\tau}\frac{1}{2}\varphi^2, \tag{5.59}$$

we may readily re-write the Lagrangian and action in terms of the centralized coordinates, damping (5.51) and frequency (5.45) as

$$\begin{array}{rcl} \mathcal{L}(\eta,\xi) & = & \dfrac{M_J}{2}\left(\dot{\eta}\dot{\xi}-\gamma(\dot{\eta}\xi-\eta\dot{\xi})-\omega(\tau)\eta\xi+2i_x\xi\right), \\ S(\eta,\xi) & = & \displaystyle\int_0^t \mathcal{L}(\eta,\xi)d\tau-\frac{M_J\gamma}{2}\eta\xi\,\Big|_0^t. \end{array} \tag{5.60}$$

To analyze the effect of fluctuations, we represent coordinates η, ξ as a sum of classical trajectories which satisfy equations of motion (5.57, 5.58) and small deviations $\tilde{\eta}$, $\tilde{\xi}$

$$\begin{array}{rcl} \eta & = & \eta(\tau)+\tilde{\eta}(\tau), \\ \xi & = & \xi(\tau)+\tilde{\xi}(\tau), \end{array} \tag{5.61}$$

where

$$\begin{array}{rcl} \tilde{\eta}(0)=\tilde{\eta}(t) & = & 0, \\ \tilde{\xi}(0)=\tilde{\xi}(t) & = & 0. \end{array} \tag{5.62}$$

Expanding the action in terms (5.61), we see it may represented as a sum of

the classical action and variations

$$S = S_{\rm cl}(\eta,\xi) + \tilde{S}(\tilde{\eta},\tilde{\xi}) - 2i_x t\left(\tilde{\xi}_f - \tilde{\xi}_i\right), \tag{5.63}$$

where

$$\begin{aligned} S_{\rm cl}(\eta,\xi) &= \int_0^t \mathcal{L}(\eta,\xi)d\tau - \frac{M_J\gamma}{2}\left(\eta_f\xi_f - \eta_i\xi_i\right) && (5.64)\\ \tilde{S}(\tilde{\eta},\tilde{\xi}) &= \frac{M_J}{2}\int_0^t \left[\dot{\tilde{\eta}}\dot{\tilde{\xi}} - \gamma\left(\tilde{\xi}\dot{\tilde{\eta}} - \tilde{\eta}\dot{\tilde{\xi}}\right) - \omega^2(\tau)\tilde{\eta}\tilde{\xi}\right]d\tau, && (5.65)\end{aligned}$$

where η_i, ξ_i, η_f, ξ_f are the initial and final points of the classical trajectory.

Expanding the thermal contribution Eq. (5.52) in terms of the classical path and variation,

$$\Theta_T(\eta,\xi) = \frac{2M_J\gamma}{\pi}\int_0^\Omega \nu\coth\left(\frac{\hbar\nu}{2T}\right)\int_0^t\int_0^\tau \left(\xi_\tau + \tilde{\xi}_\tau\right)\left(\xi_s + \tilde{\xi}_s\right) \times\cos\left(\nu(\tau-s)\right)\, dsd\tau d\nu. \tag{5.66}$$

Writing only terms involving classical trajectories we may write the thermal contribution as sum $\Theta_T(\eta,\xi) = \Theta(\eta,\xi) + \Theta_F(\tilde{\eta},\tilde{\xi},\eta,\xi)$, where

$$\Theta(\eta,\xi) = \frac{2M_J\gamma}{\pi}\int_0^\Omega \nu\coth\left(\frac{\hbar\nu}{2T}\right)\int_0^t\int_0^\tau \xi(\tau)\xi(s)\cos\left(\nu(\tau-s)\right)\, dsd\tau d\nu. \tag{5.67}$$

To evaluate the path integral, we write the differential paths as

$$\begin{aligned} D\varphi D\varphi' &= D\varphi\wedge D\varphi',\\ &= \tfrac{1}{4}\left[(D\eta + D\xi)\wedge(D\eta - D\xi)\right], \end{aligned} \tag{5.68}$$

where we have written the paths in the Grassmann algebra as the wedge product ($\wedge$) of differential 1-forms.

Expanding over the classical path and variation, and using

$$\begin{aligned} D\tilde{\eta}\wedge D\tilde{\eta} &= 0,\\ D\tilde{\xi}\wedge -D\tilde{\xi} &= 0, \end{aligned} \tag{5.69}$$

the differential path may be written

$$\begin{aligned} D\varphi D\varphi' &= \tfrac{1}{4}\left[(D\tilde{\eta}\wedge -D\tilde{\xi}) + (D\tilde{\xi}\wedge D\tilde{\eta})\right], \\ &= \tfrac{1}{4}\left[(D\tilde{\xi}\wedge D\tilde{\eta}) + (D\tilde{\xi} + D\tilde{\eta})\right], \end{aligned} \tag{5.70}$$

or

$$D\varphi D\varphi' = \frac{1}{2} D\tilde{\eta} D\tilde{\xi}. \tag{5.71}$$

Collecting terms (5.63), (5.67), (5.49), the system propagator may be written

$$J(\eta,\xi) = F^2(t)\exp\left[\frac{\mathrm{i}}{\hbar}S_{\mathrm{cl}}(\eta,\xi) - \frac{1}{\hbar}\Theta(\eta,\xi)\right], \tag{5.72}$$

where

$$F^2(t) = \frac{1}{2}\oint\oint D\tilde{\eta}D\tilde{\xi}\exp\left[\frac{\mathrm{i}}{\hbar}\tilde{S}(\tilde{\eta},\tilde{\xi}) - \frac{1}{\hbar}\Theta_F(\tilde{\eta},\tilde{\xi},\eta,\xi)\right] \tag{5.73}$$

is simply a normalization factor to the final density matrix.

To evaluate the action, we solve equations of motion (5.57, 5.58). By coincidence, the same numerical conditions as section 3.1.3 apply here. Namely, we represent the solution as a Dirichlet boundary value problem with solutions

$$\begin{aligned} \eta(\tau,t) &= \eta_i a_1(\tau,t) + \eta_f a_2(\tau,t) + 2i_x a(\tau,t), \\ \xi(\tau,t) &= \xi_i b_1(\tau,t) + \xi_f b_2(\tau,t), \end{aligned} \tag{5.74}$$

where functions $a_{1,2}$, $b_{1,2}$ are solutions of the homogeneous equations with boundary conditions

$$\begin{array}{l|l} a_1(0) = b_1(0) = 1 & a_2(0) = b_2(0) = 0, \\ a_1(0) = b_1(0) = 0 & a_2(0) = b_2(0) = 1, \end{array} \tag{5.75}$$

and function $a(\tau,t)$ is the solution to Eq. (5.57) with unit right-hand side and boundary conditions $a(0,t) = a(t,t) = 0$.

Plugging solutions (5.74) into Lagrangian (5.53), and evaluating the action (5.64) we find

$$S_{\mathrm{cl}} = K_1\eta_i\xi_i + K_2\eta_f\xi_f - L\eta_i\xi_f - N\eta_f\xi_i + 2i_x\left(Q_1\xi_i + Q_2\xi_f\right), \tag{5.76}$$

where

$$\begin{pmatrix} K_1 \\ K_2 \end{pmatrix} = \frac{E_J}{\omega_p^2 \hbar} \int_0^t d\tau \left\{ \begin{pmatrix} \dot{a}_1 \dot{b}_1 \\ \dot{a}_2 \dot{b}_2 \end{pmatrix} - \omega^2(\tau) \begin{pmatrix} a_1 b_1 \\ a_2 b_2 \end{pmatrix} + \gamma \begin{pmatrix} a_1 \dot{b}_1 - \dot{a}_1 b_1 + 1 \\ a_2 \dot{b}_2 - \dot{a}_2 b_2 - 1 \end{pmatrix} \right\}, \tag{5.77}$$

$$\begin{pmatrix} N \\ L \end{pmatrix} = -\frac{E_J}{\omega_p^2 \hbar} \int_0^t d\tau \left\{ \begin{pmatrix} \dot{a}_2 \dot{b}_1 \\ \dot{a}_1 \dot{b}_2 \end{pmatrix} - \omega^2(\tau) \begin{pmatrix} a_2 b_1 \\ a_1 b_2 \end{pmatrix} + \gamma \begin{pmatrix} a_2 \dot{b}_1 - \dot{a}_2 b_1 \\ a_1 \dot{b}_2 - \dot{a}_1 b_2 \end{pmatrix} \right\}, \tag{5.78}$$

$$\begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} = \frac{E_J}{\omega_p^2 \hbar} \int_0^t d\tau \left\{ \begin{pmatrix} \dot{a}\dot{b}_1 \\ \dot{a}\dot{b}_2 \end{pmatrix} - \omega^2(\tau) \begin{pmatrix} a b_1 \\ a b_2 \end{pmatrix} + \gamma \begin{pmatrix} a \dot{b}_1 - \dot{a} b_1 \\ a \dot{b}_2 - \dot{a} b_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\}. \tag{5.79}$$

Using solutions (5.74),the term representing thermal contributions to the fluctuations (5.67), may be expressed as

$$\Theta(\eta, \xi) = C\xi_i^2 + B\xi_i\xi_f + A\xi_f^2, \tag{5.80}$$

where

$$\begin{pmatrix} A \\ B \\ C \end{pmatrix} = \frac{2E_J\gamma}{\pi\omega_p^2\hbar} \int_0^\Omega d\nu \nu \coth\left(\frac{\hbar\nu}{2T}\right) \int_0^t \int_0^\tau d\tau ds \cos\left[\nu\left(\tau - s\right)\right] \times \begin{pmatrix} b_2(\tau,t) b_2(s,t) \\ b_1(s,t) b_2(\tau,t) + b_1(\tau,t) b_2(s,t) \\ b_1(\tau,t) b_1(s,t) \end{pmatrix}. \tag{5.81}$$

Collecting all terms (5.77), (5.78), (5.79), (5.81), the system propagator is expressed as

$$J(\eta_f, \xi_f, t | \eta_i, \xi_i, 0) = F^2(t) \times \exp\left\{ \mathrm{i}\left[K_1\eta_i\xi_i + K_2\eta_f\xi_f - L\eta_i\xi_f - N\eta_f\xi_i + i_x\left(Q_1\xi_i + Q_2\xi_f\right) \right] - \left[A\xi_f^2 + B\xi_f\xi_i + C\xi_i^2 \right] \right\}. \tag{5.82}$$

Equation (5.82), represents the generalization of the harmonic oscillator propagator (see, Eq. 6.26 of Ref. [121]) to the case of arbitrary time dependence of the potential curvature $\mu(t)$. It shows that since the initial density matrix is Gaussian, with average phase $\langle\varphi_i\rangle$ and variance $\langle\tilde{\varphi}_i^2\rangle$, the final density matrix will remain Gaussian. In the centralized coordinate representation

$\xi_f = 0$, so the final density matrix for the system is entirely determined in terms of K_1, N, Q_1 and C. The other terms may affect the intermediate phase velocity, which is unimportant for determination of the final state gray zone width.

5.3.1 Numerical Evaluation of Parameters

For practical devices, the inversion function may be calculated numerically from circuit schematics using the Personal Superconducting Circuit Analyzer (PSCAN) package [123]. PSCAN is a CAD tool, developed at Stony Brook for the numeric simulation and analysis of superconducting circuits, focusing on RSFQ logic.

Functions $a_{1,2}(\tau, t)$, $b_{1,2}(\tau, t)$ and $a(\tau, t)$ are found using the same method as section 3.1.3, namely, the differential equation is solved numerically with a tri-diagonal system of equations. The integrals for parameters K_1, N, Q_1 are then calculated using standard Romberg's extension to trapezoidal quadrature [77]. To evaluate parameter C, we first notice that the limits of integration of the inside integral $[0, t]$, $[0, \tau]$ are just integration over half space $[0, t]$ and changing variables $\epsilon = \nu/\omega_p$

$$C = \frac{2E_J\gamma}{\pi\hbar} \int_0^{\Omega/\omega_p} \epsilon \coth\left(\frac{\hbar\omega_p}{2T}\epsilon\right) \int_0^t \int_0^t b_1(s,t) b_1(\tau,t) \cos\left[\epsilon\omega_p(\tau - s)\right] \, ds d\tau d\epsilon. \tag{5.83}$$

An integration by parts helps numerically with the discontinuity near $\nu \to 0$, and using boundary values $b_1(0) = 1$, $b_1(t) = 0$ we find

$$C = \frac{E_J\gamma}{\pi\hbar} \int_0^{\Omega/\omega_p} d\epsilon \, \coth\left(\frac{\hbar\omega_p}{2T}\epsilon\right) \left\{ \int_0^t b_1(\tau,t) \sin(\nu\tau) \, d\tau + \int_0^t \int_0^t b_1(\tau) \dot{b}_1(s) \sin\left[\omega_p\epsilon(\tau - s)\right] \, ds d\tau \right\}. \tag{5.84}$$

The 2-D and 3-D integrals of Eq. (5.84) are evaluated using Monte Carlo integration techniques. The shape of functions $b_{1,2}$ allows importance sampling algorithms to improve convergence times, so a variant of the VEGAS algorithm [77] is used. To insure high quality random sampling of the integral space, we use a multi-stream linear congruential (Lehmer) random number generator [124].

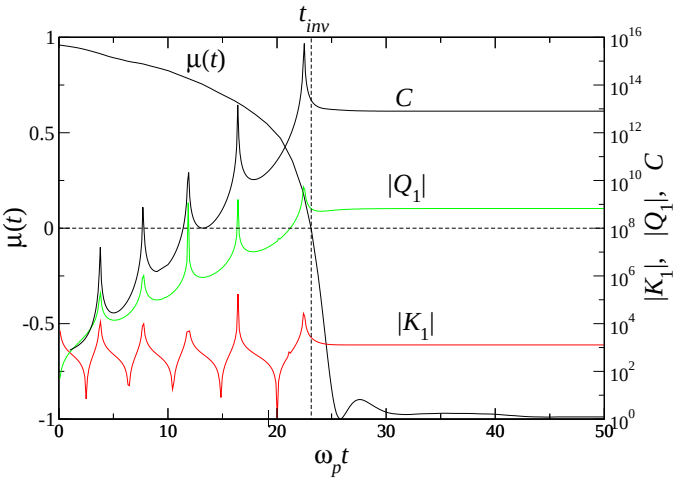

Figure 5.8: Calculated function $\mu(t)$ and numeric factors to the propagator and damping parameter $\beta_c = \omega_c^2/\omega_p^2 = 1$. The temperature for parameter C is taken $2T/\hbar\omega_p = 0.01$. Dashed lines represent the moment of inversion $t_{\rm inv}$.

It is more conventional to quote damping in terms of Stewart-McCumber [125, 126] damping parameter

$$\beta_c \equiv \left(\frac{\omega_c}{\omega_p}\right)^2 = \frac{\omega_p}{2}\gamma^{-1}, \tag{5.85}$$

which is proportional to the shunting resistance R^2. For devices with small capacitance ($\omega_p \gg \omega_c$, high damping devices) $\beta_c \gg 1$. The calculated parameters K_1, Q_1, C for critical damping $\beta_c = 1$ are shown in Fig. 5.8 for numerically calculated switching $\mu(t)$. Before inversion time $t_{\rm inv}$, the parameters represent the integrated effect of free oscillations in the non-inverted potential well. The peaks correspond to times of evaluation t when the phase has reached the potential minimum. Thermal term C and the term representing the inhomogeneous solution of the tilted potential Q_1 acquire exponential amplitude through their respective interactions. The system dynamics are resolved quickly on the time scale of ω_p, for times after potential inversion $t - t_{\rm inv} > 0$. After inversion and this short transient the final phase state is effectively decided and the parameters become independent of the evaluation time.

5.3.2 Quantum and Thermal Limits

In the quantum and thermal limits, parameter C takes on much simpler expressions. When $T \gg \hbar\omega_p$, we may write

$$\coth\left(\frac{\hbar\nu}{2T}\right) \to \frac{2T}{\hbar\nu}, \tag{5.86}$$

and Eq. (5.81) becomes

$$C = \frac{4E_J\gamma T}{\pi\omega_p^2\hbar^2}\int_0^t\int_0^\tau b_1(s)b_1(\tau)\int_0^\Omega \cos\left[\nu(\tau-s)\right]\, d\nu ds d\tau. \tag{5.87}$$

Using the relation (see appendix A.5)

$$\lim_{\Omega\to\infty}\int_0^\Omega \cos\left[\nu(\tau-s)\right] d\nu = \pi\delta(\tau-s), \tag{5.88}$$

and writing in terms of natural scale of quantum fluctuations $2E_J\gamma/\omega_p^2\hbar = 1/(2\gamma_Q\sqrt{\beta_c})$, integral (5.87) is evaluated as

$$\lim_{T\gg\hbar\omega_p} C = \frac{T}{\hbar\gamma_Q\beta_c^{1/2}}\int_0^\infty b_1^2(\tau,t)\, d\tau. \tag{5.89}$$

In the opposite limit, as $T \to 0$

$$\coth\left(\frac{\hbar\nu}{2T}\right) \to 1, \tag{5.90}$$

and the integration over ν in Eq. (5.81) readily yields

$$C = \frac{2E_J\gamma}{\pi\omega_p^2\hbar}\int_0^t\int_0^\tau b_1(s)b_1(\tau)\left[\frac{\cos[\Omega(\tau-s)]-1}{(\tau-s)^2} + \Omega\frac{\sin[\Omega(\tau-s)]}{(\tau-s)}\right] ds d\tau. \tag{5.91}$$

Integrating the ds integral by parts and again using $b_1(0) = 1$, Eq. (5.91) is

evaluated

$$\lim_{T\to 0} C = \frac{1}{2\pi\gamma_Q\beta_c^{1/2}}\left\{\int_0^t \frac{b_1(\tau)\left[\cos(\Omega\tau)-1\right]}{\tau} + \int_0^t\int_0^\tau \frac{b_1(\tau)\dot{b}_1(s)\left(\cos[\Omega(\tau-s)]-1\right)}{\tau-s}\,dsd\tau\right\}. \quad (5.92)$$

5.4 Grey Zone Width

The initial density matrix may be found from Wigner's transformation of the classical Gibbs distribution for a system in thermal equilibrium [118, 119],

$$\rho(\eta_i,\xi_i,0) = \frac{1}{\sqrt{4\pi\langle\varphi_i\rangle^2}}\exp\left[-\frac{(\eta_i-2i_x/\mu_i)^2}{16\langle\varphi_i\rangle^2} - \beta_c\gamma_Q^{-2}\langle\dot{\varphi}_i^2\rangle\xi_i^2\right], \quad (5.93)$$

where we have used quantum fluctuation scale (5.41), and initial distributions

$$\begin{aligned}\langle\tilde{\varphi}_i^2\rangle &= \frac{\gamma_Q}{2\pi\beta_c^{1/2}}\int_0^{\Omega/\omega_p}\coth\left(\frac{\hbar\omega_p}{2T}x\right)\frac{x}{\left(\mu_i-x^2\right)^2+\beta_c^{-1}x^2}\,dx,\\ \langle\dot{\varphi}_i\rangle &= \frac{\gamma_Q}{2\pi\beta_c^{3/2}}\int_0^{\Omega/\omega_p}\coth\left(\frac{\hbar\omega_p}{2T}x\right)\frac{x^3}{\left(\mu_i-x^2\right)^2+\beta_c^{-1}x^2}\,dx,\end{aligned} \quad (5.94)$$

found from the frequency response of a driven harmonic oscillator

$$\alpha''(\nu)\propto\frac{\nu}{(\mu_1-\nu)^2+\beta_c^{-1}\nu^2}, \quad (5.95)$$

within the fluctuation-dissipation theorem (FDT) [127].

Using expressions (5.93), (5.82) and carrying out Gaussian integral (5.48), the density may be expressed in terms of coordinate phase φ as

$$\rho(\varphi,\varphi,t) = \frac{F^2(t)}{(4\pi\langle\tilde{\varphi}^2\rangle)^{1/2}}\exp\left\{-\frac{(\varphi-\langle\varphi\rangle)^2}{4\langle\tilde{\varphi}^2\rangle}\right\}, \quad (5.96)$$

with average phase and variance

$$\begin{aligned} \langle\varphi\rangle &= \frac{K_1\mu_i^{-1}+Q_1}{2N} i_x, \\ \langle\tilde{\varphi}^2\rangle &= \frac{C+4K_1^2\langle\tilde{\varphi}_i^2\rangle+\beta_c\hbar^2\gamma_Q^{-2}\langle\dot{\varphi}_i^2\rangle}{4N^2}, \end{aligned} \tag{5.97}$$

where we have used the relations $\eta_f = 2\varphi$, $\eta_i = 0$, and $\mu_i\varphi_i = 2i_x$.

The probability of switching may then be calculated

$$P_2 = \pi^{-1/2} \int_{-\infty}^{-\langle\varphi\rangle/2\langle\tilde{\varphi}^2\rangle} e^{-y^2} dy, \tag{5.98}$$

with expression for the width of smearing

$$\Delta I_x = \lim_{t\to\infty} \frac{(2\pi\langle\tilde{\varphi}^2\rangle)^{1/2}}{\left|\frac{d}{dI_x}\langle\varphi\rangle\right|}, \tag{5.99}$$

which gives final expression

$$\Delta I_x = 2\pi^{1/2} I_c \frac{\left[C+4K_1^2\langle\tilde{\varphi}_i^2\rangle+\beta_c\hbar^2\gamma_Q^{-2}\langle\dot{\varphi}_i^2\rangle\right]^{1/2}}{\left|K_1\mu_i^{-1}+Q_1\right|}. \tag{5.100}$$

In the limit of low damping ($\beta_c \gg 1$) and the potential is inverted instantaneously, the solutions to equations of motion (5.57, 5.58) are trivial and Eq. (5.100) is readily reduced to the following limiting expressions [119]

$$\Delta I_x = (\pi\gamma_T)^{1/2}\left(\frac{\mu_i\mu_f}{\mu_i+\mu_f}\right)^{1/2}, \quad T \gg \hbar\omega_p, \tag{5.101}$$

$$\Delta I_x = (\pi\gamma_Q)^{1/2}\left(\frac{\mu_i\mu_f}{\mu_i+\mu_f}\right)^{1/2}, \quad T \to 0. \tag{5.102}$$

The temperature dependence of the gray zone is shown in Fig. 5.9(a) for several damping parameters β_c. The dotted line shows the limit of $\beta_c \to \infty$, for instantaneous switching of the potential. At high temperatures $\Delta I_x \propto T^{1/2}$, shown by the dashed lines, Eq. (5.89) due to thermal fluctuations. In the opposite limit, quantum fluctuations of the phase saturate the gray zone width as $T \to 0$. This effect may be made more clear by interpretation of the terms of the Gaussian propagator which contribute to the smearing width (5.100). K_1

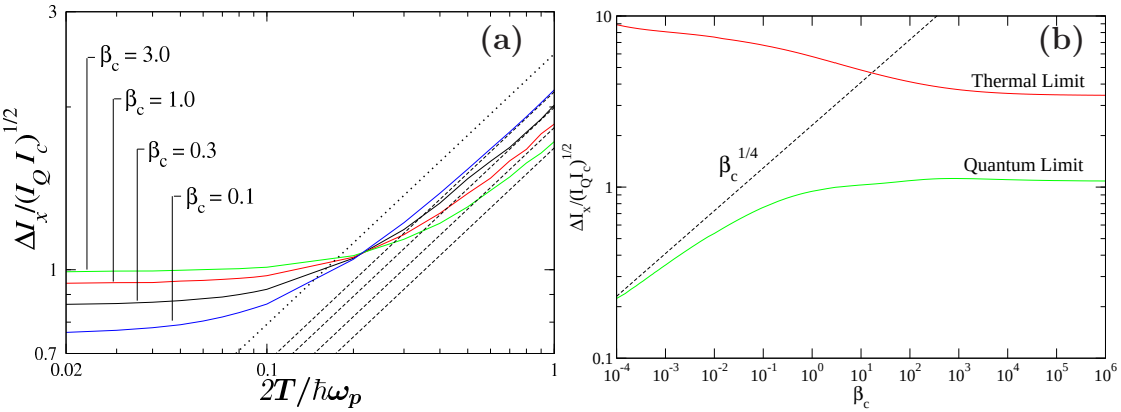

Figure 5.9: Dependence of ΔI_x on temperature (a) for $\mu(t)$ shown in Fig. 5.8 for several values β_c. Dashed lines represent the thermal limit. Dotted black line shows the thermal limit for instantaneous switching, Eq. (5.101). ΔI_x versus damping (b) β_c in the thermal and quantum limits. Dashed line represents limiting expression $\beta_c^{1/4}$.

represents the impact of the initial state of the system, Q_1 the evolution of the average phase in the tilted potential and C fluctuations from environmental coupling. When the point of potential inversion $t_{\rm inv}$ and the measurement time $t - t_{\rm inv}$ are much longer than the oscillator's reciprocal bandwidth

$$\Delta\omega^{-1} = \max\left[\omega_c^{-1}, \gamma^{-1}\right], \tag{5.103}$$

then $C, |Q_1|^2 \gg K_1^2$ and the final switching probability is independent of the initial state of the system and the gray zone may be expressed very simply as

$$\Delta I_x = 2I_c\pi^{1/2}\frac{C^{1/2}}{|Q_1|}. \tag{5.104}$$

The quantum and thermal limits also display different behaviors with damping parameter β_c. Panel (b) of Fig. 5.9 shows this dependence explicitly for the quantum and thermal limits, Eqs. (5.92, 5.89). In the limit of high damping case ($\beta_c \to 0$), thermal fluctuations saturate with parameter β_c while quantum fluctuations tend to grow as $\beta_c^{1/4}$, shown by the dashed line while in the opposite limit of low dissipation ($\beta_c \to \infty$), both temperature limits saturate with the damping parameter. These effects may be summarized in the following simple model: Eq. (5.99) suggests that sufficiently far from the inversion time $t_{\rm inv}$, the average phase equals the root-mean square of the thermal equilibrium phase noise, which roughly estimates gray zone width ΔI_x.

The phase noise may be estimated as an equivalent current noise source (*e. g.*, see Eq. (1.59) of Ref. [6]) with equilibrium spectral density calculated from the fluctuation-dissipation theorem [127]

$$S_I(\nu) = \frac{4}{\pi R}\frac{\hbar\nu}{2}\coth\left(\frac{\hbar\nu}{2T}\right), \tag{5.105}$$

which acts on a time-independent oscillator within available bandwidth (5.103). In the thermal ($\coth() \to 2T/\hbar\nu$) and quantum ($\coth() \to 1$) limits this yields the Johnson-Nyquist [6] $S_I(\nu) = 4T/\pi R = \text{const}$ and quantum $S_I(\nu) = 2\hbar\nu/\pi R$ spectral densities, respectively. In the latter case this yields in the high damping limit

$$\Delta I_x \propto \left[\frac{1}{\omega_c}\int_0^{\omega_c}\nu d\nu\right]^{1/2} \propto \beta_c^{1/4}. \tag{5.106}$$

5.4.1 Comparison with Experiment

To compare these results with experimental data, we need to account for the practical dependence of $I_c(T)$ on the system temperature. We scale the junction critical current according to the Ambegaokar-Baratoff theory [128]

$$I_c(T) = \frac{\pi\Delta(T)}{2eR}\tanh\left(\frac{\Delta(T)}{2T}\right). \tag{5.107}$$

The calculated smearing width, scaled by Eq. (5.107) is shown by the dashed lines in Fig. 5.10 compared with two sets of experimental devices. Both experimental devices are critically damped ($\beta_c = 1$) niobium-trilayer ($Nb/AlO_x/Nb$) Josephson junctions with critical current $I_c|_{T=4.2K} = 145\mu A$. The triangles represent junctions with critical current density $j_c = 1$ kA/cm^2 and plasma frequency $\omega_p^{-1} \approx 1.01$ ps [115]. The squares represent devices with higher critical current density $j_c = 5.5$ kA/cm^2 ($\omega_p^{-1} \approx 0.47$ ps) [116]. Contributions from external noise sources was ruled out by separate comparator circuits on the same chip driven with softer driver pulses. The deviation of the two points circled in red are likely caused by local self-heating of the sample chip. Hence, without the benefit of a single fitting parameter, the calculated gray zone width, Eq. (5.100) agrees nearly perfectly with the experimental data. While the system temperatures for the data shown were not taken low enough to truly saturate in the quantum limit, we believe this is strong evidence to indicate that signal resolution of the comparator circuit, when operated in the sub-1 Kelvin regime is limited only by the natural quantum mechanical fluctuations within this device. In fact, recent experiments [129] have demonstrated

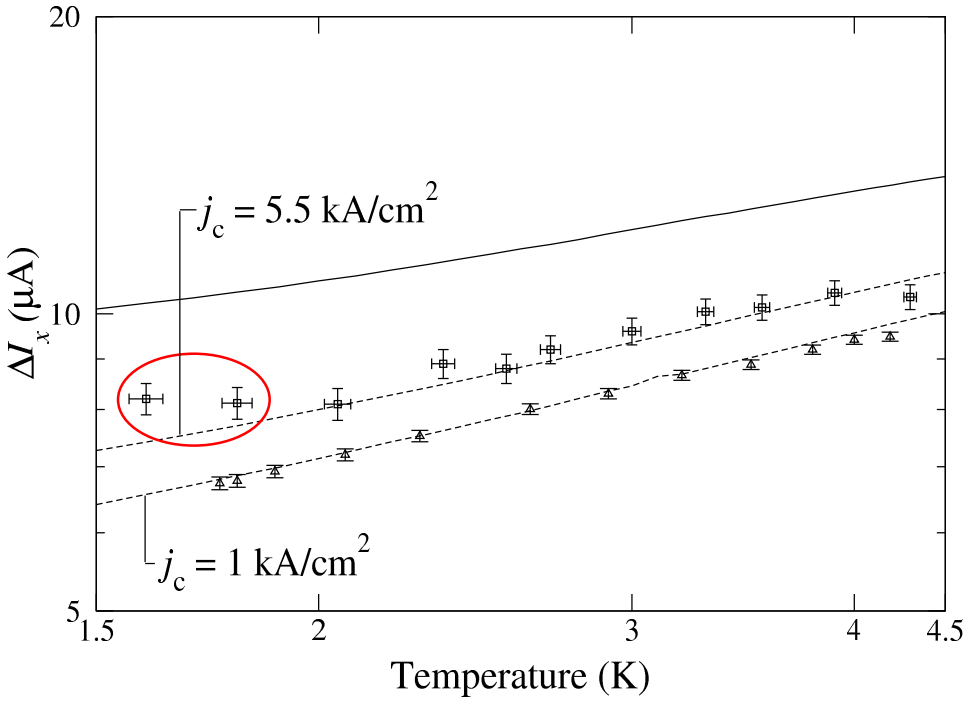

Figure 5.10: Temperature dependence of ΔI_x. Points show experimental data [115, 116] and dashed lines show numeric results including temperature dependence of I_c. Solid line represents the limit of instantaneous change from $\mu_i = 1$ to $-\mu_f = -1$.

the quantum saturation limit predicted by this theory.

Furthermore, the resolution may be increased by driving the device with softer pulses. The solid black line indicates the results calculated in the limit of instantaneous inversion of the potential. This result is natural because the fluctuations are essentially determined by their equilibrium initial values, frozen at the moment of inversion. For softer pulses, the phase is allowed more time to evolve within the slope defined by signal I_x. In fact, it may be shown that in the limit $\mu_f \to 0$, the width of the gray zone is reduced $\Delta I_x \to 0$ [119]. This limit is a bit artificial however because the inflation stage required for the creation of the output SFQ is destroyed and no output signal could be measured.

Chapter 6

Comparator for Flux Qubit Readout

The results of chapter 5 give a quantitatively correct description of the dynamics within the comparator circuit. Unfortunately, calculation of the parameters K_1, Q_1 and especially C are a bit bulky. These results do not extend readily for a description of the comparator system when the measurement may impact the signal source. For example, when the comparator may be used for measurement while inductively coupled to a magnetic flux qubit, shown schematically in Fig. 6.1(a).

In it's most basic configuration, a flux qubit is a superconducting loop interrupted by a single junction. Including an externally applied bias flux Φ_b, the qubit potential is written [130]

$$U_q = -E_J \cos\left(\frac{2\pi}{\Phi_0}\Phi\right) + \frac{(\Phi - \Phi_b)^2}{2L_q}, \tag{6.1}$$

with the self-inductance of the loop L_q. When the loop inductance is large $L_q \gg E_J/(\Phi_0/2\pi)^2$ and the applied bias $\Phi_b \approx \Phi_0/2$, then this potential forms a double well potential near $\Phi \approx \Phi_0/2$. In the quantum limit, only the lowest level of each well is occupied and the system reduces to a two level system with tunable interaction strength between the states.

Comparator fluctuations are represented as a contribution from equilibrium noise sources $I_{f1}(t)$, $I_{f2}(t)$. Such solid state qubit implementations are being actively pursued experimentally [108, 131] where a common device acts as a superposition of distinct flux states [132]. These individual elements will likely be inductively coupled [133], possibly in a controllable way [134, 135], to potentially build a quantum information processor [4]. Readout of the quantum states of each qubit will also likely be done through inductive coupling and the

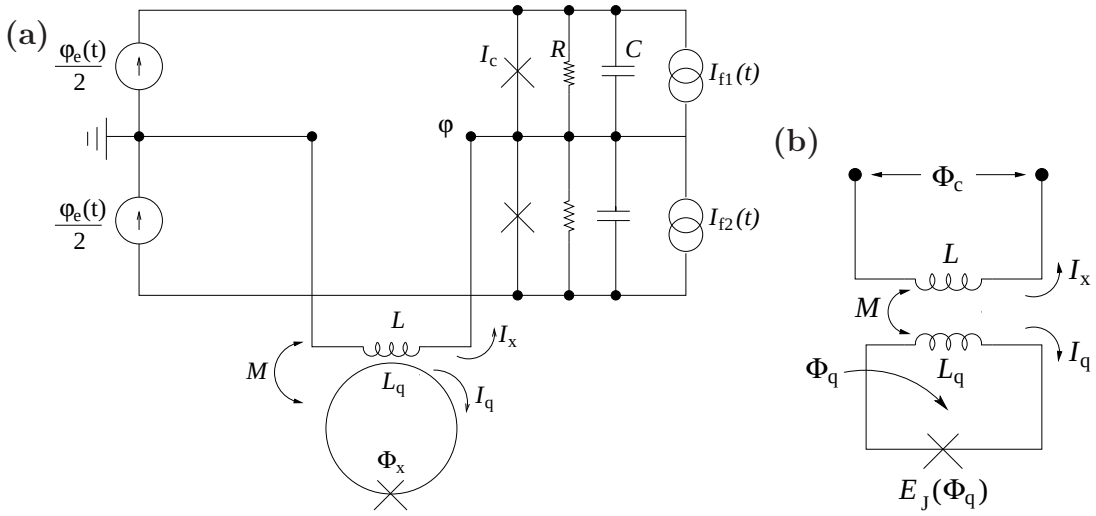

Figure 6.1: Equivalent circuit of a comparator-qubit system

potential quantum limited sensitivity of the balanced comparator makes it an attractive option for high accuracy, rapid, single-shot measurement [112].

To derive the Hamiltonian for the coupled system, we consider the simplified circuit 6.1(b). The total Hamiltonian may be represented as a sum of components

$$\mathcal{H} = \mathcal{H}_c + \mathcal{H}_q + \mathcal{H}_{\text{int}}, \tag{6.2}$$

where $\mathcal{H}_c$ describes the comparator and $\mathcal{H}_q + \mathcal{H}_{\text{int}}$ the qubit and interaction terms. The energy components of the mutual interaction may be presented in the standard way as

$$\mathcal{H}_{\text{coupled}} = E_J(\Phi_q) + \frac{LI_x^2}{2} + \frac{L_q I_q^2}{2} + MI_xI_q. \tag{6.3}$$

The flux in the comparator and qubit parts is then

$$\begin{aligned} \Phi_c &= LI_x + MI_q, \\ \Phi_q &= L_qI_q + MI, \end{aligned} \tag{6.4}$$

which can be presented as linear system

$$\begin{pmatrix} L & M \\ M & L_q \end{pmatrix} \begin{pmatrix} I_x \\ I_q \end{pmatrix} = \begin{pmatrix} \Phi_c \\ \Phi_q \end{pmatrix}, \tag{6.5}$$

whose solution yields device currents

$$\begin{aligned} I_x &= \frac{1}{LL_q(1-k^2)}\left(\Phi_c L_q - M\Phi_q\right), \\ I_q &= \frac{1}{LL_q(1-k^2)}\left(\Phi_q L - M\Phi_c\right). \end{aligned} \tag{6.6}$$

Here,

$$k^2 \equiv M^2/LL_q, \tag{6.7}$$

is the comparator-qubit mutual coupling parameter. Plugging these solutions into Eq. (6.3) we find

$$\mathcal{H}_{\text{coupled}} = \underbrace{E_J(\Phi_q) + \frac{\Phi_q^2}{2L_q(1-k^2)}}_{*} - \underbrace{\frac{M}{LL_q(1-k^2)}\Phi_c\Phi_q}_{**} + \underbrace{\frac{\Phi_c^2}{2L(1-k^2)}}_{***}. \tag{6.8}$$

The last term (***) may be seen as simply an addition term in the comparator's part of the total Hamiltonian and the first two terms (*), (**) are the sum of the qubit and interaction energies which may be written as [132]

$$\mathcal{H}_q + \mathcal{H}_{\text{int}} = \epsilon\sigma_z + \Delta\sigma_x - \kappa\sigma_z\Phi_q\varphi, \tag{6.9}$$

where σ_i are the Pauli matrices, Δ is the tunnel coupling between the two qubit states and ϵ is their energy bias. The qubit flux amplitude $\Phi_x = \sigma_z\Phi_q < \Phi_0/2$ and κ is the coupling coefficient

$$\kappa = \left(\frac{\Phi_0}{2\pi}\right)\frac{M}{LL_q}. \tag{6.10}$$

We have restricted ourselves to the case of low coupling ($k \ll 1$), so that the renormalization of the terms due to the mutual interaction may be ignored. For a symmetric qubit $\epsilon \to 0$ and since the natural frequency scale of the qubit evolution

$$\hbar\omega_q = \epsilon^2 + \Delta^2, \tag{6.11}$$

is much lower than that of the comparator, we may take $\Delta = 0$. Thus, the Heisenberg equation of motion, for generic operator $\mathcal{A}$

$$\dot{\mathcal{A}} = \frac{1}{\mathrm{i}\hbar}\left[\mathcal{A}, \mathcal{H}\right], \tag{6.12}$$

which for the qubit gives

$$\begin{aligned} \mathrm{i}\hbar\dot{\sigma}_z &= [\sigma_z, \mathcal{H}], \\ &= [\sigma_z, \mathcal{H}_{\mathrm{int}}], \\ &= [\sigma_z, \kappa\sigma_z\Phi_q\varphi], \\ &= 0. \end{aligned} \tag{6.13}$$

Independent of the measurement, the qubit remains in a definite flux state $\langle\sigma_z\rangle = \pm 1$ and the qubit signal $\kappa\sigma_z\Phi_q \propto \pm\Phi_x$ is fixed. These assumptions do not prevent of the comparator from affecting the qubit through the measurement back-action. The comparator may still alter the σ_x and σ_y components of the qubit energy. Defining

$$\sigma_\pm \equiv \sigma_x \pm \mathrm{i}\sigma_y, \tag{6.14}$$

we find equations of motion for the off-diagonal terms

$$\begin{aligned} \mathrm{i}\hbar\dot{\sigma}_\pm &= [\sigma_\pm, \mathcal{H}], \\ &= \kappa\Phi_q\varphi, ([\sigma_x, \sigma_y] \pm \mathrm{i}[\sigma_y, \sigma_z]), \\ &= \mp 2\kappa\Phi_q\varphi\sigma_\pm, \end{aligned} \tag{6.15}$$

or time evolution

$$\sigma_\pm(t) = \sigma_\pm(0) \exp\left\{ \pm\mathrm{i}\frac{2\kappa\Phi_q}{\hbar} \int_0^t \varphi(t')\, dt' \right\}. \tag{6.16}$$

6.1 Langevin-Heisenberg-Lax Model

The kinetic energy of the comparator is $K = Q^2/4C$, half of Eq. (5.26) due to the two Josephson junctions. The equation of motion of the comparator flux $\dot{\Phi} = \partial\mathcal{H}_c/\partial Q$ yields relations

$$\begin{aligned} \dot{\Phi} &= \frac{Q}{2C}, \\ \frac{Q^2}{4C} &= C\left(\frac{\Phi_0}{2\pi}\right)^2 \dot{\varphi}^2, \\ \frac{Q^2}{4C} &= \omega_p^{-2} E_J \dot{\varphi}^2. \end{aligned} \tag{6.17}$$

The Hamiltonian of the comparator may be written as

$$\mathcal{H}_c = \omega_p^{-2} E_J \dot{\varphi}^2 + U(\varphi, t) + \left(\frac{\Phi_0}{2\pi}\right) \varphi I_R(\boldsymbol{q}) + \mathcal{H}_d(\boldsymbol{q}), \tag{6.18}$$

where $I_R(\boldsymbol{q})$ is the total dissipative current that couples to the environment $\mathcal{H}_d(\boldsymbol{q})$ of $\boldsymbol{q}$ degrees of freedom. The potential energy may be written

$$U(\varphi, t) = -2E_J \cos(\varphi) \cos\left(\frac{\varphi_e(t)}{2}\right) + E_J \left(\frac{\Phi_0}{2\pi}\right)^2 \frac{\varphi^2}{2L}. \tag{6.19}$$

Defining inductive parameter for the comparator

$$\lambda \equiv E_J \left(\frac{\Phi_0}{2\pi}\right)^2 L, \tag{6.20}$$

we see that if $\lambda > 1/2$, the SFQ pulse ($\Delta\varphi_e = 2\pi$) injected from the driver circuit again inverts the potential near the initial phase equilibrium point $\varphi = \varphi_i$, and the transient process which creates the SFQ pulse at the output stage $J_{1,2}$ is determined by the sign of signal current $I_x = \sigma_z \Phi_x M / L L_q$.

In the same manner as section 5.2, we assume the driver circuit providing the pulse for inversion of the potential is sufficiently fast so that near the initial phase equilibrium point the potential is approximately quadratic, and the potential energy written

$$U(\varphi, t) = E_J \varphi^2 \mu(t). \tag{6.21}$$

The comparator inductance may then be seen as nothing more than a renormalization of the time dependent driver pulse (5.43)

$$\mu(t) \equiv \cos\left(\frac{\varphi_e(t)}{2}\right) + \frac{1}{2\lambda}. \tag{6.22}$$

In the case of the comparator, the time derivative $\dot{\varphi} \propto p$ may be seen as the equivalent momentum operator, hence the equation of motion for the phase operator (6.12) is linear and may be found from general commutator properties

$$[\mathcal{A}, \mathcal{B}\mathcal{C}] = [\mathcal{A}, \mathcal{B}]\,\mathcal{C} + \mathcal{B}\,[\mathcal{A}, \mathcal{C}]\,. \tag{6.23}$$

Including the external noise source, the equation of motion for the phase operator may be written as [136, 137]

$$\ddot{\boldsymbol{\varphi}} + 2\gamma\dot{\boldsymbol{\varphi}} + \omega^2(t)\boldsymbol{\varphi} = i_x + i_f(t). \tag{6.24}$$

Here, the dot represents differentiation over time t, $\gamma \equiv \omega_p^2/2\omega_c$ (5.51), $\omega^2(t) \equiv \omega_p^2\mu(t)$ (5.45), $i_x \equiv I_x/2I_c$ (5.44), and

$$i_f(t) \equiv \frac{I_{f1}(t) + I_{f2}(t)}{2I_c}. \tag{6.25}$$

While the $\boldsymbol{\varphi}$, i_x, $i_f(t)$ are operators in the equation of motion (6.24), the linearity of the equation allows us to represent the phase as a sum of its statistical average $\langle\varphi\rangle$ and fluctuations $\tilde{\varphi}$, with equations of motion:

$$\langle\ddot{\varphi}\rangle + 2\gamma\langle\dot{\varphi}\rangle + \omega^2(t)\langle\varphi\rangle = i_x, \tag{6.26}$$

$$\ddot{\tilde{\varphi}} + 2\gamma\dot{\tilde{\varphi}} + \omega^2(t)\tilde{\varphi} = i_f(t). \tag{6.27}$$

The general solution to (6.26) may be presented as

$$\langle\varphi(t)\rangle = i_x a(t) + \int_0^t i_x K(t,\tau), \tag{6.28}$$

or

$$\langle\varphi(t)\rangle = \langle\varphi_i\rangle \left[a(t) + \mu_i \int_0^t K(t,\tau)d\tau \right], \tag{6.29}$$

for constant initial phase $\langle\varphi_i\rangle = \langle\varphi(0)\rangle$,

$$\langle\varphi_i\rangle = i_x/\mu_i. \tag{6.30}$$

Function $a(t)$ is the solution to the homogeneous form of (6.26) which obeys boundary conditions

$$\begin{aligned} a(0) &= 1, \\ \dot{a}(0) &= 0, \end{aligned} \tag{6.31}$$

and kernel $K(t,\tau)$ is obtained from integration of the equation with right-hand part $\omega_p^2\delta(t-\tau)$, and may be seen as the solution to the homogeneous equation with zero phase for time $t < \tau$ and unit derivative at time $t = \tau$,

$$\begin{aligned} K(t \le \tau, \tau) &= 0, \\ \dot{K}(\tau,\tau) &= \omega_p^2, \end{aligned} \tag{6.32}$$

shown schematically in Fig. 6.2. In the same manner we may write the solution

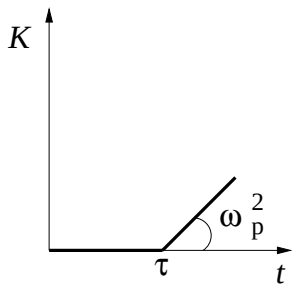

Figure 6.2: Kernel function $K(t,\tau)$ shown schematically.

for the fluctuations

$$\tilde{\varphi}(t) = \tilde{\varphi}_i a(t) + \dot{\tilde{\varphi}}_i b(t) + \int_0^t K(t,\tau) i_f(\tau) d\tau, \tag{6.33}$$

for the same functions $a(t)$, $K(t,\tau)$ and solution of the homogeneous equation $b(t)$ with boundary conditions

$$\begin{aligned} b(0) &= 0, \\ \dot{b}(0) &= 1. \end{aligned} \tag{6.34}$$

Since $\tilde{\varphi}_i$ and $\dot{\tilde{\varphi}}_i$ are uncorrelated, the variance of the phase fluctuation term is then calculated

$$\langle \tilde{\varphi}^2(t) \rangle = \langle \tilde{\varphi}_i^2 \rangle a^2(t) + \langle \dot{\tilde{\varphi}}_i^2 \rangle b^2(t) + \iint_0^t K(t,\tau) K(t,s) \langle i_f(\tau) i_f(s) \rangle \, d\tau ds. \tag{6.35}$$

The term $\langle i_f(\tau) i_f(s) \rangle$ is interpreted to mean

$$\langle i_f(\tau) i_f(s) \rangle = \frac{1}{2} \langle i_f(\tau) i_f(s) + i_f(s) i_f(\tau) \rangle. \tag{6.36}$$

For the case of a linear distribution of environment oscillators (Ohmic dissipation), the statistical equilibrium of the fluctuation current may be obtained from the FDT [137, 138]

$$\frac{1}{2} \langle i_f(\tau) i_f(s) + i_f(s) i_f(\tau) \rangle = \frac{1}{2\pi I_c^2 R} \int_{-\infty}^{\infty} \frac{\hbar\nu}{2} \coth\left(\frac{\hbar\nu}{2T}\right) \cos\left[\nu(\tau - s)\right] \, d\nu. \tag{6.37}$$

The pre-factor term may be written

$$\frac{\hbar}{4\pi I_c^2 R} = \left(\frac{\hbar\omega_i}{4E_J\mu_i}\right)\frac{\mu_i^{1/2}}{\omega_p^2\pi\beta_c^{1/2}}, \tag{6.38}$$

and the average fluctuation may be written

$$\iint_0^t K(t,\tau)K(t,s)\langle i_f(\tau)i_f(s)\rangle\, d\tau ds = \left(\frac{\hbar\omega_i}{4E_J\mu_i}\right)\Xi^2, \tag{6.39}$$

where

$$\Xi^2 \equiv \Xi_0 \iint_0^t ds d\tau \int_0^\infty d\nu\nu \coth\left(\frac{\hbar\nu}{2T}\right) K(t,\tau)K(t,s)\cos\left[\nu(\tau-s)\right], \tag{6.40}$$

and

$$\Xi_0 \equiv \frac{2\mu_i^{1/2}}{\omega_p^2\pi\beta_c^{1/2}}. \tag{6.41}$$

The amplitude of the equilibrium initial fluctuations may also be found from the fluctuation-dissipation theorem [139]

$$\langle E\rangle = \Theta(\omega_i, T) = \frac{\hbar\omega_i}{2}\coth\left(\frac{\hbar\omega_i}{2T}\right), \tag{6.42}$$

where we use the notation

$$\begin{aligned} \omega_i^2 &\equiv \omega_p^2\mu_i, \\ \omega_f^2 &\equiv \omega_p^2\mu_f. \end{aligned} \tag{6.43}$$

Using the potential and kinetic energy terms $\langle U\rangle = 2E_J\mu_i\langle\tilde{\varphi}_i^2\rangle/2$, $\langle K\rangle = 2E_J\omega_p^{-2}\langle\dot{\tilde{\varphi}}_i^2\rangle/2$, and in equilibrium $\langle E\rangle = 2\langle U\rangle = 2\langle K\rangle$, we find

$$\begin{aligned} \langle\tilde{\varphi}_i^2\rangle &= \frac{\hbar\omega_i}{4\mu_i E_J}\coth\left(\frac{\hbar\omega_i}{2T}\right), \\ \langle\dot{\tilde{\varphi}}_i^2\rangle &= \omega_p^2\frac{\hbar\omega_i}{4E_J}\coth\left(\frac{\hbar\omega_i}{2T}\right), \\ &= \omega_i^2\langle\tilde{\varphi}_i^2\rangle. \end{aligned} \tag{6.44}$$

Collecting terms (6.35), (6.39), and (6.44) the variance of the phase fluc-

tuations at arbitrary time t may be written

$$\langle\tilde{\varphi}^2\rangle = \left(\frac{\hbar\omega_i}{4E_J\mu_i}\right)\left\{\coth\left(\frac{\hbar\omega_i}{2T}\right)\left[a^2(t)+\omega_i^2 b^2(t)\right]+\Xi^2\right\}, \tag{6.45}$$

and in the limit of low damping $\beta_c \to \infty$,

$$\langle\tilde{\varphi}^2\rangle = \left(\frac{\hbar\omega_i}{4E_J\mu_i}\right)\coth\left(\frac{\hbar\omega_i}{2T}\right)\left[a^2(t)+\omega_i^2 b^2(t)\right]. \tag{6.46}$$

6.1.0.1 Evaluation of Ξ^2

Formally, evaluation of Ξ^2 (6.40) is very similar to the evaluation of the thermal parameter C (5.84) contributing to the density matrix. In the development of the Heisenberg-Langevin-Lax theory above, we also developed a method of evaluation of Ξ^2 that is vastly superior to the Monte Carlo techniques previously employed. By expanding the cos() term, we may write the integral as

$$\Xi^2 = \Xi_0\int_0^\infty \nu\coth\left(\frac{\hbar\nu}{2T}\right)\left[\left(\int_0^t\cos(\nu\tau)K(t,\tau)d\tau\right)^2+\left(\int_0^t\sin(\nu\tau)K(t,\tau)d\tau\right)^2\right]d\nu. \tag{6.47}$$

Defining functions

$$\begin{aligned} K_c(t,\nu) &\equiv \int_0^t\cos(\nu\tau)K(t,\tau)d\tau, \\ K_s(t,\nu) &\equiv \int_0^t\sin(\nu\tau)K(t,\tau)d\tau, \end{aligned} \tag{6.48}$$

Ξ^2 may be written compactly as a one dimensional integral

$$\Xi^2 = \Xi_0\int_0^\infty \nu\coth\left(\frac{\hbar\nu}{2T}\right)\left[K_c^2(t,\nu)+K_s^2(t,\nu)\right]\,d\nu. \tag{6.49}$$

Using $\int_0^\infty\cos[\nu(\tau-s)]\,d\nu = \pi\delta(\tau-s)$, function Ξ^2 has thermal and quantum limits

$$\begin{aligned} \Xi_{\text{therm}} &= 2T\int_0^t K^2(\tau)d\tau, \\ \Xi_{\text{quan}} &= 2\left[\int_0^\infty \nu K_s^2(t,\nu)\,d\nu+\int_0^\infty \nu K_c^2(t,\nu)\,d\nu\right]. \end{aligned} \tag{6.50}$$

The key is to notice that internal integrals (6.48) may be calculated rapidly to high accuracy using Filon quadrature [98, 140, 141]. The thermal contribution (6.49) may then be evaluated rapidly using standard trapezoidal numerical integration without the problems normally associated with multidimensional numeric quadrature and without the need for time consuming Monte Carlo methods.

This technique should also be useful for evaluation of Eq. (5.84) even though functions $K(t,\tau)$, and $b_1(\tau,t)$ differ slightly. However, the utility of the Lax model has made calculation of the density matrix in the Caldeira-Leggett somewhat obsolete for the problem at hand.

6.1.1 Signal Resolution

The sensitivity of the detector to the qubit flux is proportional to the fluctuations of signal current created through the inductive coupling, proportional to the resolution of the initial comparator phase ($\phi_i = i_x/\mu_i$)

$$\delta\Phi_x^2 \;=\; k^{-2}LL_q\delta I_x^2 \;=\; 4I_c^2k^{-2}LL_q\mu_i^2\delta\phi_i^2. \tag{6.51}$$

After the moment of potential inversion and a short transient, both the average phase and the fluctuations begin to grow exponentially $\langle\varphi\rangle^2 \propto \langle\tilde{\varphi}^2\rangle \propto \exp[2\omega_f(t-t_{\rm inv})]$. It is this exponential growth before the phase settles into the final equilibrium state which produces the large SFQ pulse at the output stage (see section 5.2) allowing analysis of the signal resolution to be limited to within the comparator circuit. Hence, after sufficient time after inversion ($\omega_f(t-t_{\rm inv}) \gg 1$), the signal resolution may be determined at such a value that the signal to noise ratio is equal to one,

$$\langle\varphi\rangle^2 = \langle\tilde{\varphi}^2\rangle. \tag{6.52}$$

For average phase $\langle\varphi\rangle$ determined from initial position $\langle\varphi_i\rangle f(t)$ the signal resolution may be calculated as

$$\delta\varphi_i = \frac{\langle\tilde{\varphi}_i^2\rangle}{f(t)}. \tag{6.53}$$

Collecting terms (6.29) and (6.45) we may write the flux resolution of the comparator as

$$\delta\Phi_x^2 = \lambda k^{-2}L_q\hbar\omega_i\mu_i\frac{\coth\left(\frac{\hbar\omega_i}{2T}\right)\left[a^2(t)+\omega_i^2b^2(t)\right]+\Xi^2}{\left[a(t)+\mu_i\int\limits_0^t K(t,\tau)\,d\tau\right]^2}. \tag{6.54}$$

We note that in the limit of low damping $\Xi^2 \to 0$ and in contrast with the previous results, the flux sensitivity in the Heisenberg-Langevin-Lax model is completely determined by the initial thermal equilibrium phase fluctuations.

To estimate the flux resolution for realistic devices, we may calculate $\delta\Phi_x$ in the upper limit of instantaneous inversion of the potential and low damping $\beta_c \gg 1$. The solution for the component functions may be expressed analytically

$$\begin{aligned} a(t > t_{\rm inv}) &= \cosh(\omega_f \Delta t) - r_\omega \sinh[\omega_f \Delta t] \sin(\omega_i t_{\rm inv}), \\ b(t > t_{\rm inv}) &= \omega_f^{-1} \left[\sinh(\omega_f \Delta t) \cos(\omega_i t_{\rm inv}) + r_\omega^{-1} \cosh(\omega_f \Delta t) \sin(\omega_i t_{\rm inv})\right], \\ \mu_i \int_0^\infty K(t > t_{\rm inv}, \tau)\, d\tau &= \left\{ \begin{array}{l} r_\omega \sinh(\omega_f \Delta t) \sin(\omega_i t_{\rm inv}) \\ \quad + \cosh(\omega_f \Delta t) \left[1 + r_\omega^2 - \cos(\omega_i t_{\rm inv})\right] - r_\omega^2 \end{array} \right\}, \end{aligned} \tag{6.55}$$

with time after inversion $\Delta t = t - t_{\rm inv}$ and ratio of initial and final frequencies

$$r_\omega \equiv \frac{\omega_i}{\omega_f}. \tag{6.56}$$

The analytical solutions yield flux resolution

$$\delta\Phi_x^2 = \lambda k^{-2} L_q \hbar \omega_i \frac{\mu_i \mu_f}{\mu_i + \mu_f}. \tag{6.57}$$

Using parameters for realistic devices, for example the typical fabrication technology for the Stony Brook qubit [108] ($L_q \approx 250$ pH, $\omega_p \approx 3 \times 10^{10}$ s^{-1}, $E_J \approx 76$ K, $I_c \approx 3$ μA) and values $T \ll \hbar\omega_p$, $\lambda = 1$, we find quite good measurement properties $\delta\Phi_x/\Phi_x \approx 0.1$ at a measurement time $\Delta t \approx \ln(\Phi_0/\delta\Phi_x)/\omega_f \approx 10$ ps. Experimental results from a different group also confirm that the comparator is an excellent candidate for inductive measurements of the RF-SQUID qubit [129].

Perhaps a physically more meaningful measure of the detector resolution is the energy "output noise" figure-of-merit:

$$E_{\rm out} \equiv k^2 \frac{\delta\Phi_x^2}{2\omega_f L_q} = \frac{\hbar}{2} \lambda \mu_i r_\omega \frac{\langle\varphi\rangle^2}{\langle\tilde{\varphi}^2\rangle}. \tag{6.58}$$

This measure is identical to ϵ_ν used for characterization of SQUIDs used for continuous measurements with the replacement of the narrow bandwidth Δf with ω_f the reciprocal time scale of the single-shot measurement. Plugging in

solutions for the phase and variance this becomes similar to (6.54),

$$E_{\text{out}} = \frac{\hbar}{2}\lambda\mu_i r_\omega \frac{\coth\left(\frac{\hbar\omega_i}{2T}\right)\left[a^2(t)+\omega_i^2 b^2(t)\right]+\Xi^2}{\left[a(t)+\mu_i\int_0^t K(t,\tau)\,d\tau\right]^2}. \tag{6.59}$$

For instantaneous switching of the potential and the most important case $\beta_c \gg 1$, E_{out} takes a simple form

$$E_{\text{out}} = \frac{\hbar}{2}\lambda r_\omega \coth\left(\frac{\hbar\omega_i}{2T}\right)\frac{\mu_i\mu_f}{\mu_i+\mu_f}. \tag{6.60}$$

In the case of quantum-limited fluctuations and low damping, Eq. (6.60) may be written explicitly in terms of λ ($\mu_i = 1+1/2\lambda$, $\mu_f = 1-1/2\lambda$) as

$$E_{\text{out}} = \frac{\hbar}{4}\lambda\left[\left(1+\frac{1}{2\lambda}\right)^3\left(1-\frac{1}{2\lambda}\right)\right]^{1/2}. \tag{6.61}$$

The results of Eq. (6.60) versus inductive parameter $\lambda \propto L$ are shown in Fig. 6.3(a) for decreasing temperature, while panel (b) shows the quantum limit of Eq. (6.59) for decreasing damping parameter β_c to the critical value. The results in the quantum limit and low damping are shown in panel (c) compared to the calculated switching in experimental devices of section 5.3.1 (see Fig. 5.8). As discussed in section 5.4.1, the sensitivity of the device is increased by driving the comparator with "softer" pulses. In all cases, when λ is not too close to 1/2, $E_{\text{out}} \propto \lambda \propto L$ and $E_{\text{out}} \to \infty$ as $\lambda \to \infty$. The interpretation of the this result is clear, increasing the comparator's coupling inductance L decreases the signal current $I_x \propto \Phi_x/L$, and for a fixed current resolution, the sensitivity to the flux is reduced. More interestingly is the energy resolution in the opposite limit $\lambda \to 1/2$. Figure 6.3 shows that as $\lambda \to 1/2$, E_{out} tends to zero, a result easily confirmed by inspection of Eq. (6.61). In fact, E_{out} may fall below the apparent quantum limit $\hbar/2$ for even modest values of λ. A similar situation occurs with the parameter ϵ_ν for the characterization of dc SQUIDS, and is a reflection of the fact that neither ϵ_ν nor E_{out} take into account the back-action fluctuations of the comparator measurement onto the source signal. In the case of SQUIDs, the back-action noise ϵ_i has been analyzed, and when properly accounted for with the correlation term $\epsilon_{\nu i}$, the normalized energy sensitivity

$$\epsilon \equiv \left(\epsilon_\nu\epsilon_i - \epsilon\nu i^2\right)^{1/2}, \tag{6.62}$$

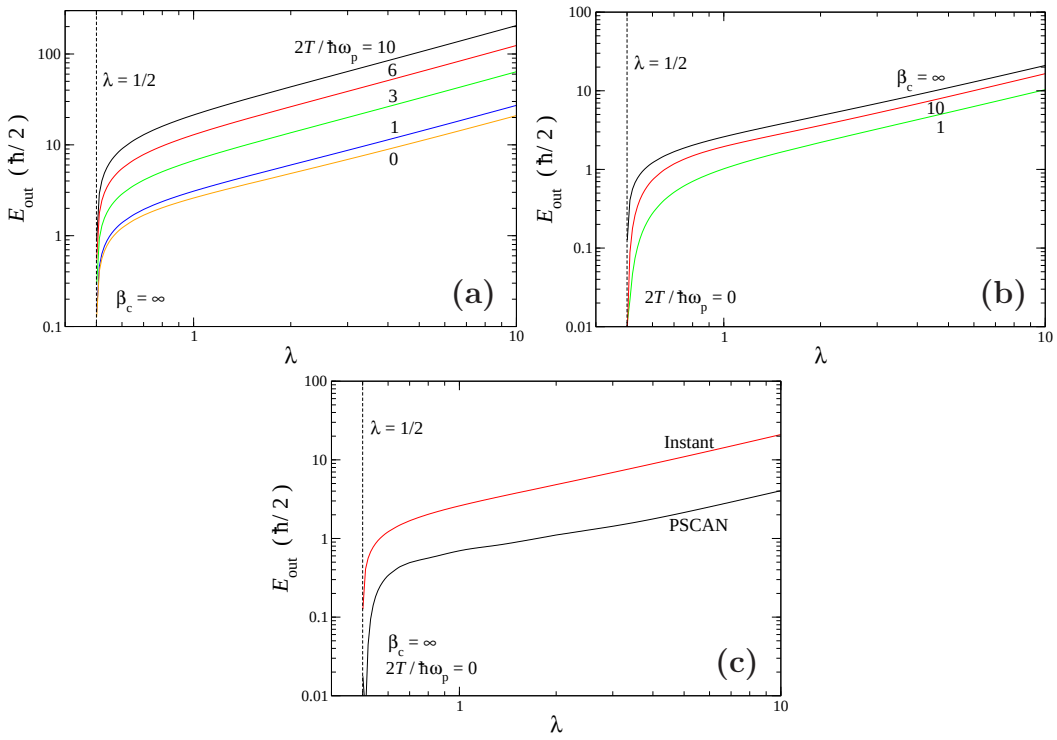

Figure 6.3: "Output noise" resolution E_{out} of the comparator with instantly inverted potential versus λ for: (a) decreasing temperatures and low damping, (b) decreasing damping parameter in the quantum limit, and (c) low-damping, quantum versus switch with actual RSFQ pulse calculated with PSCAN.

is indeed limited by the lower bound $\epsilon > \hbar/2$ [142]. However, no clear analog to ϵ_i has been found for the comparator-qubit system. The next section of the text explores the quantum mechanical limits of the measurement with consideration for the comparator's back-action effects.

6.2 Information versus Back-action

The goal of the comparator is to distinguish between two distinct quantum states of the qubit with high probability within a few, ideally one, measurements. Hence, a good measurement of the detector's fidelity is the ratio of the rate at which information is acquired to the dephasing of the states caused by the back-action of the measurement.

To quantify the information obtained in a measurement we note that distinguishing between qubit states $|1\rangle$,$|2\rangle$ can be seen as distinguishing between probability distributions $p_{1,2}(n)$ of possible outcomes n. The fidelity of the measurement may then be quantified in terms of the overlap between these two distributions [143]. For quantum information systems, the most appropriate way to characterize this over lap is [4]

$$\sum_n \left[p_1(n)p_2(n)\right]^{1/2}, \tag{6.63}$$

and the information obtained by the measurement (*i. e.* the ability of the comparator to discriminate between states $|1\rangle$ and $|2\rangle$) may be defined as [144, 145]

$$I = -\ln \sum_n \left[p_1(n)p_2(n)\right]^{1/2}. \tag{6.64}$$

Assuming that the comparator measurement can distinguish between all values of the phase φ at sufficiently long times after potential inversion then the confidence that the comparator my differentiate between the two qubit states in a single measurement is then given by

$$I = \ln \int dx \left[p_1(x)p_2(x)\right]^{1/2}. \tag{6.65}$$

Any dissipation in the comparator will couple directly to qubit, so we may restrict ourselves to the dissipation free limit $T \to 0$, $\beta_c \to \infty$. The qubit flux state will remain fixed through the measurement, see section 6, so the back-action of the comparator may be characterized as suppression of the off-diagonal elements of the total density matrix, which is expressed as the product

of density matrix of the comparator and qubit,

$$\rho_{\rm tot}(t) = \rho_c(t) \times \rho_q(t). \tag{6.66}$$

Equation 6.16 shows that the off-diagonal elements of the qubit may be expressed in terms of the time evolution of the comparator phase φ. The dephasing may then be expressed as the overlap of *comparator* wavefunctions $\Psi_{1,2}$ corresponding to qubit states $|1\rangle$,$|2\rangle$:

$$\Gamma = -\ln\left[\langle\Psi_1|\Psi_2\rangle\right]. \tag{6.67}$$

The results of chapter 5 show that the comparator wavefunctions are Gaussian throughout the measurement dynamics so $\Psi_{1,2}$ may be represented by Gaussian wavefunctions

$$\Psi_i = C\exp\left\{-\zeta\varphi_i^2/2 + \varrho\varphi_i\right\}, \tag{6.68}$$

shifted in opposite directions through the coupling to the qubit, shown schematically in Fig. 6.4. Parameters ζ, ϱ are in general complex numbers so defining

$$\begin{aligned} \zeta &\equiv \eta + \mathrm{i}\xi, \\ \varrho &\equiv \gamma + \mathrm{i}\delta, \end{aligned} \tag{6.69}$$

the constant C is found from the normalization requirement $|\Psi_i|^2 = 1$ through general Gaussian integral [98]

$$\int e^{-ax^2+bx+c} = \sqrt{\frac{\pi}{a}}e^{\frac{b^2}{4a}+c}\,dx \tag{6.70}$$

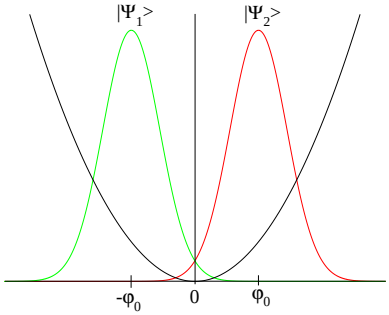

Figure 6.4: Schematic representation comparator wavefunctions in a quadratic potential.

to be

$$|C| = e^{-\gamma^2/\eta} \left(\frac{\eta}{\pi}\right)^{1/4}. \tag{6.71}$$

The off-diagonal matrix elements may then be calculated as

$$\begin{aligned} \langle \Psi_1 | \Psi_2 \rangle &= C_2 C_1^* \int d\varphi \exp\left\{-\zeta_1 \varphi^2/2 + \varrho^* \varphi - \zeta_2 \varphi^2/2 + \varrho\varphi\right\}, \\ &= \left(\tfrac{4\eta_1\eta_2}{\bar{\zeta}}\right)^{1/4} \exp\left\{-\tfrac{\gamma_2^2\eta_1 + \gamma_1^2\eta_2}{2\eta_1\eta_2} + \tfrac{\bar{\varrho}^2}{2\bar{\zeta}}\right\}, \end{aligned} \tag{6.72}$$

where we have defined

$$\begin{aligned} \bar{\zeta} &\equiv \zeta_1^* + \zeta_2, \\ \bar{\varrho} &\equiv \varrho_1^* + \varrho_2. \end{aligned} \tag{6.73}$$

The parameters γ, η may be related to the average and variance variables of the comparator dynamics. Calculating the average and variance of the phase we find the following relations:

$$\begin{aligned} \langle \varphi \rangle &= \langle \Phi | \varphi | \Phi \rangle, \\ &= \int d\varphi \, \varphi \, |\Phi|^2, \\ &= \frac{\gamma}{\eta}, \end{aligned} \tag{6.74}$$

and

$$\begin{aligned} \langle \tilde{\varphi}^2 \rangle &= \langle \Phi \, | (\varphi - \langle \varphi \rangle)^2 | \, \Phi \rangle, \\ &= \frac{1}{2\eta}. \end{aligned} \tag{6.75}$$

To find the equivalent "momentum" variable conjugate to the phase

$$[q, \varphi] = \mathrm{i}\hbar, \tag{6.76}$$

we begin with kinetic energy term $K = Q^2/4C$. The comparator charge (Q) and flux (Φ) have commutation relations

$$[\Phi, Q] = \mathrm{i}\hbar, \tag{6.77}$$

which yields relations for the phase ($\varphi = (2\pi/\Phi_0)\Phi$)

$$[\varphi, q] = \mathrm{i}\hbar. \tag{6.78}$$

Variable q may be seen as a normalized charge variable

$$q = \left(\frac{\Phi_0}{2\pi}\right) Q = 2\omega_p^{-2} E_J \dot{\varphi}, \tag{6.79}$$

where factor $2E_J/\omega_p^2$ may be replaced with the oscillator's equivalent mass (5.36), yielding kinetic term $K = q^2/2M_J$. This associated momentum has average value and fluctuation terms

$$\begin{aligned} \langle q \rangle &= \tfrac{\hbar}{\mathrm{i}} \langle \Psi | (\varrho - \zeta\varphi) | \Psi \rangle, \\ &= \hbar \left(\delta - \frac{\xi\gamma}{\eta} \right), \end{aligned} \tag{6.80}$$

and

$$\begin{aligned} \langle \tilde{q}^2 \rangle &= \langle q^2 \rangle - \langle q \rangle^2, \\ &= \hbar^2 \left(\frac{\eta^2 + \xi^2}{2\eta} \right). \end{aligned} \tag{6.81}$$

Combining solutions (6.74), (6.75), (6.80), (6.81) we may express the fundamental parameters in terms of the average phase, momentum and fluctuations,

$$\begin{aligned} \eta &= \frac{1}{2\langle \tilde{\varphi}^2 \rangle}, \\ \gamma &= \frac{\langle \varphi \rangle}{2\langle \tilde{\varphi}^2 \rangle}, \\ \xi &= -\left(\frac{4\langle \tilde{q} \rangle \langle \tilde{\varphi}^2 \rangle - \hbar^2}{4\hbar^2 \langle \tilde{\varphi}^2 \rangle^2} \right), \\ \delta &= \frac{\langle q \rangle}{\hbar} - \xi \langle \varphi \rangle. \end{aligned} \tag{6.82}$$

Note that the sign of ξ is negative because the initial Gaussian wavefunctions evolve in an *inverted* potential and the sign should be chosen to maintain the proper sign of the momentum [146]. This yields results for the back-action dephasing

$$\Gamma = \frac{\gamma_2^2 \eta_1 + \gamma_1^2 \eta_2}{2\eta_1 \eta_2} - \frac{\bar{\varrho}^2}{2\bar{\zeta}} - \frac{1}{4} \ln \left(\frac{4\eta_1 \eta_2}{\bar{\zeta}} \right), \tag{6.83}$$

and the ratio $\bar{\zeta}/2\bar{\varrho}$ may be expanded

$$\frac{\bar{\zeta}}{2\bar{\varrho}} = \frac{\left(\langle \varphi_1 \rangle \langle \tilde{\varphi}_2^2 \rangle + \langle \varphi_2 \rangle \langle \tilde{\varphi}_1^2 \rangle + \frac{2\mathrm{i}}{\hbar} \langle \tilde{\varphi}_1^2 \rangle \langle \tilde{\varphi}_2^2 \rangle \left[\langle q_2 \rangle - \langle q_1 \rangle + \zeta_2 \langle \varphi_2 \rangle - \zeta_1 \langle \varphi_1 \rangle \right] \right)^2}{4\langle \tilde{\varphi}_1^2 \rangle \langle \tilde{\varphi}_2^2 \rangle \left[\langle \tilde{\varphi}_1^2 \rangle + \langle \tilde{\varphi}_2^2 \rangle + 2\mathrm{i} \langle \tilde{\varphi}_1^2 \rangle \langle \tilde{\varphi}_2^2 \rangle (\zeta_2 - \zeta_1) \right]} \tag{6.84}$$

The information measure is calculated from the same comparator wavefunctions

$$\begin{aligned}\sqrt{p_1(x)p_2(x)} &= |C_1|\,|C_2| \exp\left\{-\frac{(\eta_1+\eta_2)^2}{2}\varphi^2 + (\gamma_1+\gamma_2)\,\varphi\right\}, \\ \int d\varphi \sqrt{p_1(x)p_2(x)} &= \left(\frac{4\eta_1\eta_2}{(\eta_1+\eta_2)^2}\right)^{1/4} \times \\ & \exp\left\{-\gamma_1^2/2\eta_1 - \gamma_2^2/2\eta_2 + (\gamma_1+\gamma_2)^2/2(\eta_1+\eta_2)\right\}.\end{aligned} \tag{6.85}$$

or

$$I = \frac{\gamma_2^2\eta_1 + \gamma_1^2\eta_2}{2\eta_1\eta_2} - \frac{(\gamma_1+\gamma_2)^2}{2(\eta_1+\eta_2)} - \frac{1}{4}\ln\left(\frac{4\eta_1\eta_2}{(\eta_1+\eta_2)^2}\right). \tag{6.86}$$

The comparator evolution for the two qubit states is symmetric, $\varphi(t) \propto \varphi_i \propto \sigma_z \propto \pm 1$, so phase $\langle\varphi_1\rangle$, $\langle\varphi_2\rangle$ and their relative momenta differ only by a sign

$$\begin{array}{rcccl}\langle\varphi_1\rangle &=& -\langle\varphi_2\rangle &\equiv& \langle\varphi\rangle, \\ \langle q_1\rangle &=& -\langle q_2\rangle &\equiv& \langle q\rangle.\end{array} \tag{6.87}$$

The fluctuation terms are independent of the initial shift

$$\begin{array}{rcccl}\langle\tilde{\varphi}_1^2\rangle &=& \langle\tilde{\varphi}_2^2\rangle &\equiv& \langle\tilde{\varphi}^2\rangle, \\ \langle\tilde{q}_1^2\rangle &=& \langle\tilde{q}_2^2\rangle &\equiv& \langle\tilde{q}^2\rangle,\end{array} \tag{6.88}$$

so we find relations for the Gaussian parameters

$$\begin{array}{rcccl}\eta_1 &=& \eta_2 &\equiv& \eta, \\ \gamma_1 &=& -\gamma_2 &\equiv& \gamma, \\ \zeta_1 &=& \zeta_2 &\equiv& \zeta.\end{array} \tag{6.89}$$

The information and dephasing terms then take very simple forms

$$\begin{aligned} I &= \frac{\gamma^2}{\eta}, \\ \Gamma &= \frac{\gamma^2}{\eta} - \frac{\langle\tilde{\varphi}^2\rangle}{2}\left(\frac{2\mathrm{i}}{\hbar}\left[\langle q\rangle + \hbar\xi\langle\varphi\rangle\right]\right)^2.\end{aligned} \tag{6.90}$$

This is a key point. The total dephasing may then be seen as the information acquired in a single shot measurement plus an additional backaction term

$$\Gamma = I + \tilde{\Gamma}. \tag{6.91}$$

If the measurement time is sufficiently far from inversion, when the average phase and fluctuations have reached the stage of exponential growth then ξ

takes the simple form

$$\xi \approx -\frac{1}{\hbar}\left(\frac{\langle\tilde{q}^2\rangle}{\langle\tilde{\varphi}^2\rangle}\right)^{1/2}, \tag{6.92}$$

and the information and dephasing terms may be expressed in terms of the comparator's phase and momentum variables using relations (6.82) as

$$\begin{aligned} I &= \frac{1}{2}\frac{\langle\varphi\rangle^2}{\langle\tilde{\varphi}^2\rangle}, \\ \tilde{\Gamma} &= \frac{2}{\hbar^2}\left[\langle q\rangle\langle\tilde{\varphi}^2\rangle^{1/2} - \langle\varphi\rangle\langle\tilde{q}^2\rangle^{1/2}\right]^2. \end{aligned} \tag{6.93}$$

Since $\tilde{\Gamma}$ is positive semi-definite we can achieve an ideal quantum measurement when $\tilde{\Gamma} = 0$ or

$$\langle q\rangle^2\langle\tilde{\varphi}^2\rangle = \langle\varphi\rangle^2\langle\tilde{q}^2\rangle. \tag{6.94}$$

6.2.1 Measurement Optimization

The result (6.93) may be analyzed analytically for the case of instantaneous inversion of the comparator potential from $\omega_i \to \omega_f$ at time t_{inv}. The Heisenberg equation of motion for the phase operator with shifted initial values is

$$\ddot{\boldsymbol{\varphi}} + \omega^2(t)\boldsymbol{\varphi} + \frac{\lambda\sigma_z}{M_J} = 0, \tag{6.95}$$

where for notational convenience we have used equivalent effect mass (5.36). The solution before inversion may be written

$$\boldsymbol{\varphi}(t < t_{\text{inv}}) = \boldsymbol{\varphi}_i \cos\left[\omega_i(t + t_{\text{inv}})\right] + \frac{\mathbf{q}_i}{M_J}\sin\left[\omega_i(t + t_{\text{inv}})\right] - \langle\varphi_i\rangle, \tag{6.96}$$

with average initial values

$$\begin{aligned} \langle\varphi_i\rangle &= \frac{\lambda\sigma_z}{M_J\omega_i^2}, \\ \langle q_i\rangle &= 0. \end{aligned} \tag{6.97}$$

The general solution after switching is

$$\boldsymbol{\varphi}(t > t_{\text{inv}}) = A\cosh\left[\omega_f(t - t_{\text{inv}})\right] + B\sinh\left[\omega_f(t - t_{\text{inv}})\right] + \varphi_\lambda, \tag{6.98}$$

where

$$\varphi_\lambda \equiv \frac{\lambda\sigma_z}{M_J\omega_f^2} = r_\omega^2\langle\varphi_i\rangle. \tag{6.99}$$

Expressing the operators at the moment of inversion as the average value plus small fluctuation

$$\begin{aligned} \boldsymbol{\varphi}(t_{\text{inv}}) &= \varphi_i + \tilde{\varphi}_i, \\ \mathbf{q}(t_{\text{inv}}) &= \tilde{q}_i, \end{aligned} \tag{6.100}$$

and matching the solutions at t_{inv} we find solutions for the phase and momentum operators

$$\begin{aligned} \boldsymbol{\varphi}(t) &= (\varphi_i + \tilde{\varphi}_i)\,\mathcal{C} + \varphi_\lambda\,(1-\mathcal{C}) + \frac{\tilde{q}_i}{M_J\omega_f}\mathcal{S}, \\ \mathbf{q}(t) &= M_J\omega_f\,(\varphi_i + \tilde{\varphi}_i - \varphi_\lambda)\,\mathcal{S} + \tilde{q}_i\mathcal{C}, \end{aligned} \tag{6.101}$$

where we have defined

$$\begin{aligned} \mathcal{C} &\equiv \cosh\left[\omega_f(t-t_{\text{inv}})\right], \\ \mathcal{S} &\equiv \sinh\left[\omega_f(t-t_{\text{inv}})\right], \end{aligned} \tag{6.102}$$

for notational convenience. The average phase, momentum and variances for instantaneous inversion of the potential may then be expressed

$$\begin{aligned} \langle\varphi\rangle &= \langle\varphi_i\rangle\left[(1+r_\omega^2)\,\mathcal{C} - r_\omega^2\right], \\ \langle q\rangle &= M_J\omega_f\langle\varphi_i\rangle\,(1+r_\omega^2)\,\mathcal{S}, \\ \langle\tilde{\varphi}^2\rangle &= \langle\tilde{\varphi}_i^2\rangle\,(\mathcal{C}^2 + r_\omega^2\mathcal{S}^2), \\ \langle\tilde{q}^2\rangle &= M_J^2\omega_f^2\langle\tilde{\varphi}_i^2\rangle\,(\mathcal{S}^2 + r_\omega^2\mathcal{C}^2). \end{aligned} \tag{6.103}$$

Plugging solutions (6.103), we find that the fidelity of the measurement becomes constant at large times ($\omega_f(t-t_{\text{inv}}) \gg 1$,

$$I = \frac{1+r_\omega^2}{2}\frac{\langle\varphi_i\rangle^2}{\langle\tilde{\varphi}_i^2\rangle}, \tag{6.104}$$

and is determined entirely from the equilibrium initial fluctuations (6.44). On the other hand, the dephasing grows exponentially

$$\tilde{\Gamma} = \frac{1}{4}\frac{\langle\varphi_i\rangle^2}{\langle\tilde{\varphi}_i^2\rangle}\exp\left[2\omega_f(t-t_{\text{inv}})\right]. \tag{6.105}$$

This exponential growth continues until the comparator reaches the non-linear part of the potential and settles into final equilibrium state $\varphi_f = \varphi_i \pm \pi$ while the back-action saturates at some value $\tilde{\Gamma} \gg I$.

This result shows that even though the comparator has very high (potentially quantum limited) flux resolution, it imparts back-action noise much higher than fundamental quantum mechanical limits. From this stand-point, the detector is actually rather poor, and would be greatly improved if the effect of the measurement could be minimized for constant flux resolution. We propose two ways to achieve this goal.

6.2.1.1 Quenched Coupling

The first option is to note that the comparator may be decoupled from the qubit at time t_0, slightly before the measurement pulse preventing the feedback between the two components (see Fig. 6.5). The coupling parameter κ may be made externally controllable through the use of Josephson-junction circuitry [147]. At the moment that coupling is quenched, the oscillator is free to evolve in the comparator potential well until the moment of inversion of the potential ($t_{\rm inv} = 0$), and the solution to Eq. (6.95) may be written

$$\begin{aligned} \boldsymbol{\varphi}(t>0) &= (\varphi(0)+\tilde{\varphi}_i)\,\mathcal{C} + \frac{q(0)+\tilde{q}_i}{M_J\omega_f}\mathcal{S}, \\ \mathbf{q}(t>0) &= M_J\omega_f\,(\varphi(0)+\tilde{\varphi}_i)\,\mathcal{S} + (q(0)+\tilde{q}_i)\,\mathcal{C}. \end{aligned} \tag{6.106}$$

Since the state is fixed until the coupling is suppressed ($q(t \le -t_0) = 0$) , we may express the average phase and momentum at the moment of inversion as

$$\begin{aligned} \varphi(0) &= \varphi_i \mathcal{C}_0, \\ q(0) &= -M_J\omega_i\varphi_i\mathcal{S}_0, \end{aligned} \tag{6.107}$$

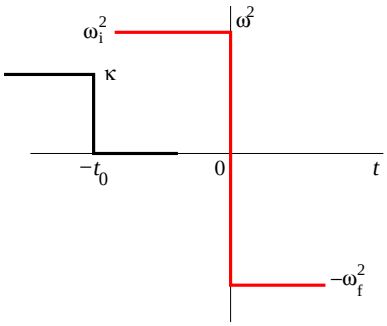

Figure 6.5: Instantaneous inversion of the comparator potential preceded by rapid quenching of the coupling coefficient κ.

where we have defined

$$\begin{aligned} \mathcal{C}_0 &\equiv \cos(\omega_i t_0), \\ \mathcal{S}_0 &\equiv \sin(\omega_i t_0). \end{aligned} \tag{6.108}$$

This gives values for the statistical properties

$$\begin{aligned} \langle\varphi\rangle &= \varphi_i \left(\mathcal{C}_0\mathcal{C} - r_\omega \mathcal{S}_0\mathcal{S}\right), \\ \langle q\rangle &= M_J\omega_f\varphi_i \left(\mathcal{C}_0\mathcal{S} - r_\omega \mathcal{S}_0\mathcal{C}\right), \\ \langle\tilde{\varphi}^2\rangle &= \langle\tilde{\varphi}_i^2\rangle \left(\mathcal{C}^2 + r_\omega^2\mathcal{S}^2\right), \\ \langle\tilde{q}^2\rangle &= \left(M_J\omega_f\right)^2 \langle\tilde{\varphi}_i^2\rangle \left(\mathcal{S}^2 + r_\omega^2\mathcal{C}^2\right). \end{aligned} \tag{6.109}$$

Plugging these expressions into Eq. (6.93), we again find that the information measure evolves to a constant

$$I = \frac{1}{1+r_\omega^2}\frac{\langle\varphi_i\rangle^2}{2\langle\tilde{\varphi}_i^2\rangle}\left[\mathcal{C}_0 - r_\omega\mathcal{S}_0\right]^2, \tag{6.110}$$

as do the dephasing fluctuations

$$\tilde{\Gamma} = \frac{1}{1+r_\omega^2}\frac{\langle\varphi_i\rangle^2}{2\langle\tilde{\varphi}_i^2\rangle}\left[r_\omega\mathcal{C}_0 + \mathcal{S}_0\right]^2. \tag{6.111}$$

The dependencies of I and $\tilde{\Gamma}$ are shown in Fig. 6.6 along with the total dephasing, which is independent of the delay time

$$\Gamma = \frac{\langle\varphi_i\rangle^2}{2\langle\tilde{\varphi}_i^2\rangle}. \tag{6.112}$$

If the comparator and qubit are decoupled at the moment the comparator makes the measurement, then the backaction dephasing is greater than the received information by a factor of $1 + r_\omega^2$. In this case we may have an ideal quantum measurement only in the impractical limit $\omega_f \gg \omega_i$. If however the comparator is allowed to undergo free oscillations, the phase and momentum will reach values such that amplitude of the comparator wavepacket which reaches the inversion stage is maximized and we get an optimal measurement. These periodically repeating optimal delay times are given for an arbitrary value of r_ω as

$$\tan(\omega_i t_0) = -r_\omega, \tag{6.113}$$

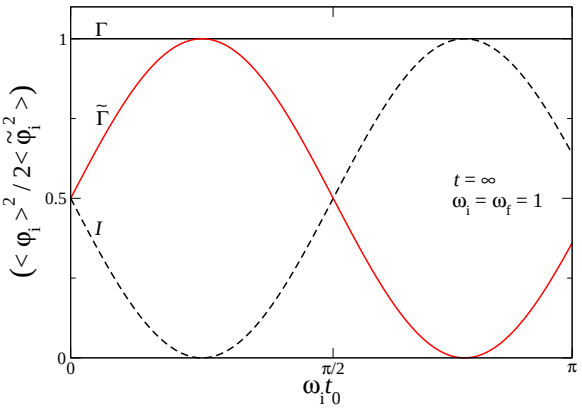

Figure 6.6: Information (dashed line), dephasing fluctuations (red line), and total back-action (black line) versus delay time between suppression of the coupling and potential inversion.

and the information is indeed maximized to value

$$I_{\max} = \Gamma. \tag{6.114}$$

At this delay time, the comparator represents an "ideal" quantum detector.

We may calculate the flux resolution in this limit, again from the requirement of unitary signal-to-noise ratio ($\langle\varphi\rangle^2 = \langle\tilde{\varphi}^2\rangle$) to be

$$\delta\varphi_i^2 = \langle\tilde{\varphi}_i^2\rangle \frac{1 + r_\omega^2}{\left(\mathcal{C}_0 - r_\omega \mathcal{S}_0\right)^2}. \tag{6.115}$$

The flux resolution is optimized at the same delay time (6.113) and the energy resolution of the optimal measurement is calculated in terms of inductive parameter λ

$$E_{\text{out}} = \frac{\hbar}{2}\lambda \left[\left(1 + \frac{1}{2\lambda}\right)^3 \left(1 - \frac{1}{2\lambda}\right)^{-1}\right]^{1/2}. \tag{6.116}$$

The output noise resolution for the ideal quantum measurement is shown Fig. 6.7 along with the previous result for constant coupling (6.61). For all inductive parameters, the decoupling to achieve the ideal measurement decreases the flux resolution of the comparator. In fact, as $\lambda \to 1/2$, the resolution has the exact inverse behavior $E_{\text{out}} \to \infty$ from the constant coupling case. The ideal measurement reaches a minimum in flux resolution $E_{\text{out}} = 3\hbar/4\sqrt{3}$ at $\lambda = 1$.

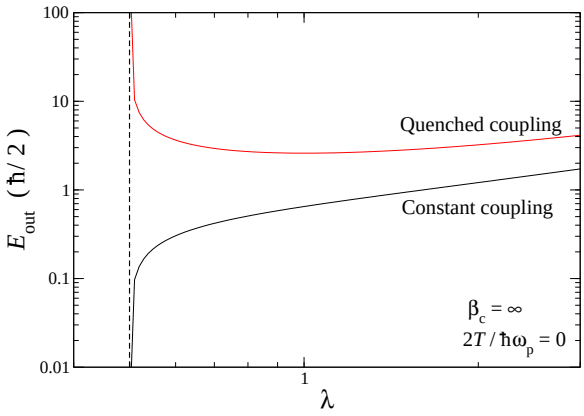

Figure 6.7: Output noise resolution for the instantaneously inverted, dissipation free comparator for constant (black line) and optimally quenched (red line) comparator-qubit coupling.

6.2.1.2 Feedback Circuit

The second way we might compensate for the negative effects of the comparator's backaction noise is to compensate for the coupled signal through an external circuit. Because the signal $\langle\varphi\rangle(t)$ exhibits exponential growth, we may neglect additional noise at this stage of the circuit and simply the amplified signal, modified by an arbitrary function, as feedback to the comparator. Taking the measurement time t_m to be in the stage of exponential growth, the feedback pulse applies an additional phase

$$\varphi_f(t) = f(t) \times \varphi(t_m) \tag{6.117}$$

shown in Fig. 6.8. For all times before the measurement $f(t) \equiv 0$.We note that the signal itself comes with exponentially growth noise signal $\langle\tilde{\varphi}^2\rangle$ which will also be multiplied by the feedback function. In fact, the conjugate momentum

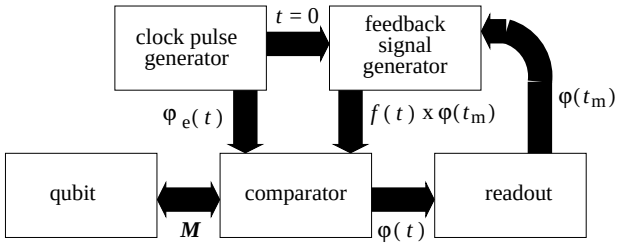

Figure 6.8: Measurement circuit with the comparator's backaction compensated by a feedback loop.

variable is simply $q_c(t) = E_J\dot{\varphi}_c(t) \propto \exp(2\omega_f t)$. So if we compensate for the back-action signal $\langle\varphi\rangle(t)$ exactly then we automatically compensate for the momentum part of the signal as well. Hence, $\langle\varphi\rangle(t)$, $\langle q\rangle(t)$ may be made vanishingly small simultaneously for a single, appropriate choice of function $f(t)$. In this way we may reach the ideal limit $\Gamma = I$ while maintaining quantum-limited flux resolution (6.59).

6.3 Conclusions

We have developed two methods to analyze quantum fluctuations for arbitrary time inversion of a damped harmonic oscillator potential. The first method calculates the time evolution of the system density matrix coupled to bath of linear oscillators. The second method calculates the particle motion of the statistical properties within the Heisenberg-Langevin-Lax equations of motion. The results show that when the quantum wavepacket is prepared in an initially Gaussian state, the envelope remains Gaussian throughout its evolution. When applied to the balanced Josephson junction comparator these results may used for numerical prediction of the gray zone width defining the accuracy of the detector. The numerical calculations show nearly perfect agreement with experimental results for typical niobium tri-layer devices and predict than when the temperature of the system is reduced below ≈ 1 K, the signal resolution may be limited only by quantum fluctuations within the device itself.

We have also calculated the flux resolution, energy output noise and back-action strength of the comparator taking into account the inductive coupling with a magnetic flux qubit. For a system where the coupling between the two components remains constant, the flux sensitivity of the comparator may again be quantum limited and increased by tuning of the inductive parameter λ. In fact, the resolution of the device may be made nearly arbitrarily high but the measurement disturbs the qubit state quite significantly. This back-action effect may be reduced to the quantum mechanical limit of an ideal measurement by at least two methods. An ideal measurement may be obtained by decoupling of the comparator qubit system, although at reduced flux sensitivity. The back-action noise may also be compensated by an additional optimized feedback circuit on the exponentially amplified signal while maintaining high flux resolution.

6.3.1 Possible future work

An interesting extension to this work remains a subject for interesting future work. Namely, how does the minimal back-action of the comparator manifest itself in the qubit after measurement? For example, Eq. (6.16) suggests that the off-diagonal elements of the qubit two state Hamiltonian will experience an exponentially strong perturbation during the measurement without compensation, in agreement with Eq. (6.105). Combined with the arbitrarily high precision of the flux measurement, this state represents a kind of "squeezed" detector measurement. We have demonstrated that in the sense of information and dephasing we may reach a quantum mechanically ideal measurement. We have not, however, calculated the affect that this measurement has on the charge variable conjugate to the qubit flux. In fact, a proper definition of the qubit charge noise remains unclear. Such a definition may be found, and along with the calculated flux noise compared against the fundamental quantum limit.

This problem has been solved for continuous measurement of a harmonic oscillator via an "ideal" linear amplifier [148]. For that case, there is a minimum limit such that the signal-to-noise ratio at the output precisely doubles the input signal-to-noise ratio for a signal with quantum mechanical uncertainty. The interpretation of this doubling is that the fundamental uncertainty of the detector multiplies the uncertainty of the input exactly doubling it. Such a fundamental relation is more elusive for the comparator-qubit system however because the qubit flux state remains fixed (see Eq. (6.13))throughout the interaction with the detector.

Bibliography

[1] G. E. Moore. Progress in digital integrated electronics. In *IEEE Electron Devices Meeting: IEDM Tech. Digest.*, pages 11–13, 1975.

[2] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8), Apr. 1965.

[3] International Technology Roadmap for Semiconductors. Process integration, devices and structures, 2007. URL http://www.itrs.net/Links/2007ITRS/Home2007.htm.

[4] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge, 2000. ISBN 0-521-63503-9.

[5] O. Mukhanov, V. K. Semenov, and K. K. Likharev. Ultimate performance of the RSFQ logic circuits. *IEEE Trans. Magn.*, 23(2):759–762, Mar. 1987.

[6] K. K. Likharev. *Dynamics of Josephson Junctions and Circuits*. Gordon and Breach, 1986. ISBN 2-88124-042-9.

[7] International Technology Roadmap for Semiconductors. Executive summary, 2007. URL http://www.itrs.net/Links/2007ITRS/Home2007.htm.

[8] R. J. Baker, H. W. Li, and D. E. Boyce. *CMOS: Circuit design, layout, and simulation*. IEEE Press, 1998. ISBN 0-7803-3416-7.

[9] S. M. Sze. *Physics of Semiconductor Devices*. Wiley, 2nd edition, 1981. ISBN 0-471-05661-8.

[10] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc. Design of ion-implanted MOSFETs with very small physical dimensions. *IEEE J. Solid-State Circuits*, 9(5):256–258, Oct. 1974.

[11] J. D. Jackson. *Classical Electrodynamics*. Wiley, 3rd edition, 1999. ISBN 0-471-30932-X.

[12] D. J. Frank, Y. Taur, and H. P. Wong. Generalized scale length for two-dimensional effects in MOSFET's. *IEEE Elec. Device Lett.*, 19(10): 385–387, Oct. 1998.

[13] Y. Taur. CMOS scaling to nanometer lengths. In H. Morkoc, editor, *Advanced Semiconductor and Organic Nano-Techniques (Part I)*, pages 211–237. Elsevier, 2003.

[14] Y. Naveh and K. K. Likharev. Shrinking limits of silicon MOSFETs: numerical study of 10 nm scale devices. *Superlatt. and Microstruct.*, 27 (2/3):111–123, 2000.

[15] Y. Naveh and K. K. Likharev. Modeling of 10-nm-Scale ballistic MOSFETs. *IEEE Elec. Device Lett.*, 21(5):242–244, May 2000.

[16] V. A. Sverdlov, T. J. Walls, and K. K. Likharev. Nanoscale silicon MOSFETs: A theoretical study. *IEEE Trans. Elec. Devices*, 50(9):1926–1933, Sept. 2003.

[17] Press release Freescale Inc., 2008. URL http://www.freescale.com/webapp/sps/site/overview.jsp?nodeId=0ST287482180CAE.

[18] H. P. Wong. Beyond the conventinal transistor. *IBM J. Res. Dev.*, 46 (2/3):133–168, Mar./May 2002.

[19] M. V. Fischetti, S. E. Laux, and E. Crabbe. Understanding hot-electron transport in silicon devices - is there a shortcut. *J. Appl. Phys.*, 78(2): 1058–1087, July 1995.

[20] T. Sjodin, H. Petek, and H. Dai. Ultrafast carrier dynamics in silicon: A two-color transient reflection grating study on a (111) surface. *Phys. Rev. Lett.*, 81(25):5664–5667, Dec. 1998.

[21] M. V. Fischetti. Private communication, 1999.

[22] D. Esseni, M. Mastrapasqua, G. K. Celler, F. H. Baumann, C. Fiegna, L. Selmi, and E. Sangiorgi. Low field mobility of ultra-thin SOI N- and P-MOSFETs: Measurements and implications on the performance of ultra-short MOSFETs. In *IEEE Electron Devices Meeting: IEDM Tech. Digest.*, pages 671–674, San Francisco, CA, Dec. 2000.

[23] D. Esseni, M. Mastrapasqua, C. Fiegna, G. K. Celler, L. Selmi, and E. Sangiorgi. An experimental study of low field electron mobility in double-gate, ultra-thin SOI MOSFETs. In *IEEE Electron Devices Meeting: IEDM Tech. Digest.*, pages 19.7.1–19.7.4, Washington, DC, Dec. 2001.

[24] K. Uchida, J. Koga, and S. Takagi. Experimental study on carrier transport mechanisms in double- and single-gate ultrathin-body mosfets - coulomb scattering, volume inversion, and /spl delta/t/sub SOI/-induced scattering. In *IEEE Electron Devices Meeting: IEDM Tech. Digest.*, pages 33.5.1–33.5.4, Washington, DC, Dec. 2003.

[25] K. Uchida and S. Takagi. Carrier scattering induced by thickness fluctuation of silicon-on-insulator film in ultrathin-body metal-oxide-semiconductor field-effect transistors. *Appl. Phys. Lett.*, 82(17):2916–2918, Apr. 2003.

[26] D. Esseni. On the modeling of surface roughness limited mobility in SOI MOSFETs and its correlation to the transistor effective field. *IEEE Trans. Elec. Devices*, 51(3):394–401, Mar. 2004.

[27] H. van Houten and C. W. J. Beenakker. Quantum point contacts and coherent electron focusing. In W. van Haeringen and D. Lenstra, editors, *Analogies in Optics and Microelectronics*. Lavoisier, 1990. URL arXiv:cond-mat/0512611.

[28] H. van Houten and C. W. J. Beenakker. Princeples of solid state electron optics. In E. Burstein and C. Weisbuch, editors, *Confined Electrons and Photons*. Plenum, 1995. ISBN 0-306-44990-0.

[29] K. Natori. Ballistic metal-oxide-semiconductor field effect transistor. *J. Appl. Phys.*, 76(8):4879–4890, Oct. 1994.

[30] D. J. Frank, S. E. Laux, and M. V. Fischetti. Monte carlo simulation of a 30 nm Dual-Gate MOSFET: How short can we go? In *IEEE Electron Devices Meeting: IEDM Tech. Digest.*, pages 553–556, Dec. 1992.

[31] M. Lundstrom. Elementary scattering theory of the Si MOSFET. *IEEE Elec. Device Lett.*, 18(7):361–363, July 1997.

[32] F. G. Pikus and K. K. Likharev. Nanoscale field-effect transistors: An ultimate size analysis. *Appl. Phys. Lett.*, 71(25):3661–3663, Dec. 1997.

[33] K. K. Young. Short-channel effect in fully depleted SOI MOSFETs. *IEEE Trans. Elec. Devices*, 36(2):399–402, Feb. 1989.

[34] K K. Young. Analysis of conduction in fully depleted SOI MOSFETs. *IEEE Trans. Elec. Devices*, 36(3):504–506, Mar. 1989.

[35] K. Akarvardar, A. Mercha, S. Christoloveanu, P. Gentil, E. Simoen, V. Subramanian, and C. Claeys. A two-dimensional model for interface coupling in triple-gate transistors. *IEEE Trans. Elec. Devices*, 54(4): 767–775, Apr. 2007.

[36] Y. Taur, X. Liang, W. Wang, and H. Lu. A continuous, analytic drain-current model for DG MOSFETs. *IEEE Elec. Device Lett.*, 25(2):107–109, Feb. 2004.

[37] A. Ortiz-Conde, F. J. Garcia Sanchez, and J. Muci. Rigorous analytic solution for the drain current of undoped symmetric dual-gate MOSFETs. *Sol. St. Elec.*, 49(6):640–647, 2005.

[38] A. S. Roy, J. M. Sallese, and C. C. Enz. A closed-form charge-based expressions for drain current in symmetric and asymmetric double gate MOSFET. *Sol. St. Elec.*, 50:687–693, 2006.

[39] O. Moldovan, D. Jimenez, R. Guitart, A. Chaves, and B. Iniguez. Explicit analytical charge and capacitance models of undoped double-gate MOSFETs. *IEEE Trans. Elec. Devices*, 54(7):1718–1724, Jul. 2007.

[40] A. Svizhenko and M. P. Anantram. Role of scattering in nanotransistors. *IEEE Trans. Elec. Devices*, 50(6):1459–1466, Jun. 2003.

[41] S. Datta. Nanoscale device modeling: the green's function method. *Superlatt. and Microstruct.*, 28(4):253–278, 2000.

[42] S. Datta. Quantum devices. *Superlatt. and Microstruct.*, 6(1):83–93, 1989.

[43] M. Lundstrom and Z. Ren. Essential physics of carrier transport in nanoscale MOSFETs. *IEEE Trans. Elec. Devices*, 49(1):133–141, Jan. 2002.

[44] J. Wang and M. Lundstrom. Does source-to-drain tunneling limit the ultimate scaling of MOSFETs? In *IEEE Electron Devices Meeting: IEDM Tech. Digest.*, pages 707–710, San Francisco, CA, Dec. 2002.

[45] J. H. Rhew, Z. Ren, and M. S. Lundstrom. A numerical study of ballistic transport in a nanoscale MOSFET. *Sol. St. Elec.*, 46:1899–1906, 2002.

[46] A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal. Two-dimensional quantum mechanical modeling of nanotransistors. *J. Appl. Phys.*, 91(4):2343–2354, Feb. 2002.

[47] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic. Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches. *J. Appl. Phys.*, 92(7):3730–3739, Oct. 2002.

[48] A. Martinez, M. Bescond, J. R. Barker, A. Svizhenko, M. P. Anantram, C. Millar, and A. Asenov. A self-consistent full 3-D real-space NEGF simulator for studying nonperturbative effects in nano-MOSFETs. *IEEE Trans. Elec. Devices*, 54(9):2213–2222, Sept. 2007.

[49] T. J. Walls, V. A. Sverdlov, and K. K. Likharev. Nanoscale SOI MOSFETs: a comparison of two options. *Sol. St. Elec.*, 48:857–865, 2004.

[50] S. E. Laux, A. Kumar, and M. V. Fischetti. Ballistic FET modeling using QDAME: Quantum Device Analysis by Modal Evaluation. *IEEE Trans. Nanotechnol.*, 1(4):255–259, Dec. 2002.

[51] S. E. Laux, A. Kumar, and M. V. Fischetti. Analysis of quantum ballistic electron transport in ultrasmall silicon devices including space-charge and geometric effects. *J. Appl. Phys.*, 95(10):5545–5582, May 2004.

[52] T. J. Walls, V. A. Sverdlov, and K. K. Likharev. Quantum mechanical modeling of advanced sub-10-nm MOSFETs. In *Proc. of IEEE-Nano Conf.*, pages 338 (1–4), San. Francisco, CA, Aug. 2003.

[53] T. J. Walls, V. A. Sverdlov, and K. K. Likharev. MOSFETs below 10 nm: quantum theory. *Physica E*, 19(1-2):23–27, 2003.

[54] J. C. Slater. A simplification of the Hartree-Fock method. *Phys. Rev.*, 81(3):385–390, 1951.

[55] R. Sasajima, K. Fujimaru, and H. Matsumura. A metal-insulator tunnel transistor with 16 nm channel length. *Appl. Phys. Lett.*, 74(21):3215–3217, May 1999.

[56] H. Kawaura, T. Sakamoto, and T. Baba. Observation of source-to-drain direct tunneling in 8 nm gate electrically variable shallow junction MOSFETs. *Appl. Phys. Lett.*, 77(25):3810–3812, Jun. 2000.

[57] L. Kang, Lee B. H, W. J. Qi, Y. Jeon, R. Nieh, S. Gopalan, K. Onishi, and J. C. Lee. Electrical characteristics of highly reliable ultrathin hafnium oxide gate dielectric. *IEEE Elec. Device Lett.*, 21(4):181–183, Apr. 2000.

[58] M. Vinet, T. Poiroux, J. Widiez, J. Lolivier, B. Previtali, C. Vizoiz, V. Guillaumot, Y. L. Tiec, P. Besson, B. Biasse, F. Allain, M. Casse, D. Lafond, J. M. Hartmann, Y. Morand, J. Chiaroni, and S. Deleonibus. Bonded planar Double-Metal-Gate NMOS transistors down to 10 nm. *IEEE Elec. Device Lett.*, 26(5):317–319, May 2005.

[59] J. Widiez, J. Lolivier, T. Poiroux, B. Previtali, F. Dauge, M. Mouis, and S. Deleonibus. Experimental evaluation of gate architecture influence on DG SOI MOSFETs performance. *IEEE Trans. Elec. Devices*, 52(8): 1772–1779, Aug 2005.

[60] V. Barral, T. Poiroux, M. Vinet, J. Widiez, B. Previtali, P. Grosgeorges, G. L. Carval, S. Barraud, J. L. Autran, D. Munteanu, and S. Deleonibus. Experimental determination of the channel backscattering coefficient on 10-70 nm-metal-gate Double-Gate transistors. *Sol. St. Elec.*, 51:537–542, 2007.

[61] Y. Li, H. M. Chou, and J. W. Lee. Investigation of electrical characteristics on surrounding-gate and omega-shaped-gate nanowire FinFETs. *IEEE Trans. Nanotechnol.*, 4(5):510–516, Sept. 2005.

[62] A. Gokirmak and S. Tiwari. Accumulated body ultranarrow channel silicon transistor with extreme threshold voltage tunability. *Appl. Phys. Lett.*, 19(243504), 2007.

[63] R. J. Zaman, K. Mathews, W. Xiong, and S. Banerjee. Trigate FET device characteristics improvement using a hydrogen anneal process with a novel hard mask approach. *IEEE Elec. Device Lett.*, 28(10):916–918, 2007.

[64] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H. P. Wong. Device scaling limits of Si MOSFETs and their application dependencies. *Proc. IEEE*, 89(3):259–288, Mar. 2001.

[65] K. K. Likharev. Electronics below 10-nm. In J. Greer et al., editors, *Giga and Nano Challenges in Microelectronics*, pages 27–68. Elsevier, 2003.

[66] V. K. Khanna. Physics ofcarrier-transport mechanisms and ultra-small scale phenomena for theoretical modelling of nanometer MOS transistors from diffusive to ballistic regimes of operation. *Phys. Rep.*, 398(2):67–131, Aug. 2004.

[67] W. Haensch, E. J. Nowak, R. H. Dennard, P. M. Solomon, A. Bryant, O. H. Dokumaci, A. Kumar, X. Wang, J. B. Johnson, and M. V. Fischetti. Silicon CMOS devices beyond scaling. *IBM J. Res. Dev.*, 50 (4/5):339–361a, Jul./Sept. 2006.

[68] J. S. Blakemore. Approximations for fermi-dirac integrals, especially the function F1/2(eta) used to describe electron density in a semiconductor. *Sol. St. Elec.*, 25(11):1067–1076, 1982.

[69] S. Datta. *Electronic Transport in Mesoscopic Systems*. Cambridge, 1995. ISBN 0-521-41604-3.

[70] B. Nikolic and P. B. Allen. Electron transport through a circular constriction. *Phys. Rev. B*, 60(6):3963–3969, Aug. 1999.

[71] Y. V. Sharvin. A possible method for studying Fermi surfaces. *Sov. Phys. JETP*, 21(3):655–656, 1965.

[72] S. Luryi. Quantum capacitance devices. *Appl. Phys. Lett.*, 52(6):501–503, Feb. 1988.

[73] L. D. Landau and E. M. Lifshitz. *Quantum Mechanics*, volume 3. Pergamon, 3rd edition, 1999. ISBN 0-08-029140-6.

[74] V. A. Sverdlov, Y. Naveh, and K. K. Likharev. Nanoscale SOI ballistic MOSFETs: An impending power crisis. In *Proc. of IEEE SOI Conf.*, pages 151–152, Durango, CO, Oct. 2001.

[75] V. A. Sverdlov, Y. Naveh, and K. K. Likharev. Temperature scaling of nanoscale silicon MOSFETs. *J. Phys. IV*, 12, 2002.

[76] V. A. Sverdlov, Y. Naveh, and K. K. Likharev. Temperature scaling of CMOS circuit power consumption. *Physica E*, 18:151–152, 2002.

[77] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge, 2nd edition, 1997. ISBN 0-521-43108-5.

[78] S. Humphries. *Field Solutions on Computers*. CRC Press, 1997. ISBN 0-8493-1668-5.

[79] R. Pregla and W. Pascher. The method of lines. In T. Itoh, editor, *Numerical Techniques for Microwave and Millimeter-wave Passive Structures*, chapter 6, pages 381–446. Wiley, 1989. ISBN 0-471-62563-9.

[80] S. F. Helfert and R. Pregla. Finite difference expressions for arbitrarily positioned dielectric steps in waveguide structures. *J. Light. Tech.*, 14: 2414–2421, Oct. 1996.

[81] M. A. Celia and W. G. Gray. *Numerical Methods for Differential Equations*, chapter 2. Prentice Hall, 1992. ISBN 0-13-626961-3.

[82] B. Carnahan. *Applied Numerical Methods*. Wiley, 1969. ISBN 471-13507-0.

[83] H. B. Keller. *Numerical Methods for Two-Point Boundary-Value Problems*. Dover, 1991. ISBN 0-486-66925-4.

[84] H. E. Green. The numerical solutions of some important transmission-line problems. *IEEE Trans. Microwave Theory Tech.*, 23:676–692, Sept. 1965.

[85] J. S. Hornsby and A. Gopinath. Numerical analysis of a dielectric-loaded waveguide with a microstrip line - finite difference methods. *IEEE Trans. Microwave Theory Tech.*, 17:684–690, Sept. 1969.

[86] D. D. Johnson. Modified broyden's method for accelerating convergence in self-consistent calculations. *Phys. Rev. B*, 38(18):12807–12813, Dec. 1988.

[87] V. Eyert. A comparative study on methods for convergence acceleration of iterative vector sequences. *J. Comp. Phys.*, 124(2):271–285, Mar. 1996.

[88] C. G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Math. Comp.*, 19(92):577–593, Oct. 1965.

[89] B. J. McCartin. A model-trust region algorithm utilizing a quadratic interpolant. *J. Comp. Appl. Math.*, 91(2):249–259, May 1998.

[90] D. R. Bowler and M. J. Gillan. An efficient and robust technique for achieving self consistency in electronic structure calculations. *Chem. Phys. Lett.*, 324(4):473–476, Jul 2000.

[91] P. Wolfe. The secant method for simultaneous nonlinear equations. *Comm. ACM*, 2(12):12–13, 1959.

[92] D. Anderson. Iterative procedures for nonlinear equations. *J. Assn. Comp. Mach.*, 12(4):547–560, Oct. 1965.

[93] G. P. Srivastava. Broyden's method for self-consistent field convergence. *J. Phys. A*, 17(6):L317–L321, 1984.

[94] P. H. Dederichs and R. Zeller. Self-consistency iterations in electronic structure calculations. *Phys. Rev. B*, 28(10):5462–5472, Nov. 1983.

[95] D. Vanderbilt and S. G. Louie. Total energies of diamond (111) surface reconstructions by a linear combination of atomic orbitals method. *Phys. Rev. B*, 30(10):6118–6130, 1984.

[96] E. Pop, R. Dutton, and K. Goodson. Thermal analysis of ultra-thin body device scaling. In *IEEE Electron Devices Meeting: IEDM Tech. Digest.*, pages 36.6.1–36.6.4, Washington, DC, Dec. 2003.

[97] V. A. Sverdlov, X. Oriols, and K. K. Likharev. Effective boundary conditions for carriers in ultrathin SOI channels. *IEEE Trans. Nanotechnol.*, 2(1):59–63, Mar. 2003.

[98] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover, 1972. ISBN 0-486-61272-4.

[99] V. A. Sverdlov, E. Ungersboeck, H. Kosina, and S. Selberherr. Current transport models for nanoscale semiconductor devices. *Mat. Sci. Eng.*, 58:228–270, 2008.

[100] J. Li, T. J. Walls, and K. K. Likharev. Nanoscale SOI MOSFETs: In search for the best geometry. In G. K. Celler, editor, *SOI Technology and Devices XII*, pages 11–20. ECS, Pennington, NJ, 2005.

[101] L. I. Glazman, G. B. Lesovik, D. E. Khmel'nitskii, and R. I. Shekhter. Reflectionless quantum transport and fundamental ballistic-resistance steps in microscopic constrictions. *JETP Lett.*, 48(4):238–241, Aug. 1988.

[102] A. Szafer and A. D. Stone. Theory of quantum conduction through a constriction. *Phys. Rev. Lett.*, 62(3):300–303, Jan. 1989.

[103] M. V. Fischetti. Effect of the electron-plasmon interaction on the electron mobility in silicon. *Phys. Rev. B*, 44(11):5527–5534, Sept. 1991.

[104] C. Ahn and M. Shin. Ballistic quantum transport in nanoscale schottky-barrier tunnel transistors. *IEEE Trans. Nanotechnol.*, 5:278–283, May 2006.

[105] Press release IBM, 2007. URL http://www-03.ibm.com/press/us/en/pressrelease/21142.wss.

[106] M. Balog, M. Schieber, M. Michman, and S. Patai. Chemical vapor deposition and characterization of HfO2 films from organo-hafnium compounds. *Thin Sol. Films*, 41(3):247–259, Mar. 1977.

[107] International Technology Roadmap for Semiconductors. Process integration, devices and structures, 2003. URL http://www.itrs.net/Links/2003ITRS/Home2003.htm.

[108] J. R. Friedman, V. Patel, W. Chen, S. K. Tolpygo, and J. E. Lukens. Quantum superposition of distinct macroscopic states. *Nature*, 406:43–46, Jul. 2000.

[109] C. A. Hamilton, C. J. Burroughs, and S. P. Benz. Josephson voltage standard - a review. *IEEE Trans. Appl. Supercond.*, 7(2):3756–3761, Jun. 1997.

[110] P. Bunyk, K. K. Likharev, and D. Zinoviev. RSFQ technology: Physics and devices. *Int. J. of High Speed Elec.*, 11(1):257–305, 2001.

[111] T. J. Walls, T. V. Filippov, and K. K. Likharev. Quantum fluctuations in Josephson junction comparators. *Phys. Rev. Lett.*, 89(12):217004, Nov. 2002.

[112] T. J. Walls, D. V. Averin, and K. K. Likharev. Josephson junction comparator as a quantum-limited detector for flux qubit readout. *IEEE Trans. Appl. Supercond.*, 17(2):136–141, June 2007.

[113] N. W. Ashcroft and N. D. Mermin. *Solid State Physics*. Brooks/Cole, 1976. ISBN 0-03-083993-9.

[114] B. D. Josephson. Possible new effects in superconductive tunnelling. *Phys. Lett.*, 1(7):251–253, July 1962.

[115] T. V. Filippov, Y. A. Polyakov, V. K. Semenov, and K. K. Likharev. Signal resolution of RSFQ comparators. *IEEE Trans. Appl. Supercond.*, 5(2):2240–2243, June 1995.

[116] V. K. Semenov, T. V. Filippov, Y. A. Polyakov, and K. K. Likharev. SFQ balanced comparators at a finite sampling rate. *IEEE Trans. Appl. Supercond.*, 7(2):3617–3621, June 1997.

[117] A. M. Savin, J. P. Pekola, T. Holmqvist, J. Hassel, L. Gronberg, P. Helisto, and A. Kidiyarova-Shevchenko A. High-resolution superconducting single-flux quantum comparator for sub-kelvin temperatures. *Appl. Phys. Lett.*, 89(13):133505, Sept. 2006.

[118] T. V. Filippov. Quantum dissipation properties of a Josephson junction comparator. *JETP Lett.*, 61(10):858–864, May 1995.

[119] T. V. Filippov. The quantum dissipative properties of a Josephson junction comparator. *Russian Microelec.*, 25(4):250–256, 1996.

[120] U. Fano. Description of states in quantum mechanics by density matrix and operator techniques. *Rev. Mod. Phys.*, 29(1):74–93, Jan. 1957.

[121] A. O. Caldeira and A. J. Leggett. Path integral approach to quantum brownian motion. *Physica A*, 121(3):587–616, Sept. 1983.

[122] A. O. Caldeira and A. J. Leggett. Quantum tunnelling in a dissipative system. *Ann. Phys.*, 149(2):374–456, Sept. 1983.

[123] S. V. Polonsky, V. K. Semenov, and P. N. Shevchenko. PSCAN - personal superconductor circuit analyzer. *Supercond. Sci. Tech.*, 4(11):667–670, Nov. 1991.

[124] S. K. Park and K. W. Miller. Random number generators - good ones are hard to find. *Comm. ACM*, 31(10):1192–1201, Oct. 1988.

[125] W. C. Stewart. Current-voltage characteristics of Josephson junctions. *Appl. Phys. Lett.*, 12(8):277–280, Apr. 1968.

[126] D. E. McCumber. Effect of AC impedance on dc voltage-current characteristics of superconductor weak-link junctions. *J. Appl. Phys.*, 39(7): 3113–3118, Jun. 1968.

[127] H. B. Callen and T. A. Welton. Irreversibility and generalized noise. *Phys. Rev.*, 83(1):34–40, 1951.

[128] L. D. Landau and E. M. Lifshitz. *Statistical Physics - Part II*, volume 5. Pergamon, 3rd edition, 1999.

[129] T. Ohki, A. Savin, J. Hassel, L. Gronberg, T. Karminskaya, and A. Kidiyarova-Shevchenko. Balanced comparator for RSFQ qubit readout. *IEEE Trans. Appl. Supercond.*, 17(2):128–131, Jun. 2007.

[130] Y. Makhlin, G. Schön, and A. Shnirman. Quantum-state engineering with Josephson junction devices. *Rev. Mod. Phys.*, 73:357–400, Apr. 2001.

[131] C. H. van der Wal, A. C. J. ter Haar, F. K. Wilhelm, R. N. Schouten, C. J. P. M. Harmans, T. P. Orlando, S. Lloyd, and J. E. Mooij. Quantum superposition of macroscopic persistent-current states. *Science*, 290:773–777, Oct. 2000.

[132] D. V. Averin, J. R. Friedman, and J. E. Lukens. Macroscopic resonant tunneling of magnetic flux. *Phys. Rev. B*, 62(17):11802–11811, Nov. 2000.

[133] A. Izmalkov, M. Grajcar, E. Ilichev, T. Wagner, H. G. Meyer, A. Y. Smirnov, M. H. S. Amin, A. M. van den Brink, and A. M. Zagoskin. Evidence for entangled states of two coupled flux qubits. *Phys. Rev. Lett.*, 93(3):037003, Jul. 2004.

[134] C. Granata, B. Ruggiero, M. Russo, and A. Vettoliere. Josephson devices for controllable flux qubit and interqubit coupling. *Appl. Phys. Lett.*, 87 (17):172507, Oct. 2005.

[135] L. Yu-xi, L. F. Wei, J. S. Tsai, and F. Nori. Controllable coupling between flux qubits. *Phys. Rev. Lett.*, 96(6):067003, Feb. 2006.

[136] P. Langevin. A fundamental formula of kinetic theory. *Annales de Chimie et de Physique*, 5:245–288, 1905. In French.

[137] M. Lax. Quantum noise 4. quantum theory of noise sources. *Phys. Rev.*, 145(1):110–129, 1966.

[138] D. V. Averin, A. B. Zorin, and K. K. Likharev. Bloch oscillations in small-size Josephson junctions. *Sov. Phys. JETP*, 61:407, 1985.

[139] L. D. Landau and E. M. Lifshitz. *Statistical Physics - Part I*, volume 5. Pergamon, 3rd edition, 1999.

[140] E. A. Flinn. A modification of Filon method of numerical integration. *J. Assn. Comp. Mach.*, 7(2):181–184, 1960.

[141] K. Petras. Error-estimates for Filon quadrature-formulas. *BIT*, 30(3): 529–541, 1990.

[142] V. Danilov, K. K. Likharev, and A. B. Zorin. Quantum noise in SQUIDs. *IEEE Trans. Magn.*, 19(3):572–575, May 1983.

[143] W. K. Wootters. Statistical distance and Hilbert space. *Phys. Rev. D*, 23(2):357–362, Jan. 1981. doi: 10.1103/PhysRevD.23.357.

[144] D. V. Averin and E. V. Sukhorukov. Counting statistics and detector properties of quantum point contacts. *Phys. Rev. Lett.*, 95(12):126803, Sept. 2005. doi: 10.1103/PhysRevLett.95.126803.

[145] D. V. Averin, K. Rabenstein, and V. K. Semenov. Rapid ballistic readout for flux qubits. *Phys. Rev. B*, 73(9):094505, Mar. 2006. doi: 10.1103/PhysRevB.73.094504.

[146] D. V. Averin. Private communication, 2006.

[147] T. V. Filippov, S. K. Tolpygo, J. Mannik, and J. E. Lukens. Tunable transformer for qubits based on flux states. *IEEE Trans. Appl. Supercond.*, 13(2):1005–1008, Jun. 2003. doi: 10.1109/TASC.2003.814125.

[148] H. A. Haus and J. A. Mullen. Quantum noise in linear amplifiers. *Phys. Rev.*, 128(5):2407–2413, Dec. 1962.

[149] J. G. Simmons. Generalized formula for the electric tunnel effect between similar electrodes separated by a thin insulating film. *J. Appl. Phys.*, 34 (6):1793–1803, Jun. 1963.

[150] J. G. Simmons. Electric tunnel effect between dissimilar electrodes separated by a thin insulating film. *J. Appl. Phys.*, 34(9):2581–2590, Sept. 1963.

[151] Y. C. Yeo, Q. Lu, W. C. Lee, T. J. King, C. Hu, X. Wang, X. Guo, and T. P. Ma. Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric. *IEEE Elec. Device Lett.*, 21(11): 540–542, 2000.

[152] H. Wu, Y. Zhao, and M. H. White. Quantum mechanical modeling of MOSFET gate leakage for high-k dielectrics. *Sol. St. Elec.*, 50:1164–1169, 2006.

[153] L. Yan, S. H. Olsen, M. Kanoun, R. Agaiby, and A. G. O'Neill. Gate leakage machanisms in strained Si devices. *J. Appl. Phys.*, 100:104507, 2006.

[154] B. Brar, G. D. Wilk, and A. C. Seabaugh. Direct extraction of the electron tunneling effective mass in ultrathin SiO_2. *Appl. Phys. Lett.*, 69:2728, 1996.

Appendix A

Auxiliary Calculations

A.1 Gate Leakage

In the WKB approximation, the tunneling probability is given by

$$T(E_x) \approx e^{-2\gamma} \tag{A.1}$$

where

$$\gamma = \frac{1}{\hbar} \int_0^{t_{ox}} |p(x)|\, dx, \tag{A.2}$$

an approach used by Simmons to derive the current density for a general potential barrier for both similar and dissimilar gate and drain electrode materials [149, 150]. The WKB result shows that to a good approximation, the current is dominated simply by the average value of the potential in the film. The current can be expressed for electrodes with identical work functions as

$$J = J_0 \left[\bar{\varphi} e^{-A\bar{\varphi}^{1/2}} - (\bar{\varphi} + eV) e^{-A(\bar{\varphi}+eV)^{1/2}} \right] \tag{A.3}$$

where

$$\begin{aligned} J_0 &\equiv \frac{e}{4\pi^2 \hbar (\beta t_{ox})^2} \\ A &\equiv \frac{2\beta t_{ox}}{\hbar} (2m_x)^{1/2}. \end{aligned}$$

Here, β is a higher order correction factor that in most cases $\beta \approx 1$. $\bar{\varphi}$ is the average value of the potential in the insulator, where in Simmons derivation is measured from the Fermi energy of the gate electrode.

A.1.1 Trapezoidal Barrier

If the exact potential in the insulator is not known, the first approach is to approximate it as linear with potential

$$U(x) = \Phi_{ox} - V_g - V_d \frac{x}{t_{ox}}. \tag{A.4}$$

A.1.1.1 Simmons' Equations

For the case of a trapezoidal barrier, the average value of the potential is given simply in terms of φ_0, the height of the potential barrier above the Fermi level of the gate electrode.

$$\bar{\varphi} = \begin{cases} \varphi_0 & V \to 0 \\ \varphi_0 - V_d/2 & V < \varphi_0/e \\ \varphi_0/2 & V > \varphi_0/e \end{cases},$$

where in the last case, the effective barrier width $\Delta t_{ox} = t_{ox}\varphi_0/V_d$. Limiting expressions for equation A.3 are derived as

$$J = \begin{cases} \frac{3(2m_x\varphi_0)^{1/2}}{2t_{ox}} \left(\frac{e}{2\pi\hbar}\right)^2 V \exp\left[-2t_{ox}/\hbar\left(2m_x\varphi_0\right)^{1/2}\right] & V \to 0 \\ J_0\left[(\varphi_0 - V_d/2)e^{-A(\varphi_0 - V_d/2)^{1/2}} - (\varphi_0 + V_d/2)e^{-A(\varphi_0 + V_d/2)^{1/2}}\right] & V < \varphi_0/e \\ \left(\frac{e^3(F/\beta)^2}{8\pi^2\hbar\varphi_0}\right)\left[e^{-A_1\varphi_0^{3/2}} - \left(1 + \frac{2V_d}{\varphi_0}\right)e^{-A_1\varphi_0^{3/2}(1+V_d/\varphi_0)^{1/2}}\right] & V > \varphi_0/e \end{cases} \tag{A.5}$$

where $F \equiv V_d/t_{ox}$ is the field strength in the insulator and constant

$$A_1 \equiv \frac{2\beta}{\hbar e F} m_x^{1/2}.$$

Figure A.1 shows the results of Simmons equations for a trapezoidal barrier.

The full WKB solution for a trapezoidal barrier is given by

$$\gamma = \frac{2\sqrt{2m_{ox}}t_{ox}}{3\hbar V_d} \Re\left\{(\Phi_0 - V_g - E_x)^{3/2} - (\Phi_0 - V_g - V_d - E_x)^{3/2}\right\} \tag{A.6}$$

Figure A.2 shows the same thin film considered by Simmons. The red line is the result of the full WKB calculation with $E_{\mathrm{F}} = 0.151$ eV. The green line is the full numerical solution of the 1-D transmission problem. The oscillations in the numerical solution are the result of over barrier reflections which are not accounted for in the WKB approximation.

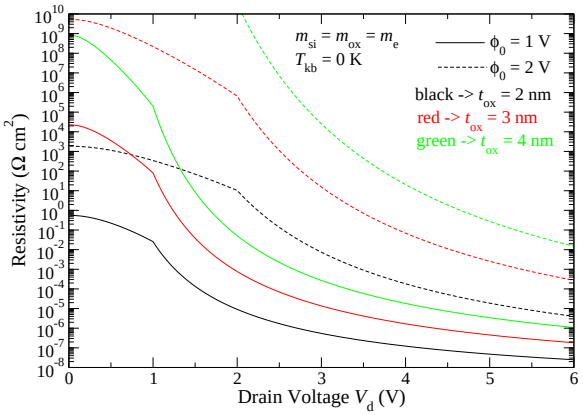

Figure A.1: Resistivity results for transmission through a thin film barrier calculated using Simmons approximation.

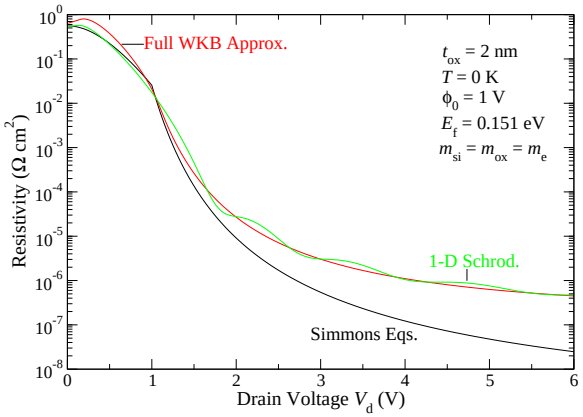

Figure A.2: Resistivity results for transmission through a thin film barrier calculated using WKB approximation.

A.1.2 Applications

Many attempts have been made to apply the WKB approximation to gate leakage current in MOSFET devices [151–153]. Figure A.3 shows the results of equation A.6 to a generic device with parabolic electron mass and an electrode doping density $n_D = 0.3$ nm^{-3} and silicon dioxide band gap $\Phi_{ox} = 3.2$ eV. Here we have used an effective electron mass in the range $0.3 - 0.4m_0$ [154]. A slightly more useful result is obtained by plugging in the effective mass for Si oriented in the (100) plane,shown in Fig. A.4.

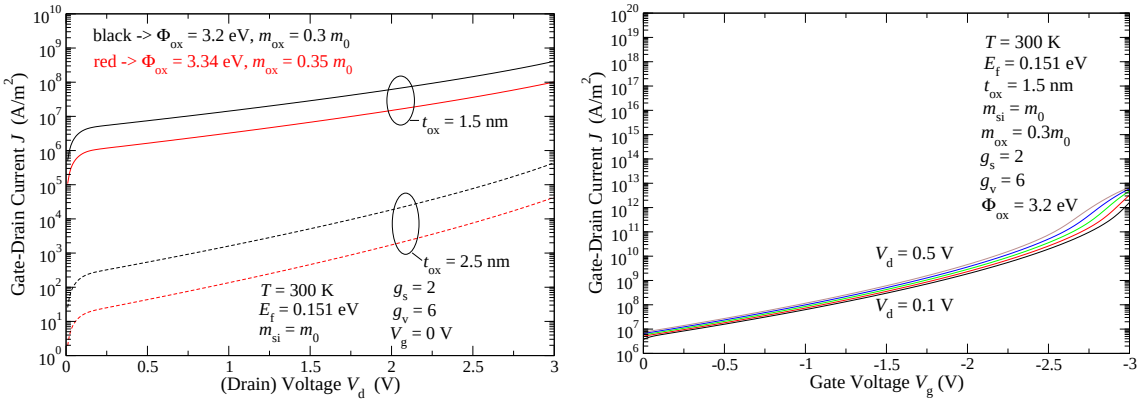

Figure A.3: The gate-drain current density for a generic device as a function of (a) V_d and (b) and V_g.

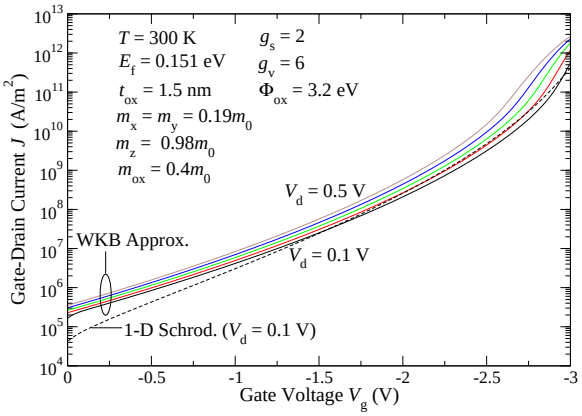

Figure A.4: The gate-drain current density for Si (100) effective mass.

A.1.3 Full Potential

For full account of the actual MOSFET potential, the current density flowing from the gate to the drain is given by

$$J = \int_0^{g_w} J(x,z)dz \tag{A.7}$$

where $J(x,z)$ is identical to Eq. 3.28 and the potential between the gate and drain electrodes is taken from the full self-consistent solution of the Poisson and Schrodinger equations.

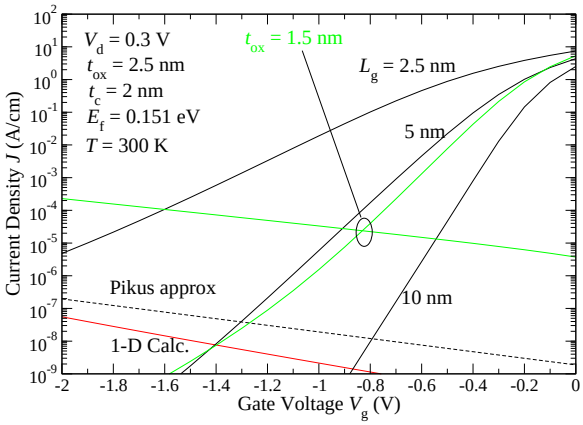

Figure A.5: The gate-drain current density for bulk electrode FET device.

Figure A.5 shows the results for real nano-FET potential profiles. The potential is calculated using a 1-D approximation to the Schrodinger equation for electrons in the channel.

The black lines represent the source-drain current density and the dashed black line represented the gate leakage based on the WKB approximation. The gate-drain currents from the full numeric calculation where nearly identical for all gate lengths and the results are given by the red line. The green lines are the source-drain current and gate-drain current for an oxide thickness $t_{ox} = 1.5$ nm and gate length $L_g = 5$ nm..

A.2 Average Potential

In this appendix we will derive the result for the average potential summed over all wavefunction indicies n. We will assume a generic two dimensional potential $U(x,z)$ that is symmetric in $\hat{z}$ around $U(x,0)$. The average potential defined as

$$\bar{U}_n \equiv \sum_{n'} \int_{-t_c/2}^{t_c/2} U(x,z)\psi_{n'}(z)\psi_n^*(z)dz \tag{A.8}$$

where $\psi(z)$ is the infinite square well wavefunction

$$\psi_n(z) = \left(\frac{2}{t_c}\right)^{1/2} \begin{cases} \cos\left(\frac{n\pi}{t_c}z\right) & n = \text{odd} \\ \sin\left(\frac{n\pi}{t_c}z\right) & n = \text{even.} \end{cases} . \tag{A.9}$$

A.2.1 Example: Quadratic Potential

We begin by looking at a potential of the form

$$U(z) = U_0 + \alpha z^2. \tag{A.10}$$

From equation A.8, the average potential is given by

$$\bar{U}_n = \sum_{n'} \left[U_0 \int_{-t_c/2}^{t_c/2} \psi_{n'}(z)\psi_n^*(z)dz + \alpha \int_{-t_c/2}^{t_c/2} z^2\psi_{n'}(z)\psi_n^*(z)dz \right] . \tag{A.11}$$

The first integral is clearly $\delta_{n,n'}$ so we can write the average potential as

$$\bar{U}_n = U_0 + \bar{U}'_n \tag{A.12}$$

where

$$\begin{aligned} \bar{U}'_n = \frac{2\alpha}{t_c} \Bigg[& \sum_{n'=\text{odd}} \int_{-t_c/2}^{t_c/2} z^2 \cos\left(\frac{n\pi}{t_c}z\right) \cos\left(\frac{n'\pi}{t_c}z\right) dz \\ & + \sum_{n'=\text{even}} \int_{-t_c/2}^{t_c/2} z^2 \cos\left(\frac{n\pi}{t_c}z\right) \sin\left(\frac{n'\pi}{t_c}z\right) dz \Bigg] . \end{aligned} \tag{A.13}$$

By odd-even functionality, the second summation vanishes and using the solution

$$\frac{2}{t_c}\int_{-t_c/2}^{t_c/2} z^2 \cos\left(\frac{n\pi}{t_c}z\right)\cos\left(\frac{n'\pi}{t_c}z\right)dz = \begin{cases} \frac{t_c^2(n^2\pi^2-6)}{12n^2\pi^2} & n=n' \\ -\frac{8(-1)^{(n+n')/2}t^2nn'}{\pi^2(n^2-n'^2)^2} & n\neq n' \end{cases} \tag{A.14}$$

the variation of the average potential becomes

$$\bar{U}'_n = \frac{\alpha t_c^2}{\pi^2}\left[\frac{n^2\pi^2-6}{12n^2} - \sum_{n'=\text{odd},n\neq n'} \frac{8(-1)^{(n+n')/2}nn'}{(n^2-n'^2)^2}\right]. \tag{A.15}$$

The restrictions on the summation are a little awkward, so we define

$$k \equiv \frac{n'-1}{2} \tag{A.16}$$

and this can be re-written as

$$\begin{aligned}\bar{U}'_n = \frac{\alpha t_c}{\pi^2}\Bigg[\frac{n^2\pi^2-6}{12n^2} &- \sum_{k=0}^{(n-3)/2} \frac{8(-1)^{(n+2k+1)/2}n(2k+1)}{(n^2-(2k+1)^2)^2} \\ &- \sum_{k=(n+1)/2}^{\infty} \frac{8(-1)^{(n+2k+1)/2}n(2k+1)}{(n^2-(2k+1)^2)^2}\Bigg],\end{aligned} \tag{A.17}$$

Using

$$\sum_n \frac{(-1)^n}{(n+1)^2} = \frac{1}{12}\pi^2,$$

the two summations in equation A.17 reduce to

$$\begin{aligned}\sum_{k=0}^{(n-3)/2} \frac{8(-1)^{(n+2k+1)/2}n(2k+1)}{(n^2-(2k+1)^2)^2} + \sum_{k=(n+1)/2}^{\infty} \frac{8(-1)^{(n+2k+1)/2}n(2k+1)}{(n^2-(2k+1)^2)^2} \\ = \frac{1}{12}\pi^2 - \frac{1}{2n^2}\end{aligned} \tag{A.18}$$

so the final result is

$$\bar{U}'_n = 0, \tag{A.19}$$

and

$$\bar{U}_n = U_0. \tag{A.20}$$

A.2.2 Fourier Series

For fixed x, we can represent the potential as a Fourier series over the channel thickness t_c.

$$U_x(z)|_{-t_c/2}^{t_c/2} = \frac{a_0}{2} + \sum_{m=1}^{\infty} \left[a_m \cos\left(\frac{2m\pi}{t_c} z\right) + b_m \sin\left(\frac{2m\pi}{t_c} z\right) \right] \tag{A.21}$$

where the Fourier coefficients are given by

$$\begin{aligned} a_m &= \frac{2}{t_c} \int_{-t_c/2}^{t_c/2} U_x(z') \cos\left(\tfrac{2m\pi}{t_c} z'\right) dz' \quad m = 0, 1, \ldots \\ b_m &= \frac{2}{t_c} \int_{-t_c/2}^{t_c/2} U_x(z') \sin\left(\tfrac{2m\pi}{t_c} z'\right) dz' \quad m = 1, \ldots \end{aligned} \tag{A.22}$$

Since $U_x(z)$ is an even function, all $b_m = 0$. Hence, the potential is given by the Fourier series

$$U_x(z)|_{-t_c/2}^{t_c/2} = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos\left(\frac{2m\pi}{t_c} z\right) \tag{A.23}$$

(a) **(b)**

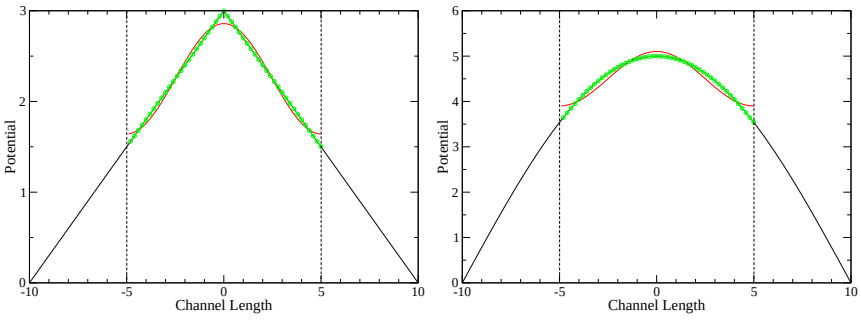

Figure A.6: Potential profiles and Fourier approximations for (a) triangle potential and (b) cos potential. The red line is the Fourier approximation for $N = 1$ and the green dots for $N = 100$.

A.2.3 Average Potential

A.2.3.1 Case: n = odd

Plugging the Fourier expansion of $U_x(z)$ into equation A.8 we get

$$\begin{aligned}\bar{U}_n(x) \;=\; & \sum_{n'=\text{odd}}^{\infty} \left\{ \tfrac{a_0}{2}\left(\tfrac{2}{t_c}\right) \int_{-t_c/2}^{t_c/2} \cos\left(\tfrac{n\pi}{t_c}z\right)\cos\left(\tfrac{n'\pi}{t_c}z\right) dz \right. \\ & \left. + \sum_{m=1}^{\infty} a_m \left(\tfrac{2}{t_c}\right) \int_{-t_c/2}^{t_c/2} \cos\left(\tfrac{2m\pi}{t_c}z\right)\cos\left(\tfrac{n\pi}{t_c}z\right)\cos\left(\tfrac{n'\pi}{t_c}z\right) dz \right\} \\ & + \sum_{n'=\text{even}}^{\infty} \left\{ \tfrac{a_0}{2}\left(\tfrac{2}{t_c}\right) \int_{-t_c/2}^{t_c/2} \cos\left(\tfrac{n\pi}{t_c}z\right)\sin\left(\tfrac{n'\pi}{t_c}z\right) dz \right. \\ & \left. + \sum_{m=1}^{\infty} a_m \left(\tfrac{2}{t_c}\right) \int_{-t_c/2}^{t_c/2} \cos\left(\tfrac{2m\pi}{t_c}z\right)\cos\left(\tfrac{n\pi}{t_c}z\right)\sin\left(\tfrac{n'\pi}{t_c}z\right) dz \right\}\end{aligned}$$

Using basic trig results A.2.4, and collecting the terms

$$\bar{U}_n(x) = \frac{a_0}{2} + \frac{1}{2}\sum_{n'=\text{odd}} \left(a_{\frac{n+n'}{2}} + a_{\frac{|n-n'|}{2}, n\neq n'} \right). \tag{A.24}$$

This can be re-written as

$$\bar{U}_n(x) = \frac{1}{2}\left[a_0 + \lim_{N\to\infty}\left(2\sum_{m=1}^{N} a_m - a_N \right)\right]. \tag{A.25}$$

Since $a_N \to 0$ as $N \to \infty$, the average potential is simply

$$\bar{U}_n(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \tag{A.26}$$

or

$$\bar{U}_n(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos\left(\frac{2m\pi}{t_c}z\right)\Big|_{z=0}. \tag{A.27}$$

Thus

$$\bar{U}_n(x) = U(x,0) \tag{A.28}$$

Figure A.7 shows the results for $n = 1$. The black line is the value of the average potential calculate by brute force evaluation of equation A.8. The red circles are the result of evaluating equation A.24 and the dashed line is the limiting expression A.28.

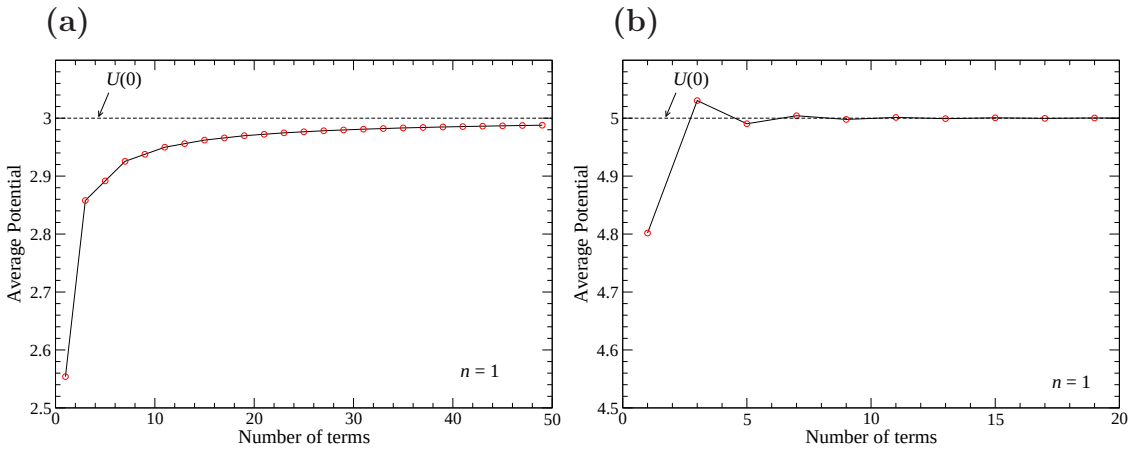

Figure A.7: Average potential for $n = 1$ for (a) linear potential and (b) cos potential.

A.2.3.2 Case: $n =$ even

Plugging the Fourier expansion of $U_x(z)$ into equation A.8 we get

$$\begin{aligned}\bar{U}_n(x) \;=\; & \sum_{n'=\text{odd}}^{\infty} \left\{ \frac{a_0}{2}\left(\frac{2}{t_c}\right) \int_{-t_c/2}^{t_c/2} \sin\left(\frac{n\pi}{t_c}z\right)\cos\left(\frac{n'\pi}{t_c}z\right)dz \right. \\ & \left. + \sum_{m=1}^{\infty} a_m \left(\frac{2}{t_c}\right) \int_{-t_c/2}^{t_c/2} \cos\left(\frac{2m\pi}{t_c}z\right)\sin\left(\frac{n\pi}{t_c}z\right)\cos\left(\frac{n'\pi}{t_c}z\right)dz \right\} \\ & + \sum_{n'=\text{even}}^{\infty} \left\{ \frac{a_0}{2}\left(\frac{2}{t_c}\right) \int_{-t_c/2}^{t_c/2} \sin\left(\frac{n\pi}{t_c}z\right)\sin\left(\frac{n'\pi}{t_c}z\right)dz \right. \\ & \left. + \sum_{m=1}^{\infty} a_m \left(\frac{2}{t_c}\right) \int_{-t_c/2}^{t_c/2} \cos\left(\frac{2m\pi}{t_c}z\right)\sin\left(\frac{n\pi}{t_c}z\right)\sin\left(\frac{n'\pi}{t_c}z\right)dz \right\}\end{aligned}$$

Again using A.2.4, and collecting the terms

$$\bar{U}_n(x) = \frac{a_0}{2} + \frac{1}{2}\sum_{n'=\text{odd}} \left(a_{\frac{|n-n'|}{2}, n\neq n'} - a_{\frac{n+n'}{2}} \right). \tag{A.29}$$

Which can be re-written as

$$\bar{U}_n(x) = \frac{a_0}{2} + \frac{1}{2}\left(2\sum_{m=1}^{(n-2)/2} a_m + a_{n/2} \right) \tag{A.30}$$

Figure A.8 shows the results for $n = 2$. The black line is the value of the

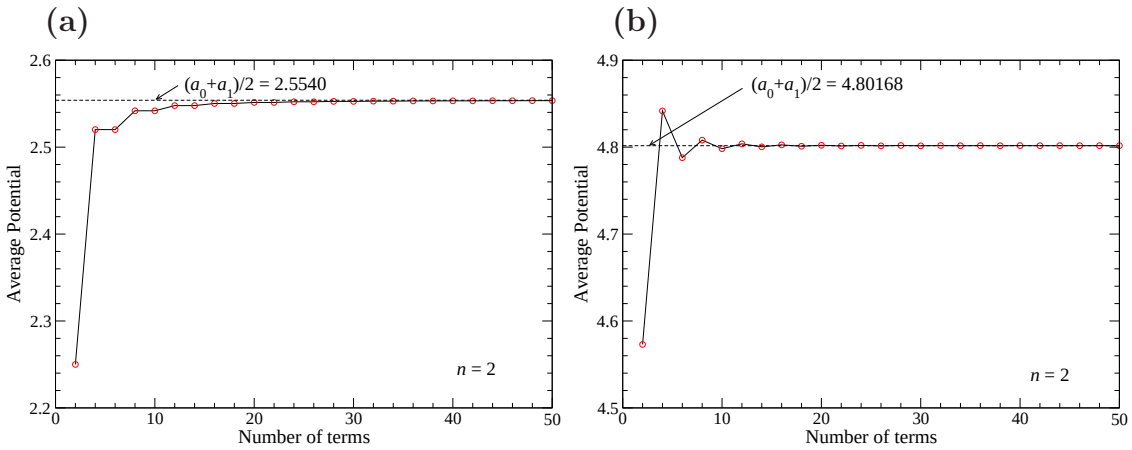

Figure A.8: Average potential for $n = 2$ for (a) linear potential and (b) cos potential.

average potential calculate by brute force evaluation of equation A.8. The red circles are the result of evaluating equation A.29 and the dashed line is the limiting expression A.30.

A.2.4 Basic Trig Integrals

For reference, we will derive some basic results for trigonometric integrals. Here, we assume all n,m,l are integers.

A.2.4.1 Two terms

Here we calculate integrals of the form

$$\int_{-\pi}^{\pi} \sin(nx)\sin(mx)dx \tag{A.31}$$

A.2.4.2 $m \neq n$

$$\begin{aligned}
\int_{-\pi}^{\pi} \sin(nx)\sin(mx)dx &= \frac{1}{2}\int_{-\pi}^{\pi} \left[\cos((n-m)x) - \cos((n+m)x)\right] dx \\
&= \frac{1}{2}\left[\frac{\sin((n-m)x)}{n-m} - \frac{\sin((n+m)x)}{n+m}\right]_{-\pi}^{\pi} \\
&= \left[\frac{\sin((n-m)\pi)(n+m) - \sin((n+m)\pi)(n-m)}{n^2-m^2}\right] \\
&= 0, \quad (n \neq m)
\end{aligned}$$

A.2.4.3 $m = n$

$$\begin{aligned}
\int_{-\pi}^{\pi} \sin^2(nx)dx &= \frac{1}{2}\int_{-\pi}^{\pi} \left(1 - \cos(2nx)\right) dx \\
&= \frac{1}{2}\left[x - \frac{\sin(2nx)}{2n}\right]_{-\pi}^{\pi} \\
&= \pi, \quad (m = n)
\end{aligned}$$

A.2.4.4 Three terms

$$\begin{aligned}
&\int_{-\pi}^{\pi} \cos(nx)\sin(mx)\sin(lx)dx = \\
&= \tfrac{1}{2}\int_{-\pi}^{\pi} \cos(nx)\left(\cos((m-l)x) - \cos((m+l)x)\right) dx \\
&= \frac{1}{4}\int_{-\pi}^{\pi} \left[\cos((n-m+l)x) + \cos((n+m-l)x) - \cos((n-m-l)x)\right. \\
&\qquad\qquad \left. - \cos((n+m+l)x)\right] dx \\
&= \tfrac{1}{4}\left\{ \left.\tfrac{\sin((n-m+l)x)}{n-m+l}\right|_{-\pi}^{\pi} + \left.\tfrac{\sin((n+m-l)x)}{n+m-l}\right|_{-\pi}^{\pi} - \left.\tfrac{\sin((n-m-l)x)}{n-m-l}\right|_{-\pi}^{\pi} - \left.\tfrac{\sin((n+m+l)x)}{n+m+l}\right|_{-\pi}^{\pi} \right\}
\end{aligned}$$

Expanding we see

$$\int_{-\pi}^{\pi} \cos(nx)\sin(mx)\sin(lx) = \begin{cases} -\frac{\pi}{2}, & n = m+l \\ \frac{\pi}{2}, & n = \pm(m-l) \\ 0, & \text{else} \end{cases} \tag{A.32}$$

$$\int_{-\pi}^{\pi} \sin(nx)\sin(mx)\sin(lx)dx =$$
$$= \tfrac{1}{2}\int_{-\pi}^{\pi} \sin(nx)\left(\cos((m-l)x) - \cos((m+l)x)\right)dx$$
$$= \frac{1}{4}\int_{-\pi}^{\pi} [\sin((n-m+l)x) + \sin((n+m-l)x) - \sin((n-m-l)x) - \sin((n+m+l)x)]\,dx$$
$$= \tfrac{1}{4}\left\{ \left.\tfrac{\cos((n-m-l)x)}{n-m-l}\right|_{-\pi}^{\pi} + \left.\tfrac{\cos((n+m+l)x)}{n+m+l}\right|_{-\pi}^{\pi} - \left.\tfrac{\cos((n-m+l)x)}{n-m+l}\right|_{-\pi}^{\pi} - \left.\tfrac{\cos((n+m-l)x)}{n+m-l}\right|_{-\pi}^{\pi} \right\}$$

Expanding we see

$$\int_{-\pi}^{\pi} \sin(nx)\sin(mx)\sin(lx) = 0 \tag{A.33}$$

$$\int_{-\pi}^{\pi} \sin(nx)\cos(mx)\cos(lx)dx =$$
$$= \tfrac{1}{2}\int_{-\pi}^{\pi} \sin(nx)\left(\cos((m-l)x) + \cos((m+l)x)\right)dx$$
$$= \frac{1}{4}\int_{-\pi}^{\pi} [\sin((n-m+l)x) + \sin((n+m-l)x) + \sin((n-m-l)x) + \sin((n+m+l)x)]\,dx$$
$$= \tfrac{1}{4}\left\{ \left.\tfrac{\cos((n-m-l)x)}{n-m-l}\right|_{-\pi}^{\pi} + \left.\tfrac{\cos((n+m+l)x)}{n+m+l}\right|_{-\pi}^{\pi} + \left.\tfrac{\cos((n-m+l)x)}{n-m+l}\right|_{-\pi}^{\pi} + \left.\tfrac{\cos((n+m-l)x)}{n+m-l}\right|_{-\pi}^{\pi} \right\}$$

Expanding we see

$$\int_{-\pi}^{\pi} \sin(nx)\cos(mx)\cos(lx) = 0. \tag{A.34}$$

$$
\begin{aligned}
&\int_{-\pi}^{\pi} \cos(nx)\cos(mx)\cos(lx)dx = \\
&= \tfrac{1}{2}\int_{-\pi}^{\pi} \cos(nx)\left(\cos((m-l)x) + \cos((m+l)x)\right)dx \\
&= \frac{1}{4}\int_{-\pi}^{\pi} [\cos((n-m+l)x) + \cos((n+m-l)x) + \cos((n-m-l)x) \\
&\qquad\qquad + \cos((n+m+l)x)]\,dx \\
&= \tfrac{1}{4}\left\{ \tfrac{\sin((n-m-l)x)}{n-m-l}\Big|_{-\pi}^{\pi} + \tfrac{\sin((n+m+l)x)}{n+m+l}\Big|_{-\pi}^{\pi} + \tfrac{\sin((n-m+l)x)}{n-m+l}\Big|_{-\pi}^{\pi} + \tfrac{\sin((n+m-l)x)}{n+m-l}\Big|_{-\pi}^{\pi} \right\}
\end{aligned}
$$

Expanding we see

$$
\int_{-\pi}^{\pi} \cos(nx)\cos(mx)\cos(lx) = \begin{cases} \frac{\pi}{2}, & n = m+l,\ \pm(m-l) \\ 0, & \text{else.} \end{cases} \tag{A.35}
$$

A.3 Classical Approach

First derived by Pikus and Likharev [32, 65], these compact expressions yield an analytical expression for channel electrons which was the impetus behind the numerical approaches explored here.

In the classical approximation, electrons enter the channel with unit probability for all energies greater than the conduction band minimum. If there is a potential barrier between the source and drain regions, then the electrons will be reflected from the barrier adding an additional term to the electron density.

We begin by considering the case when there is a single potential maximum ϕ_0 between the source and drain regions at position x_0. We can then consider the channel as two distinct regions. For the case when $x < x_0$, there are contributions from three different electron sources.

1. **Electrons incident from the source.** The electrons incident from the source are calculated as

$$
n_{2D}^{(1)}(x) = \frac{g_s g_v}{(2\pi)^2}\iint_{k_x>0} d^2k f(E), \tag{A.36}
$$

or

$$
n_{2D}(x) = \frac{\sqrt{m_x m_y}T}{\pi\hbar^2}\ln\left[1+\exp((E_{\mathrm{F}}-\Phi(x))/T)\right]. \tag{A.37}
$$

2. **Electrons incident from the drain.**

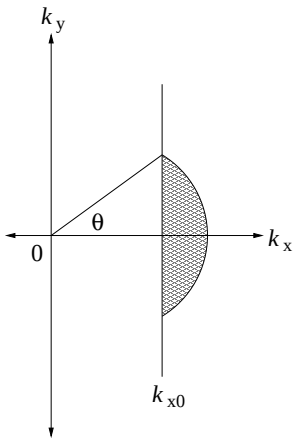

Figure A.9: Effective classical integration region.

The density of electrons incident from the drain is again given by

$$n_{2D}^{(2)}(x) = \frac{g_s g_v \sqrt{m_x m_y}}{(2\pi)^2} \int d\theta \int k'\, dk'\, f(E + V_D). \tag{A.38}$$

But we now only consider electrons with sufficient energy $E_x > \Phi_0$ to pass the potential barrier. To evaluate the integral, we will assume a fixed energy in the $\hat{x}$ direction incident from the drain $k'_x = \sqrt{\Phi_0}/\hbar$. At position $x < x_0$, the particle's energy is

$$k'_{x0} = \frac{\sqrt{\Phi_0 - \Phi(x)}}{\hbar}, \tag{A.39}$$

yielding maximum angle

$$\theta_0 = \arctan\left(\frac{k'_y}{k'_{x0}}\right). \tag{A.40}$$

The integral over theta is trivial, and we can now integrate only along the line

$$k'^2 = k'^2_y + k'^2_{x0} \tag{A.41}$$

$$k'\, dk' = k'_y\, dk'_y, \tag{A.42}$$

giving expression

$$n_{2D}^{(2)}(x) = \frac{g_s g_v \sqrt{m_x m_y}}{(2\pi)^2} \int_{-\infty}^{\infty} dk'_y \, k'_y f(E_y + \Phi_0 + V_D). \tag{A.43}$$

A graphical representation of the integral is shown in Fig. A.9. Expressing this integral in terms of dimensionless energy variable $\epsilon \equiv E/T$ yields

$$n_{2D}^{(2)}(x) = \frac{2 g_s g_v \sqrt{m_x m_y} T}{(2\pi\hbar)^2} \int_0^{\infty} d\epsilon_y \arctan\left(\sqrt{\frac{\epsilon_y T}{\Phi_0 - \Phi(x)}}\right) f\left(\epsilon_y T + \Phi_0 + V_D\right) \tag{A.44}$$

3. **Electrons reflected from the barrier.** All electrons incident from the source for $x < x_0$ and energy $E_x < \Phi_0$ are reflected. These electrons contribute an additional

$$n_{2D}^{(3)}(x) = \frac{g_s g_v}{(2\pi)^2} \iint_{k_x>0}^{E_x<=\Phi_0} d^2k \, f(E).$$

This integral becomes easier to express if the upper limit $E_x \to \infty$. To achieve this we add and subtract a term with limits $E_x = [\Phi_0, \infty]$. Hence, we have two terms with forms equivalent to equations A.37, A.44, but adjusted for the drain voltage. This gives

$$n_{2D}^{(3)}(x) = \frac{g_s g_v \sqrt{m_x m_y} T}{4\pi\hbar^2} \left\{ \ln\left[1 + \exp((E_F - \Phi(x))/T)\right] - \frac{2}{\pi} \int_0^{\infty} d\epsilon_y \arctan\left(\sqrt{\frac{\epsilon_y T}{\Phi_0 - \Phi(x)}}\right) f\left(\epsilon_y T + \Phi_0\right) \right\} \tag{A.45}$$

The total electron density is then given by a summation of the three different contributions $n_{2D}(x) = n_{2D}^{(1)} + n_{2D}^{(2)} + n_{2D}^{(3)}$. It now becomes convenient to

define function

$$Z(\alpha;\beta) \equiv \frac{2}{\pi}\int_0^\infty d\epsilon_y \frac{\arctan\left(\sqrt{\frac{\epsilon_y}{\beta}}\right)}{1+\exp\left(\epsilon_y+\beta-\alpha\right)} \tag{A.46}$$

$$Z(\alpha;0) \equiv \ln\left(1+e^{\alpha}\right) \tag{A.47}$$

Summing all three terms, and using degeneracy factors $g_s = 2, g_v = 2$ the total electron density for particles to the left of the barrier ($x < x_0$) is

$$n_{2D}(x<x_0) = \frac{2\sqrt{m_x m_y}T}{\pi\hbar^2}\left[Z\left(\frac{E_F-\Phi(x)}{T};0\right) - \frac{1}{2}Z\left(\frac{E_F-\Phi(x)}{T};\frac{\Phi_0-\Phi(x)}{T}\right) + \frac{1}{2}Z\left(\frac{E_F-V_D-\Phi(x)}{T};\frac{\Phi_0-\Phi(x)}{T}\right)\right]. \tag{A.48}$$

For the electrons to the right of the potential maximum ($x > x_0$), the derivation is exactly the same, with the replacement that source and drain Fermi levels should be reversed. The expression becomes

$$n_{2D}(x>x_0) = \frac{2\sqrt{m_x m_y}T}{\pi\hbar^2}\left[Z\left(\frac{E_F-V_D-\Phi(x)}{T};0\right) - \frac{1}{2}Z\left(\frac{E_F-V_D-\Phi(x)}{T};\frac{\Phi_0-\Phi(x)}{T}\right) + \frac{1}{2}Z\left(\frac{E_F-\Phi(x)}{T};\frac{\Phi_0-\Phi(x)}{T}\right)\right]. \tag{A.49}$$

To calculate the current density in the classical approach, we note that transmission probability $T(E_x) = 1$ if $E_x > \Phi_0$ and $T(E_x) = 0$, if $E_x < \Phi_0$. Thus using relations

$$\begin{aligned} k_x dk_x &= \frac{m_x}{\hbar^2}dE_x \\ dk_y &= \frac{\sqrt{2m_y}}{2\hbar}E_y^{-1/2}dE_y \end{aligned}$$

and dimensionless energy variables $\epsilon \equiv E/T$, summation 2.34 may be written as

$$J_{\rm cl} = \frac{\sqrt{2m_y}T^{3/2}}{\pi^2\hbar^2}\int_0^\infty \epsilon_y^{-1/2}\,d\epsilon_y \int_{\Phi_0/T}^\infty f(E)\,d\epsilon_x. \tag{A.50}$$

Shift the second integral by potential maximum and evaluating we find the one direction classical current density

$$J_{\rm cl}(\Phi_0) = \frac{J_0}{\pi}\int_0^\infty \epsilon_y^{-1/2}\ln\left[1+\exp\left(-\Phi_0/T-\epsilon_y\right)\right] \tag{A.51}$$

where J_0 is defined by Eq. 3.29.

A.4 $\bar{A}$ Approximations

The formalism used in section 4.1.3 for the 2-D solution of the Schrodinger equation was originally developed by Szafer and Stone to formulate a so-called "mean-field approximation" (MFA) valid at higher sub-band energies [102]. It is motivated by the observation that when $t_B \gg t_c$, $|a_{nw}|^2$ is strongly peaked around confined state wavefunctions $q_{w,n} = [w, n]\pi/t_{B,c}$ such that $q_w \approx q_n$. Hence, the true coupling strength may approximated as uniform coupling to all modes w within one step of the confinement energy E_n.

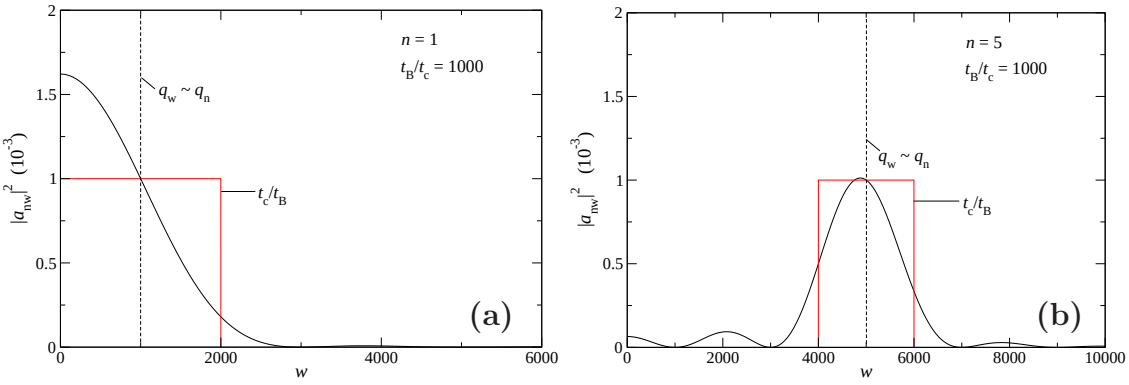

Figure A.10: Overlap $|a_{nw}|^2$ versus wavenumber for (a) $n = 1$ and (b) $n = 5$. Dashed lines represents the point $q_w = q_n$ and the red box shows the MFA assumption.

The calculated $|a_{nw}|^2$ versus wavenumber are shown in Fig. A.10. The dashed lines show the point $q_w = q_n$. Noting that the coupling between different channel sub-bands may be assumed small at high mode indexes, the product of two arbitrary overlaps may be written

$$a_{nw}a_{mw} \approx \delta_{nm}\left(\frac{t_c}{t_B}\right)\left[\Theta(q_w - q_{n-1}) - \Theta(q_{n+1} - q_w)\right]. \tag{A.52}$$

This approximation is shown as the red boxes in Fig A.10.

Since the value $\bar{A}_{nm}$ relies only the product of two a_{nw}, the kernel elements may be written

$$\bar{A}_{nm} \approx \delta_{nm}\sum_w a_{nw}a_{mw}k_w \tag{A.53}$$

to a good approximation. This limit is valid when $t_B/t_c \gg 1$, and the sum-

mation can be expressed as the integral

$$\bar{A}_{nm} \approx \delta_{nm} \frac{t_c}{t_B} \sum_{\nu} k_\nu \rightarrow \delta_{nm} \frac{t_c}{t_B} \frac{1}{2} \int k_\nu d\nu.$$

The factor of 1/2 comes from the fact that only either even or odd modes contribute to the summation. The differential $d\nu$ is evaluated by defining angle θ such that

$$\begin{aligned} q_\nu &= |k| \cos(\theta) \\ k_\nu &= |k| \sin(\theta), \end{aligned}$$

or

$$\theta_\nu \equiv \frac{q_v}{|k|} \tag{A.54}$$

for particle of fixed energy $E = \hbar^2 |k|^2 / 2m$. Thus

$$\begin{aligned} q_\nu &= \frac{\nu \pi}{t_B} \\ &= |k| \cos(\theta) \\ \frac{\pi}{t_B} d\nu &= -|k| \sin(\theta) d\theta. \end{aligned}$$

Plugging into $\bar{A}_{nm}$ we find

$$\bar{A}_{nm} \approx \delta_{nm} \frac{t_c}{2\pi} |k|^2 \int_{\theta_{n-1}}^{\theta_{n+1}} -\sin^2(\theta) d\theta$$

or

$$\bar{A}_{nm} \approx \delta_{nm} \frac{t_c}{8\pi} |k|^2 \left[\sin(2\theta) - 2\theta \right]_{\theta_{n-1}}^{\theta_{n+1}}. \tag{A.55}$$

When the energy of the next confined level E_{n+1} is greater than the particle energy, $\bar{A}_{nn}$ will have an imaginary component. We in general will write $\bar{A}_{nm} = K_{nm} + \mathrm{i} J_{nm}$ where

$$K_{nm} \equiv \Re(\bar{A}_{nm}) \tag{A.56}$$

$$J_{nm} \equiv \Im(\bar{A}_{nm}). \tag{A.57}$$

Figure A.11 shows the full calculation if $\bar{A}_{nm}$ versus index and energy compared to the MFA approximation.

When the masses in the $\hat{x}$ and $\hat{z}$ are not equal, the calculation of $\bar{A}$ needs

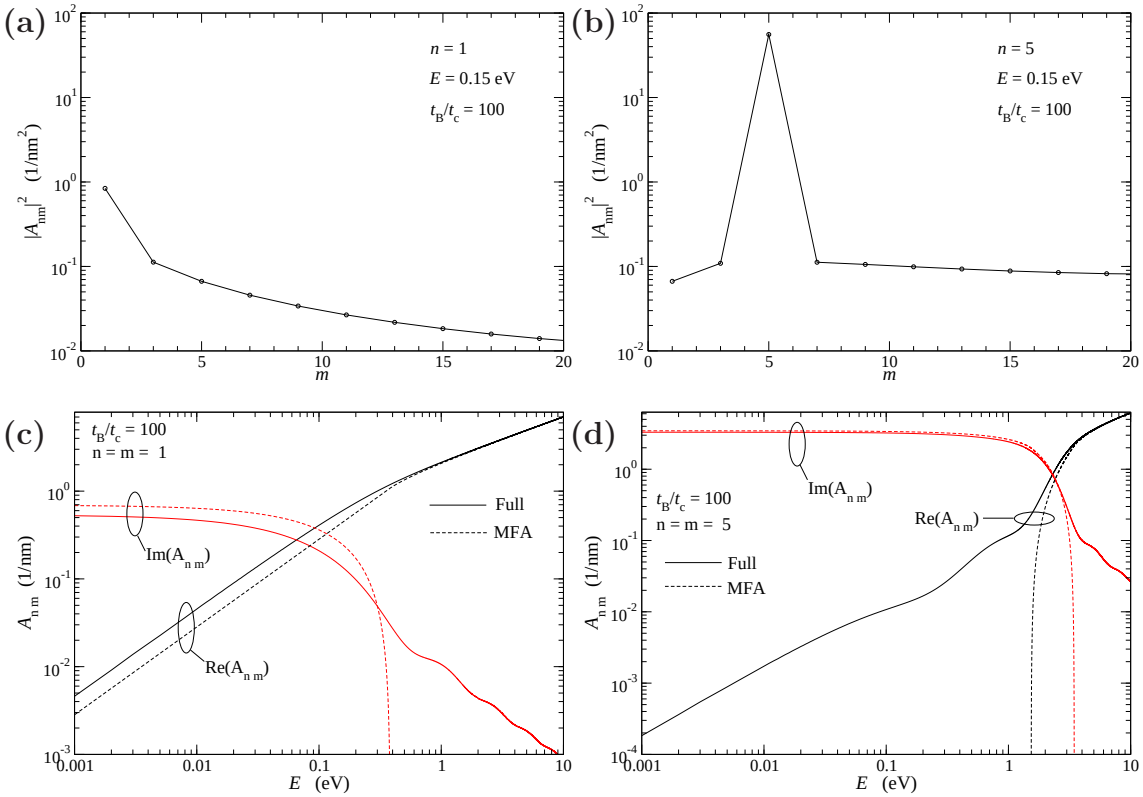

Figure A.11: Coupling strength $\bar{A}$ versus mode index m (panels (a),(b)) and energy (panels (c),(d)) for two indexes $n = 1$, $n = 5$.

a slight adjustment. Using the relation

$$E = \frac{\hbar}{2}\left(\frac{k_x^2}{m_x} + \frac{k_z^2}{m_z}\right) \tag{A.58}$$

we see the definitions

$$\begin{aligned} k_x' &= \frac{\sqrt{2m_x}}{\hbar}k_x \\ k_z' &= \frac{\sqrt{2m_z}}{\hbar}k_z \end{aligned}$$

create the uniform density surface $E = k_x'^2 + k_z'^2$ and the expression for $\bar{A}$ in the MFA becomes

$$\boxed{\bar{A}_{nm} \approx \delta_{nm}\frac{t_c}{4\pi}\frac{\sqrt{m_x m_z}}{\hbar^2}E\left[\sin(2\theta) - 2\theta\right]_{\theta_{n-1}}^{\theta_{n+1}}}. \tag{A.59}$$

A.4.1 Transmission at a Step

Applying approximation A.59, to the single mode transmission probability at a wide narrow junction

$$\bar{t}_{nw}k_n + \sum_m \bar{A}_{nm}\bar{t}_{mw} = 2k_w a_{nw}, \tag{A.60}$$

the un-normalized transmission may be written

$$\bar{t}_{nw} = \frac{2k_w a_{nw}}{k_n + \bar{A}_{nn}}. \tag{A.61}$$

and the full transmission from all available modes w into mode n is given by (with proper normalization $t_{nw} = \sqrt{k_n/k_w}\bar{t}_{nw}$)

$$T_n = \sum_w^{(E)} |t_{nw}|^2 = \frac{4k_n K_{nn}}{(k_n + K_{nn})^2 + J^2}. \tag{A.62}$$

where the notation (E) signifies that only modes allowed by particle energy E are summed. Expression A.62 has obvious parallels with the analogous solution for the 1-D problem of transmission over a step

$$T_{1\mathrm{d}}(E) = \frac{4k_w k_n}{(k_w + k_n)^2}. \tag{A.63}$$

A.4.2 δ-$\bar{A}$ Approximation

While the mean field approximation is very handy, another very useful approximation is obtained by

$$\bar{A}_{nm} \approx \delta_{nm}\bar{A}_{nm} \tag{A.64}$$

while maintaining a full calculation of the $\bar{A}_{nn}$ terms. Since the full overlap strength is still being calculated, this approximation is equivalent to simply turning off the coupling between different channel sub-bands. In both the MFA and δ_{A} approximations, the calculation time is greatly reduced because $\bar{A}$ is now diagonal. This removes the numerical complexities of the inversion of matrix M_4 (see section 4.1.3) as the inversion may be done analytically beforehand. These results may be used to obtain a relatively good approximate 2-D answer while maintaining a computational speed close to the 1-D approximation.

For the case of unit boundary conditions (4.61), the weighting coefficients

may be written explicitly as

$$\begin{aligned} C_{nw} &= \frac{2\mathrm{i}k_w a_{nw}(\mathrm{i}A_n^\nu - g'_{nL})}{((\mathrm{i}A_n^w + f'_{n0})(\mathrm{i}A_n^\nu - g'_{nL}) + g'_{n0}f'_{nL})}, \\ D_{nw} &= \frac{2\mathrm{i}k_w a_{nw} f'_{nL}}{((\mathrm{i}A_n^w + f'_{n0})(\mathrm{i}A_n^\nu - g'_{nL}) + g'_{n0}f'_{nL})}, \\ F_{\nu w} &= \sum_n \frac{2\mathrm{i}k_w a_{nw} a_{n\nu} f'_{nL}}{((\mathrm{i}A_n^w + f'_{n0})(\mathrm{i}A_n^\nu - g'_{nL}) + g'_{n0}f'_{nL})}, \\ r_{ww'} &= \sum_n C_{nw} a_{nw'} - \delta_{ww'}, \end{aligned} \tag{A.65}$$

and compact expressions are possible because the cross terms vanish.

The total transmission coefficient $\mathcal{D}(E)$ is found to be

$$\mathcal{D}(E) = \sum_n \frac{4K_n^{(w)} K_n^{(\nu)} f'^2_{nL}}{\left|((\mathrm{i}A_n^w + f'_{n0})(\mathrm{i}A_n^\nu - g'_{nL}) + g'_{n0}f'_{nL})\right|^2}, \tag{A.66}$$

and the normalized sum of the wavefunction $\psi_{n,II}(x) = C_n f_n(x) + D_n g_n(x)$ is given by

$$\sum_w k_w^{-1} \left|\Psi_{II}\right|^2 = \sum_n \frac{4K_n^{(w)} \left|(\mathrm{i}A_n^{(\nu)} - g'_{nL})f(x) + f'_{nL}g(x)\right|^2}{\left|((\mathrm{i}A_n^w + f'_{n0})(\mathrm{i}A_n^\nu - g'_{nL}) + g'_{n0}f'_{nL})\right|^2} \varphi_n^2(z). \tag{A.67}$$

Again the vector $K_n = \Re(A_{nn})$ has been used.

Plugging the results into the expression for the wavefunction in the source bulk yields

$$\Psi_I = 2\mathrm{i}\left[\chi_w(z)\sin(k_w x) + k_w \sum_n \frac{a_{nw}\left(\mathrm{i}\bar{A}_n^\nu - g'_{nL}\right)}{\left((\mathrm{i}\bar{A}_n^w + f'_{n0})(\mathrm{i}\bar{A}_n^\nu - g'_{nL}) + g'_{n0}f'_{nL}\right)} \bar{\Lambda}_n^w\right]. \tag{A.68}$$

For the drain region we get wavefunction

$$\Psi_{III} = 2\mathrm{i}k_w \sum_n \frac{a_{nw} f'_{nL}}{\left((\mathrm{i}\bar{A}_n^w + f'_{n0})(\mathrm{i}\bar{A}_n^\nu - g'_{nL}) + g'_{n0}f'_{nL}\right)} \bar{\Lambda}_n^\nu. \tag{A.69}$$

For brevity of notation we define

$$\begin{aligned}\Xi_n^w &= \frac{(\mathrm{i}A_n^\nu - g'_{nL})}{\left((\mathrm{i}\bar{A}_n^w + f'_{n0})(\mathrm{i}\bar{A}_n^\nu - g'_{nL}) + g'_{n0}f'_{nL}\right)}\bar{\Lambda}_n^w, \\ \Xi_n^\nu &= \frac{f'_{nL}}{\left((\mathrm{i}\bar{A}_n^w + f'_{n0})(\mathrm{i}\bar{A}_n^\nu - g'_{nL}) + g'_{n0}f'_{nL}\right)}\bar{\Lambda}_n^\nu,\end{aligned} \tag{A.70}$$

and the normalized summations are calculated as

$$\sum_w k_w^{-1}\left|\Psi_I\right|^2 = 4\left[\sum_w\left(k_w^{-1}\chi_w^2(z)\sin^2(k_w x) + 2\chi_w(z)\sin(k_w x)\sum_n a_{nw}\Re(\Xi_n^w)\right) + \sum_n \Re(\bar{A}_n^w)\left|\Xi_n^w\right|^2\right], \tag{A.71}$$

$$\sum_w k_w^{-1}\left|\Psi_{III}\right|^2 = 4\sum_n \Re(\bar{A}_n^w)\left|\Xi_n^\nu\right|^2. \tag{A.72}$$

A.4.3 Comparison of Transmission Probabilities

The transmission probability versus energy is shown in Fig. A.12 for the classical approximation (dashed line), WKB approximation (red line), 1-D approximation to the Schrodinger equation (green line), MFA approximation (blue line), δ_{A} (orange line with diamonds), and 2-D solution (black line). The adiabatic transport assumed by the WKB and 1-D solutions over-estimate the transmission by neglect of the back-scattering at the source electrode. The MFA approximation tends to underestimate the transmission by assuming equal coupling and the δ_{A} tends to slightly under-estimate transmission by neglect of channel sub-band coupling.

A.5 Integral of $\cos(\nu x)$

The limiting integral of the cosine function may be found rather straightforwardly from the definition of the Dirac delta function [98]

$$\delta(x) \equiv \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\mathrm{i}\nu x}\, d\nu, \tag{A.73}$$

and the expansion of cosine in the complex plane

$$\cos(\nu x) = \frac{1}{2}\left(e^{\mathrm{i}\nu x} + e^{-\mathrm{i}\nu x}\right). \tag{A.74}$$

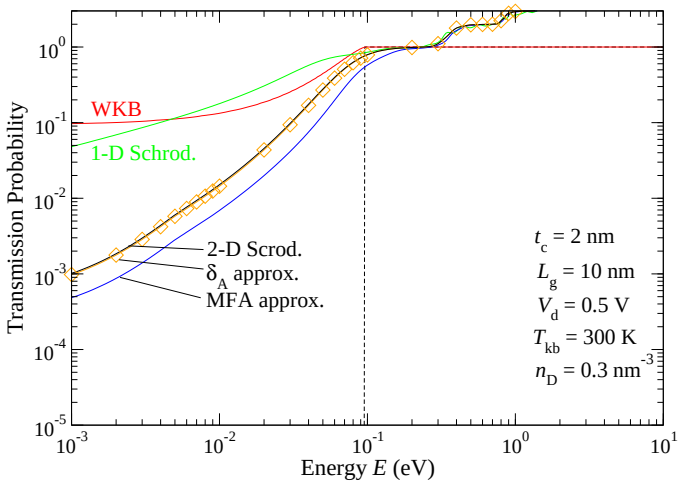

Figure A.12: Transmission probability for various wavefunction approximations.

Thus, we may replace each term in the integral $\int_0^\infty \cos(\nu x)\, d\nu$ with

$$\int_0^\infty e^{\mathrm{i}\nu x}\, d\nu = \pi\delta(x), \tag{A.75}$$

and we arrive at the fundamental result

$$\int_0^\infty \cos(\nu x)\, d\nu = \pi\delta(x). \tag{A.76}$$

Appendix B

General Numeric Methods

B.1 Energy Integral

A key component to the speed of the self-consistent algorithm is the rapid evaluation of the integral over energy states required for calculation of the electron density. Because this integral must be calculated at each channel node point, the reduction of redundant calculations of the wavefunction is essential. The main idea of the algorithm, known as "Richardson's deferred approach", is to calculate successive numeric approximations to the integral in decreasing step size Δ_i, then extrapolate the value to $\Delta \to 0$. For the special case of trapezoidal quadrature of and spacing steps

$$\Delta_i = \frac{\Delta_0}{2^i} \tag{B.1}$$

this algorithm is known as Romberg integration. This is a special case because the leading error term in successive calculations cancels and the numeric result converges as Δ^2 [77]. In the usual case, the extrapolation to zero is done using a fifth order polynomial fitting the last five calculation steps.

We have adopted this algorithm for use in all integrations over energy. When calculating the channel electron density, the range of integration is initially divided into 2^n parts. The wavefunction at each energy E_i is evaluated and the array of wavefunction values for each energy is stored for each node point. Value of the integral at each node point is the calculated for the stored arrays using the extrapolation to zero and the error of each integral is estimated. If the error term in the result is not within a specified tolerance, the number of energy points is doubled. This is equivalent to placing a new node point halfway between E_i and E_{i+1}. Thus we may reuse previously calculated wavefunctions only introducing more as greater accuracy is needed. In the

way we combine the power of Romberg's integration scheme and eliminate the waste of redundant wavefunction evaluations.

B.2 Parallelization and Run Times

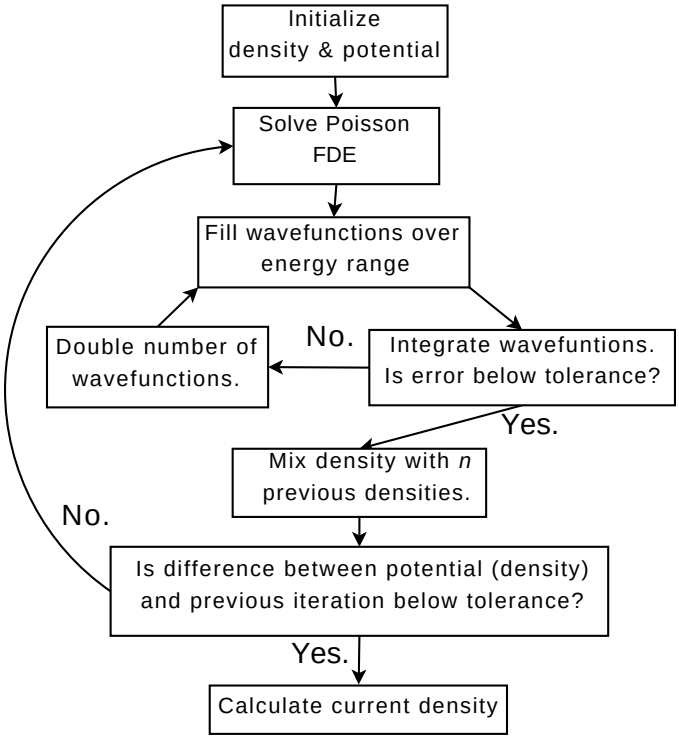

Figure B.1: Decision graph of the self-consistent algorithm.

The algorithm for the self-consistent calculation is shown as a decision graph in Fig. B.1. It is the algorithm used to calculate a single heavy mass valley for all 1-D results. It may be immediately parallelized in two ways. The first is at the result level, were all individual data points may be calculated simultaneously. The second step to parallelize the code may be done during the integration over energy states required for calculation of the electron density. The integration calculated as a summation of energy states, may be split into m parts, and calculated in parallel on slave nodes. This is shown as the red box in Fig. B.2.

The calculations times of a standard system with $L_g = 10$ nm, $V_d = V_g = 0$ V are shown in Table B.1. The calculations include only heavy mass orientation and are performed on a 1.73 GHz $i686$ laptop processor. The number of

Approx.	**# Iters**	**Run-time: 9-point FDE (s)**
Classical	24	23
WKB	36	49
1-D Schrodinger	25	67
2-D MFA	24	269
2-D δ_{A}	24	367
2-D Schrodinger	24	783

Table B.1: Calculation times for a benchmark system ($L_g = 10$ nm, $t_c = 2$ nm, $t_{ox} = 1.5$ nm, $n_D = 0.3$ nm^{-3}, $T = 300$ K, $V_g = 0$ V, $V_d = 0$ V) and various quantum mechanical channel wavefunction approximations.

iterations required for the system reach self-consistency is shown in the second column while the total run time is shown in column three.

The full distributed algorithm is shown in Fig. B.2. The final parallelization is performed in the 2-D solution where all three heavy mass orientations are included. The blue and green parenthesis show the calculation of the electron density performed for each orientation at the same time.

The total run time on various computing architectures is shown in Table B.2. The benckmark system was set with $L_g = 10$ nm and $V_d = V_g = 100$ meV. The serial computation was done using only a single electron valley. For the 2-D calculation, parallelization of the energy integral incurs a large communication cost because the wavefunction arrays for **all** the node points in the channel must be passed between nodes. This cost can actually slow down the overall computation time on machines with faster processors, hence no parallelization of the energy was used. The serial task was used only a single electron valley, so this job has no parallelization at all even on the cluster machines. The parallel job was calculated using all three heavy mass orientation so the run utilized 4 processors simultaneously: three for each valley and one overall master node.

Machine	**CPU Type**	**Ethernet**	**# Nodes**	**# CPUs**	**Max CPUs per Job**	**Serial Job Time (s)**	**Parallel Job Time (s)**
Njal	2.0 GHz Xeon	100 Mb Ethernet	14	56	56	1209	1498
Seawulf	3.4 GHz Xeon	1 Gb Ethernet	235	470	120	1215	1345
Kraken	1.7 GHz Power4+	4 Gb/sec switch	368	2994	512	1821	1352
Babbage	1.9 GHz Power5+	4 Gb/sec switch	192	3072	512	1634	1169
NY Blue	700 MHz PPC 440	10 Gb/sec switch	18432	36864	2048	5783	4435
Laptop	1.73 GHz i686	N/A	N/A	1	1	2236	N/A

Table B.2: Calculation times for a benchmark 2-D system ($L_g = 10$ nm, $t_c = 2$ nm, $t_{ox} = 1.5$ nm, $n_D = 0.3$ nm^{-3}, $T = 300$ K, $V_g = 0$ V, $V_d = 0$ V) on serial and parallel systems.

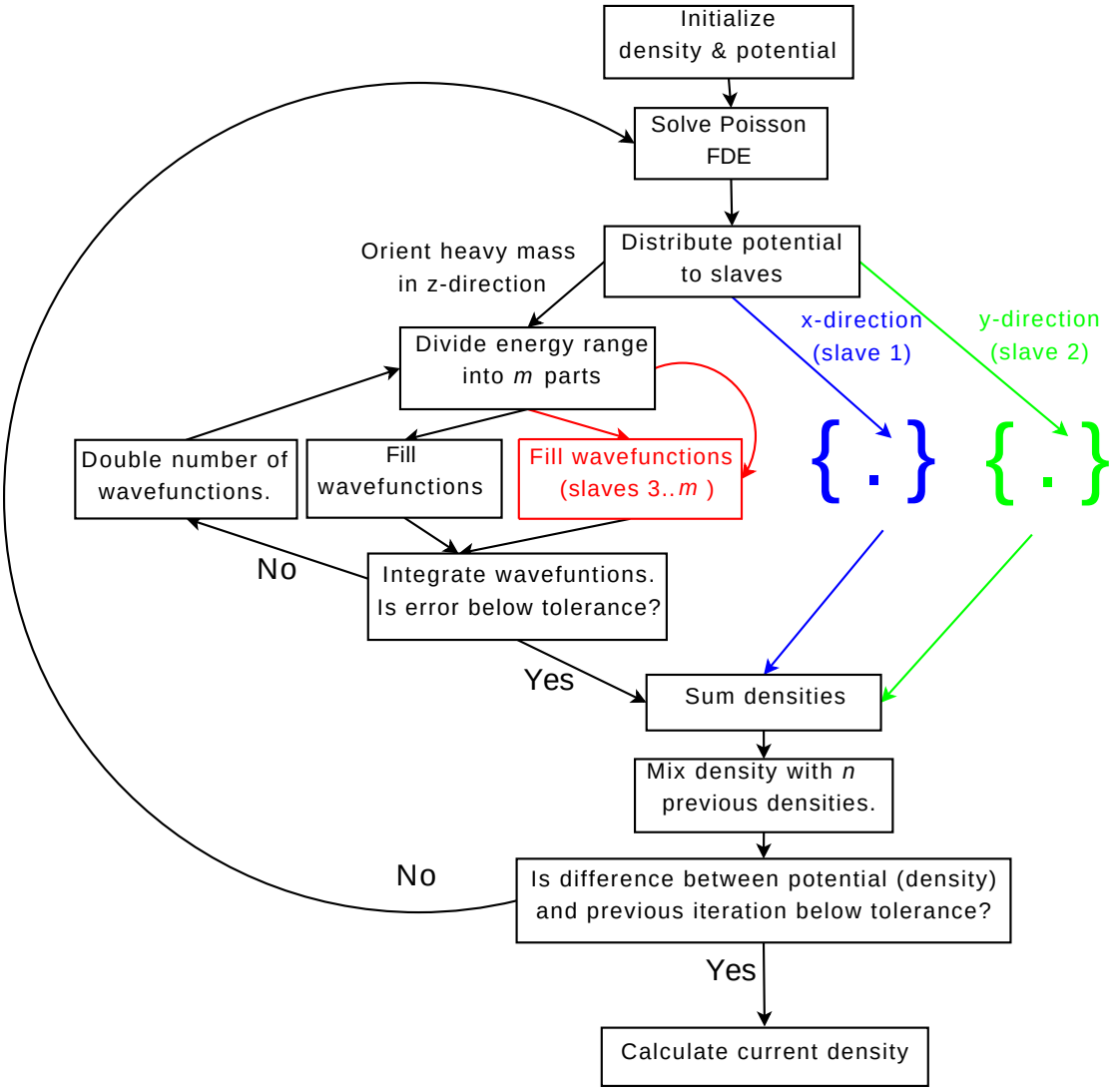

Figure B.2: Decision graph of the parallel algorithm. Black boxes indicate work done by the primary calculation node. Colored boxes are simultaneous work done by slave nodes.